SYSTEMATIC REVIEW*

Application of AI-based Models for Online Fraud Detection and Analysis

Antonis Papasavva[1*], Shane Johnson[1], Ed Lowther[3], Samantha Lundrigan[2], Enrico Mariconti[1], Anna Markovska[2] and Nilufer Tuptuk[1]

This manuscript has been accepted in the Crime Science Journal. Please cite accordingly.
DOI: 10.1186/s40163-025-00248-8

Abstract

Background: Fraud is a prevalent offence that extends beyond financial loss, impacting victims emotionally, psychologically, and physically. Advances in online communication technologies continue to create new opportunities for fraud, and fraudsters increasingly using these channels for deception. With the progression of technologies like Generative Artificial Intelligence (GenAI), there is a growing concern that fraud will increase in scale using these advanced methods, with offenders employing deep-fakes in phishing campaigns, for example. However, the application of AI to detect and analyse patterns of online fraud remains understudied. This review addresses this gap by investigating the potential role of AI in analysing online fraud using text data.

Methods: We conducted a Systematic Literature Review (SLR) to investigate the application of AI and Natural Language Processing (NLP) techniques for online fraud detection. The review adhered to the PRISMA-ScR protocol, with eligibility criteria including language, publication type, relevance to online fraud, use of text data, and AI methodologies. Out of $2,457$ academic records screened, 350 met our eligibility criteria, and 223 were analyzed and included herein.

Results: We discuss the state-of-the-art NLP techniques used to analyse various online fraud categories; the data sources used for training the NLP models; the NLP algorithms and models built; and the performance metrics employed for model evaluation. We find that the current state of research on online fraud is broken into the various scam activities that take place, and more specifically, we identify 16 different frauds that researchers focus on. Finally, we present the most recent and best-performing AI methods employed for detecting online scams and fraud activities.

Conclusions: This SLR enhances academic understanding of AI-based detection methods for online fraud and offers insights for policymakers, law enforcement, and businesses on safeguarding against such activities. We conclude that existing approaches focusing on specific scams are unlikely to generalise effectively, as they will require new models to be developed for each fraud type.

Furthermore, we conclude that the evolving nature of scams limits the effectiveness of models trained on outdated data. We also identify that researchers often omit discussions of the limitations of their data or training biases. Finally, we find issues in the consistency with which the performance of models is reported, with some studies selectively presenting metrics, leading to potential biases in model evaluation.

Keywords: Artificial Intelligence; Natural Language Processing; Online Fraud; Systematic Literature Review

1 Introduction

Online fraud has emerged as one of the most pervasive and challenging threats in the digital age, affecting individuals of all ages, businesses of different sizes, and governments. Defined broadly, online fraud is an umbrella term that involves acts of deception or deliberate impersonation on the Internet for the personal gain of the fraudster, often resulting in a financial loss for the victim [1]. In addition to financial losses, fraud can have a wide range of impacts on victims. These include emotional and psychological effects such as anger, fear, shame, depression, loss of confidence, and trauma; im-

[*] Correspondence: antonis.papasavva@ucl.ac.uk
[1] Security and Crime Science, University College London, London, United Kingdom
Full list of author information is available at the end of the article

pacts on physical and mental well-being; it can harm relationships and lead to loneliness and isolation; and cause negative changes in behaviour [2]. Although evidence suggests that certain sociodemographic groups face higher risks of fraud (e.g., women aged 25-44 and those in the highest income bracket), fraud affects individuals across all demographics [2] and sometimes in different ways. For example, in a UK study, victims earning £20,000 or less, those aged 65 and over, and female victims reported that fraud impacted their self-confidence more than did victims in general.

For the year ending March 2023, the Crime Survey for England and Wales estimated that 3.5 million fraud offences, including online fraud, took place that year [3]. In that year, compared to the year ending March 2020, advance fee fraud increased significantly, from $60,000$ to $391,000$ offences. This increase is largely due to society's growing reliance on the Internet and digital platforms for everyday services, transactions, and communications. According to The Office of Communications (Ofcom) [4], 92% of adults in the UK use the Internet for a wide variety of activities, including communication, education, and entertainment. Activities such as banking, shopping, and socialising are increasingly happening via online platforms, expanding the landscape for fraudsters to exploit vulnerabilities or use these platforms to deceive victims. In 2020, online shopping scams made up 38% of all reported scams worldwide, an increase of 6% compared to the pre-Covid-19 outbreak [5].

Online fraud encompasses a wide range of deceptive activities, including identity theft, phishing, advance fee fraud, romance scams, fraudulent investment scams, and more. It is important to highlight that there is no universally accepted definition of "online fraud," and the term is often used interchangeably with the term "scam". Legally, "fraud is defined as false representation to cause loss to another or to expose another to a risk of loss" [6], and scam is the process where criminals gain the trust of victims to deceive or cheat them [7] through false representation and other means, so that the victim trusts them, which in turn results in various kinds of losses.

The National Fraud Authority of UK published a literature review [8], in which they compared the distinction of the term fraud as defined by the amended Fraud Act 2006 [6] and the typology produced by Levi [9]. They found that fraud embraces a broad scope of crimes, whereas scams often focus on fraud against individuals and small firms. For example, different scams like advance fee, romance, tech support, etc., all fall under the fraud umbrella, but they are also deception methods, which are, in part, scams. Hence, in this work, we use both terms as various scams represent the different deception methods scammers use to trick victims, while the term fraud includes all scams.

Online frauds exploit the virtual nature of the Internet and the anonymity it provides to reach victims. This virtual environment, coupled with jurisdictional challenges (where offenders and victims may be in different regions of the world), makes fraud difficult to detect and prevent using traditional policing techniques. The complexity of online fraud is further heightened by its evolving nature, as fraudsters continuously adapt their techniques to bypass new security measures and exploit emerging technologies to target new victims [10].

Given the scale and impact of online fraud, there is a need for new methods to detect and prevent such activities. The use of Natural Language Processing (NLP), in combination with other Artificial Intelligence (AI), has been proposed for identifying, characterising, and detecting fraudulent patterns in applications like phishing [11], fake job advertisement [12], and for the purposes of analysing scam patterns [13] which could help develop preventive measures and mitigate risks of online fraud. However, understanding the current state of AI techniques in combating online fraud, the data sources used, the evaluation methods for AI models, and the specific types of fraud that are most prevalent, remains a significant challenge. This is due to the constant emergence of new fraud activities that use various communication mediums and social engineering attacks, in an attempt by fraudsters to remain undetected. Therefore, there is a pressing need to shift from detecting and analysing the effects of fraud to the early detection of emerging fraudulent activities online and new methods of social engineering.

This study aims to address these challenges by conducting a comprehensive review of the state-of-the-art AI techniques used to detect fraudulent online activities. Specifically, we examine the data sources widely used by researchers to study online fraud, the methods researchers use to evaluate the developed AI models, and the most popular types of online fraud targeted in their studies. By synthesising findings from academic papers, this review aims to provide a thorough understanding of the current landscape of online fraud detection and prevention, highlighting gaps in existing research, and proposing directions for future studies.

Manuscript Structure. The rest of the paper is organised as follows. The next section (§2) introduces various well-known types of online fraud and provides a detailed discussion of the latest and most widely used AI methodologies, including how they are evaluated. The background review conducted for this section helped us to formulate and refine our research questions.

Section 3 outlines the methodology followed in this SLR, including the protocol, the criteria used to filter eligible papers, and the data extracted from each study. The results of the SLR are then presented in Section 4. In Section 5, we discuss the findings of our literature review, categorized by the various types of online fraud identified, and provide detailed insights into how each of our research questions were addressed.

Finally, Section 6 offers a deeper analysis of our findings, highlighting limitations and shortcomings in the reporting of AI models, particularly regarding performance and data sources. We also propose recommendations for researchers developing detection models for online fraud, before concluding in Section 7.

2 Online Fraud and AI

Online fraud refers to any deliberate act of deception conducted over the Internet to cause an unlawful or unfair loss [6]. It involves exploiting online platforms, services, and technologies to deceive individuals or organisations for financial, personal, or material gain. Online fraud can take many forms, each characterised by the method of deception and the medium used.

2.1 Fraud Categories

The list of online fraud activities is extensive and constantly evolving, with new types and sub-types emerging [10]. To conduct our SLR, it was important first to identify the most prevalent types of offences likely to be analysed by the studies included. To briefly discuss online fraud, we studied various taxonomies, studies, and reports published or discussed by UK government bodies [2, 8], financial services [1], telecommunication providers [14], policing think tanks [10], and academics [9, 15, 16].

Developing a comprehensive taxonomy or classification for all online fraud activities requires special attention, which is beyond the scope of this work. Note that many taxonomies, especially the ones published by UK government bodies [2, 8] also discuss fraud and crime that potentially can take place offline, which we omit in the following discussion of online fraud as offline crime falls beyond the focus of our SLR. Below, we outline some of the well-known and most discussed online fraud types we encountered while performing our preliminary research on online fraud, aided by the reports discussed above. We note that the following is not intended as a complete taxonomy of online fraud, nor does it represent the findings of this SLR. Instead, it is intended to briefly discuss popular online frauds and scams for the reader.

- **Phishing** is the process where fraudsters impersonate representatives of legitimate organisations or acquaintances of the targeted victim to trick them into providing personal information such as usernames, passwords, credit card details, or bank account details. This activity can be done through various mediums, like email, phone calls (aka Vishing), SMS (aka Smishing), and any other way of online communication. Various phishing scams have surfaced over the years, including the Royal Mail scam [17], banking scams [18], and HMRC scams [19]. Notably, phishing scams often include deceptive web addresses created by cybercriminals to trick victims into believing they are visiting legitimate websites. The primary goal of these URLs is to steal personal data, including usernames, passwords and credit card details, for financial gain.
- **Fake Reviews** are deceptive or fraudulent reviews created to mislead potential customers about the quality, reliability, or legitimacy of a product, service, or app. On fraudulent e-commerce websites and app stores, fake reviews play a crucial role in tricking victims into trusting and using fraudulent apps or purchasing substandard or non-existent products. This leads to potential victims trusting fraudulent websites, services, or apps, providing them with their credit card details for a purchase, which leads the victim to a vulnerable position [20].
- **Recruitment Fraud** is a type of online scam where fraudsters pose as legitimate employers or recruiters to deceive job seekers. The primary goal of these scams is to receive “fees” for a job application, steal personal information, extort money, or exploit the victim in some other way. This type of fraud preys on individuals seeking employment, often targeting those who are most vulnerable or desperate for work [21].
- **Romance Fraud** (aka romance scams or dating scams) involves fraudsters creating fake profiles on dating websites, social media, or other online platforms to deceive victims into believing they are in a genuine romantic relationship. The primary objective is to exploit the victim’s emotions to extort money, personal information, or other benefits. This elaborate scam is extremely difficult to detect since it is also under-reported due to victims feeling ashamed and hurt for being victimised by someone they considered to be a romantic partner [22]. In these scams, fraudsters communicate with victims for a long time before presenting them with an “investment opportunity” or requesting their financial aid. Romance scams are closely related to *Cryptocurrency Pig Butchering scams* [23], where victims are gradually lured into making increasing contributions over a long period of time, usually in cryptocurrency, to a fraudulent scheme [24].
- **Fraudulent Investment** includes scams where fraudsters promise victims significant winnings or

lucrative opportunities [25]. These scams are usually associated with the romance scams discussed above. Once the victims try to withdraw their "winnings," the scammers will extort them by asking for "fees" and "taxes" to be paid in advance. The promised benefits and winnings never materialize, and the initial investment sums and fees are lost [26]. Fraudulent investment is the umbrella that covers Cryptocurrency Pig Butchering scams explained above, and various *Ponzi schemes* [27] where early investors greatly benefit from the investments of later investors, also known as *pyramid schemes*.

- **Crypto Market Manipulation** involves artificially increasing or decreasing the price of cryptocurrencies to achieve financial gain. It often involves coordinated efforts by individuals or groups to manipulate the market to create false perceptions of supply, demand, or market sentiment. Some common techniques used in crypto market manipulation include: *Pump and Dump*, which inflates the price of a cryptocurrency through misleading or false statements (pumping), encouraging others to buy, and then selling off the cryptocurrency at a profit once the price has been pumped up (dumping); *Wash Trading* occurs when a trader buys and sells the same cryptocurrency simultaneously to create deceptive activity on the market; *Spoofing* involves placing significant buy or sell orders to withdraw them before execution to mislead perceptions related to the market demand or supply; *Front-running* involves placing orders ahead of a large trade that is known to occur, to benefit from the subsequent price movement caused by the large trade; and many others [28]. These scams are also similar to *Stock Market Manipulation*.
- **Fraudulent E-Commerce** involves deceptive practices or scams conducted through online e-commerce platforms. These scams aim to exploit digital payment systems to deceive consumers or businesses by paying for a fraudulent product or service.
- **Fraudulent Crowdfunding** refers to the misuse of crowdfunding platforms to deceive donors or backers, often by providing false or misleading information about a crowdfunding campaign's nature, purpose, or outcome. Crowdfunding is a method of raising money from many people via online platforms to fund projects, products, or causes [29]. A fraud similar to crowdfunding is *Charity Fraud* and *Disaster Scams*, where scammers seek donations for organisations that do not exist or do little work. These scams are particularly common after high-profile disasters as criminals often use tragedies to exploit people who are looking to donate [30].
- **Gambling Fraud** is any illegal activity that is intended to cheat players or an online gambling platform. Fraudsters manage to trick victims and platforms in different ways, including rigged games, fake websites (phishing URLs described above), account takeovers (via stealing legitimate users' access codes), and creating fake apps with fake reviews, as discussed above, to gain the trust of users. Online gambling fraud can happen on multiple platforms and involve a wide variety of games, including *casino scams, sports betting scams*, and *lottery scams* [31].
- **Tax Scams** occur when scammers falsify information regarding pending tax money or maliciously impersonate tax officials to trick individuals or business entities into wilfully paying them "fees" [32]. Scams similar to tax scams are *Council Tax Scams*, various *Utility bill scams*, *Insurance Scams*, etc. These scams fall under the umbrella of *Phishing* as they often take place via SMS, phone calls, or emails.
- **Pension Scams** are similar to Tax Scams. Scammers aim to make money through fees, direct access to pension savings, or by receiving investments [33].

The complexity and interconnected nature of scams and frauds make categorising them under a single typology challenging [10, 34]. Phishing scams, for instance, serve as a broad umbrella that covers phishing conducted using various methods like vishing (voice) and smishing (SMS). Yet, they can also be integral parts of investment scams when scammers develop phishing websites to gain victims' trust. Similarly, most scams often involve the scammer impersonating an authority (government, law enforcement), friend, organisation (e.g., bank), or other entity (e.g., delivery service), making impersonation scams difficult to break down as they are an integral part of other scams (e.g., a delivery scam is also an impersonation scam as the offender is clearly not an Amazon representative, for example).

To summarise, different scam types frequently overlap, blurring the lines between distinct categories and demonstrating today's intricate web of fraudulent activities. The multifaceted nature of these scams highlights the difficulty in creating a comprehensive classification system that can effectively encompass all types of fraudulent schemes. We discuss this challenge and limitation in detail before concluding our SLR, in our Discussion (Section 6).

2.2 AI techniques

This study investigates AI-based techniques for processing unstructured text data to analyse fraud. Much of this text data, like news articles, research papers, government reports, books, social media posts (tweets and Facebook comments), communications (such as emails, SMS messages, and chat logs), and web content

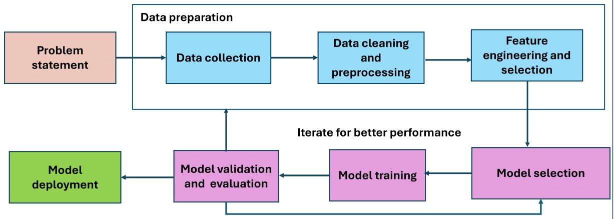
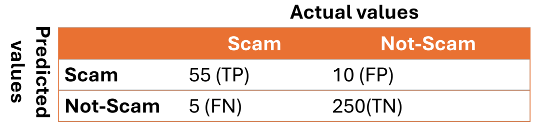

Data preparation
Problem statement
Data collection
Data cleaning and preprocessing
Feature engineering and selection
Iterate for better performance
Model deployment
Model validation and evaluation
Model training
Model selection

Figure 1 Common pipeline for NLP-based models

(reviews on online marketplaces, travel and hospitality platforms, and comments on video sharing platforms), is inherently unstructured. Statista [35] estimated that the global open data that is accessible on the entire Web was 64.2 zettabytes in 2020, and it is expected to exceed 180 zettabytes by 2025. With each new digital platform or communication channel, this data is increasing. Most of the data created is unstructured text that provides opportunities for understanding human behaviour, habits, opinions and experiences. It contains information about users' experiences, events, themes, opinions, and sentiments that can be important for deriving meaningful insight from their experience related to fraudulent activities. Manual traditional data analysis techniques, like keyword searches and the coding of themes, are often limited, and the extraction of meaningful insights at scale is unachievable, making advanced computer-driven automated techniques necessary.

However, there are often significant challenges associated with the analysis of this data due to the diversity of natural (human) language used. This includes dealing with noise (irrelevant or useless data), a wide array of linguistic variations of human language due to regional or cultural nuances, the use of slang or jargon, abbreviations, spelling errors, typos and grammatical mistakes, which often pose challenges for the efficient analysis of text data. NLP techniques were designed to effectively understand the structure (syntax) and comprehend (semantic) spoken and written human language the way humans do. Advancements in AI, including machine and deep learning, along with improvements in technology (such as increased computing power) and software (such as the availability of programming tools and libraries), have significantly improved the ability to process and understand large volumes of unstructured text. These tools have been widely used in many fields, from sales and marketing to spam detection. The process of collecting, preprocessing, and training AI-based models using text data often involves the pipeline shown in Figure 1:

– **Problem statement**: Using domain knowledge, a suitable research question is formulated for AI to address. This could be a classification problem (e.g. to classify text into a number of categories), or explanatory analysis involving the identification of patterns within a text.
– **Data preparation**: AI-based models require the collection of appropriate data towards the building of a *corpus* (a collection of structured sets of *documents* such as emails, news articles, social media posts, or transcripts) used to train models to analyse the research question. Often, the acquired data comes as unstructured data and requires cleaning and pre-processing. The data cleaning and pre-processing involve removing unwanted or redundant data to reduce the noise in the data. This may include removing duplicates or incomplete entries, symbols, punctuations, numbers, stopwords, converting acronyms to full words, and handling non-English words, slang, or jargon. Further pre-processing may involve text normalisation techniques like *stemming* or *lemmatisation* to reduce words to their root or base form to improve the accuracy of text analysis.
– **Feature engineering and selection**: Feature engineering involves preparing data for machine learning models. It consists of extracting and selecting predictive features in supervised learning or finding patterns in unlabeled data in unsupervised learning. This task requires using domain knowledge to develop and select appropriate features. Common text features often used are *n-grams* (sequences of n consecutive words); *Term Frequency-Inverse Document Frequency (TF-IDF)* (a statistical method that weights the importance of a word/term in a document within a corpus) matrix; sentiments and emotions present; lexical features (e.g. presence of certain words, Keyword-in-Context and lexical diversity); syntactic features (e.g. Part-of-Speech tags); semantic features (e.g. entities mentioned, word-embedding); readability scores; structural features (e.g. length of the text, number of paragraphs); and domain-specific features (e.g. presence of specialised terms).
– **AI technique/algorithm selection**: This step involves selecting an appropriate AI algorithm for building the AI-based model. Tasks associated with text often involve two main categories of AI-based models: *supervised* and *unsupervised* machine learning. The choice of the algorithm will depend on the learning, and type of AI required. Supervised machine learning algorithms are often used for text classification problems. The learning algorithm is fed with input features (training data) and labels



		Actual values	
		Scam	Not-Scam
Predicted values	Scam	55 (TP)	10 (FP)
	Not-Scam	5 (FN)	250(TN)

Figure 2 Confusion Matrix

(discrete outputs). The supervised machine learning algorithm aims to map input features to discreet outputs. Traditional supervised machine learning algorithms include Logistic Regression (LR), Naive Bayes (NB), Decision Trees (DT), Random Forest (RF - multiple DTs), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). New supervised machine learning techniques include neural networks (NN) and deep learning-based models. Unsupervised machine learning models, often used in exploratory data analysis, involve working with unlabelled data to discover hidden patterns and themes. Unsupervised machine learning algorithms include clustering techniques using algorithms like K-means, hierarchical clustering and Density-Based Spatial Clustering (DBSCAN); and topic modelling, achieved by algorithms like Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA).

- **Model training**: Often, machine learning algorithms will have parameters that need to be tuned before learning begins, known as hyperparameters. The tuning process involves re-training the model multiple times using different values for these hyperparameters and selecting the best combination of values based on model performance on a metric of interest. In the case of supervised machine learning, the hyperparameters might be tuned using model performance on different "folds" of the data in an approach known as cross-validation. With this approach, a randomly selected proportion of the data is kept separate from the training data, and used for final model evaluation. This approach provides the best indication of how the model will likely perform on new, unseen data. In the case of unsupervised modelling, heuristics are used to identify the optimal number of clusters or topic modelling.
- **Model evaluation**: The model's performance needs to be evaluated. In the case of supervised modelling, this will involve measuring the model's performance on the test data. The classic supervised machine learning algorithms can be evaluated using performance metrics such as a confusion matrix (Figure 2), accuracy, precision, recall, F1-score, sensitivity, specificity, Receiver Operating Characteristic (ROC) curve, and the Area Under the Curve (AUC) curve:

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

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

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

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Where:
True Positives (TP): The model correctly predicts a positive class (e.g., those that were classified as scam.)
True Negatives (TN): The model correctly predicts a negative class (e.g., those classified as not-scam).
False Positives (FP): The model incorrectly predicts the positive class (e.g. not-scam is predicted as a scam).
False Negatives (FN): The model incorrectly predicts the negative class (e.g. scam predicted as not scam).

Sensitivity is the same as recall or the true positive rate, and it captures the model's ability to identify the positive class (i.e. scam cases) correctly:

$$Sensitivity(TPR) = Recall = \frac{TP}{TP + FN}$$

Specificity, also known as false positive rate (FPR), measures the proportion of true negatives, and it captures the model's ability to identify negative class (i.e. not-scam cases) correctly:

$$Specificity = \frac{TN}{FP + TN}$$

The ROC curve illustrates the performance of one or more binary classifiers. It plots the sensitivity against the 1-specificity for various thresholds. The AUC is calculated as the area under the ROC curve.

- **Deploy model**: Once the models work well, they can be deployed. When considering deployment of the model, one must address questions regarding why others should trust the model, how the model arrived at its conclusions and usability, and carefully assess the ethical implications of AI to ensure its suitability for deployment and that it is not biased.

In unsupervised machine learning models, due to a lack of ground truth labels, the performance of the model evaluation may involve subjective interpretation to interpret the outputs (e.g. clusters or topics) generated by the model.

2.3 Advanced NLP techniques

This section briefly introduces advanced NLP techniques, aiming to familiarise the reader with these concepts as they are later referred to during the SLR findings.

Word embeddings is an important technique in NLP that involves encoding words as vectors of real numbers that are designed to capture their similarities.

Words closer together in the vector space are expected to have similar meanings or relationships. Two of the widely used word embedding techniques are Word2Vec and GloVe. Word2Vec uses a simple neural network trained on large text datasets iteratively to predict either context words or target words. Word2Vec uses two approaches [36]: Continuous Bag of Words (CBOW) predicts the target word based on its surrounding context words, whereas Skip-gram predicts surrounding context words based on a given target word. In the sentence ‘The quick brown fox jumps over the lazy dog’, if ‘fox’ is used as the target word, the CBOW model uses ‘The’, ‘quick’, ‘brown’, ‘jumps’, ‘over’, ‘the’, ‘lazy’, and ‘dog’ as context and predicts the word ‘fox.’ In Skip-gram, ‘fox’ is used to predict the surrounding words like ‘The’, ‘quick’, ‘brown’, ‘jumps’, ‘over’, ‘the’, ‘lazy’, and ‘dog’. *Global Vectors for Word Representation* (GloVe) [37] learns the vector representation of words using global word-word co-occurrence statistics obtained from the training data to show the semantic relationships between words.

Large Language Models use word embeddings to generate responses to natural language inputs.

Large Language Models (LLMs) [38] are advanced NLP tools trained on billions of words from a wide variety of sources and are designed to perform complex tasks like translations, summarisation and the performance of human-like conversational abilities. Most LLMs are developed using a transformer-based architecture (*transformers*) [39], and billions of parameters are used for training. Transformers are a type of deep-learning neural network model, and they are more efficient compared with predecessor state-of-the-art models based on Recurrent Neural Networks (RNN). Transformers use a complicated architecture with encoder and decoder layers to understand sequences of words and provide an output [39]. While the encoder layer processes input text data, extracting hierarchical representations through mechanisms like self-attention, the decoder layers generate output sequences based on the input received from the encoder. Transformer-based LLMs include GPT models like Generative Pre-trained Transformer 3 (GPT-3), GPT-4 and GPT-4o developed by OpenAI. GPT-4 and GPT-4o are multimodal models that accept text and image and produce text [40]. Other transformers include *Bidirectional Encoder Representations from Transformers* (BERT) and its smaller and lighter version of *DistilBERT*, designed for applications with limited computational resources. BERT and DistilBERT are also designed to understand context in language processing and are suitable for NLP tasks like text classification, answering questions, and named entity recognition. The difference between models like BERT and GPT is the way their architecture is designed and the intended learning objectives.

LLMs can assist in analysing large amounts of text data and identify patterns automatically, which can be helpful when dealing with fraud and other crime-related data. LLMs have been successfully applied in various areas of human communications, including chatbots in customer support systems, by generating human-like text, content generation, and performing language translation. *Generative AI* (GenNAI or GAI) refers to AI techniques that create new text, audio, images and video that closely resemble human-generated content. On the other hand, criminals can misuse these resources to generate content for fraudulent activities, such as fake websites, targeted phishing emails, and scam advertisements, to deceive potential victims.

2.4 The Use of AI in Fraud Detection

Although some literature reviews explore the application of AI for fraud and crime, to the best of our knowledge, no reviews currently aim to understand the state-of-the-art in detecting online fraud in general. The literature reviews we found, discuss the detection of specific online fraud or scams, such as credit card fraud [41] and phishing SMS [42], among others.

In more detail, our preliminary analysis of literature reviews finds that specific AI models work best towards detecting specific types of fraud (e.g., phishing URLs, smishing, etc.), as researchers perform literature reviews to analyse specific offences and not analyse the general task of fraud overall. In addition, a single/universal model does not perform well at classifying various types of fraud. Hence, researchers must constantly develop and update their trained models to detect specific fraud types. In this work, we aim to understand whether there are *universal* AI methodologies that attempt to detect online fraud, in general, focusing on textual data.

Research Questions:

Search String	Library	Notes
("Online Fraud*" OR "scam*")	ACM	All text
AND (("machine learning")	ProQuest	All text, Journals, Conferences
OR "NLP" OR ("natural language processing")	Web of Science	Topic
OR "classifier" OR ("Large Language Models") OR "LLM"	IEEE Xplore	Abstract, Journals, Conferences
OR ("Generative Artificial Intelligence")	arXiv	Abstract, Computer Science, Jan 2023-Mar 2024
OR "GenAI" OR "GAI")	Google Scholar	All text, Review articles, 2023-2024

Table 1 Search query for the literature selection in various academic libraries. Notes depict the advanced search filters applied to each library.

- **RQ1**: What is the state-of-the-art of AI techniques used to detect online fraud?
- **RQ2**: What are the data sources researchers use to analyse online fraud?
- **RQ3**: How do researchers evaluate their AI models?
- **RQ4**: What are the most popular fraud activities that researchers studied?

Although a wide number of studies have explored the application of AI for fraud detection and other types of cybercrime, we are unaware of any systematic literature reviews that have examined the application of AI models using text data. This SLR focuses on AI-based models that study textual data to detect and gain insights about online fraud. Thus, this study identifies NLP models used to detect online fraud.

3 Systematic Review Methodology

Systematic reviews differ from traditional literature reviews as they aim to identify all relevant studies that address a set of research questions using a structured methodology that can be replicated [43].

3.1 Methods

We use the following methodology to conduct the SLR and address the selection process to identify relevant publications and avoid biases.

3.1.1 Protocol

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis extensions for Scoping Reviews (PRISMA), as proposed by Moher et al. [44]. In a nutshell, this provides a comprehensive framework for conducting and reporting systematic reviews and meta-analyses. The process includes a checklist and flow diagram to ensure transparency, reproducibility, and rigour in summarizing research evidence, to improve the quality of reviews in various fields and to standardise how a literature review should be reported.

3.1.2 Eligibility

This review focuses on studies that use AI-based models, specifically NLP models, including Machine Learning (ML) and Deep Learning (DL) techniques using text data. For example, studies that employ AI to detect fraudulent bank accounts, fraudulent credit card transactions, or fraudulent networks of users online were out of scope. For a study to be considered for inclusion in the SLR, we used the following eligibility criteria:

- **Peer-Reviewed Studies**: We focused on peer-reviewed studies published in English between January 2019 and March 2024. Our search was restricted to academic records found in journals and conference proceedings. We excluded theses, legal documents, patents, and citations.
- **Grey Literature**: To capture the latest AI-based models, we also included grey literature, specifically pre-prints from ArXiv published between January 2023 and March 2024 that have not yet been incorporated into conference proceedings or academic journals. We also conducted searches on Google Scholar between January 2023 and March 2024.
- **Search Strategy**: Table 1 shows the search string used to query related literature in ACM Library, ProQuest, Web of Science, IEEE Xplore, arXiv, and Google Scholar. This was finalised after trying various searches in these libraries. Due to the different functionalities of each library, we had to adjust our search accordingly: for ACM library, we queried our search string across the entire text and adjusted the time range; for ProQuest, we queried our search string across the entire text, adjusted the time range, included only papers from conferences and journals, included only full text and peer-reviewed results, and filtered the subjects to exclude medical terms; for Web of Science we queried our search string on the topic (title, abstract, and keywords) of the paper and adjusted the time range; for IEEE Xplore, we queried our search string on the abstract of the papers, adjusted the time range, and filtered results for papers published in conference proceedings and journals; for arXiv, we queried our search string on the abstract of papers, published in computer science between January 2023 and March 2024; and finally, for Google Scholar, we queried our search string and filtered our results on review articles published between 2023 and 2024. The adjustments mentioned above were implemented to better capture literature related to the scope of our study, and a consensus was reached after various iterations and discussions between all of the authors.

Label	Description	Example
Title	The title of the manuscript	Detecting Phishing URLs Using NLP
Author	Author's full name plus the abbreviation et al. if applicable	Smith, John or Smith et al.
Year	The year the work got published (YYYY)	2020
Fraud Type	The type of scam the authors try to detect, analyse, or discuss in their manuscript	Phishing URLs
Data Type	The type of data the authors use for their analysis	URLs
Data Quantity	The amount of data used for the analysis	100 phishing URLs, 100 legitimate URLs
Models Used	All models the authors experimented with	RF, LDA, W2V
Best Model	The model with the best accuracy	Random Forest
Model Stats	All performance metrics	A=0.95, P=0.81, R=0.9, AUC=0.89

Table 2 Data items and characteristics extracted from the literature.

- **Scope and Focus**: Studies must address fraud performed online and use AI-based methodologies for analysing online fraudulent activities. The focus was on studies using *textual data*, whether from scammers, victims, or victim reports, to understand, detect, or analyse online fraud activities. Our goal was to understand the state-of-the-art models designed to prevent and detect scams before the victim gets defrauded. Studies that analysed money transactions, credit card transactions, and cryptocurrency transactions were *excluded* from this review, as they do not use text data.
- **AI-Methodology**: Papers had to include a methodology or similar section where the authors discuss their AI implementation and fine-tuning along with the accuracy of their model. Finally, we considered studies published after 2019.
- **Publication Time Frame**: Papers published between January 2019 and March 2024 were included in this review. This period was selected to manage the overwhelmingly large volume of online fraud papers and align with our available resources. Also, we believe that studies conducted before 2019 are less likely to reflect recent advancements in AI methods. Given the rapid evolution of AI-based models, our time frame ensures the inclusion of the most up-to-date and relevant research.

3.1.3 Data extraction

Next, the authors agreed on the data to be extracted from the included studies. Only one of the three reviewers carried out the task of extracting data. This was deemed sufficient since the reviewer's role involved extracting the required details from the papers, and a second reviewer did not have to check accuracy. The only aspect of the data extraction that the reviewer had to conceptualize was the specific *Fraud Type* analysed by the study under question. For example, if a study analysed Phishing URLs, it was labelled as *Phishing URLs*.

Not all papers explicitly specified the type of fraud analysed. Due to the diversity of scams, there is no agreed way of labelling fraud types. As a result, we used an umbrella term to categorise them. For example, online scam campaigns made by bot users on various social media platforms to advertise fraudulent phishing URLs include *fake users* and *phishing URLs* analysis; hence, we agreed to label papers of broad online scam campaigns as *social media scams*.

We did not classify a paper with more than one fraud type to ease the representation of our findings. Instead, the reviewer recorded the paper's primary goal when identifying the fraud analysed. For example, if a paper used phishing emails to extract phishing URLs towards detecting phishing URLs, then that paper would be labelled as *Phishing URLs*, as that was the study's main goal. The final version of the data extracted from each record is depicted in Table 2, along with relevant examples. A thematic analysis was conducted on the extracted data of the studies included for qualitative analysis, and we present our findings in Section 5.

4 Results

4.1 Study selection and characteristics

The PRISMA-ScR flow diagram in Figure 3 summarises the study selection process. We identified 2, 617 studies for eligibility screening. The ACM Digital Library returned 389 documents, IEEE Xplore returned 712 documents, Web of Science returned 253 documents, ProQuest returned 783 documents, Google Scholar returned 399 documents, and ArXiv returned 47 documents. Experts in the area recommended an additional 34 papers. After removing duplicates, 2, 457 papers remained for further review.

At this stage, 10% of these papers were selected ($N = 242$) for Inter-Rater Reliability to calculate the multi-annotator agreement between the three annotators of this review. The Fleiss Kappa score between all three annotators was 0.83, indicating almost perfect agreement. The Cohen Kappa Agreement was also calculated between each pair of annotators. The agreement between annotators AP and NT was 0.65 (substantial agreement), between AP and EM was 0.66, and between NT and EM was 0.52 (moderate agree-

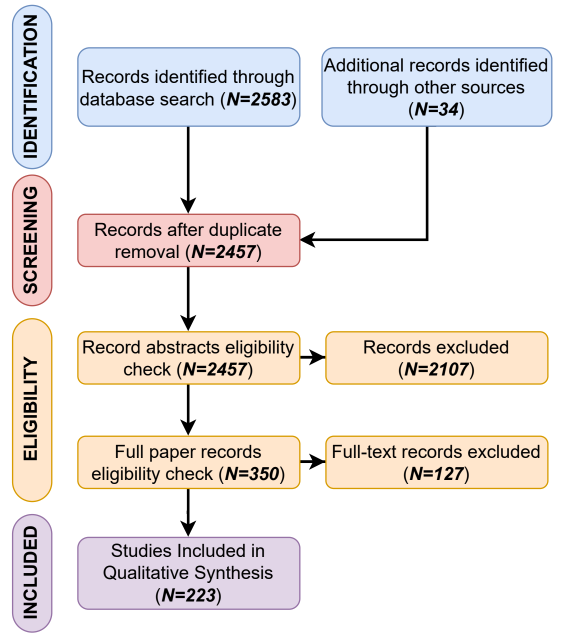

IDENTIFICATION
Records identified through database search (*N=2583*)
Additional records identified through other sources (*N=34*)
SCREENING
Records after duplicate removal (*N=2457*)
ELIGIBILITY
Record abstracts eligibility check (*N=2457*)
Records excluded (*N=2107*)
Full paper records eligibility check (*N=350*)
Full-text records excluded (*N=127*)
INCLUDED
Studies Included in Qualitative Synthesis (*N=223*)

Figure 3 PRISMA Chart

ment).[1] The three annotators compared their annotation process and reviewed this SLR's eligibility criteria and goals. Then, the lead annotator performed the rest of the annotations of the papers included in this review.

Reviewing the abstract of those papers resulted in 2, 107 papers being excluded from the study as they did not fit the eligibility criteria discussed in Section 3.1.2. This resulted in 350 full-papers passing to eligibility screening, out of which 127 did not fit the eligibility criteria and were excluded. Overall, this process resulted in 223 full-text papers being included for qualitative analysis.

4.2 Types of Online Fraud Identified in the Literature

Figure 4 shows the types of fraud analysed in the full-text papers included in the qualitative analysis. The reviewer of these studies manually coded each paper with the *scam type* that the study focuses on, based on its title, abstract, and methodology. The majority of studies focus on *phishing* detection, with about a third (29%) of the studies analysing phishing URLs online ($N = 64$). More specifically, these works tackle the problem of automatically detecting whether a URL is likely fraudulent. Many papers were related to detecting phishing emails ($N = 29$), followed by studies on SMS phishing detection ($N = 20$). Other studies on phishing include phone call transcripts towards understanding and detecting voice call phishing ($N = 12$), and a few studies attempt to understand phishing methods via victim reports ($N = 4$).

Moving on to other types of fraud, we found many studies that detect fake reviews on various platforms like the Google Play Store, Apple App Store, Yelp, and TripAdvisor ($N = 23$). Another widely studied scam was *recruitment fraud* ($N = 20$). We also found several studies that employed AI techniques to detect *fake accounts* on Facebook, Instagram, and Twitter ($N = 18$).

Similarly, 3 studies focus on *romance fraud* via analysing profiles on social media and discussions with victims. We also identified 3 studies that analysed and detected fraudulent *cryptocurrency investment* scams and 2 studies that attempted to detect the likelihood of cryptocurrency manipulation. Finally, we identified 2 studies that analysed fraudulent crowdfunding online and 1 that studied fraudulent e-commerce websites.

Studies on other emerging types of fraud. A few studies have used Generative AI (GenAI) to analyse and better understand existing and emerging fraud techniques, especially ones where GenAI or LLMs are misused towards social engineering attacks. We identified 3 studies that employed GenAI models to automatically interact with scammers to waste their resources and time while gathering information on the methods they used to defraud users, thereby disrupting their operations. This approach is defined as *scam baiting*: the process of using generative AI models to deceive and engage with scammers. Notably, this is a *countermeasure* against online fraud ("GenAI Scam Bait" in the Figure 4).

Our search also returned many studies that discuss the exploitation of *GenAI towards social engineering attacks* ($N = 13$), where scammers use these advanced models to create legitimate-looking targeted emails or SMSs to earn the trust of potential victims. Although GenAI models offer numerous benefits, these studies show the significant risks they pose when used for malicious purposes, particularly in social engineering. GenAI can generate coherent, contextually relevant, and grammatically correct emails that mimic the style and tone of professional communication. This increases the likelihood of victims perceiving fraudulent emails as legitimate and trusting the message [45].

Regarding other scams, we found 6 studies that detect *social media scams*. These scams included various fraudulent activities, like fake user accounts and online groups, advertisements of fraudulent apps, or phishing

[1] AP stands for author Antonis Papasavva, NT for author Nilufer Tuptuk, and EM for author Enrico Mariconti.

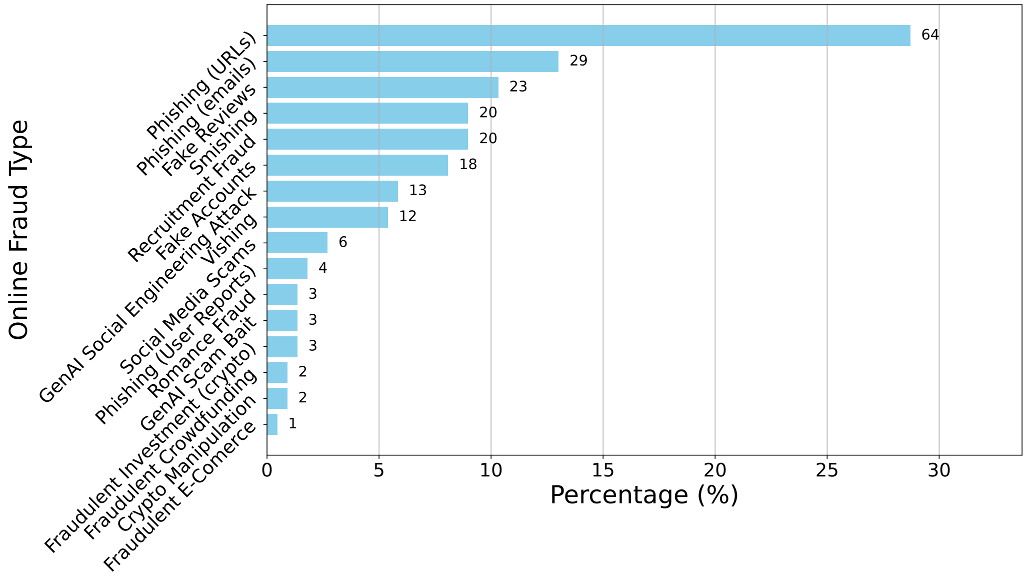


Figure 4 Percentage of scam types analyzed in the studies included for qualitative analysis

websites that aim to trick users into exposing their personal information or paying money.

The above three groups of studies were not identified in the literature discussed in Section 2; hence, we grouped and discussed them briefly here.

5 Summary of Findings

This section summarises our findings, categorized per fraud activity analysed within the papers included in our SLR for qualitative analysis.

5.1 Data Sources

First, we report the most popular data sources used, and the datasets analysed.

Data Used for Phishing URL Detection. We start by understanding the chosen data sources for analysing and detecting phishing URLs; the most popular scam-type category we have detected in our SLR. We analysed the data sources and detection methodologies of the identified 63 papers that focused on this issue. Table 3 summarizes these.

Researchers used various websites that offer information on URLs for the analysis of malicious and legitimate domains. This information may be web page rankings (how trusted the webpage is), phishing reports, and historical data. By far, the most popular data source used was PhishTank[2], a website that allows users to report webpages that might be malevolent or suspicious, with 25 studies using it as already labelled malicious websites [46–71].

Another website that offers a list of phishing URLs is OpenPhish[3] and it was used in 3 studies [47, 64, 65]. Two studies used URLhaus[4], a project for sharing malicious URLs, for the collection and analysis of phishing URLs [52, 70]. We also find one study that used SpamHaus[5] for the collection of IP and domain reputation [72], and one that used URLscan[6] [73]. Interestingly, a study [73] also collected user-reported domains from ScammerInfo[7], a forum where users post and discuss various scams. Lastly, the webpage WhoIs[8], a webpage that offers historical data on webpages, was used for feature collection in two studies [57, 74].

The most used data source for the collection of legitimate webpages was "Alexa," (a global ranking system that estimated a website's popularity that shut down in May 2022) with 10 studies using it to collect legitimate annotated webpages [46, 58, 61, 62, 68–70, 75–77]. Google's search engine was used for one study [78], while another used the Majestic Million site[9] for legitimate webpage collection [50], a site similar to Alexa.

Many studies used existing publicly available datasets for their analysis. More specifically, 11 studies [71, 79, 80, 80–87] used publicly available datasets published on Kaggle (a repository for researchers to pub-

[2] https://phishtank.org/

[3] https://openphish.com/

[4] https://urlhaus.abuse.ch/

[5] https://www.spamhaus.org/

[6] https://urlscan.io/

[7] https://scammer.info/

[8] https://who.is/

[9] https://majestic.com/reports/majestic-million

lish data). Similarly, 5 studies [51, 75, 88–90] used the UCI Phishing Dataset.[10] Other studies used publicly available datasets from other sources [71, 91–96].

The most recent studies (published in 2023) that attempted to detect phishing URLs automatically collected data from alternative sources like social networks and user-reported phishing URLs [76, 76, 97, 97–99], while others used datasets from telecom and Security organizations [77, 81, 100]. Alas, we failed to detect the data source used by 9 studies, as the authors did not report how or from where they acquired the dataset used in their study [101–109].

Data Used for Phishing Email Detection. We now discuss the data sources used in the 29 works that tackle phishing email detection. The data extracted from the literature and presented in this section are shown in Table 4.

The overwhelming majority of papers used datasets made available in previous work [110–117], or used datasets published on Kaggle [112, 117–125], or datasets published at UCI ML repository [126–130].

Two studies [131, 132] used emails received in the author's personal or professional email spam folder. Another study that included datasets from alternative sources was by Mehdi et al. [133]. They used various techniques to develop their dataset, including GPT2 generated synthetic phishing emails made available in previous research [134], along with TextAttack[11], a Python framework for adversarial attacks, data augmentation, and model training in NLP, Textfooler[12], a Model for Natural Language Attack on Text Classification, and Probability Weighted Word Saliency (PWWS) [135], a technique for generating adversarial text.

Another alternative data source for phishing email detection was used by Janez et al. [117] who used data from SPAM Archive[13], a website that publishes spam email repositories at the end of every month and is constantly updated. Gallo et al. [136] analysed user-reported emails. Lastly, the data source used in 3 studies was not clearly stated within the manuscript [137–139].

Data Used for Phishing SMS Detection. Regarding Phishing SMS (*smishing*), we included and analysed 20 papers in this SLR. For the detailed data, refer to Table 5.

Similarly to previous analyses, the overwhelming majority of works opted for using already publicly available datasets to analyse and train a model. More specifically, a variety of subsets from a publicly available dataset on Kaggle[14] was used by 14 studies [140–153]. Another study [154] used the Kaggle dataset but incorporated Fake Base Station data and made it available to researchers.[15] Similarly, this work [140] used a subset of the Kaggle dataset in combination with emails and YouTube comments for spam content detection, while Lai et al. [155] used data provided by users.[16] Tang et al. [156] collected tweets where users reported smishing for their analysis. Two other works used data from the Korean Internet and Security Agency [157] and 360 Mobile Safe [158]. Lastly, Timko et al. [159] proposed a platform where users can freely post Phishing SMS for researchers to use.[17]

Data Used for Phishing Phone Call Detection. We identified 12 studies that used phone call transcripts to understand voice call-enabled phishing (*vishing*).

Derakhshan et al. [160] used the CallHome dataset, which includes 120 unscripted 30-minute telephone conversations between native speakers of English.[18] Another study [161] used AI-generated deepfake voice recordings (Tacotron 2[19], Deepvoice 3[20], and FastSpeech 2 [162]). For authentic voice recordings, they used the synplaflex dataset [163], a corpus of audiobooks in French.

Some studies used telecommunication operator datasets, like [164] using fraudulent caller IDs and phone transcripts [165] from telecommunication operators in China, Hu and Yuan [166] used data from the Public Security Bureau in Zhejiang Province, China, and [167] used data from the Korean Financial Supervisory Service. Kale et al. [168] developed their dataset via questionnaires and victim testimonies. Other authors collected data from various social networks, including YouTube transcripts [169], Facebook, online blogs and forums, public datasets, as well as some that were developed based on studies of scammers' activities and behaviours [170]. Others opted for using previously analysed and publicly available data [171, 172], while the data Zhong et al. [173] used was unclear.

Data Used for Phishing (User Reports) Detection. Four studies used user reports to understand phishing activities. First, one study [174] constructed a

[10] https://archive.ics.uci.edu/dataset/327/phishing+websites

[11] https://github.com/QData/TextAttack

[12] https://github.com/jind11/TextFooler

[13] http://untroubled.org/spam/

[14] https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

[15] https://github.com/Cypher-Z/FBS_SMS_Dataset

[16] https://www.datafountain.cn/competitions/508

[17] https://smishtank.com/

[18] https://catalog.ldc.upenn.edu/LDC97S42

[19] https://pytorch.org/hub/nvidia_deeplearningexamples_tacotron2/

[20] https://r9y9.github.io/deepvoice3_pytorch/

fraud complaint dataset from the Internet finance service in China. Similarly, another [175] used court documents from Chinese online judgement records, while a third [176] used incident record forms from victims and interviews in the Philippines. In the final study, the authors [13] launched and introduced a website operated by the National Crime Prevention Council (NCPC) in Singapore, where users could report and receive information on the latest phishing activities. [21]

Data Used for Fake Review Detection. For this type of scam, 23 studies were included in our analysis, and the data extracted from them is depicted in Table 8

The overwhelming majority of papers used a previously published YELP dataset[22] [177–184], or the OTT publicly available dataset on Kaggle[23] [181, 184].

Other studies used application reviews from the Google Play Store or Apple's App Store [185–187]. Other studies used previously available Amazon product reviews[24], or collected Amazon reviews [188–195].

Some studies collected reviews of Amazon Hotel and Holiday packages [196, 197], or TripAdvisor review data [180] And, lastly, one study [198] used YouTube transcripts to interpret false review exaggeration. The data source used by Ganesh et al. [199] was unclear.

Data Used for Recruitment Fraud Detection. We found 19 studies that focused on the detection of fraudulent job postings (see Table 7).

The overwhelming majority ($N = 16$) of papers, used the same publicly available dataset from Kaggle,[25], which holds about 18K job postings of which 800 are fraudulent. Notably, this dataset includes data from 2012 to 2014 [12, 200–214].

The other three studies developed custom crawlers to collect data from various job posting sites in the UK (SEEK, Glassdoor, Indeed, and Gumtree) [215], in Bangladesh (job.com.bd, bdjobstoday, deshijob) [216], and in China (China-Boss, Zhipin, Liepin, and 51job) [217]

Data Used for Fake Account Detection. Some studies attempted to tackle the automated detection of fake profiles online, and Table 6 shows the data extracted from them.

We found that many authors collected user profile data from online social networks including Twitter [218–224], Instagram [225–227], Facebook [228–230], YouTube [231], and Sina Weibo [232]. A different approach used in one study [233] was to collect real names from various web pages, schools, and other sources to detect fake names online automatically.

Other authors have used previously published and openly accessible datasets that included user data from various social networks [234, 235].

Data Used for GenAI Social Engineering Attack Detection. Under this category, we found many works that investigated how generative AI models can be misused to defraud people.

Some studies [45, 236, 237] develop and discuss an initial taxonomy for which they discuss how scammers can misuse AI-generated content. At the same time, Carlini et al. [238] test various membership inference attacks – which is when someone attempts to figure out whether a specific piece of data was used to train a machine learning model – on OpenAI's GPT2 model and confirm that the model is vulnerable to this kind of attack which poses risks to privacy. Similarly, Kumar et al. [239] discuss the significant implications for cybersecurity, privacy, and ethical considerations that should be considered when developing and using these models.

Apropos misuse cases of these models, Ayoobi et al. [240] discuss how LLMs and GenAI can be used to create fake professional profile bios to trick users into believing that the account is legitimate. Similarly, DiResta and Goldstein [241] show that scammers can use these models to create AI-generated images to be posted on social networks. Their case study shows that these images tend to receive high volumes of engagement on Facebook as many users do not seem to recognize that the images are synthetic. Other research shows how these models can be *jailbroken*[26] to produce code to imitate legitimate webpages (phishing URLs) [242], malware code, phishing emails, phishing SMSs, SQL injection attacks, and other potentially malicious material [243–245].

Other research suggests that humans may be able to accurately detect phishing AI-generated content [246], while Roy et al. [247] discuss and experiment with countermeasures to prevent malicious prompts (jailbreaking) for GPT and provide insights into how the model can become more robust against this vulnerability.

Data Used for Social Media Scam Detection. Six studies examined various scams and spammers facilitated by Social Networks. Xu et al. [248] used data from WeChat (a Chinese messaging, social media, and mobile payment app) and Konect repository to detect users that use WeChat to defraud people. La Morgia

[21] https://www.scamalert.sg/

[22] http://odds.cs.stonybrook.edu/yelpzip-dataset/

[23] https://www.kaggle.com/discussions/general/281540

[24] https://snap.stanford.edu/data/web-Amazon.html

[25] https://www.kaggle.com/datasets/amruthjithrajvr/recruitment-scam

[26] Jailbreaking a generative AI model means bypassing its safety rules or restrictions to make it produce responses it's not supposed to.

et al. (2023) [249] and La Morgia et al. (2021) [250] used Telegram data to characterize and detect Fake Telegram channels, while Shah et al. [251] collected data from Telegram and compared it to Twitter data to understand and detect fake users. Similarly, Al-Hassan et al. [252] collected and analysed Twitter and Institute of Informatics and Telematics data to detect scammers on Twitter. Finally, Tripathi et al. [253] collected YouTube data to detect scammers attempting to lure victims using comments posted alongside YouTube videos.

Data Used for Romance Fraud Detection. In their study, He et al. [254] attempted to automatically detect malicious accounts on Momo,[27] a dating website. Similarly, Suarez-Tangil et al. [255] collected data from datingnmore.com and `scamdigger.com` to develop automated methods to understand fraudulent profiles within dating social networks. Lastly, Lokanan [256] analysed the sentiment of tweets with the hashtag #tinderswindler to provide an understanding of users sharing their experiences regarding romance fraud.

Data Used for GenAI Scam Baiting. We identified three studies. [257–259] that used LLMs and Generative AI to automatically engage with scammers online to waste their resources and collect data on various fraud activities. For these studies, the researchers collected data from their own baiting accounts and emails and said data was not made publicly available.

Data Used for Fraudulent Investment Detection. Studies that attempted to understand fraudulent investment scams employed various datasets and methodologies. First, Siu et al. [260] analyse investment scam advertisements found in Bitcointalk.[28] Li et al. collected YouTube comments to detect bots that advertise fraudulent investment content automatically [261]. Lastly, Kuo and Tsang [262] develop a scam detection model based on emotional fluctuations of user discussions collected from one of Taiwan's most popular instant messaging applications.

Data Used for Fraudulent Crowdfunding Detection. Two studies were identified that analyse fraudulent crowdfunding [263, 264]. These collected the descriptions and metadata from hundreds of Kickstarter campaigns.[29]

Data Used for Crypto Manipulation Detection. Market, and more specifically, cryptocurrency coin manipulation, is when users collectively attempt to alter investor interactions towards manipulating the price of a coin.

Nizzoli et al. [265] discuss this process via data acquired from Twitter, Telegram, and Discord channels. Similarly, Mirtaheri et al. [266] identify and analyse cryptocurrency manipulations from user activity collected from Telegram and Twitter.

Data Used for Fraudulent E-commerce Detection. The only study for this category [126] analysed the terms and conditions of websites that sell (a variety of) products to inform understanding and the detection of obscured financial obligations in online agreements.

5.2 Methodologies employed

We now discuss the most popular AI and NLP methodologies employed to study each type of online fraud.

Methods Applied for Phishing URL Detection. The studies included in our SLR attempted to automatically detect phishing URLs using a variety of NLP and AI methodologies. These included classic supervised machine learning algorithms such as Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), Support Vector Machines (SVM), XGBoost (Extreme Gradient Boosting), KNN (k-Nearest Neighbors), as well as Artificial Neural Networks (ANN) and more advanced deep learning approaches such as Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN). LSTMs are Recurrent Neural Networks (RNN) designed to capture long-dependencies in sequential data, making them suitable for handling and predicting text sequences. On the other hand, CNNs aim to identify key features in the text by capturing local patterns. In the research reviewed, these models were developed to complete the binary classification task of determining whether a URL was fraudulent or not.

NLP techniques related to text mining have been used to extract features from URLs, which are then used as features to train an AI-based classification model. For example, such features include the counts of the characters, special characters, and n-grams of the URLs. Also, the authors collected other URL features, such as whether the URL had a secure scheme or not (e.g. https), domain (e.g. amazon.co.uk) and top-level domain (e.g. /kitchen), and sub-directories (e.g. /appliances). All the above and more features were used to fit and train a malicious URL detection model. Some studies used hybrid or a combination of methodologies, including more advanced techniques. For example, Li et al. [46] used a Bidirectional Long Short-Term Memory (Bi-LSTM) recurrent neural network, that could process sequences of text in both

[27] `https://www.immomo.com/aboutus.html`

[28] `https://bitcointalk.org/`

[29] `https://www.kickstarter.com/`

forward and backward directions, along with a Visual Geometry Group (VGG) which is a type of CNN. Vo Quang et al. [47] used a Convolutional Neural Network (CNN), along with features extracted using Word2Vec (W2V), a Gated Recurrent Unit (GRU), which is a more simplified type of RNN than LSTM, and a Bi-LSTM. Nakano et al. [98] used BERT with RF; Bitaab et al. [97] developed a hybrid system that uses RF, SVM, FNN, and XGBoost; Alswailem et al. [53] used Linear Regression (LR) and DT; and, Villanueva et al. [92] used LSTM and GRU.

Other stand-alone AI methodologies, like LightGBM,[30] RF, NB, and ANN, also seem to work well on detecting phishing URLs, but the approach with the best performance seems to be RF.

Methods Applied for Phishing Email Detection. The methodologies used for phishing email recognition focus more on NLP analysis. The majority of studies used various NLP methodologies for feature extraction, including, but not limited to topic modeling (LDA, BERT, BERTLARGE), text representation (TF-IDF, BOW, Clustering, W2V), and sentiment analysis (VADER, WordNet).

Studies that also aimed to automate the detection of phishing emails employed LLM analysis, RF, NB, SVM, CatBoost, LSTM, RNN, and many more.

Methods Applied for Phishing SMS Detection. Like phishing email detection, phishing SMS detection relies on state-of-the-art NLP methodologies, including LLMs, LDA, BERT and W2V. The existing literature also used AI methodologies like LR, SVM, CNN, GNN, LSTM, NB, and KNN for automated detection.

Methods Applied for Phishing Phone Call Detection. Studies on vishing detection used various AI methods for automated detection. Most used transcript text data for their analysis, except for Djiré et al. [161] who analysed deepfake voice analysis. In that study, the authors found that RNNs performed best.

Overall, various NLP and AI techniques were used on text data, including but not limited to SVM, NB, LSTM, CNN, RF, BERT, W2V, LR, and KNN.

Methods Applied for Phishing (User Reports) Detection. We found that the use of BERT, Sequential Minimal Optimization (SMO), J48 (an implementation of decision tree), NB, RF, XGBoost, Doc2Vec (D2v - an extension of Word2Vec), Jaccard, NER (Named Entity Recognition), and TF-IDF was applied to analyse user reports of various phishing activities.

Methods Applied for Fake Review Detection. The most popular NLP technique for fake review detection rely heavily on sentiment detection techniques, including VADER and WordNet. Similar to previous fraud analyses, the AI methods applied included various neural network models, such as CNN.

Methods Applied for Recruitment Fraud Detection. The techniques employed for the automated detection of Fraudulent Job Postings included standalone machine learning algorithms used for classification tasks, including LR, SVM, KNN, RF, XBoost, and ANN, and deep learning models, including Bi-LSTM.

Methods Applied for Fake Account Detection. The studies included in our SLR leveraged various NLP and AI-based techniques to detect fake accounts on social media platforms such as Twitter, Facebook, Instagram, and YouTube. Classic supervised learning approaches, including NB, RF, DT, SVM, LR, KNN, and ANN, were widely adopted. Ensemble learning methods such as AdaBoost and stacking models were also explored, along with advanced methods like Gradient Boosting techniques (e.g., CatBoost). Notably, RF was often found to deliver the best performance in several studies, e.g., [225, 227, 234].

Deep learning approaches were also employed, particularly when tackling larger datasets. For example, Na et al. [231] used RoBERTa to detect fake accounts involved in scam campaigns on YouTube, while Alhosseini et al. [222] leveraged Graph Convolutional Neural Networks (GCNN) for spam bot detection on Twitter. Studies such as [229] adopted unsupervised learning techniques like HDBSCAN for anomaly detection in social networks.

Across the studies reviewed, researchers have extracted diverse features, including user profile characteristics (e.g., number of followers, account age), content-based features (e.g., hashtags, posts), domain-based features, and behavioural patterns.

Deep learning models such as Bi-LSTM, GRU, and CNN were less frequently applied but showed promise. Fathima et al. [226] developed an ANN-based system to categorise fake profiles on Instagram. Ensemble learning approaches, such as combining ANN, SVM, and RF [230], were also utilised to enhance detection performance.

The performance of these methods varied depending on the dataset and feature selection, but overall, RF emerged as the most consistent and accurate classifier for fake account detection across multiple platforms and studies.

Methods Applied for GenAI Social Engineering Attack Detection. GPT-3.5 and GPT-4 were

[30] Light Gradient Boosting Machine - an ensemble learning technique designed for handling large datasets with large features

the most commonly used models, particularly in generating phishing emails, malicious websites, and smishing campaigns. Studies by Roy et al. [247] and Shibli et al. [243] demonstrated how these models could be exploited to craft highly convincing phishing content, leveraging sophisticated language capabilities. Ayoobi et al. [240], utilised models like BERT, RoBERTa, and Flair to detect fake LinkedIn profiles generated by ChatGPT. Defensive mechanisms were also explored; for instance, Roy et al. [247] proposed BERT-based countermeasures to mitigate malicious prompt exploitation. Despite the promising results in detecting and preventing misuse, other research [244] [245] has highlighted vulnerabilities in GPT-3.5, particularly its susceptibility to jailbreak attempts, which enable the generation of harmful content such as SQL injections, malware, and phishing scams. Overall, LLMs demonstrated advanced capabilities for deception, and their susceptibility to misuse necessitates robust detection and prevention strategies. Studies under this category did not report performance metrics as they tested the limits of LLMs and GenAI models using qualitative approaches.

Methods Applied for Social Media Scam Detection. Xu et al. [248] proposed the BREAD framework, which uses bidirectional k-hop reachability query processing over dynamic graphs to extract fraud-related features. La Morgia et al. (2023) [249] studied fake Telegram channels employing a Multilayer Perceptron (MLP) model. Shah et al. [251] applied techniques like W2V, D2V, P2V, and TF-IDF to detect illicit activity. Tripathi et al. [253] examined monetised scam videos on YouTube using RF and W2V. Al-Hassan et al. [252] developed DSpamOnto, an ontology-based model for social spammers on Twitter, and benchmarked it against classifiers such as NB, SVM, and RF. Finally, La Morgia et al. (2021) [250] used LDA for topic modelling.

Due to the diversity in the types of fraud analysed across social media platforms and the distinct datasets and methodologies employed, identifying a single best-performing model for this category is not applicable.

Methods Applied for Romance Fraud Detection. Studies of Romance Fraud detection used various NLP methodologies, including sentiment detection, which uses BOW and textBlob, and statistical methods like TF-IDF, for feature extraction. When testing different models, researchers have found that Random Forest (RF) performed best for the detection of this offence [256]. He et al. [254] found that LSTM is most effective at detecting malicious accounts in dating applications, while another study [255] showed that Ensemble Machine Learning (EML) also works well.

Methods Applied for GenAI Scam Baiting. One of the Scam Baiting studies [257] used OpenAI's ChatGPT to reply to scammer emails. Similarly, Bajaj and Edwards [258] experimented with OpenAI's ChatGPT and DistillBERT to categorize scam emails they received and provided a qualitative analysis of how well the two models performed. Chen et al. [259] set up an email server as a "honeypot" from which they sent emails to scammers (to encourage those scammers to send phishing emails to them) identified in data from the Scambaiter mail server, Enron Email Dataset, and ScamLetters.Info. They then employed their own semi-unsupervised DistillBERT model to engage with scammers automatically and used their model filtering to categorise and analyse the emails received.

Methods Applied for Fraudulent Investment Detection. Three of the studies related to Fraudulent Investment used data from different sources (forums, messaging apps, and YouTube). Two of the studies [260, 262] reviewed under this SLR used models to detect emotional fluctuations in discussions between victims and found that DT was the best-performing machine learning model for this task. Siu et al. [260] also concluded that XGBoost performed well in terms of detecting malicious advertisements for fraudulent investment websites.

Methods Applied for Fraudulent Crowdfunding Detection. One of the papers on fraudulent crowdfunding detection applied NLP methodologies, including Named Entity Recognition and other NLP features detected in the descriptions of Kickstarter campaigns, and built an LR model that performed well [263]. The other study [264] developed an LSTM-LDA topic detection model that analyses the crowdfunding campaign and people's comments with the aim of estimating whether a campaign was a scam.

Methods Applied for Crypto Manipulation Detection. One study on cryptocurrency market manipulation used pre-existing methods for detecting fake users, along with CorEx Topic Analysis [265]. The other study found that SVM with SGD and TF-IDF worked best for detecting discussions that aimed to manipulate the market [266].

Methods Applied for Fraudulent E-Commerce Detection. Finally, the only study that we identified that used text data to inform understanding of Fraudulent E-commerce activities, used OpenAI's GPT-4 model to automatically detect obscure financial obligations in the terms and conditions of the websites sampled [126].

5.3 Key Findings

We now summarise our findings in relation to the research questions listed in Section 2. To remind the reader, these questions were: to detect the state-of-the-art AI techniques used to detect online fraud (RQ1); the data sources used (RQ2); how researchers evaluate their AI models (RQ3), and what the most studied fraud activities were (RQ4). The answers to RQs 1-3 are organised by fraud type, while the listing of these fraud crime types addresses RQ4.

Takeaways: Phishing URLs. While established datasets have played a major role in developing phishing URL detection models, there is a clear need to incorporate more dynamic and current data sources. Leveraging user-reported phishing URLs from social networks and data from telecom and security organizations would offer a more effective approach to combating phishing attacks. These sources provide real-time, diverse, and relevant data that enhance the robustness and accuracy of detection models, keeping pace with the evolving nature of phishing threats. By combining the strengths of both traditional and modern data sources, researchers can develop more comprehensive and adaptive phishing detection systems, better protecting users from phishing URLs.

Regarding the methodologies used, we find that the existing literature used state-of-the-art methodologies to analyse and detect phishing URLs. Of these, RF seems to be the stand-alone model that works best, while other hybrid methodologies also report promising performance. Regarding performance reporting, authors often fail to adequately report all of the performance metrics of their model. Although the Accuracy of the model is reported in all but seven studies, other metrics like Precision, Recall, F1, and AUC are omitted in more than half of the studies we analysed for this type of online fraud.

Takeaways: Phishing Emails. While publicly available datasets have laid the groundwork for phishing email detection research, the rapidly evolving nature of phishing attacks requires the use of more dynamic and up-to-date data sources. Leveraging user-reported emails, real-time spam collections, and advanced synthetic data generation techniques could significantly enhance the robustness and accuracy of phishing detection models. By combining traditional datasets with innovative data sources, researchers could develop more comprehensive and adaptive phishing detection systems that are better equipped to detect phishing activities via email.

All of the studies that developed automated approaches to phishing email detection reported very good performance, with RF, BERT, LSTM, RNN, and SVM being the most popular. Similar to the Phishing URLs above, accuracy was the metric reported most often. Only two studies reported AUC, and Precision, Recall, and F1 were rarely reported.

Takeaways: Phishing SMS. This type of scam was also found to rely on existing datasets. Alas, the rapidly evolving nature of smishing requires a more dynamic and diversified approach to data collection. Researchers could develop more effective and resilient smishing detection models by integrating publicly available datasets with real-time, user-reported data and specialized security sources. This approach would ensure that models remain relevant and capable of addressing new and sophisticated smishing threats as they arise.

Contrary to phishing email detection, we find that in the case of phishing SMS detection, which tends to involve much less text, SVM and various applications of Neural Networks performed best. Again, the Precision, Recall, F1-score, and AUC metrics were underreported, with Accuracy being the metric most studies report.

Takeaways: Phishing Phone Calls. Regarding the data sources used for automated vishing detection, most studies used text data obtained by transcribing voice recordings. Other research also used caller ID information from various telecommunication operators. Some researchers collected data from user reports, and only one attempted to detect deepfake signals in voice recordings. The best-performing technique used for automated vishing detection was SVM. Although Accuracy was also the most reported performance metric for this category, many studies failed to provide any metrics.

Takeaways: Phishing (User Reports). Reviewing the four studies that focused on phishing via user reports, we found that the researchers used data from various sources, including court judgements, public data from forums, and user reports from Financial Institutions. The applied methodologies varied and included Named Entity Recognition, various NLP and statistical techniques (D2V, Jaccard, TF-IDF), and ML techniques (SMO, J48, NB, RF, XGBoost). Only one study [174] provided performance metrics, and this was for their BERT model that performed best.

Takeaways: Fake Reviews. Although many studies used previously available datasets to establish and test their detection models, we noticed a clear trend where more recent studies tended to collect data from platforms like Amazon, Google Play, the Apple App Store, and YouTube. This is very encouraging as the data used for these detection models need to be constantly updated. Hybrid models, including LR, SVM, CNN, and LSTM, seemed to perform best in the detection of fake reviews. Again, the Precision, Recall, F1, and AUC metrics were underreported, with Accuracy being the metric most studies report.

Takeaways: Recruitment Fraud. While the Kaggle dataset has been pivotal in advancing research on fraudulent job postings, the rapidly evolving nature of job scams necessitates the use of more current and diverse data sources. Custom data collection methods, which tap into active job posting sites, represent a critical step forward in enhancing the effectiveness of detection models. Researchers can develop more comprehensive systems to effectively combat fraudulent job postings by leveraging a mix of established and new data sources. The models that performed best for this fraudulent activity varied. However, we find that Bi-LSTM, KNN, RF, DNN, and LightGBM performed well. All but one study reported the accuracy performance metric of their best-performing model, while the other four metrics (Precision, Recall, F1, and AUC) remain underreported.

Takeaways: Fake Accounts. Most studies utilised user profile data from popular social networks such as Twitter, Instagram, Facebook, YouTube, and Sina Weibo. Several works relied on openly accessible datasets published from previous studies, and others collect user data through APIs, web scraping, or manual curation. Advanced techniques such as RoBERTa, CatBoost, and Graph Convolutional Neural Networks (GCNN) have been employed alongside traditional classifiers like Random Forest, Support Vector Machines, and Neural Networks, with Random Forest being one of the most popular and frequently high-performing models.

Performance across studies is generally strong, with reported accuracy often exceeding 90%, though other metrics like Precision, Recall, and F1 are underreported in some works. However, the rapidly changing strategies used by fake account creators highlight the need for more dynamic datasets and adaptive models to tackle this challenge effectively.

Takeaways: GenAI Social Engineering Attacks. Many studies discuss how GenAI models might affect cybersecurity and privacy, the ethical issues they pose, and how they could be misused to create fraudulent content automatically. However, only two studies identified in our review used data collected from real use cases. These were Ayoobi et al. [240], who discussed fake profile AI-generated content on professional social networks, and Diresta et al. [241], who examined deepfakes posted on Facebook. No performance metrics were reported in these studies as they were not applicable. Authors studying this offence were not building models but rather evaluating or experimenting with existing tools, including OpenAI's ChatGPT.

Takeaways: Social Media Scams. The detection of social media scams presents unique challenges due to the diversity of platforms, fraud types, and datasets. Most studies leverage platform-specific data sources such as Telegram [249–251], WeChat [248], Twitter [251, 252], and YouTube [253] to build detection models.

The methodologies employed in these studies vary widely, encompassing advanced NLP techniques (W2V, P2V, and LDA) [250, 251], as well as machine learning classifiers (RF, NB, and SVM) [252, 253].

While these studies report promising results, this category involved studies on various kinds of social media scams. Hence, the lack of standardisation across models and datasets limits the generalisability of these findings. Notably, these studies also often underreported evaluation metrics like AUC and F1.

Takeaways: Romance Fraud. Romance fraud detection has primarily focused on analysing user-generated content on dating platforms and social media. Studies utilised diverse datasets, including user profiles from dating platforms [254, 255] and tweets tagged with #tinderswindler [256]. Methods employed feature extraction techniques like BOW, TF-IDF, and sentiment analysis with textBlob [256]. Among machine learning models, RF consistently performed well [256], while LSTM achieved the best performance in detecting malicious accounts in dating apps [254], and EML yielded high accuracy in identifying fraudulent profiles [255]. Due to the limited number of studies under this category, we cannot make conclusions regarding the models' overall performance reporting and generalisability.

Takeaways: GenAI Scam Baiting. The use of GenAI for scam baiting has shown promising potential in wasting scammers' resources and collecting data for fraud analysis. Studies employed LLMs such as ChatGPT [257, 258] and semi-unsupervised DistillBERT [258, 259] to engage with scammers and categorise phishing emails. These studies highlighted the potential of GenAI tools in automating scam baiting, but future work should focus on creating publicly available datasets and refining engagement strategies to improve scalability and efficacy.

Takeaways: Fraudulent Investment Detection. The detection of fraudulent investment scams has been explored using a variety of data sources, including Bitcointalk [260], YouTube [261], and instant messaging apps [262]. However, the small sample of just three studies limits the ability to conclusively identify the best-performing model for this category. The studies reported varied performance metrics, with Siu et al. [260] highlighting XGBoost as the top performer for detecting fraudulent investment advertisements on Bitcointalk. Kuo and Tsang [262] found DT to be the best model for identifying emotional fluctuations in scam discussions on a popular Taiwanese messaging app, and Li et al. [261] focused on arbitrage bot scams and utilised NNs for their analysis, without reporting performance metrics. As the three studies under this category analysed different aspects of a fraudulent investment, we cannot draw any conclusions regarding the best-performing model.

Takeaways: Fraudulent Crowdfunding. Fraudulent crowdfunding detection has been explored using a small set of two studies, both focusing on Kickstarter campaigns [263, 264]. LR, in combination with many NLP features, was employed to identify fraudulent campaigns, achieving an accuracy of 87.3%. A combination of LSTM and LDA based on user comments was also reported, without reporting performance metrics. Given the limited scope of these studies, further research is required to assess the robustness of these models in detecting fraudulent crowdfunding activities. Both studies highlight the effectiveness of NLP-based approaches for feature extraction and classification in this domain.

Takeaways: Crypto Manipulation. The detection of cryptocurrency market manipulation was investigated in two studies, which used data collected from Twitter, Telegram, and Discord [265, 266]. Methodologies used included network analysis, topic analysis, and SVM. The limited number of studies made it difficult to assess the robustness of the models, and further research is needed to identify the best-performing AI technique for this type of fraud.

Takeaways: Fraudulent E-Commerce. The only study identified for fraudulent e-commerce detection [126] used OpenAI's GPT-4 model to analyze the terms and conditions of e-commerce websites and detect obscure financial obligations, such as shipping, subscription, and refund fees. However, the study did not provide performance metrics or compare different models. Due to the limited nature of this single study, it is impossible to draw conclusions about the approach's effectiveness or the model's general applicability in detecting fraudulent e-commerce websites.

6 Discussion

We now discuss our findings, discussing recognised limitations and shortcomings identified in the reporting of AI models related to the performance and data sources used. We also provide recommendations for researchers developing detection models for online fraud.

6.1 Data Sources

Overall, we analysed the data sources used and the detection methodologies employed in 223 papers that aimed to address a range of online fraud problems. Although our findings reveal a preference for well-established datasets, especially in the automated detection of various phishing and fake reviews detection, more recent studies (published after 2023) seem to shift towards more dynamic and recent sources.

We found that the overwhelming majority of studies concerned with the detection of phishing relied on well-known and extensively studied datasets from websites like PhishTank, OpenPhish, and SpamHaus. In contrast, others used datasets made available from previous studies or publicly available repositories like Kaggle, the University of California Irvine (UCI) Machine Learning Repository, and GitHub. This was also the case for studies that focused on phishing email detection, fake review detection, and fraudulent recruitment detection. While these established datasets provide a valuable foundation for research, they come with limitations.

Online fraud is dynamic, with new scam techniques continually evolving, or building over older scam methods. Studies show that LLM-empowered bots, or scammers could be deployed to generate and automate sophisticated and targeted fraudulent and phishing content online, either in the form of email, a professional profile, or deceptive terms and conditions for fake e-commerce websites [45, 236, 237]. Hence, relying on outdated datasets may limit the effectiveness of detection models when applied to current or evolving threats. The historical datasets used do not capture the latest trends and variations in the different online fraud techniques and activities we see today (and will see tomorrow).

Recent studies have started to leverage more dynamic and real-time data sources to address these limitations. Regarding automated phishing detection, recent studies used user-reported phishing URLs from social networks and data from telecom and security organizations. For instance, studies published in 2023 utilized data from sources like Twitter, Facebook, and specialized security organisations [76, 76, 77, 81, 97, 97–100].

At the same time, recent research on automated phishing email detection has utilised user-reported emails, providing a real-time perspective on phishing threats [136]. For example, Genc and Jiang[131, 132] used emails from their personal or professional spam folders, capturing a more realistic and up-to-date snapshot of phishing attacks. Mehdi et al.[133] took an innovative approach by incorporating various techniques to develop their dataset; GPT-2 generated synthetic phishing emails and tools like TextAttack, TextFooler and PWWS. This approach provides a diverse dataset and ensures the model is robust against sophisticated phishing techniques. In their study, Janez et al. [117] used data from the SPAM Archive[31], which is a continuously updated repository of spam emails.

Similarly, several studies concerned with automated phishing SMS detection have extended the datasets used by combining pre-existing publicly available datasets with additional sources to improve the robustness and generalizability of their models. This has included the use of additional data from Fake Base Stations [154], emails, YouTube comments [140], user-reported data for research [155] or content posted on Twitter [156]. In particular, Timko et al. [159] proposed a platform, SmishTank, where users can post phishing SMS messages, creating an ongoing and up-to-date repository for researchers. We also observed studies using data from specialised security agencies [157] and mobile security services [158], which offer a more targeted collection of smishing examples.

On Fake Review detection, one study [198] used YouTube transcripts to interpret false review exaggeration, showcasing an innovative approach to identifying fraudulent content in multimedia contexts. Others have adopted similar techniques and developed their own data collection methodologies to collect reviews from sources like App stores, e-commerce websites, and location and travel research platforms.

An overwhelming number of studies focused on fraudulent recruitment detection using the same dataset from Kaggle.[32] Only three studies employed custom crawlers to gather data from various job posting websites in different countries, providing a more diverse and up-to-date perspective. Mahbub et al. [215] collected data from job posting sites in the UK, Tabassum et al. [216] gathered job postings from Bangladeshi sites, and Zhang et al. [217] sourced data from Chinese job sites. The use of up-to-date data collection from these sources offers some advantages. For one, these studies can capture the most recent and relevant data by scraping data from active job posting sites, reflecting current fraudulent practices. Second, collecting data from multiple sources across different regions provides a richer and more varied dataset, which can enhance the robustness and generalizability of detection models. Finally, custom datasets often include a wider variety of job postings, including niche or less common types of employment scams, which can be critical for developing more comprehensive detection systems.

Overall, combining publicly available datasets with *recent* data from other sources, such as social media, user reports, and specialized agencies, may significantly enhance the robustness and relevance of detection models. This approach could ensure that models are exposed to a wider variety of fraud tactics and can adapt to new threats more effectively.

There are various advantages of using such dynamic data sources:

- Real-time Updates: Social networks and security organizations provide continuously updated data. This ensures that the detection models are trained on the most recent phishing URLs, making them more robust against new and emerging threats.
- Diverse Data: User-reported data from social networks and institutions often include a wide variety of phishing techniques and strategies. This diversity enhances the model's ability to generalize and detect a broader range of phishing attacks.
- Early Detection: These sources can help in the early detection of new phishing campaigns. Social networks, in particular, can act as early warning systems where new phishing URLs are often first reported.
- Enhanced Relevance: Data from telecom and security organizations are often more relevant to current threats and can include targeted phishing attacks that are not present in older datasets.

However, despite the advancements in data collection methods, there are still gaps. For instance, the data sources used in studies for several types of online fraud were not clearly stated within the manuscripts. This lack of transparency can hinder reproducibility and the ability to compare results across different studies.

6.2 Methodologies

The methodologies employed across various studies of phishing and fraudulent activities involved a wide range of AI that incorporate NLP, machine learning and deep learning techniques. Most of these techniques involved the extraction of features using NLP techniques and then applying supervised machine learning (i.e. models that use labelled data) and deep learning algorithms to build binary classifiers.

For phishing URLs, Machine Learning and Deep learning algorithms such as CNN, ANN, KNN, LSTM, NB, RF, DT, SVM, and XGBoost were commonly used, often with URL feature extraction through NLP methods like character counts and n-grams. We also found various studies that applied hybrid approaches combining multiple techniques, demonstrating strong detection capabilities. Similarly, phishing email detection heavily relied on NLP for feature extraction (e.g.,

[31] http://untroubled.org/spam/

[32] https://www.kaggle.com/datasets/amruthjithrajvr/recruitment-scam

LDA, BERT) followed by AI models like RF, NB, and SVM for classification. Similar techniques were applied for phishing SMS detection.

For vishing (phishing phone calls), transcript analysis was primarily conducted using NLP and AI models, with some studies examining deepfake voice detection. The analysis of user reports about phishing activities utilized models like BERT and RF, while studies of fake reviews often involved sentiment analysis using methods like VADER.

Fraudulent job posting detection has involved both the use of machine learning and deep learning models such as LR and Bi-LSTM, while romance fraud detection has used sentiment detection methods in combination with machine learning models (e.g. RF) and deep learning models (e.g. LSTM). Classic machine learning models like DT, XGBoost, and SVM were employed for fraudulent investment and crypto manipulation. Studies on fraudulent e-commerce and crowdfunding leveraged advanced NLP and machine learning techniques, including GPT-4 and LR, respectively.

Although the research and detection methodologies applied in the reviewed literature performed well, they are not without limitations. Overall, popular machine learning algorithms like RF and SVM often rely heavily on the quality of extracted features, which can be labour-intensive to generate and may miss out on subtle indicators when dealing with large data sets.

In addition, complex models based on deep learning techniques like Bi-LSTM with VGG or hybrid approaches can be computationally expensive and difficult to implement in real-time systems due to the large amount of data processing resources that they require.

Notably, models developed that work well for one kind of fraud might not be generalised well to other fraud activities. For example, unlike emails, SMS messages are typically short, providing limited data for accurate feature extraction and classification. In addition, natural language processing techniques used may struggle to capture those features necessary to understand the context (semantics) or syntax of phishing content, leading to potential false positives/negatives (i.e., misclassifications).

Models trained on specific languages or datasets may not perform well on emails in other languages or different styles. Many of the challenges that may arise regarding the trained AI models are often the result of poor data quality. More specifically, effective feature extraction is critical but can be difficult due to the varied nature of how natural language is used. At the same time, the textual content used in online fraud activities - such as fraudulent emails, SMSs, or job postings - keeps evolving, making it challenging for static models to remain effective over time.

Although collecting data from various websites, forums, social networks, and telecommunication operators for online fraud detection is invaluable, at the same time, different platforms hold data inconsistently or have unique features and user behaviours, complicating model generalization. Also, identifying fraudulent activities in real-time is challenging due to the dynamic nature of these scams and the lack of real-time occurrences in the various data sources.

In short, while these methodologies offer powerful tools for detecting phishing and fraudulent activities, there are challenges related to feature extraction, model complexity, generalizability, and data quality. Finally, one of the major issues highlighted in many of these studies was that the models use supervised machine learning models, which require labelled data. Creating labelled data is often challenging and time-consuming. This is one reason why so many researchers use existing labelled data, but, as discussed, such data will become less useful as fraud evolves over time.

6.3 Recommendations

Datasets. The reliance on older, established datasets for training AI-based models is a double-edged sword. While they offer a solid foundation for model development and facilitate the comparative analysis of different models, their static nature may limit their effectiveness in detecting evolving or emerging fraud trends, hence limiting their effectiveness in detecting new types of fraud in the real world. Therefore, a strong case exists for incorporating more dynamic and diverse data sources. Recent studies that use custom crawlers to gather data from various online platforms that focus on a range of fraud types exemplify best practices in this area. These approaches provide real-time, relevant data that can significantly improve the adaptability and accuracy of detection models. Going forward, it is recommended that researchers consider combining established datasets with freshly collected data to create more robust and resilient AI models.

Methodologies Used. Overall, most studies reviewed used stand-alone machine learning and deep learning models to detect online fraud. In many cases, NLP techniques used for feature extraction were under-reported or ignored. Working on online fraud activities that involve textual data should utilise more sophisticated NLP techniques such as transformer-based models (e.g., BERT, GPT) for deeper semantic understanding and better context handling. Although LLMs have limitations, such as generating hallucinated or inconsistent results, they are extremely powerful for context extraction.

Recent research demonstrates the utility of hybrid models, which combine different AI and NLP techniques to leverage their strengths. These hybrid models appear to perform very well. Most existing studies have used supervised machine learning models that require labelled data to detect fraudulent activities. Due to the challenges of obtaining new labelled data, researchers often rely on existing datasets that may not capture the content of new techniques and tactics that scammers employ.

To address these challenges, further exploration of active learning, semi-supervised learning and anomaly-based models that rely on small amounts of labelled data or no labelled data is needed. For example, unsupervised or semi-supervised anomaly detection techniques could be studied to identify outliers and novel fraud patterns that may not be present in the training data. Finally, we observed that almost none of the models reported had real-time applications. There should be a shift in focus where researchers attempt to optimize models for real-time processing to ensure the timely detection and mitigation of fraudulent activities.

Model Performance Reporting. While assessing the literature reviewed in this SLR, we observed many studies that only reported a subset of model performance metrics, and authors frequently relied on the Accuracy performance metric alone. This was the case for all of the online fraud types identified and analysed herein. However, using accuracy on its own, especially when the dataset is unbalanced, can be misleading. In a dataset where one class has more observations than another (for example, having fewer phishing emails compared to not-phishing emails), a model could achieve very high accuracy simply by predicting the majority class (i.e. not-phishing emails) without doing a good job of detecting the phishing emails.

Overall, it is essential that researchers report a more comprehensive range of performance metrics beyond accuracy alone. These should include Precision, Recall, the F1-score, AUC score, or ROC curve. These metrics provide a more complete and nuanced picture of a model's performance, especially when dealing with imbalanced datasets. In addition, there is a need to conduct and report detailed error analysis to identify common failure cases and the reasons behind them. This can help understand the limitations of the model and areas for improvement. Finally, models need to be cross-validated to ensure the robustness of the reported performance metrics. For example, reporting results from multiple folds of data samples can provide a more reliable estimate of model performance.

Reproducibility. Many studies failed to explain key and critical aspects of their model development. This included the features engineered and selected, methods used for extracting features, the data used, the size of the dataset, the partition of the data into training-test sets, and the hyper-parameters used for tuning and training the models.

To this end, we recommend researchers provide access to the code, datasets, and pre-trained models used in their studies through platforms like GitHub, GitLab, or institutional repositories. This would help improve the reproducibility of their work. Researchers should also ensure that the methodology section of their paper is sufficiently detailed to allow others to replicate their model and critically assess it. This should include a clear description of pre-processing steps, feature engineering and selection, model hyper-parameters, the training-testing data split and training protocols. Furthermore, the use of standardized frameworks and libraries for model implementation (e.g., TensorFlow, PyTorch) could improve reproducibility. Comprehensive documentation and setup instructions will help others understand and reproduce the work more easily.

Usability. Most of the papers reviewed were proof-of-concept studies, so the usability of AI-based models has not been addressed. The effective application and use of AI-based approaches depend on successful usability studies that enable users to develop these models into toolkits and provide user feedback. Usability goals are generally determined by efficiency, effectiveness, engagement, error tolerance and ease of use. It is thus also imperative to ensure collaboration between the developers of AI-based tools and practitioners. However, while the field of technology usability assessment in front-line policing is growing [267, 268], there is a lack of usability studies considering the use of AI in preventive policing, including its application to cybercrimes like online fraud.

Bias. The majority of studies do not discuss the limitations of their models or data. Researchers should clearly identify and report such limitations, including any assumptions made, potential biases in the data, and methodologies' limitations. The overuse of existing labelled datasets could impact the performance of the models, potentially leading to issues such as overfitting. However, issues like this were not discussed in most of the studies reviewed here. Researchers should examine the distribution of different classes and any potential sources of bias in the data used for training and testing. Lastly, there is limited discussion on the generalisability of the models across different datasets, contexts, and evolving fraud patterns. Hence, researchers should conduct sensitivity experiments to evaluate the generalisation performance of their models.

Limitations. As with every study, ours comes with potential limitations. One that stands out lies in the possibility of missed studies. Although we took great care in designing and refining our search string to capture as many relevant publications as possible, the diversity and rapid growth of literature in this field make it likely that some studies were inadvertently omitted. Furthermore, grey literature sources such as pre-prints may lack consistent metadata, which could have further limited our ability to identify all relevant studies.

Finally, the inherent subjectivity in labelling fraud types and selecting the primary focus of each paper poses a challenge. While efforts were made to standardise the categorisation process via discussions between all authors, bias or misclassification could have introduced slight inconsistencies into the analysis. In more detail, the challenges associated with classifying online fraud have been extensively discussed in various reports and academic papers. Rabitti et al. [15] and Eling et al. [269] explain that the field of cyber risks is rapidly expanding, and many taxonomy-based systems have been proposed. At the same time, the partial taxonomies produced by various bodies have resulted in various reports, often not in harmony with one another, which can lead to many frauds being misclassified or the introduction of grey areas and uncertainty. This lack of uniformity in the academic literature renders the identification of online fraud a challenge that is still to be addressed, calling for further research. Similarly, Cohen et al. [34] highlight the lack of consistency across said taxonomies and models, emphasizing the need to understand this risk better.

Moving on to literature and taxonomies produced outside academia, a review from the UK's National Fraud Authority [8] explains that online frauds are diverse and can be differentiated further, highlighting the evolving nature of online fraud. That review demonstrates how diverse online fraud can be, taking many forms. It compares how different taxonomies and literature reviews use different umbrella terms to classify specific scams and online fraud. Lastly, a white paper by the UK Police Foundation [10] highlights that "Fraud is daunting in terms of its scale and variety". The report discusses in detail the various methods adopted by fraudsters, the exploited criminal opportunities, and the experiences of the victims. Notably, the author explains that the online fraud landscape is continuously evolving as fraud methodologies and fraudsters adapt to new technological, social, and commercial opportunities before explaining that online fraud is not defined concretely and is often underreported by victims. We acknowledge that while undertaking this systematic literature review, we also faced and confirmed the above challenges regarding the evolving nature and inherent issue of classifying online fraud, which likely affected how we classified the studies selected for qualitative analysis and, hence, the presentation of our analysis and final results.

7 Conclusion

In this systematic literature review, we have examined a wide range of studies focusing on the detection of various fraudulent activities using AI-based models and Natural Language Processing techniques. Our goal was to examine the current state-of-the-art models and techniques used for development and training, investigate the sources of data used, and assess how these models are evaluated. Due to resource limitations, we restricted the SLR data collection to 2019-2024. The studies we identified covered a wide range of fraudulent activities, but there was a particular focus on phishing attacks. However, there is growing interest in using more advanced, Generative AI content to create deceptive content and tools that can be used for scam-baiting.

Significant attention has been given to building classification models that could be used to detect fraudulent activities. In particular, hybrid models that combine advanced NLP techniques with deep learning, including LLMs, have been developed. However, there remains considerable room for improvement. The key AI-model development areas that require attention include performance reporting, reproducibility, and transparency. Providing detailed performance reporting will help us to compare and evaluate different models. Improving reproducibility is important and requires researchers to provide sufficient details about what they did and how they did it. Increasing transparency means providing clear information on how the AI-based models work and make decisions. This will help fraud practitioners to interpret and understand the models, and mitigate any biases in AI-based models.

Furthermore, most existing models rely heavily on labelled data and supervised machine-learning techniques. Future studies should give some attention to the application of unsupervised and semi-supervised machine learning for detecting fraud. Similarly, the data sources used for training these models are unsuitable for capturing the dynamic nature of fraud. Future studies should, therefore, investigate building AI-based models that can capture emerging fraudulent patterns and their usability in fraud prevention.

Addressing these gaps is crucial for creating more robust, reliable, fair, and ethical AI-based systems to detect fraud. By considering the recommended practices, the research community can help to better understand, prevent, detect, and mitigate fraudulent activities and reduce victimisation, ultimately contributing to a safer and more secure digital environment.

Appendix

Here we list tables, based on specific scam types, discussed in the body of this manuscript. For the complete data extracted from the academic works included in this review, please refer to the public repository.[33]

[33] https://osf.io/nrx7y/?view_only=ca1050d48c4c4a969817c6d5f677cb87

#	URL Source	Collection Method	Models Used	Best Model	Performance				
					P	R	A	F1	AUC
[53](2019)	phishtank.org*	Custom crawler	RF	RF			0.98		
[62](2019)	phishtank.org and Alexa*	UNK	J48, RF, SMO, LR, MLP, BN, SVM, AdaBoost	RF	0.99		0.99		
[64](2019)	phishtank.org and openphish.com*	Previous work [65]	BNET, NB, J48, LR, RF, MLP	RF			0.98		
[65](2019)	Alexa, phishtank.org, openphish.com, and commoncrawl.org*	UNK	RF, SVM, NB, C4.5, JRip, PART	RF			0.94		
[66](2019)	Alexa and phishtank.org*	UNK	SVM, KNN, DT, RF, GBDT, XGBoost, LGB, Hybrid (GBoost, XGBoost, and LightGBM)	Hybrid (GBoost, XGBoost, and LightGBM)			0.97		
[107](2019)	UNK	UNK	C4.5, AdaBoost, KNN, RF, SMO, NB	Hybrid (XCS/UNK)	0.98		0.98	0.98	0.99
[67](2019)	phishtank.org, Yandex Search API, and GitHub	Open dataset [270] and custom crawler	DT, AdaBoost, Kstar, KNN, RF, SMO, NB	DT	0.96		0.97	0.97	
[77](2019)	NetLab360 and Alexa*	UNK	LR, SVM, LSTM	LSTM	0.98		0.98		
[87](2020)	Kaggle	Open dataset [271]	NB, KNN, SVM, RF, Bagging, NN	Hybrid (NN, RF, and Bagging)	0.95	0.98	0.97	0.96	
[68](2020)	Alexa and phishtank.org*	Previous work [62]	DNN, LSTM, CNN	LSTM			0.99		
[69](2020)	Alexa, phishtank.org, Mendeley, openphish.com, and commoncrawl.org*	Open dataset [272] and custom crawler	NB, SVC, KNN	SVC, KNN					
[70](2020)	phishtank.org, URLHaus, Majestic, Kaggle	Open datasets [273–275] and custom crawler	SVM, RF	RF	0.98	0.97	0.99		
[94](2020)	Refer to open dataset	GitHub open dataset [276]	NB	NB	1	0.95		0.97	
[83](2020)	Kaggle*	UNK open dataset	MLP	MLP			0.93		
[48](2020)	UNK*	Previous works [277, 278]	RF, RNN, CNN	CNN			0.94		0.91
[90](2020)	UCI ML Repository	Open dataset [279]	RF, DT, ANN, KNN	RF			0.95		
[59](2020)	phishtank.org and Google Search*	UNK	SVM, DT, LR, RF, XGBoost, AdaBoost, ET	Hybrid (RF, XGBoost and ET)			0.98		
[104](2020)	phishtank.org*	UNK	KNN	KNN			0.98		
[106](2021)	Kaggle and Canadian Institute of Cybersecurity*	UNK	SVC, LR, KNN, NB, RF	KNN	0.98	0.98	0.98	0.98	
[73](2021)	scammer.info and urlscan.io*	Custom crawler	LGBM	LGBM	1	0.96	0.98	0.97	0.98
[109](2021)	UNK*	Previous work [280]	RF, DT, NB, LR	RF			0.83		
[102](2021)	UNK*	UNK	LR, DT, NB	LR, DT	1	1	1	1	
[105](2021)	Alexa and cryptoscamdb.org*	Custom crawler	NB, SVM, KNN, RF	RF	0.98	0.95	0.97	0.96	
[60](2021)	Alexa, phishtank.org, and Mendeley*	Custom crawler and open dataset [281]	XBoost, RF, SVM, KNN, ANN, LR, DT, NBB	ANN	0.96	0.97	0.97	0.97	
[63](2021)	phishtank.org, relbanks.com, and millersmiles.co.uk*	Custom crawler	ANFIS, NB, PART, J48, JRip	PART	0.98	0.99	0.99	0.99	0.98
[78](2022)	Google Rankings and whoscall.com	Custom crawler	RF, DNN	RF	1	0.99	0.99		

[46](2022)	Alexa and phishtank.org*	Custom crawler	Bi-LSTM, Hybrid (Bi-LSTM and CNN), Hybrid (Bi-LSTM and VGG)	Hybrid (Bi-LSTM and VGG)		0.96	0.96	0.96	
[74](2022)	who.is*	Custom crawler	BPNN, RBFN, SVM, NB, DT, RF, KNN	NB			0.96		
[101](2022)	UNK*	UNK	DT, RF	RF			0.8		
[75](2022)	Alexa, UCI, phishtank.org and Kaggle*	UNK open dataset	KNN, RF, DT, CBoost, LGBM, ABoost, VC	CBoost			0.98	0.98	
[51](2022)	phishtank.org and UCI ML Repository	Custom Crawler and open dataset [279]	Hybrid (CNN)	Hybrid (CNN)			0.97		
[93](2022)	Canadian Institute for Cybersecurity	Open dataset [282]	LSTM	LSTM	0.99	0.99	0.99		
[79](2022)	Kaggle*	UNK	RF, KNN, XGBoost	XGBoost			0.96	0.96	
[54](2022)	pishitank.org*	Custom crawler	RF,DT	RF			0.87		
[92](2022)	GitHub	Open dataset [270]	LR, NB, LSTM, GRU	LSTM or GRU			0.95		
[89](2022)	UCI ML Repository	Open datasets [279] and UNK	AdaBoost, CART, GBoost, MLP, SVM, RF, NB, SEM	SEM			0.98		
[56](2022)	phishtank.org and Alexa*	Custom crawler	DT, RF	RF			0.87		
[72](2022)	Farsight SIE [283], spamhaus.org, and surbl.org*	UNK	J48, RF	RF					
[58](2022)	UNK*	Previous work [280]	Hybrid (CNN and LSTM)	(CNN and LSTM)	0.98	0.99	0.99	0.99	0.99
[91](2022)	Mendeley and previous works	Mendeley open dataset [284] and previous works [285, 286]	Hybrid DLM, Stack model, URLNet	Hybrid DLM			0.93	0.93	
[61](2022)	phishtank.org and Alexa*	UNK	SVM, NB	Hybrid (UNK)	0.99	0.99		0.99	0.99
[71](2022)	Canadian Institute of Cybersecurity, phishtank.org, and Kaggle*	UNK	SVM, RF	RF	0.99	0.99	0.99		
[99](2022)	Twitch*	Twitch API	XGBoost, RF, NB	RF	0.93	0.93		0.93	
[47](2023)	phishtank.org and openphish.com	Custom crawler	CNN	SharkEyes (CNN, W2V, GRU, Bi-LSTM)	0.94	0.94	0.95	0.94	
[98](2023)	Tweets, spamhunter.io, and tweetfeed.live*	Twitter APi and custom crawler	Hybrid (BERT and RF)	Hybrid (BERT and RF)	0.96		0.95	0.95	
[76](2023)	Twitter and Meta's crowdtangle.com*	Twitter API and custom crawler	UNK	Pre-trained model [66]	0.96	0.97	0.97	0.96	
[80](2023)	Kaggle*	Data no longer available	RF, LR, KNN	RF	0.97	0.99		0.97	
[97](2023)	reddit.com/r/Scams/ and Paolo Alto Networks*	Custom crawler	RF, XGBoost, SVM, FFNN	BeyondPhish (RF and XGBoost and SVM and FFNN)			0.98		
[81](2023)	Kaggle and Canadian Institute of Cybersecurity	Open datasets [282, 287]	KNN, LR	KNN			0.9		
[82](2023)	Kaggle*	UNK open dataset	DT, KNN, RF, SVM	SVM	0.99	0.96	0.98	0.97	
[95](2023)	Mendeley	Open dataset [272]	RF, J48, NB, KNN, LR	RF	0.97	0.9	0.94		
[84](2023)	Kaggle	Open dataset [288]	DT, KNN, RF, GBoost	UNK hybrid			0.98		
[49](2023)	Mendeley	Open dataset [272]	MLP, RF, RT, KNN, SVM	RF	0.98	0.98	0.98		

[96](2023)	UNK*	Previous work [96]	DT, KNN, SVM, NB, LR, XBoost, Aboost	Hybrid (DT, SVM, LR)	0.99	0.98	0.99	0.99	
[50](2023)	phishTank, Kaggle, and Majestic*	UNK	BERT	BERT	0.97	0.96	0.97	0.97	
[52](2023)	URLHaus and phishtank.org	Custom crawler	Custom rule based	Custom rule based	0.93	0.93	0.93	0.93	
[88](2023)	UCI ML Reposiroty and Mendeley	Open dataset [272, 279, 281]	LGBM, XGBoost, AdaBoost, CatBoost, GB, Hybrid (BMLSELM)	Hybrid (BMLSELM)	0.97	0.97	0.97	0.97	
[85](2023)	Kaggle*	UNK	LR, NB, DT, SVM, RF, KNN	KNN			0.99		
[86](2023)	Kaggle*	UNK	RF, XGBoost, LightGBBM	RF	0.99	0.94	0.96	0.96	
[103](2023)	UNK*	UNK	RF, AdaBoost, XGBoost, GBoost, KNN	RF			0.91		
[55](2023)	phishtank.org*	UNK	LR, RF	RF	0.93	0.79	0.96	0.85	
[108](2023)	PubMed	Open dataset [289]	RF, NB, LSTM, CNN	UNK					
[57](2023)	phishtank.org and who.is*	Custom crawler	LSTM, CNN, Hybrid (LSTM and CNN)	Hybrid (LSTM and CNN)			0.93		
[100](2024)	Zhejiang Mobile Innovation Research Institute*	UNK	MBERT, XGBoost, LBoost, LSTM, NB, LR, RF, SVM, KNN	MBERT	0.94	0.94		094	

Table 3: Data sources and Detection Methods used for Phishing URL detection. A single asterisk (*) indicates that the data is not publicly available. *UNK* indicates *Unclear/Unknown/Unspecified* details. Empty cells indicate missing values. *P: Precision, R: Recall, A: Accuracy, F1: F1 Score, AUC: Area under the Curve.*

#	URL Source	Collection Method	Models Used	Best Model	Performance				
					P	**R**	**A**	**F1**	**AUC**
[126](2019)	UNK*	UNK	RF, KNN, ANN, SVM, LR, NB	RF			0.97		
[136](2019)	Spam emails received by a company*	UNK	GNB, DT, SVM, NN, RF	RF, SVM	0.92	0.97	0.89		
[119](2019)	cs.cmu.edu	Open dataset [290]	RF, KNN, SVM, DT	RF	0.92	0.94	0.91		
[120](2020)	aclweb.org and previous work	Open dataset [291] and previous work [292]	NB, Dt, RF, SVM	SVM	0.98	0.97	0.98	0.97	
[139](2020)	Spam emails received by a company*	UNK	Clustering	Clustering			0.89		
[121](2021)	Open datasets*	UNK	LR, SVM, RF, XGBoost	XGBoost					
[111](2021)	cs.cmu.edu	Open dataset [290]	LSTM	LSTM			0.97		
[113](2021)	UNK	UNK	Various topic modelling	N/A	N/A	N/A	N/A	N/A	N/A
[131](2021)	Author's spam folder*	Custom	LDA, Jaccard	N/A	N/A	N/A	N/A	N/A	N/A
[116](2021)	UNK*	UNK	RNN, LSTM, CNN, BERT	UNK					
[117](2021)	Questionnaires and untroubled.org/spam	Open dataset[34]	NB, SVM, RF, LR	NB			0.88	0.8	

[34] https://untroubled.org/spam/

[138](2022)	UNK*	Custom crawler	BOW (Rule based)	BOW (Rule based)			0.99		%
[123](2022)	Kaggle	Open dataset [293]	CBoost, LR, DT, RF, GNB, SVM, KNN, XGBoost, LGBM, AdaBoost	CBoost	0.97		0.96	0.97	
[125](2022)	Kaggle	Open dataset [294]	RF, NB, SVM, AdaBoost, LR	RF			0.99		
[110](2022)	Previous work*	Previous work [295]	RF, LR, SVM, MNB	RF, LR, SVM,	0.95	0.95	0.95		
[114](2022)	GitHub, monkey.org, cs.cmu.edu	Open datasets [290, 296, 297]	K-Means, DBSCAN, and Agglomerative Clustering	Agglomerative Clustering	N/A	N/A	N/A	N/A	
[137](2022)	UNK	UNK	SVM, DT, LR, DNN, RF	DT	1	1	1	1	
[127](2022)	UCI ML Reposiroty*	UNK	NB, SVM, KNN, J48, DT	DT			0.98		
[112](2023)	Previous work, cs.cmu.edu, spamassassin.apache.org, and csmining.org[35]	Previous work [298, 299], open datasets [290, 300], and UNK	NB, SVM	UNK					
[133](2023)	Synthetic data	Data generated using various techniques [301]	ALBERT, RoBERTa, BERT, DBERT, SQ, YOSO	ALBERT			0.94	0.95	
[122](2023)	Kaggle*	UNK	RNN, LSTM, CNN	RNN	0.99	0.92	0.99	0.95	
[128](2023)	UCI ML Repository	Open dataset[302]	BERT	BERT	0.95	0.93	0.98	0.94	
[129](2023)	UCI ML Repository*	Previous work [129]	SVM, RF, NB	RF			0.95		
[115](2023)	Previous work*	Previous work [303]	KNN, NB, DT, RF, SVM, LR, XGBoost, BERT	BERT	0.97	0.97	0.97	0.97	
[130](2023)	UCI ML Repository*	UNK	CatBoost	CatBoost	0.97	0.96	0.96	0.97	0.99
[304](2023)	Previous work, Kaggle, and monkey.org	Previous work [305–307] and open datasets [297, 308]	Various topic modelling	N/A	N/A	N/A	N/A	N/A	N/A
[124](2023)	Kaggle	Open dataset [309]	MLP, DT, LR, RF, KNN, SVM	MLP, SVM		0.99	0.99	0.99	0.99
[118](2024)	Kaggle	Open dataset [310]	GPT-3.5, GPT-4, Custom (CyberGPT)	Custom (CyberGPT)			0.97		
[132](2024)	UNK*	UNK	GPT-3.5, GPT-4	UNK					

Table 4: Data sources and Detection Methods used for Phishing Email detection. A single asterisk (*) indicates that the data is not publicly available. *UNK* indicates *Unclear details*. Empty cells indicate missing values. *P: Precision, R: Recall, A: Accuracy, F1: F1 Score, AUC: Area under the Curve.*

#	**URL Source**	**Collection Method**	**Models Used**	**Best Model**	**Performance**				
					P	**R**	**A**	**F1**	**AUC**
[142](2019)	Previous work*	Previous work [292]	SVM, LR, NN, NB, RF	RF			0.98		
[154](2020)	360 Mobile Safe*	UNK	SVM, NB, LR, RF	SVM	0.96	0.96		0.96	
[158](2021)	360 Mobile Safe*	UNK	LR, DT, NB, SVM	LR	0.93	0.93		0.93	
[152](2021)	UCI ML repository	Open dataset [302]	CNN, GRU, MLP, SVM, XGBoost, Hybrid (CNN, GRU)	Hybrid (CNN, GRU)	0.99	0.96		0.98	

[35] Data link broken

[155](2022)	https://www.datafountain.cn/*	Custom crawler	CNN, BERT, RoBERTa, ChineseBERT	Hybrid (Semorph/UNK)	0.96	0.84		0.89	
[156](2022)	Twitter*	Twitter API	Custom/UNK	Custom/UNK	0.98	0.97	0.97		
[153](2022)	UCI ML repository*	UNK	SVM, NB, LR, DT	SVM	0.96	0.93	0.98	0.95	
[140](2022)	Kaggle and YouTube*	UNK	NB, DT, KNN	NB			0.97		
[144](2022)	Kaggle	Open dataset [311]	LR, SVC, RF, NB, GBM	RF	0.99	0.95	0.99	0.97	
[159](2023)	smishtank.com*	Custom crawler	Various NLP methods	N/A					
[151](2023)	Kaggle and inaccessible website	Open dataset [311] and custom crawler	LinearDA, QDA, SVM, PCA, NB	SVM			0.97		
[150](2023)	Kaggle	Open dataset [311]	KNN, NB, RF, SVC, ETC, LR, XGBoost, AdaBoost, GBDT, DT,	NBB	1		0.95		
[149](2023)	Previous work*	Previous work [154]	BERT-GCN	BERT-GCN		0.92	0.96	0.93	
[148](2023)	UCI ML repository*	UNK	LSTM, CNN, RF, Hybrid (various), BERT, LSTM, XGBoost	Hybrid (CNN, LSTM)	0.99	0.99	0.99	0.99	
[147](2023)	Kaggle	Open dataset [311]	DNN, LSTM	DNN			0.95		
[146](2023)	UCI ML repository	Open dataset [302]	LSTM, GRU, NB, BERT	BERT	0.99		0.99		
[145](2023)	Kaggle and Mendeley*	UNK	SNN, RNN, CNN	CNN			0.99		
[143](2023)	UNK	UNK	BPA, RF, NB, DT	BPA			0.97		0.98
[141](2023)	UNK	UNK	NB, RF, ETC	ETC	0.99		0.96		
[157](2024)	Korean Internet and Security Agency	UNK	NB, RF, LGBM, CNN, KoELECTRA	CharCNN			0.99	0.99	

Table 5: Data sources and Detection Methods used for Phishing SMS detection. A single asterisk (*) indicates that the data is not publicly available. *UNK* indicates *Unclear details*. Empty cells indicate missing values. *P: Precision, R: Recall, A: Accuracy, F1: F1 Score, AUC: Area under the Curve.*

#	Data	Collection Method	Models Used	Best Model	Performance				
					P	**R**	**A**	**F1**	**AUC**
[222](2019)	Twitter user profiles	Open dataset [312]	GCNN, MLP, BP	GCNN					0.94
[228](2019)	Facebook user profiles*	UNK	ID3, KNN, SVM	ID3	0.98	0.98	0.97		
[220](2019)	Twitter User Profiles	Open dataset [313]	SVM, RF, MADAFE (NN and LR)	MADAFE					
[229](2019)	Twitter and Facebook (UNK)*	UNK	HDBSCAN	HDBSCAN	N/A	N/A	N/A	N/A	N/A
[218](2020)	Tweets*	Twitter API	KNN, RF, NB, DT	RF					0.95
[232](2021)	Sina Weibo User profiles*	Custom crawler	CatBoost, RF	CatBoost					0.87
[235](2021)	Twitter user profiles	Open dataset [314]	NB, QDA, SVM, KNN, RF, NN	RF			0.87	0.88	0.94
[227](2021)	Instagram user profiles*	Instagram API	RF, AdaBoost, MLP, ANN, SGD	RF	0.99	0.98	0.98	0.98	0.98
[230](2022)	Facebook user profiles*	Manual collection	ANN, SVM, RF	ANN			0.96		
[225](2022)	Instagram user profiles*	UNK	LR, KNN, SVM, RF, NB	RF	0.99	0.97	0.94	0.98	
[224](2022)	Twitter user profiles	Open dataset [315]	GA, GP	GP			0.76	0.78	
[223](2022)	Twitter user profiles	Open dataset [316]	SVM, CNB, BNB, MP, DT, RF	TweezBot (Unclear)	0.99	0.93	0.98		0.99

[231](2023)	YouTube video and user metadata and comments*	YouTube API	Sentence-BERT, RoBERTa, YouTuBERT	YouTuBERT (LLM and DBSCAN)	0.63		0.81	0.90	0.71
[233](2023)	List of names*	UNK	NB, KNN, SVM, LR, RF	NB	0.94	0.94	0.95	0.94	
[219](2023)	Twitter user profiles*	UNK	NB, DT, NN, Ensemstack	Ensemstack			0.98		
[226](2023)	Instagram user profiles*	UNK	ANN	ANN				0.74	
[234](2023)	Instagram user profiles*	UNK	LR, DT, RF	RF			0.9		
[221](2023)	Twitter user profiles and tweets*	Twitter API	LR	LR	0.93	0.93	0.93	0.93	

Table 6: Data sources and Detection Methods used for Fake User detection. A single asterisk (*) indicates that the data is not publicly available. *UNK* indicates *Unclear details*. Empty cells indicate missing values. *P: Precision, R: Recall, A: Accuracy, F1: F1 Score, AUC: Area under the Curve.*

#	Job postings source	Collection Method	Models Used	Best Model	Performance				
					P	**R**	**A**	**F1**	**AUC**
[207](2019)	Kaggle	Open dataset [317]	J48, LR, RF	Ensemble			0.95	0.94	
[203](2020)	Kaggle	Open dataset [317]	GLoVE	GLoVE			0.99		
[208](2021)	Kaggle	Open dataset [317]	ANN	ANN	0.91	0.96		0.93	
[210](2021)	Kaggle	Open dataset [317]	LR	LR		0.89	0.92		0.96
[214](2021)	Kaggle	Open dataset [317]	LightGBM, LR, DT, XGBoost, AdaBoost	LightGBM	0.93	0.94	0.95	0.93	
[216](2021)	job.com.bd, bdjobstoday, and deshijob*	Custom crawler	LR, AdaBoost, DT, RF, VC, LGBM, GB	LightGBM or GBoost			0.95		
[204](2021)	Kaggle	Open dataset [317]	KNN, RF, DT, SVM, NB, DNN	DNN			0.97		
[12](2022)	Kaggle	Open dataset [317]	RF, LR, SVM, ETC, KNN, MP	ETC			0.99		
[213](2022)	Kaggle	Open dataset [317]	KNN, RF	KNN	0.79	0.73	0.98	0.76	
[215](2022)	SEEK, Glassdoor, Indeed, and Gumtree job postings*	Custom crawler	RF, JRip, NB, J48	RF	0.82	0.69	0.91		
[200](2022)	Kaggle	Open dataset [317]	GRU	GRU				0.93	
[201](2022)	Kaggle	Open dataset [317]	LR, NB, MLP, KNN, RF, DT, Adaboost, GB, NLP	RF	0.98	0.97	0.97	0.98	
[202](2022)	Kaggle	Open dataset [317]	RF, SVM, Bi-LSTM	Bi-LSTM			0.98	0.98	
[318](2023)	Kaggle	Open dataset [317]	SVM, NB, RF, Bi-LSTM, LR	RF					
[205](2023)	Kaggle	Open dataset [317]	RF, XBoost, LightGBM, CatBoost, DT	XGBoost	0.95	0.9	0.96	0.92	
[206](2023)	Kaggle	Open dataset [317]	RF, SVM, NB, Ensemble	RF			0.98		
[209](2023)	Kaggle	Open dataset [317]	RF, NB, SVM, DT, KNN	RF			0.97		
[211](2023)	Kaggle	Open dataset [317]	LR, DT, RF, NB, GLM	GLM	0.96	0.78		0.86	0.98
[212](2023)	Kaggle	Open dataset [317]	AdaBoost, XGBoost, RF, Voting	AdaBoost	0.99	0.97	0.98	0.98	

[217](2023)	Boss Zhipin, Liepin, 51job*	Custom crawler	NB, XGBoost, SVM, LightGBM, DT, RF	DRLM (DT and RF and LightGBM)		0.98	0.94	0.92	

Table 7: Data sources and Detection Methods used for Fraudulent Recruitment detection. The asterisk (*) indicates that the data is not publicly available. Empty cells indicate missing values. *P: Precision, R: Recall, A: Accuracy, F1: F1-Score, AUC: Area Under the Curve.*

#	**Data**	**Collection Method**	**Models Used**	**Best Model**	**Performance**				
					P	**R**	**A**	**F1**	**AUC**
[191](2020)	Amazon reviews*	Custom crawler	SVM, LR, RF, DT, GNBSGD, KNN, 3LP, 4LP, XGBoost	3LP	0.98	0.98		0.98	0.98
[190](2020)	Amazon reviews*	Custom crawler	SVM, KNN, NB, Ensemble	Ensemble	0.81	0.81	0.81	0.81	
[183](2021)	Yelp and JD.com reviews	Open dataset[319, 320] and JD.com custom crawler	GraphSAGE, Cluster-GCN, HGT, *C-FATH (Custom)*	C-FATH (Unclear)*				0.68-0.87	0.95-0.97
[193](2021)	Amazon reviews	Open dataset [321]	RF	RF*	1	0.85	0.98		
[177](2021)	Yelp reviews*	UNK	CNN, SVM, LR, MLP	CNN	0.93	0.92	0.92		
[181](2022)	Yelp reviews	Open dataset [322, 323]	WaveNet, LDA	WaveNet, LDA	N/A	N/A	N/A	N/A	N/A
[195](2022)	Amazon reviews*	UNK	BERT, VADER, LSTM, WordNet, SGD, SVM, LR	LR			0.81		
[196](2022)	Amazon hotel reviews*	UNK	KNN, NB, SVM	SVM*			0.93		
[186](2022)	Smartphone App reviews*	Web Scraping	LDA, keyATM	keyATM	N/A	N/A	N/A	N/A	N/A
[187](2022)	Smartphone App reviews	Open dataset [324, 325]	SVM, DT, NN, LR, GBT	SVM	0.94	0.84		0.89	
[185](2022)	Google Play reviews*	Custom crawler	DT, RF, MLP	MLP			0.97		
[188](2022)	Amazon reviews*	UNK	CNN, SVM, NB	CNN	1	1		1	
[180](2022)	Hotel reviews	Open dataset [323, 326]	SVM, KNN, LR	SKL (SVM and KNN and LR)			0.95		
[179](2022)	Yelp reviews	Open dataset [319]	Bi-LSTM	Bi-LSTM					0.89
[194](2023)	Amazon book reviews*	UNK	SVM, LR	LR			0.86		
[184](2023)	Reviews	Open dataset [323] and UNK	CNN, LSTM, KNN, NB, SVM, W2V	CNN, LSTM			0.93		
[189](2023)	Amazon reviews*	Custom crawler	AdaBoost, RF, Lr, SVM, KNN	RF	0.99	0.99	0.99	0.99	
[182](2023)	Yelp reviews*	UNK	SVM, MLP, CNN, LR	CNN	0.85	0.85	0.85	0.85	
[197](2023)	Yelp reviews*	UNK	NB, LR, SVM, DT	SVM	0.96	0.98	0.97		
[178](2023)	Yelp reviews	Open dataset [319]	GPT-3, BERT, RF, XGBoost	GPT-3	0.73	0.64		0.68	0.75
[199](2023)	Undefined reviews*	UNK	ANN, CNN, LR, SVM, NB, KNN, RF, DT, SGD	LR			0.89		
[192](2023)	Hotel reviews*	UNK	SVM	SVM					
[198](2024)	Product reviews*	YouTube API	SVM,LR	LR	0.74	0.99		0.85	0.95

Table 8: Data sources and Detection Methods used for Fake Review detection. A single asterisk (*) indicates that the data is not publicly available. *UNK* indicates *Unclear details*. Empty cells indicate missing values. *P: Precision, R: Recall, A: Accuracy, F1: F1-Score, AUC: Area Under the Curve.*

Acknowledgements
The authors would like to thank the reviewers for their feedback and reviews.

Funding
This work was funded by The Dawes Trust.

Abbreviations
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews; NLP: Natural Language Processing; AI: Artificial Intelligence; ML: Machine Learning; SLR: Systematic Literature Review.

Availability of data and materials
The data extracted from the academic papers included in this review can be found in this public repository: https://osf.io/nrx7y/?view_only=ca1050d48c4c4a969817c6d5f677cb87.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Authors' contributions
AP and NT drafted the final manuscript. AP conducted the updated academic literature review, designed and conducted the grey literature review, extracted data from the literature, and analysed and interpreted the results of all aspects of the scoping review. NT contributed to the coding process, and supervised all aspects of the study. SJ and ED provided substantial feedback on the review process and editing of the document. EM contributed to the coding process. AM and SL contributed to the editing process.

Authors' information
[1] Security and Crime Science, University College London, London, United Kingdom.
[2] Humanities and Social Sciences, Anglia Ruskin University, London, United Kingdom.
[3] Advanced Research Computing Centre, University College London, London, United Kingdom.

Author details
[1]Security and Crime Science, University College London, London, United Kingdom. [2]Humanities and Social Sciences, Anglia Ruskin University, London, United Kingdom. [3]Advanced Research Computing Centre, University College London, London, United Kingdom.

References

1. UK Finance: Over £1.2 billion stolen through fraud in 2022: Nearly 80 percent of attacks were online. Accessed: 2024-08-01 (2023). https://www.ukfinance.org.uk/news-and-insight/press-release/over-ps12-billion-stolen-through-fraud-in-2022-nearly-80-cent-app
2. UK Parliament: Social and Psychological Implications of Fraud. Accessed: 2024-07-03. https://researchbriefings.files.parliament.uk/documents/POST-PN-0720/POST-PN-0720.pdf Accessed 2024-07-03
3. Office for National Statistics: Crime in England and Wales: Year ending March 2023. Accessed: 2024-08-01 (2023). https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingmarch2023
4. Ofcom: Adults' Media Use and Attitudes Report 2023. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/media-literacy-research/adults/adults-media-use-and-attitudes-2023/adults-media-use-and-attitudes-report-2023.pdf. Accessed: 2024-07-01 (2023)
5. Statista: E-commerce Fraud - Statistics & Facts. https://www.statista.com/topics/9240/e-commerce-fraud/. Accessed: 2024-07-03 (2024)
6. UK Legislation: Fraud Act 2006. https://www.legislation.gov.uk/ukpga/2006/35/2023-02-07. Accessed: 2024-07-09 (2006)
7. The Law Society: Legal glossary. https://www.lawsociety.org.uk/public/for-public-visitors/resources/glossary. Accessed: 2024-07-09
8. National Fraud Authority: Fraud typologies and victims of fraud. https://assets.publishing.service.gov.uk/media/5a7ad8c2ed915d670dd7efad/fraud-typologies.pdf. Accessed: 2024-07-09
9. Levi, M.: Organized fraud and organizing frauds: Unpacking research on networks and organization. Criminology & Criminal Justice **8**(4), 389–419 (2008). doi:10.1177/1748895808096470. https://doi.org/10.1177/1748895808096470
10. Skidmore, M.: Perspectives on Online Fraud. Accessed: 2024-07-03. https://www.police-foundation.org.uk/wp-content/uploads/2010/10/perspectives_online_fraud.pdf Accessed 2024-07-03
11. Salloum, S., Gaber, T., Vadera, S., Shaalan, K.: A systematic literature review on phishing email detection using natural language processing techniques. IEEE Access **10**, 65703–65727 (2022). doi:10.1109/ACCESS.2022.3183083
12. Amaar, A., Aljedaani, W., Rustam, F., Ullah, S., Rupapara, V., Ludi, S.: Detection of fake job postings by utilizing machine learning and natural language processing approaches. Neural Processing Letters, 1–29 (2022)
13. Lwin Tun, Z., Birks, D.: Supporting crime script analyses of scams with natural language processing. Crime Science **12**(1), 1 (2023)
14. OFCOM: Executive Summary Report: Online Scams & Fraud Research. Accessed: 2024-07-03. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/online-fraud-and-scams/online-scams-and-fraud-research-summary-report?v=329362 Accessed 2024-07-03
15. Rabitti, G., Khorrami Chokami, A., Coyle, P., Cohen, R.D.: A taxonomy of cyber risk taxonomies. Risk Analysis (2024)
16. Zhou, S., Liu, X.F., Nah, F.F.-H., Harrison, S., Zhang, X., Zhen, S., Yeung, D., Hsiao, J.H.-w., LC, R., Chan, A.B., *et al.*: Understanding and fighting scams: Media, language, appeals and effects. In: International Conference on Human-Computer Interaction, pp. 392–408 (2024). Springer
17. Royal Mail: Typical online scams to look out for. https://www.royalmail.com/help/scam-examples. Accessed: 2024-07-09
18. Rodger, J.: Barclays issues warning to anyone with £14,000 in their bank account. https://shorturl.at/dGjpU. Accessed: 2024-07-09
19. HM Revenue & Customs: Examples of HMRC related phishing emails, suspicious phone calls and texts. https://www.gov.uk/government/publications/phishing-and-bogus-emails-hm-revenue-and-customs-examples/phishing-emails-and-bogus-contact-hm-revenue-and-customs-examples. Accessed: 2024-07-09
20. Paul, H., Nikolaev, A.: Fake review detection on online e-commerce platforms: a systematic literature review. Data Mining and Knowledge Discovery **35**(5), 1830–1881 (2021)
21. Mehboob, A., Malik, M.: Smart fraud detection framework for job recruitments. Arabian Journal for Science and Engineering **46**(4), 3067–3078 (2021)
22. Coluccia, A., Pozza, A., Ferretti, F., Carabellese, F., Masti, A., Gualtieri, G.: Online romance scams: relational dynamics and psychological characteristics of the victims and scammers. a scoping review. Clinical practice and epidemiology in mental health: CP & EMH **16**, 24 (2020)
23. Cross, C.: Romance baiting, cryptorom and 'pig butchering': an evolutionary step in romance fraud. Current Issues in Criminal Justice, 1–13 (2023)
24. Ordekian, M., Papasavva, A., Mariconti, E., Vasek, M.: A sinister fattening: Dissecting the tales of pig butchering and other cryptocurrency scams. In: 2024 APWG Symposium on Electronic Crime Research (eCrime) (2024). IEEE
25. Vasek, M., Moore, T.: There's no free lunch, even using bitcoin: Tracking the popularity and profits of virtual currency scams. In:

Financial Cryptography and Data Security: 19th International Conference, FC 2015, San Juan, Puerto Rico, January 26-30, 2015, Revised Selected Papers 19, pp. 44–61 (2015). Springer

26. Agarwal, S., Atondo Siu, J., Ordekian, M., Hutchings, A., Mariconti, E., Vasek, M.: Defi deception–uncovering the prevalence of rugpulls in cryptocurrency projects (2023)
27. Vasek, M., Moore, T.: Analyzing the bitcoin ponzi scheme ecosystem. In: Financial Cryptography and Data Security: FC 2018 International Workshops, BITCOIN, VOTING, and WTSC, Nieuwpoort, Curaçao, March 2, 2018, Revised Selected Papers 22, pp. 101–112 (2019). Springer
28. Hamrick, J., Rouhi, F., Mukherjee, A., Feder, A., Gandal, N., Moore, T., Vasek, M.: An examination of the cryptocurrency pump-and-dump ecosystem. Information Processing & Management **58**(4), 102506 (2021)
29. Cumming, D., Hornuf, L., Karami, M., Schweizer, D.: Disentangling crowdfunding from fraudfunding. Journal of Business Ethics, 1–26 (2021)
30. FBI: Charity and Disaster Fraud. https://www.fbi.gov/how-we-can-help-you/scams-and-safety/common-scams-and-crimes/charity-and-disaster-fraud. Accessed: 2024-07-09
31. Hong, G., Yang, Z., Yang, S., Liaoy, X., Du, X., Yang, M., Duan, H.: Analyzing ground-truth data of mobile gambling scams. In: 2022 IEEE Symposium on Security and Privacy (SP), pp. 2176–2193 (2022). IEEE
32. Brody, R.G., Haynes, C.M., Mejia, H.: Income tax return scams and identity theft. Accounting and Finance Research **3**(1), 90–95 (2014)
33. Mirza-Davies, J.: Pension scams. House of Commons (2023)
34. Cohen, R.D., Humphries, J., Veau, S., Francis, R.: An investigation of cyber loss data and its links to operational risk. Journal of Operational Risk **14**(3) (2019)
35. Taylor, P.: Amount of data created, consumed, and stored 2010-2020, with forecasts to 2025. https://www.statista.com/statistics/871513/worldwide-data-created/. Accessed: 2024-07-01 (2023)
36. Rong, X.: word2vec parameter learning explained. CoRR **abs/1411.2738** (2014). 1411.2738
37. Stanford NLP Group: GloVe: Global Vectors for Word Representation. https://nlp.stanford.edu/projects/glove/. Accessed: 2024-07-03
38. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., Wen, J.-R.: A Survey of Large Language Models (2023). 2303.18223. https://arxiv.org/abs/2303.18223
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017)
40. OpenAI: ChatGPT. Accessed: 2024-07-03. https://chatgpt.com/
41. Cherif, A., Badhib, A., Ammar, H., Alshehri, S., Kalkatawi, M., Imine, A.: Credit card fraud detection in the era of disruptive technologies: A systematic review. Journal of King Saud University-Computer and Information Sciences **35**(1), 145–174 (2023)
42. Barrera, D., Naranjo, V., Fuertes, W., Macas, M.: Literature review of sms phishing attacks: Lessons, addresses, and future challenges. In: International Conference on Advanced Research in Technologies, Information, Innovation and Sustainability, pp. 191–204 (2023). Springer
43. Nightingale, A.: A guide to systematic literature reviews. Surgery (Oxford) **27**(9), 381–384 (2009)
44. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Group, P., *et al.*: Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. International journal of surgery **8**(5), 336–341 (2010)
45. Schmitt, M., Flechais, I.: Digital deception: Generative artificial intelligence in social engineering and phishing. arXiv preprint arXiv:2310.13715 (2023)
46. Li, J., Wang, D., Zhao, C., Tang, J.: Mui-vb: Malicious url identification model combining vgg and bi-lstm. In: Proceedings of the 2022 3rd International Conference on Control, Robotics and Intelligent System, pp. 141–148 (2022)
47. Vo Quang, M., Bui Tan Hai, D., Tran Kim Ngoc, N., Ngo Duc Hoang, S., Nguyen Huu, Q., Phan The, D., Pham, V.-H.: Shark-eyes: A multimodal fusion framework for multi-view-based phishing website detection. In: Proceedings of the 12th International Symposium on Information and Communication Technology, pp. 793–800 (2023)
48. Al-Milli, N., Hammo, B.H.: A convolutional neural network model to detect illegitimate urls. In: 2020 11th International Conference on Information and Communication Systems (ICICS), pp. 220–225 (2020). IEEE
49. Aslam, S., Nassif, A.B.: Phish-identifier: Machine learning based classification of phishing attacks. In: 2023 Advances in Science and Engineering Technology International Conferences (ASET), pp. 1–6 (2023). IEEE
50. Jishnu, K., Arthi, B.: Enhanced phishing url detection using leveraging bert with additional url feature extraction. In: 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1745–1750 (2023). IEEE
51. Jaber, A.N., Fritsch, L., Haugerud, H.: Improving phishing detection with the grey wolf optimizer. In: 2022 International Conference on Electronics, Information, and Communication (ICEIC), pp. 1–6 (2022). IEEE
52. Rafsanjani, A.S., Kamaruddin, N.B., Rusli, H.M., Dabbagh, M.: Qsecr: Secure qr code scanner according to a novel malicious url detection framework. IEEE Access (2023)
53. Alswailem, A., Alabdullah, B., Alrumayh, N., Alsedrani, A.: Detecting phishing websites using machine learning. In: 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), pp. 1–6 (2019). IEEE
54. Mandadi, A., Boppana, S., Ravella, V., Kavitha, R.: Phishing website detection using machine learning. In: 2022 IEEE 7th International Conference for Convergence in Technology (I2CT), pp. 1–4 (2022). IEEE
55. Jha, A.K., Muthalagu, R., Pawar, P.M.: Intelligent phishing website detection using machine learning. Multimedia Tools and Applications **82**(19), 29431–29456 (2023)
56. Marimuthu, S.K., Kalampatti Gopalasamy, S., Ben-Othman, J.: Intelligent antiphishing framework to detect phishing scam: A hybrid classification approach. Software: Practice and Experience **52**(2), 459–481 (2022)
57. Adebowale, M.A., Lwin, K.T., Hossain, M.A.: Intelligent phishing detection scheme using deep learning algorithms. Journal of Enterprise Information Management **36**(3), 747–766 (2023)
58. Shaiba, H., Alzahrani, J.S., Eltahir, M.M., Marzouk, R., Mohsen, H., Hamza, M.A.: Hunger search optimization with hybrid deep learning enabled phishing detection and classification model. Computers Materials & Continua **73**(3), 6425–6441 (2022)
59. Rao, R.S., Pais, A.R.: Two level filtering mechanism to detect phishing sites using lightweight visual similarity approach. Journal of Ambient Intelligence and Humanized Computing **11**(9), 3853–3872 (2020)
60. Salloum, S., Gaber, T., Vadera, S., Shaalan, K.: Phishing website detection from urls using classical machine learning ann model. In: International Conference on Security and Privacy in Communication Systems, pp. 509–523 (2021). Springer
61. Orunsolu, A.A., Sodiya, A.S., Akinwale, A.: A predictive model for phishing detection. Journal of King Saud University-Computer and Information Sciences **34**(2), 232–247 (2022)
62. Rao, R.S., Pais, A.R.: Detection of phishing websites using an efficient feature-based machine learning framework. Neural Computing and applications **31**, 3851–3873 (2019)
63. Barraclough, P.A., Fehringer, G., Woodward, J.: Intelligent cyber-phishing detection for online. computers & security **104**, 102123 (2021)
64. Almseidin, M., Zuraiq, A.A., Al-Kasassbeh, M., Alnidami, N.: Phishing detection based on machine learning and feature selection methods (2019)
65. Chiew, K.L., Tan, C.L., Wong, K., Yong, K.S., Tiong, W.K.: A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Information Sciences **484**, 153–166 (2019)

66. Li, Y., Yang, Z., Chen, X., Yuan, H., Liu, W.: A stacking model using url and html features for phishing webpage detection. Future Generation Computer Systems **94**, 27–39 (2019)
67. Sahingoz, O.K., Buber, E., Demir, O., Diri, B.: Machine learning based phishing detection from urls. Expert Systems with Applications **117**, 345–357 (2019)
68. Somesha, M., Pais, A.R., Rao, R.S., Rathour, V.S.: Efficient deep learning techniques for the detection of phishing websites. Sādhanā **45**, 1–18 (2020)
69. Tharani, J.S., Arachchilage, N.A.: Understanding phishers' strategies of mimicking uniform resource locators to leverage phishing attacks: A machine learning approach. Security and Privacy **3**(5), 120 (2020)
70. Do Xuan, C., Nguyen, H.D., Tisenko, V.N.: Malicious url detection based on machine learning. International Journal of Advanced Computer Science and Applications **11**(1) (2020)
71. Pradeepa, G., Devi, R.: Lightweight approach for malicious domain detection using machine learning. Scientific and Technical Bulletin of Information Technologies, Mechanics and Optics **22**(2), 262–268 (2022)
72. Fernandez, S., Korczyński, M., Duda, A.: Early detection of spam domains with passive dns and spf. In: International Conference on Passive and Active Network Measurement, pp. 30–49 (2022). Springer
73. Chen, Y.-C., Chen, J.-L., Ma, Y.-W.: Ai@ tss-intelligent technical support scam detection system. Journal of Information Security and Applications **61**, 102921 (2021)
74. Shalke, C.J., Achary, R.: Social engineering attack and scam detection using advanced natural langugae processing algorithm. In: 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), pp. 1749–1754 (2022). IEEE
75. Puri, N., Saggar, P., Kaur, A., Garg, P.: Application of ensemble machine learning models for phishing detection on web networks. In: 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), pp. 296–303 (2022). doi:10.1109/CCiCT56684.2022.00062
76. Saha Roy, S., Karanjit, U., Nilizadeh, S.: Phishing in the free waters: A study of phishing attacks created using free website building services. In: Proceedings of the 2023 ACM on Internet Measurement Conference, pp. 268–281 (2023)
77. Liang, Y., Yan, X.: Using deep learning to detect malicious urls. In: 2019 IEEE International Conference on Energy Internet (ICEI), pp. 487–492 (2019). IEEE
78. Chen, S.-W., Chen, P.-H., Tsai, C.-T., Liu, C.-H.: Development of machine learning based fraudulent website detection scheme. In: 2022 IEEE 5th International Conference on Knowledge Innovation and Invention (ICKII), pp. 108–110 (2022). IEEE
79. Gu, J., Xu, H.: An ensemble method for phishing websites detection based on xgboost. In: 2022 14th International Conference on Computer Research and Development (ICCRD), pp. 214–219 (2022). IEEE
80. Jha, R., Kunwar, G.: Machine learning based url analysis for phishing detection. In: 2023 6th International Conference on Information Systems and Computer Networks (ISCON), pp. 1–5 (2023). IEEE
81. Mehndiratta, M., Jain, N., Malhotra, A., Gupta, I., Narula, R.: Malicious url: Analysis and detection using machine learning. In: 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1461–1465 (2023). IEEE
82. Jain, S., Gupta, C.: A support vector machine learning technique for detection of phishing websites. In: 2023 6th International Conference on Information Systems and Computer Networks (ISCON), pp. 1–6 (2023). IEEE
83. Saha, I., Sarma, D., Chakma, R.J., Alam, M.N., Sultana, A., Hossain, S.: Phishing attacks detection using deep learning approach. In: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 1180–1185 (2020). IEEE
84. Kumar, S., Dubey, G.P., Gupta, B.: Hybrid machine learning technique for prediction of phishing websites. In: 2023 6th International Conference on Information Systems and Computer Networks (ISCON), pp. 1–4 (2023). IEEE
85. P, A.N., V, H.V., H, S.P.: Phishing perception and prediction. In: 2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT), pp. 1–6 (2023). doi:10.1109/ICITIIT57246.2023.10068585
86. DR, U.S., Patil, A., *et al.*: Malicious url detection and classification analysis using machine learning models. In: 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), pp. 470–476 (2023). IEEE
87. Zamir, A., Khan, H.U., Iqbal, T., Yousaf, N., Aslam, F., Anjum, A., Hamdani, M.: Phishing web site detection using diverse machine learning algorithms. The Electronic Library **38**(1), 65–80 (2020)
88. Kalabarige, L.R., Rao, R.S., Pais, A.R., Gabralla, L.A.: A boosting based hybrid feature selection and multi-layer stacked ensemble learning model to detect phishing websites. IEEE Access (2023)
89. Mohammed, B.A., Al-Mekhlafi, Z.G.: Accuracy of phishing websites detection algorithms by using three ranking techniques. In: IJCSNS, vol. 22, p. 272 (2022)
90. Priya, S., Selvakumar, S., Velusamy, R.L.: Gravitational search based feature selection for enhanced phishing websites detection. In: 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), pp. 453–458 (2020). IEEE
91. Ariyadasa, S., Fernando, S., Fernando, S.: Phishrepo: a seamless collection of phishing data to fill a research gap in the phishing domain. International Journal of Advanced Computer Science and Applications **13**(5) (2022)
92. Villanueva, A., Atibagos, C., De Guzman, J., Cruz, J.C.D., Rosales, M., Francisco, R.: Application of natural language processing for phishing detection using machine and deep learning models. In: 2022 International Conference on ICT for Smart Society (ICISS), pp. 01–06 (2022). IEEE
93. Vecile, S., Lacroix, K., Grolinger, K., Samarabandu, J.: Malicious and benign url dataset generation using character-level lstm models. In: 2022 IEEE Conference on Dependable and Secure Computing (DSC), pp. 1–8 (2022). IEEE
94. Kumar, J., Santhanavijayan, A., Janet, B., Rajendran, B., Bindhumadhava, B.: Phishing website classification and detection using machine learning. In: 2020 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–6 (2020). IEEE
95. Zin, N.A.B.M., Ab Razak, M.F., Firdaus, A., Ernawan, F., Zulkifli, N.S.A.: Machine learning technique for phishing website detection. In: 2023 IEEE 8th International Conference On Software Engineering and Computer Systems (ICSECS), pp. 235–239 (2023). IEEE
96. Pathak, P., Shrivas, A.K.: Classification of phishing website using machine learning based proposed ensemble model. In: 2022 OPJU International Technology Conference on Emerging Technologies for Sustainable Development (OTCON), pp. 1–6 (2023). IEEE
97. Bitaab, M., Cho, H., Oest, A., Lyu, Z., Wang, W., Abraham, J., Wang, R., Bao, T., Shoshitaishvili, Y., Doupé, A.: Beyond phish: Toward detecting fraudulent e-commerce websites at scale. In: 2023 IEEE Symposium on Security and Privacy (SP), pp. 2566–2583 (2023). IEEE
98. Nakano, H., Chiba, D., Koide, T., Fukushi, N., Yagi, T., Hariu, T., Yoshioka, K., Matsumoto, T.: Canary in twitter mine: collecting phishing reports from experts and non-experts. In: Proceedings of the 18th International Conference on Availability, Reliability and Security, pp. 1–12 (2023)
99. Janet, B., Nikam, A., *et al.*: Real time malicious url detection on twitch using machine learning. In: 2022 International Conference on Electronics and Renewable Systems (ICEARS), pp. 1185–1189 (2022). IEEE
100. Yu, B., Tang, F., Ergu, D., Zeng, R., Ma, B., Liu, F.: Efficient classification of malicious urls: M-bert-a modified bert variant for enhanced semantic understanding. IEEE Access (2024)
101. Alkawaz, M.H., Steven, S.J., Mohammad, O.F., Johar, M.G.M.: Identification and analysis of phishing website based on machine learning methods. In: 2022 IEEE 12th Symposium on Computer Applications & Industrial Electronics (ISCAIE), pp. 246–251 (2022). IEEE
102. El-Din, A.E., Hemdan, E.E.-D., El-Sayed, A.: Malweb: An efficient malicious websites detection system using machine learning algorithms. In: 2021 International Conference on Electronic Engineering (ICEEM), pp. 1–6 (2021). IEEE

103. Kundra, D.: Identification and classification of malicious and benign url using machine learning classifiers. In: 2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), pp. 160–165 (2023). IEEE
104. Chen, J.-L., Ma, Y.-W., Huang, K.-L.: Intelligent visual similarity-based phishing websites detection. Symmetry **12**(10), 1681 (2020)
105. Ou, H., Guo, Y., Huang, C., Zhao, Z., Guo, W., Fang, Y., Huang, C.: No pie in the sky: The digital currency fraud website detection. In: International Conference on Digital Forensics and Cyber Crime, pp. 176–193 (2021). Springer
106. Raja, A.S., Vinodini, R., Kavitha, A.: Lexical features based malicious url detection using machine learning techniques. Materials Today: Proceedings **47**, 163–166 (2021)
107. Yadollahi, M.M., Shoeleh, F., Serkani, E., Madani, A., Gharaee, H.: An adaptive machine learning based approach for phishing detection using hybrid features. In: 2019 5th International Conference on Web Research (ICWR), pp. 281–286 (2019). IEEE
108. Nagy, N., Aljabri, M., Shaahid, A., Ahmed, A.A., Alnasser, F., Almakramy, L., Alhadab, M., Alfaddagh, S.: Phishing urls detection using sequential and parallel ml techniques: comparative analysis. Sensors **23**(7), 3467 (2023)
109. Geyik, B., Erensoy, K., Kocyigit, E.: Detection of phishing websites from urls by using classification techniques on weka. In: 2021 6th International Conference on Inventive Computation Technologies (ICICT), pp. 120–125 (2021). IEEE
110. Al-Ghamdi, N., Alsubait, T.: Digital forensics and machine learning to fraudulent email prediction. In: 2022 Fifth National Conference of Saudi Computers Colleges (NCCC), pp. 99–106 (2022). IEEE
111. Bhatti, P., Jalil, Z., Majeed, A.: Email classification using lstm: A deep learning technique. In: 2021 International Conference on Cyber Warfare and Security (ICCWS), pp. 100–105 (2021). IEEE
112. Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., Alegre, E.: A review of spam email detection: analysis of spammer strategies and the dataset shift problem. Artificial Intelligence Review **56**(2), 1145–1173 (2023)
113. Stojnic, T., Vatsalan, D., Arachchilage, N.A.: Phishing email strategies: understanding cybercriminals' strategies of crafting phishing emails. Security and privacy **4**(5), 165 (2021)
114. Saka, T., Vaniea, K., Kökciyan, N.: Context-based clustering to mitigate phishing attacks. In: Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security, pp. 115–126 (2022)
115. Jena, D., Kumari, A., Tejaswini, K., Ankita, A., Kumar, B.: Malicious spam detection to avoid vicious attack. In: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–7 (2023). IEEE
116. Jonker, R.A.A., Poudel, R., Pedrosa, T., Lopes, R.P.: Using natural language processing for phishing detection. In: International Conference on Optimization, Learning Algorithms and Applications, pp. 540–552 (2021). Springer
117. Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E.: Trustworthiness of spam email addresses using machine learning. In: Proceedings of the 21st ACM Symposium on Document Engineering, pp. 1–4 (2021)
118. Chataut, R., Gyawali, P.K., Usman, Y.: Can ai keep you safe? a study of large language models for phishing detection. In: 2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0548–0554 (2024). IEEE
119. Marková, E., Bajtoš, T., Sokol, P., Mézešová, T.: Classification of malicious emails. In: 2019 IEEE 15th International Scientific Conference on Informatics, pp. 000279–000284 (2019). IEEE
120. Al-Haddad, R., Sahwan, F., Aboalmakarem, A., Latif, G., Alufaisan, Y.M.: Email text analysis for fraud detection through machine learning techniques. In: 3rd Smart Cities Symposium (SCS 2020), vol. 2020, pp. 613–616 (2020). IET
121. Islam, M.K., Al Amin, M., Islam, M.R., Mahbub, M.N.I., Showrov, M.I.H., Kaushal, C.: Spam-detection with comparative analysis and spamming words extractions. In: 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), pp. 1–9 (2021). IEEE
122. Ramprasath, J., Priyanka, S., Manudev, R., Gokul, M.: Identification and mitigation of phishing email attacks using deep learning. In: 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 466–470 (2023). IEEE
123. Singh, U., Singh, V., Gourisaria, M.K., Das, H.: Spam email assessment using machine learning and data mining approach. In: 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), pp. 350–357 (2022). IEEE
124. Emmanuel, A.A., Yamazaki, T.: Information security in social media sites: Sentiment analysis of email. In: 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT), pp. 1–5 (2023). IEEE
125. Livara, A., Hernandez, R.: An empirical analysis of machine learning techniques in phishing e-mail detection. In: 2022 International Conference for Advancement in Technology (ICONAT), pp. 1–6 (2022). IEEE
126. Salihovic, I., Serdarevic, H., Kevric, J.: The role of feature selection in machine learning for detection of spam and phishing attacks. In: Advanced Technologies, Systems, and Applications III: Proceedings of the International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies (IAT), Volume 2, pp. 476–483 (2019). Springer
127. Ismail, S.S., Mansour, R.F., Abd El-Aziz, R.M., Taloba, A.I.: Efficient e-mail spam detection strategy using genetic decision tree processing with nlp features. Computational Intelligence and Neuroscience **2022**(1), 7710005 (2022)
128. Kushwaha, A., Dutta, K., Maheshwari, V.: Analysis of bert email spam classifier against adversarial attacks. In: 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), pp. 485–490 (2023). IEEE
129. Saini, A., Guleria, K., Sharma, S.: Machine learning approaches for an automatic email spam detection. In: 2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1), pp. 1–5 (2023). IEEE
130. Mittal, K., Gill, K.S., Chauhan, R., Joshi, K., Banerjee, D.: Blockage of phishing attacks through machine learning classification techniques and fine tuning its accuracy. In: 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), pp. 1–5 (2023). IEEE
131. Genc, Y., Kour, H., Arslan, H.T., Chen, L.-C.: Understanding nigerian e-mail scams: A computational content analysis approach. Information Security Journal: A Global Perspective **30**(2), 88–99 (2021)
132. Jiang, L.: Detecting scams using large language models. arXiv preprint arXiv:2402.03147 (2024)
133. Mehdi Gholampour, P., Verma, R.M.: Adversarial robustness of phishing email detection models. In: Proceedings of the 9th ACM International Workshop on Security and Privacy Analytics, pp. 67–76 (2023)
134. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., *et al.*: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
135. Ren, S., Deng, Y., He, K., Che, W.: Generating natural language adversarial examples through probability weighted word saliency. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1085–1097. Association for Computational Linguistics, Florence, Italy (2019). doi:10.18653/v1/P19-1103
136. Gallo, L., Botta, A., Ventre, G.: Identifying threats in a large company's inbox. In: Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks, pp. 1–7 (2019)
137. Mughaid, A., AlZu'bi, S., Hnaif, A., Taamneh, S., Alnajjar, A., Elsoud, E.A.: An intelligent cyber security phishing detection system using deep learning techniques. Cluster Computing **25**(6), 3819–3828 (2022)
138. Venugopal, I., Bhaskari, D.L., Seetaramanath, M.: Detection of severity-based email spam messages using adaptive threshold driven clustering. International Journal of Advanced Computer Science and Applications **13**(10) (2022)
139. Rahmad, F., Suryanto, Y., Ramli, K.: Performance comparison of anti-spam technology using confusion matrix classification. In: IOP

Conference Series: Materials Science and Engineering, vol. 879, p. 012076 (2020). IOP Publishing

140. Vinothkumar, S., Varadhaganapathy, S., Shanthakumari, R., Ramkishore, D., Rithik, S., Tharanies, K.: Detection of spam messages in e-messaging platform using machine learning. In: 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), pp. 283–287 (2022). IEEE
141. Agrawal, N., Bajpai, A., Dubey, K., Patro, B.: An effective approach to classify fraud sms using hybrid machine learning models. In: 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), pp. 1–6 (2023). IEEE
142. Jain, A.K., Gupta, B.B.: Feature based approach for detection of smishing messages in the mobile environment. Journal of Information Technology Research (JITR) **12**(2), 17–35 (2019)
143. Mishra, S., Soni, D.: Dsmishsms-a system to detect smishing sms. Neural Computing and Applications **35**(7), 4975–4992 (2023)
144. Abid, M.A., Ullah, S., Siddique, M.A., Mushtaq, M.F., Aljedaani, W., Rustam, F.: Spam sms filtering based on text features and supervised machine learning techniques. Multimedia Tools and Applications **81**(28), 39853–39871 (2022)
145. Kohilan, R., Warakagoda, H.E., Kitulgoda, T.T., Skandhakumar, N., Kuruwitaarachchi, N.: A machine learning-based approach for detecting smishing attacks at end-user level. In: 2023 IEEE International Conference on e-Business Engineering (ICEBE), pp. 149–154 (2023). IEEE
146. Siagian, W., Setiadi, M.R., Prasetyo, S.Y.: Improving sms spam detection through machine learning: An investigation of feature extraction and model selection techniques. In: 2023 International Conference on Information Management and Technology (ICIMTech), pp. 288–293 (2023). IEEE
147. Gandhi, C., Sarangi, P.K., Saxena, M., Sahoo, A.K.: Sms spam detection using deep learning techniques: A comparative analysis of dnn vs lstm vs bi-lstm. In: 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), pp. 189–194 (2023). IEEE
148. Al-Kabbi, H.A., Feizi-Derakhshi, M.-R., Pashazadeh, S.: Multi-type feature extraction and early fusion framework for sms spam detection. IEEE Access (2023)
149. Zhang, X., Huang, R., Jin, L., Wan, F.: A bert-gcn-based detection method for fbs telecom fraud chinese sms texts. In: 2023 4th International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), pp. 448–453 (2023). IEEE
150. Dharani, V., Hegde, D., *et al.*: Spam sms (or) email detection and classification using machine learning. In: 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 1104–1108 (2023). IEEE
151. Addanki, V., Durgapu, S., Dorasanaiah, K., Abhishek, S., *et al.*: Safeguarding sms: A dynamic duo approach to tackle spam using lda and qda. In: Innovations in Power and Advanced Computing Technologies (i-PACT) (2023)
152. Ulfath, R.E., Alqahtani, H., Hammoudeh, M., Sarker, I.H.: Hybrid cnn-gru framework with integrated pre-trained language transformer for sms phishing detection. In: Proceedings of the 5th International Conference on Future Networks and Distributed Systems, pp. 244–251 (2021)
153. Jain, T., Garg, P., Chalil, N., Sinha, A., Verma, V.K., Gupta, R.: Sms spam classification using machine learning techniques. In: 2022 12th International Conference on Cloud Computing, Data Science & Engineering (confluence), pp. 273–279 (2022). IEEE
154. Zhang, Y., Liu, B., Lu, C., Li, Z., Duan, H., Hao, S., Liu, M., Liu, Y., Wang, D., Li, Q.: Lies in the air: Characterizing fake-base-station spam ecosystem in china. In: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, pp. 521–534 (2020)
155. Lai, K., Long, Y., Wu, B., Li, Y., Wang, B.: Semorph: A morphology semantic enhanced pre-trained model for chinese spam text detection. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 1003–1013 (2022)
156. Tang, S., Mi, X., Li, Y., Wang, X., Chen, K.: Clues in tweets: Twitter-guided discovery and analysis of sms spam. In: Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, pp. 2751–2764 (2022)
157. Seo, J.W., Lee, J.S., Kim, H., Lee, J., Han, S., Cho, J., Lee, C.-H.: On-device smishing classifier resistant to text evasion attack. IEEE Access (2024)
158. Liu, M., Zhang, Y., Liu, B., Li, Z., Duan, H., Sun, D.: Detecting and characterizing sms spearphishing attacks. In: Proceedings of the 37th Annual Computer Security Applications Conference, pp. 930–943 (2021)
159. Timko, D., Rahman, M.L.: Commercial anti-smishing tools and their comparative effectiveness against modern threats. In: Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks, pp. 1–12 (2023)
160. Derakhshan, A., Harris, I.G., Behzadi, M.: Detecting telephone-based social engineering attacks using scam signatures. In: Proceedings of the 2021 ACM Workshop on Security and Privacy Analytics, pp. 67–73 (2021)
161. Djiré, A.E., Sabané, A., Kabore, A.-K., Kafando, R., Bissyandé, T.F.: Evaluating acoustic parameters for deepfake audio identification. In: 2023 IEEE Afro-Mediterranean Conference on Artificial Intelligence (AMCAI), pp. 1–6 (2023). IEEE
162. Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.-Y.: Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558 (2020)
163. Sini, A., Lolive, D., Vidal, G., Tahon, M., Delais-Roussarie, É.: Synpaflex-corpus: An expressive french audiobooks corpus dedicated to expressive speech synthesis. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
164. Liang, F.-Y., Li, F.-P., Xu, R.-H., Cheng, W., Deng, S.-X., Yang, Z.-R., Wang, C.-D.: Telecom fraud detection based on feature binning and autoencoder. In: 2023 IEEE International Conference on Data Mining (ICDM), pp. 368–377 (2023). IEEE
165. Huang, Z., Wu, J., Ren, L., Hu, R., Li, D.: Learning dynamic behavior patterns for fraud detection. In: 2022 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), pp. 621–627 (2022). IEEE
166. Hu, Z., Yuan, Z.: Urf4cct: A text understanding framework for chinese telecom fraud cases. In: 2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS), pp. 121–125 (2023). IEEE
167. Kim, J.-W., Hong, G.-W., Chang, H.: Voice recognition and document classification-based data analysis for voice phishing detection. Human-centric Comput. Inf. Sci **11** (2021)
168. Kale, N., Kochrekar, S., Mote, R., Dholay, S.: Classification of fraud calls by intent analysis of call transcripts. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–6 (2021). IEEE
169. Rahman, Y.M.: Phone call speaker classification using machine learning on mfcc features for scam detection. In: 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), pp. 351–356 (2022). IEEE
170. Hong, B., Connie, T., Goh, M.K.O.: Scam calls detection using machine learning approaches. In: 2023 11th International Conference on Information and Communication Technology (ICoICT), pp. 442–447 (2023). IEEE
171. Gowri, S.M., Ramana, G.S., Ranjani, M.S., Tharani, T.: Detection of telephony spam and scams using recurrent neural network (rnn) algorithm. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1, pp. 1284–1288 (2021). IEEE
172. Malhotra, S., Arora, G., Bathla, R.: Detection and analysis of fraud phone calls using artificial intelligence. In: 2023 International Conference on Recent Advances in Electrical, Electronics & Digital Healthcare Technologies (REEDCON), pp. 592–595 (2023). IEEE
173. Zhong, R., Zhang, Z., Lin, R., Zou, H.: Encoding broad learning system: An effective shallow model for anti-fraud. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 5496–5504 (2020). IEEE

174. Liu, T., Wang, S., Fu, J., Chen, L., Wei, Z., Liu, Y., Ye, H., Xu, L., Wang, W., Huang, X.: Fine-grained element identification in complaint text of internet fraud. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 3268–3272 (2021)
175. Zhou, T., Zhao, H., Zhang, X.: Keyword extraction based on random forest and xgboost-an example of fraud judgment document. In: 2022 European Conference on Natural Language Processing and Information Retrieval (ECNLPIR), pp. 17–22 (2022). IEEE
176. Palad, E.B.B., Tangkeko, M.S., Magpantay, L.A.K., Sipin, G.L.: Document classification of filipino online scam incident text using data mining techniques. In: 2019 19th International Symposium on Communications and Information Technologies (ISCIT), pp. 232–237 (2019). IEEE
177. Javed, M.S., Majeed, H., Mujtaba, H., Beg, M.O.: Fake reviews classification using deep learning ensemble of shallow convolutions. Journal of Computational Social Science, 1–20 (2021)
178. Pengqi, W., Yue, L., Junyi, C.: Unmasking deception: A comparative study of tree-based and transformer-based models for fake review detection on yelp. In: 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1848–1853 (2023). IEEE
179. Harris, C.G.: Combining linguistic and behavioral clues to detect spam in online reviews. In: 2022 IEEE International Conference on e-Business Engineering (ICEBE), pp. 38–44 (2022). IEEE
180. Tufail, H., Ashraf, M.U., Alsubhi, K., Aljahdali, H.M.: The effect of fake reviews on e-commerce during and after covid-19 pandemic: Skl-based fake reviews detection. Ieee Access **10**, 25555–25564 (2022)
181. Balakrishna, V., Bag, S., Sarkar, S.: Identifying spammer groups in consumer reviews using meta-data via bipartite graph approach. In: 2022 International Conference on Data Analytics for Business and Industry (ICDABI), pp. 650–654 (2022). IEEE
182. Ashraf, S., Rehman, F., Sharif, H., Kirn, H., Arshad, H., Manzoor, H.: Fake reviews classification using deep learning. In: 2023 International Multi-disciplinary Conference in Emerging Research Trends (IMCERT), vol. 1, pp. 1–8 (2023). IEEE
183. Wang, L., Li, P., Xiong, K., Zhao, J., Lin, R.: Modeling heterogeneous graph network on fraud detection: A community-based framework with attention mechanism. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 1959–1968 (2021)
184. Singh, D., Memoria, M., Kumar, R.: Deep learning based model for fake review detection. In: 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), pp. 92–95 (2023). IEEE
185. Yugeshwaran, G., Benitta, D.A., Eliyas, S., *et al.*: Rank fraud and malware detection in google play using fairplay. In: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 1356–1359 (2022). IEEE
186. Tushev, M., Ebrahimi, F., Mahmoud, A.: Domain-specific analysis of mobile app reviews using keyword-assisted topic models. In: Proceedings of the 44th International Conference on Software Engineering, pp. 762–773 (2022)
187. Obie, H.O., Ilekura, I., Du, H., Shahin, M., Grundy, J., Li, L., Whittle, J., Turhan, B.: On the violation of honesty in mobile apps: Automated detection and categories. In: Proceedings of the 19th International Conference on Mining Software Repositories, pp. 321–332 (2022)
188. Rangar, K.P., Khan, A.: A machine learning model for spam reviews and spammer community detection. In: 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), pp. 632–638 (2022). IEEE
189. Iqbal, A., Rauf, M.A., Zubair, M., Younis, T.: An efficient ensemble approach for fake reviews detection. In: 2023 3rd International Conference on Artificial Intelligence (ICAI), pp. 70–75 (2023). IEEE
190. Furia, R., Gaikwad, K., Mandalya, K., Godbole, A.: Tool for review analysis of product. In: 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1–6 (2020). IEEE
191. Gupta, V., Aggarwal, A., Chakraborty, T.: Detecting and characterizing extremist reviewer groups in online product reviews. IEEE Transactions on Computational Social Systems **7**(3), 741–750 (2020)
192. Thilagavathy, A., Therasa, P., Jasmine, J.J., Sneha, M., Lakshmi, R.S., Yuvanthika, S.: Fake product review detection and elimination using opinion mining. In: 2023 World Conference on Communication & Computing (WCONF), pp. 1–5 (2023). IEEE
193. Chandana, P., Sree, N.P., Ramya, V., Bhavana, G.: Analyzing the extremist reviewer groups on online products. In: 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1–4 (2021). IEEE
194. Akshara, S., Shiva, S., Kubireddy, S., Arun, T., Kanthety, V.L.: A small comparative study of machine learning algorithms in the detection of fake reviews of amazon products. In: 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), vol. 6, pp. 2258–2263 (2023). IEEE
195. Deekshan, S., PK, A.D., *et al.*: Detection and summarization of honest reviews using text mining. In: 2022 8th International Conference on Smart Structures and Systems (ICSSS), pp. 01–05 (2022). IEEE
196. Rangari, K., Khan, A.: An empirical analysis of different techniques for spam detection. In: 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1, pp. 947–953 (2022). IEEE
197. Silpa, C., Prasanth, P., Sowmya, S., Bhumika, Y., Pavan, C.S., Naveed, M.: Detection of fake online reviews by using machine learning. In: 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), pp. 71–77 (2023). IEEE
198. Bevendorff, J., Wiegmann, M., Potthast, M., Stein, B.: Product spam on youtube: A case study. In: Proceedings of the 2024 Conference on Human Information Interaction and Retrieval, pp. 358–363 (2024)
199. Ganesh, D., Rao, K.J., Kumar, M.S., Vinitha, M., Anitha, M., Likith, S.S., Taralitha, R.: Implementation of novel machine learning methods for analysis and detection of fake reviews in social media. In: 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), pp. 243–250 (2023). IEEE
200. Nessa, I., Zabin, B., Faruk, K.O., Rahman, A., Nahar, K., Iqbal, S., Hossain, M.S., Mehedi, M.H.K., Rasel, A.A.: Recruitment scam detection using gated recurrent unit. In: 2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC), pp. 445–449 (2022). IEEE
201. Prathaban, B.P., Rajendran, S., Lakshmi, G., Menaka, D.: Verification of job authenticity using prediction of online employment scam model (poesm). In: 2022 1st International Conference on Computational Science and Technology (ICCST), pp. 1–6 (2022). IEEE
202. Pandey, B., Kala, T., Bhoj, N., Gohel, H., Kumar, A., Sivaram, P.: Effective identification of spam jobs postings using employer defined linguistic feature. In: 2022 1st International Conference on AI in Cybersecurity (ICAIC), pp. 1–6 (2022). IEEE
203. Ranparia, D., Kumari, S., Sahani, A.: Fake job prediction using sequential network. In: 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), pp. 339–343 (2020). IEEE
204. Habiba, S.U., Islam, M.K., Tasnim, F.: A comparative study on fake job post prediction using different data mining techniques. In: 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 543–546 (2021). IEEE
205. Reddy, S.M., Ali, S.M., Battula, K.M., lakshmana Charan, P., Rashmi, M.: Web app for predicting fake job posts using ensemble classifiers. In: 2023 4th International Conference for Emerging Technology (INCET), pp. 1–5 (2023). IEEE
206. Santhiya, P., Kavitha, S., Aravindh, T., Archana, S., Praveen, A.V.: Fake news detection using machine learning. In: 2023 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–8 (2023). IEEE
207. Lal, S., Jiaswal, R., Sardana, N., Verma, A., Kaur, A., Mourya, R.: Orfdetector: ensemble learning based online recruitment fraud detection. In: 2019 Twelfth International Conference on Contemporary Computing (IC3), pp. 1–5 (2019). IEEE
208. Nasser, I.M., Alzaanin, A.H., Maghari, A.Y.: Online recruitment fraud detection using ann. In: 2021 Palestinian International Conference on Information and Communication Technology (PICICT), pp. 13–17

(2021). IEEE

209. Sofy, M.A., Khafagy, M.H., Badry, R.M.: An intelligent arabic model for recruitment fraud detection using machine learning. Journal of Advances in Information Technology **14**(1) (2023)
210. Vo, M.T., Vo, A.H., Nguyen, T., Sharma, R., Le, T.: Dealing with the class imbalance problem in the detection of fake job descriptions. Computers, Materials & Continua **68**(1), 521–535 (2021)
211. Nanath, K., Olney, L.: An investigation of crowdsourcing methods in enhancing the machine learning approach for detecting online recruitment fraud. International Journal of Information Management Data Insights **3**(1), 100167 (2023)
212. Ullah, Z., Jamjoom, M.: A smart secured framework for detecting and averting online recruitment fraud using ensemble machine learning techniques. PeerJ Computer Science **9**, 1234 (2023)
213. Bhatia, T., Meena, J.: Detection of fake online recruitment using machine learning techniques. In: 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 300–304 (2022). IEEE
214. Li, J., Li, Y., Han, H., Lu, X.: Exploratory methods for imbalanced data classification in online recruitment fraud detection: A comparative analysis. In: 2021 4th International Conference on Computing and Big Data, pp. 75–81 (2021)
215. Mahbub, S., Pardede, E., Kayes, A.: Online recruitment fraud detection: A study on contextual features in australian job industries. IEEE Access **10**, 82776–82787 (2022)
216. Tabassum, H., Ghosh, G., Atika, A., Chakrabarty, A.: Detecting online recruitment fraud using machine learning. In: 2021 9th International Conference on Information and Communication Technology (ICoICT), pp. 472–477 (2021). IEEE
217. Zhang, H., Wang, M., Wang, Y., Li, Y., Gu, D., Zhu, Y.: Orfpprediction: Machine learning based online recruitment fraud probability prediction. In: 2023 International Conference on the Cognitive Computing and Complex Data (ICCD), pp. 139–144 (2023). IEEE
218. Raj, R.J.R., Srinivasulu, S., Ashutosh, A.: A multi-classifier framework for detecting spam and fake spam messages in twitter. In: 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT), pp. 266–270 (2020). IEEE
219. Gangan, J., Suprith, K., Jamdar, N., Bharne, S.: Detection of fake twitter accounts using ensemble learning model. In: 2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA), pp. 1–6 (2023). IEEE
220. Yue, H., Zhou, L., Xue, K., Li, H.: Madafe: Malicious account detection on twitter with automated feature extraction. In: 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), pp. 1–6 (2019). IEEE
221. Singh, M., Singh, A.: How safe you are on social networks? Cybernetics and Systems **54**(7), 1154–1171 (2023)
222. Ali Alhosseini, S., Bin Tareaf, R., Najafi, P., Meinel, C.: Detect me if you can: Spam bot detection using inductive representation learning. In: Companion Proceedings of the 2019 World Wide Web Conference, pp. 148–153 (2019)
223. Shukla, R., Sinha, A., Chaudhary, A.: Tweezbot: An ai-driven online media bot identification algorithm for twitter social networks. Electronics **11**(5), 743 (2022)
224. Rovito, L., Bonin, L., Manzoni, L., De Lorenzo, A.: An evolutionary computation approach for twitter bot detection. Applied Sciences **12**(12), 5915 (2022)
225. Das, S., Saha, S., Vijayalakshmi, S., Jaiswal, J.: An effecient approach to detect fraud instagram accounts using supervised ml algorithms. In: 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 760–764 (2022). IEEE
226. Fathima, A.S., Reema, S., Ahmed, S.T.: Ann based fake profile detection and categorization using premetric paradigms on instagram. In: 2023 Innovations in Power and Advanced Computing Technologies (i-PACT), pp. 1–6 (2023). IEEE
227. Anklesaria, K., Desai, Z., Kulkarni, V., Balasubramaniam, H.: A survey on machine learning algorithms for detecting fake instagram accounts. In: 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 141–144 (2021). IEEE
228. Albayati, M.B., Altamimi, A.M.: Identifying fake facebook profiles using data mining techniques. Journal of ICT Research & Applications **13**(2) (2019)
229. Venkatesan, M., Prabhavathy, P.: Graph based unsupervised learning methods for edge and node anomaly detection in social network. In: 2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP), pp. 1–5 (2019). IEEE
230. Shreya, K., Kothapelly, A., Deepika, V., Shanmugasundaram, H.: Identification of fake accounts in social media using machine learning. In: 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), pp. 1–4 (2022). IEEE
231. Na, S.H., Cho, S., Shin, S.: Evolving bots: The new generation of comment bots and their underlying scam campaigns in youtube. In: Proceedings of the 2023 ACM on Internet Measurement Conference, pp. 297–312 (2023)
232. Zhang, X., Jiang, F., Zhang, R., Li, S., Zhou, Y.: Social spammer detection based on semi-supervised learning. In: 2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 849–855 (2021). IEEE
233. Haq, I., Qiu, W., Guo, J., Peng, T.: Spammy names detection in pashto language to prevent fake accounts creation on social media. In: 2023 8th International Conference on Signal and Image Processing (ICSIP), pp. 614–618 (2023). IEEE
234. Nikhitha, K.V., Bhavya, K., Nandini, D.U.: Fake account detection on social media using random forest classifier. In: 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 806–811 (2023). IEEE
235. Bebensee, B., Nazarov, N., Zhang, B.-T.: Leveraging node neighborhoods and egograph topology for better bot detection in social graphs. Social Network Analysis and Mining **11**(1), 10 (2021)
236. Janjeva, A., Harris, A., Mercer, S., Kasprzyk, A., Gausen, A.: The rapid rise of generative ai: Assessing risks to safety and security (2023)
237. Ferrara, E.: Genai against humanity: Nefarious applications of generative artificial intelligence and large language models. Journal of Computational Social Science, 1–21 (2024)
238. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., *et al.*: Extracting training data from large language models. In: 30th USENIX Security Symposium (USENIX Security 21), pp. 2633–2650 (2021)
239. Kumar, K., Bhushan, B., *et al.*: Augmenting cybersecurity and fraud detection using artificial intelligence advancements. In: 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), pp. 1207–1212 (2023). IEEE
240. Ayoobi, N., Shahriar, S., Mukherjee, A.: The looming threat of fake and llm-generated linkedin profiles: Challenges and opportunities for detection and prevention. In: Proceedings of the 34th ACM Conference on Hypertext and Social Media, pp. 1–10 (2023)
241. DiResta, R., Goldstein, J.A.: How spammers and scammers leverage ai-generated images on facebook for audience growth. arXiv preprint arXiv:2403.12838 (2024)
242. Grbic, D.V., Dujlovic, I.: Social engineering with chatgpt. In: 2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH), pp. 1–5 (2023). IEEE
243. Shibli, A.M., Pritom, M.M.A., Gupta, M.: Abusegpt: Abuse of generative ai chatbots to create smishing campaigns. In: 2024 12th International Symposium on Digital Forensics and Security (ISDFS), pp. 1–6 (2024). IEEE
244. Alotaibi, L., Seher, S., Mohammad, N.: Cyberattacks using chatgpt: Exploring malicious content generation through prompt engineering. In: 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), pp. 1304–1311 (2024). IEEE
245. Alawida, M., Abu Shawar, B., Abiodun, O.I., Mehmood, A., Omolara, A.E., Al Hwaitat, A.K.: Unveiling the dark side of chatgpt: Exploring cyberattacks and enhancing user awareness. Information **15**(1), 27 (2024)
246. Sharma, M., Singh, K., Aggarwal, P., Dutt, V.: How well does gpt

phish people? an investigation involving cognitive biases and feedback. In: 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), pp. 451–457 (2023). IEEE

247. Roy, S.S., Thota, P., Naragam, K.V., Nilizadeh, S.: From chatbots to phishbots?–preventing phishing scams created using chatgpt, google bard and claude. arXiv preprint arXiv:2310.19181 (2023)
248. Xu, Z., Luo, S., Shi, J., Li, H., Lin, C., Sun, Q., Hu, S.: Efficiently answering k-hop reachability queries in large dynamic graphs for fraud feature extraction. In: 2022 23rd IEEE International Conference on Mobile Data Management (MDM), pp. 238–245 (2022). IEEE
249. La Morgia, M., Mei, A., Mongardini, A.M., Wu, J.: It'sa trap! detection and analysis of fake channels on telegram. In: 2023 IEEE International Conference on Web Services (ICWS), pp. 97–104 (2023). IEEE
250. La Morgia, M., Mei, A., Mongardini, A.M., Wu, J.: Uncovering the dark side of telegram: Fakes, clones, scams, and conspiracy movements. arXiv preprint arXiv:2111.13530 (2021)
251. Shah, D., Harrison, T., Freas, C.B., Maimon, D., Harrison, R.W.: Illicit activity detection in large-scale dark and opaque web social networks. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 4341–4350 (2020). IEEE
252. Al-Hassan, M., Abu-Salih, B., Al Hwaitat, A.: Dspamonto: An ontology modelling for domain-specific social spammers in microblogging. Big Data and Cognitive Computing **7**(2), 109 (2023)
253. Tripathi, A., Ghosh, M., Bharti, K.: Analyzing the uncharted territory of monetizing scam videos on youtube. Social Network Analysis and Mining **12**(1), 119 (2022)
254. He, X., Gong, Q., Chen, Y., Zhang, Y., Wang, X., Fu, X.: Datingsec: Detecting malicious accounts in dating apps using a content-based attention network. IEEE Transactions on Dependable and Secure Computing **18**(5), 2193–2208 (2021)
255. Suarez-Tangil, G., Edwards, M., Peersman, C., Stringhini, G., Rashid, A., Whitty, M.: Automatically dismantling online dating fraud. IEEE Transactions on Information Forensics and Security **15**, 1128–1137 (2019)
256. Lokanan, M.E.: The tinder swindler: Analyzing public sentiments of romance fraud using machine learning and artificial intelligence. Journal of Economic Criminology **2**, 100023 (2023)
257. Cambiaso, E., Caviglione, L.: Scamming the scammers: Using chatgpt to reply mails for wasting time and resources. arXiv preprint arXiv:2303.13521 (2023)
258. Bajaj, P., Edwards, M.: Automatic scam-baiting using chatgpt. In: 2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 1941–1946 (2023). IEEE
259. Chen, W., Wang, F., Edwards, M.: Active countermeasures for email fraud. In: 2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P), pp. 39–55 (2023). IEEE
260. Siu, G.A., Hutchings, A., Vasek, M., Moore, T.: "invest in crypto!": An analysis of investment scam advertisements found in bitcointalk. In: 2022 APWG Symposium on Electronic Crime Research (eCrime), pp. 1–12 (2022). IEEE
261. Li, K., Guan, S., Lee, D.: Towards understanding and characterizing the arbitrage bot scam in the wild. Proceedings of the ACM on Measurement and Analysis of Computing Systems **7**(3), 1–29 (2023)
262. Kuo, C., Tsang, S.-S.: Constructing an investment scam detection model based on emotional fluctuations throughout the investment scam life cycle. Deviant Behavior **45**(2), 204–225 (2024)
263. Lee, S., Shafqat, W., Kim, H.-c.: Backers beware: Characteristics and detection of fraudulent crowdfunding campaigns. Sensors **22**(19), 7677 (2022)
264. Shafqat, W., Byun, Y.-C.: Topic predictions and optimized recommendation mechanism based on integrated topic modeling and deep neural networks in crowdfunding platforms. Applied Sciences **9**(24), 5496 (2019)
265. Nizzoli, L., Tardelli, S., Avvenuti, M., Cresci, S., Tesconi, M., Ferrara, E.: Charting the landscape of online cryptocurrency manipulation. IEEE access **8**, 113230–113245 (2020)
266. Mirtaheri, M., Abu-El-Haija, S., Morstatter, F., Ver Steeg, G., Galstyan, A.: Identifying and analyzing cryptocurrency manipulations in social media. IEEE Transactions on Computational Social Systems **8**(3), 607–617 (2021)
267. Farzaneh Shahini, D.W., Zahabi, M.: Usability evaluation of police mobile computer terminals: A focus group study. International Journal of Human–Computer Interaction **37**(15), 1478–1487 (2021). doi:10.1080/10447318.2021.1894801
268. Zahabi, M., Kaber, D.: Identification of task demands and usability issues in police use of mobile computing terminals. Applied Ergonomics **66**, 161–171 (2018). doi:10.1016/j.apergo.2017.08.013
269. Eling, M., McShane, M., Nguyen, T.: Cyber risk management: History and future research directions. Risk Management and Insurance Review **24**(1), 93–125 (2021)
270. Ebubekirbbr: Phishing Detection. GitHub. https://github.com/ebubekirbbr/pdd/tree/master/input (2018)
271. Akash Kumar: Phishing website dataset. https://www.kaggle.com/datasets/akashkr/phishing-website-dataset#dataset.csv (2017)
272. Choon Lin Tan: Phishing Dataset for Machine Learning: Feature Evaluation. https://data.mendeley.com/datasets/h3cgnj8hft/1 (2018)
273. Antony J: Malicious n Non-Malicious URL. https://www.kaggle.com/datasets/antonyj453/urldataset (2017)
274. Majestic: Majestic Dataset. http://downloads.majestic.com/majestic_million.csv
275. URLhaus: URLhaus Database Dump. https://urlhaus.abuse.ch/downloads/csv/
276. Lilo, J.: Detecting Malicious URL Using Pyspark. https://github.com/rlilojr/Detecting-Malicious-URL-Machine-Learning/tree/master (2018)
277. Mohammad, R.M., Thabtah, F., McCluskey, L.: Predicting phishing websites based on self-structuring neural network. Neural Computing and Applications **25**, 443–458 (2014)
278. Mohammad, R.M., Thabtah, F., McCluskey, L.: Intelligent rule-based phishing websites classification. IET Information Security **8**(3), 153–160 (2014)
279. Mohammad, R., McCluskey, L.: Phishing Websites. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C51W2X (2015)
280. Rao, R.S., Vaishnavi, T., Pais, A.R.: Catchphish: detection of phishing websites by inspecting urls. Journal of Ambient Intelligence and Humanized Computing **11**, 813–825 (2020)
281. Vrbančič, G.: Phishing Websites Dataset. Mendeley Data. DOI: 10.17632/72ptz43s9v.1 (2020)
282. Canaadian Institute of Cybersecurity: URL dataset (ISCX-URL2016). https://www.unb.ca/cic/datasets/url-2016.html (2016)
283. Farsight Inc: Farsight SIE,. https://www.domaintools.com/resources/user-guides/?_resources_products=sie
284. Ariyadasa, S., Fernando, S., Fernando, S.: Phishrepo dataset. Mendeley Data. DOI: 10.17632/ttmmtsgbs8.4 (2022)
285. Lin, Y., Liu, R., Divakaran, D.M., Ng, J.Y., Chan, Q.Z., Lu, Y., Si, Y., Zhang, F., Dong, J.S.: Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages. In: 30th USENIX Security Symposium (USENIX Security 21), pp. 3793–3810. USENIX Association, ??? (2021). https://www.usenix.org/conference/usenixsecurity21/presentation/lin
286. Feng, J., Zou, L., Ye, O., Han, J.: Web2vec: Phishing webpage detection method based on multidimensional features driven by deep learning. IEEE Access **8**, 221214–221224 (2020)
287. Kumar, S.: Detect Malicious URL using ML. https://www.kaggle.com/code/siddharthkumar25/detect-malicious-url-using-ml (2019)
288. Satish Yadav: Phishing Dataset UCI ML CSV. https://www.kaggle.com/datasets/isatish/phishing-dataset-uci-ml-csv (2020)
289. AK., S.: Malicious and Benign Webpages Dataset. PubMed. DOI: 10.1016/j.dib.2020.106304 (2020)
290. William W. Cohen: Enron Email Dataset. https://www.cs.cmu.edu/~enron/ (2020)
291. Dragomir Radev: CLAIR collection of fraud email (Repository).

https://aclweb.org/aclwiki/CLAIR_collection_of_fraud_email_(Repository) (2008)

292. Almeida, T.A., Hidalgo, J.M.G., Yamakami, A.: Contributions to the study of sms spam filtering: new collection and results. In: Proceedings of the 11th ACM Symposium on Document Engineering, pp. 259–262 (2011)
293. M Yasser H : Spam Emails Dataset. https://www.kaggle.com/datasets/yasserh/spamemailsdataset (2021)
294. Akashsurya and Gokhan Kul: Phishing Email Collection. https://www.kaggle.com/akashsurya156/phishing-paper1 (2019)
295. Hina, M., Ali, M., Javed, A.R., Ghabban, F., Khan, L.A., Jalil, Z.: Sefaced: Semantic-based forensic analysis and classification of e-mail data using deep learning. IEEE Access **9**, 98398–98411 (2021)
296. Diegoocampoh Ocampo: MachineLearningPhishing. https://github.com/diegoocampoh/MachineLearningPhishing (2017)
297. J Nazario: Nazario Phishing Corpus. https://monkey.org/~jose/phishing/ (2005)
298. Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C.D., Stamatopoulos, P.: Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. arXiv preprint cs/0009009 (2000)
299. Cormack, G.V., Gómez Hidalgo, J.M., Sánz, E.P.: Spam filtering for short messages. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, pp. 313–320 (2007)
300. spamassasin: Index of /old/publiccorpus. https://spamassassin.apache.org/old/publiccorpus/
301. Gholampour, M.P., Verma, R.M.: IWSPA-2023-Adversarial-Synthetic-Dataset. https://github.com/ReDASers/IWSPA-2023-Adversarial-Synthetic-Dataset (2023)
302. Almeida, T., Hidalgo, J.: SMS Spam Collection. https://archive.ics.uci.edu/dataset/228/sms+spam+collection (2012)
303. Yerima, S.Y., Bashar, A.: Semi-supervised novelty detection with one class svm for sms spam detection. In: 2022 29th International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 1–4 (2022). IEEE
304. Bera, D., Ogbanufe, O., Kim, D.J.: Towards a thematic dimensional framework of online fraud: An exploration of fraudulent email attack tactics and intentions. Decision Support Systems **171**, 113977 (2023)
305. El Aassal, A., Baki, S., Das, A., Verma, R.M.: An in-depth benchmarking and evaluation of phishing detection research for security needs. Ieee Access **8**, 22170–22192 (2020)
306. Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C.D., Stamatopoulos, P.: A memory-based approach to anti-spam filtering for mailing lists. Information retrieval **6**, 49–73 (2003)
307. Metsis, V., Androutsopoulos, I., Paliouras, G.: Spam filtering with naive bayes-which naive bayes? In: CEAS, vol. 17, pp. 28–69 (2006). Mountain View, CA
308. littleRound: 19 Fall Spear Phishing Detection. https://www.kaggle.com/c/19fall-spear-phishing-detection/ (2019)
309. Abhishek Verma: Fraud Email Dataset. https://www.kaggle.com/datasets/llabhishekll/fraud-email-dataset (2018)
310. Cyber Cop: Phishing Email Detection. https://www.kaggle.com/dsv/6090437 (2023)
311. UCI Machine Learning and Esther Kim: SMS Spam Collection Dataset. https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset (2016)
312. Yang, C., Harkreader, R., Gu, G.: Empirical evaluation and new design for fighting evolving twitter spammers. IEEE Transactions on Information Forensics and Security **8**(8), 1280–1293 (2013)
313. Wu, T., Wen, S., Xiang, Y., Zhou, W.: Twitter spam detection: Survey of new approaches and comparative study. Computers & Security **76**, 265–284 (2018). doi:10.1016/j.cose.2017.11.013
314. Cresci, S., Lillo, F., Regoli, D., Tardelli, S., Tesconi, M.: $fake : Evidenceofspamandbotactivityinstockmicroblogsontwitter.$ Proceedings of the International AAAI Conference on Web and Social Media **12**(1) (2018)
315. Feng, S., Wan, H., Wang, N., Li, J., Luo, M.: Twibot-20: A comprehensive twitter bot detection benchmark. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 4485–4494 (2021)
316. Jain, C.: Training data 2 csv UTF. https://www.kaggle.com/datasets/charvijain27/training-data-2-csv-utfcsv (2018)
317. Recruitment Scam. https://www.kaggle.com/datasets/amruthjithrajvr/recruitment-scam
318. Yang, Y., Zhang, Y., Zhu, C.: Improved job scam detection methods using machine learning and resampling techniques. In: 2023 9th International Conference on Systems and Informatics (ICSAI), pp. 1–5 (2023). IEEE
319. Rayana, S., Akoglu, L.: Collective opinion spam detection: Bridging review networks and metadata. In: Proceedings of the 21th Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 985–994 (2015)
320. Rayana, S., Akoglu, L.: Collective opinion spam detection using active inference. In: Proceedings of the 2016 Siam International Conference on Data Mining, pp. 630–638 (2016). SIAM
321. Liu, W., He, J., Han, S., Cai, F., Yang, Z., Zhu, N.: A method for the detection of fake reviews based on temporal features of reviews and comments. IEEE Engineering Management Review **47**(4), 67–79 (2019)
322. Asghar, N.: Yelp dataset challenge: Review rating prediction. arXiv preprint arXiv:1605.05362 (2016)
323. Ott, M., Cardie, C., Hancock, J.T.: Negative deceptive opinion spam. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 497–501 (2013)
324. Eler, M.M., Orlandin, L., Oliveira, A.D.A.: Do android app users care about accessibility? an analysis of user reviews on the google play store. In: Proceedings of the 18th Brazilian Symposium on Human Factors in Computing Systems, pp. 1–11 (2019)
325. Obie, H.O., Hussain, W., Xia, X., Grundy, J., Li, L., Turhan, B., Whittle, J., Shahin, M.: A first look at human values-violation in app reviews. In: 2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS), pp. 29–38 (2021). IEEE
326. Yelp Open Dataset. https://www.yelp.com/dataset