IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 February 2025, accepted 17 March 2025, date of publication 26 March 2025, date of current version 3 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3555157

SURVEY

SoK: Grouping Spam and Phishing Email Threats for Smarter Security

TARINI SAKA[1,2], KAMI VANIEA[3], (Member, IEEE), AND NADIN KÖKCIYAN[2]

[1]Faculty of Computer Science, Ruhr University Bochum, 44801 Bochum, Germany
[2]School of Informatics, The University of Edinburgh, EH8 9YL Edinburgh, U.K.
[3]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada

Corresponding authors: Tarini Saka (tarini.saka@ruhr-uni-bochum.de) and Kami Vaniea (kami.vaniea@uwaterloo.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Award RGPIN-2024-06737; and in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the Cluster of Excellence Cyber Security in the Age of Large-Scale Adversaries, CASA (EXC 2092 - 390781972).

ABSTRACT Emails are a vital form of communication, owing to their open nature, which allows any individual to send emails to anyone else without centralized monitoring. While this has facilitated the widespread adoption of email, it has also inadvertently facilitated malicious activities, such as spam and phishing attacks, which pose a serious threat to the security of organizations worldwide. The volume of such emails is growing at an alarming rate, leading to security researchers finding new ways to protect their organizations. To develop effective protection, it's essential to identify commonalities among emails, such as whether they originate from the same attacker, contain similar wording, or promote nearly identical products. The commonalities used in research to group emails can vary significantly. While the range of research is laudable, the absence of consistent language, datasets, and features can make understanding the results and limitations of this field very challenging. In this systematic literature survey, we looked at 23 research articles on grouping spam and phishing emails, focusing on two foundational aspects (definition of a group and use case) and four methodological aspects (dataset, input features, clustering or grouping algorithms, and evaluation strategies). We propose three definitions of ''campaign'' representing how researchers approach the groupings: source-based, scam-based, and response-based. Furthermore, we discuss the various features and algorithms that have been utilized in relation to the goals of the researchers and highlight the key takeaways and recommendations for future work.

INDEX TERMS Phishing, spam, email security, campaigns, social engineering, email grouping, literature review.

I. INTRODUCTION

Email is a vital form of communication whose value is partially due to its open nature that allows anyone to send an email to anyone else without centralized monitoring; a feature that both enables wide adoption and malicious activities. Among the most prevalent forms of problematic emails are spam and phishing emails. *Spam* typically refers to unsolicited bulk emails, often intended for commercial purposes, which are not inherently malicious but can be annoying and violate laws [1]. *Phishing*, on the other hand, is a deliberate attempt by a malicious actor to obtain sensitive information or trick someone into installing malicious software with the goal of harming a company or individual, typically by disguising the message as a legitimate communication from a trusted source [2]. Although spam and phishing emails differ in terms of their objectives, they share certain similarities. Both are semi-adversarial in nature, with the email sender intentionally trying to bypass email filters meant to prevent the user from receiving the message. They also involve some type of brand impersonation, with spam potentially being sent by a contracted third party and phishing aiming to impersonate an existing trusted sender.

Over the past few years, there has been an alarming surge in the volume of spam and phishing emails worldwide. In 2023, over 160 billion spam emails were sent every

The associate editor coordinating the review of this manuscript and approving it for publication was Somchart Fugkeaw.

day [3]. The year 2023 also saw a record number of phishing attacks, marking the highest total ever documented by the Anti-Phishing Working Group [4]. According to Proofpoint [5], 8 in 10 organizations (84%) experienced at least one successful email-based phishing attack in 2022, with direct financial losses as a result increasing by 76% compared to 2021. In fact, phishing was the leading infection vector for attacks in 2022, with 41% of incidents remediated by X-Force using phishing to gain initial access [6]. Phishing and spam have been proven to be costly for consumers, as they require substantial resources such as time and money [7]. Thus, it is crucial to address these issues at scale in automated ways, if possible. It is important to note that both phishing and spam are often sent out in batches, and similarities can be observed in the language used, product marketed, sending servers, and linked pages [8], [9], [10]. Emails that share such similarities are often referred to as a ''campaign,'' however, there is no clear consensus on the definition of this term or the best practices for automatically detecting one. In this survey, we examine various studies that focus on grouping spam and phishing emails and explore what it means for an email to be part of a campaign, as well as the current best practices for identifying campaigns using various types of features and algorithms.

Phishing and spam are multi-disciplinary issues that are handled through a range of mitigation techniques, from technical approaches, such as automatic filtering [11], [12], to more societal approaches, such as laws and public policy [10], [13]. The most obvious technical approach is the automatic blocking of unwanted and harmful emails, which operates on the individual email level to assess whether an email should be blocked. However, many mitigation approaches can benefit from a deeper understanding of email groups, rather than individual emails. By analyzing email groups, security systems can recognize broader patterns and associations that would be missed in isolation. For example, identifying spam sent by a single author, with the goal of building evidence for prosecution. Analyzing shared characteristics—like language patterns or IP addresses—used by attackers across multiple emails strengthens the evidence trail for law enforcement and facilitates community and industry-wide threat intelligence sharing. Such email groupings are well aligned with AI approaches, in that they require emails with similar features, or possibly similar latent features, to be grouped together so that meaning can be associated with the group. The word ''campaign'' is often associated with sets of emails that are connected in some way, such as being from the same sender, having similar content, or even using the same scam wording and being sent within the same short time frame. There are many different ways in which the similarities observed in a campaign can be defined; consequently, there is a wide range of possible definitions for the term. This lack of definition is problematic for research because it makes finding related work more challenging and consequently harder to build on prior work. Researchers may also find it challenging to identify which sets of features or approaches best align with their specific subject goals, as opposed to the general goal of identifying campaigns.

In this paper, we conduct a comprehensive systematization of existing research on detecting groups of phishing and spam emails in an effort to understand the current state of the art. This study is motivated by the prevailing trend in which many email attacks manifest themselves as mass campaigns involving the dissemination of bulk emails based on templates. Our primary objective is to gain insight into the types of groups that researchers aim to identify and their motivations for doing so. In particular, we aim to identify the different meanings of *email campaign*, a widely used term in the security literature. By establishing a precise set of definitions, we can better collate similar works and devise effective detection algorithms aligned with different definitions. Additionally, we aim to examine the methodological approaches employed by researchers, including datasets, grouping algorithms, input features, and evaluation strategies. Finally, we provide useful recommendations for researchers to guide future studies. With this article, we hope to provide useful directions for future studies on spam and phishing interventions and shed light on the following research questions:

RQ1: *How do researchers define the email groups (such as email campaigns) they aim to identify?*

RQ1.1: *What are some clear definitions of these groups that could help researchers use more uniform language?*

RQ2: *What methodological approaches are commonly utilized to identify spam and phishing campaigns in existing literature?*

RQ3: *What are the limitations and research gaps identified in the existing literature and how can they inform future research?*

We identified 23 publications that focused on grouping spam or phishing emails. These papers have a wide range of research goals and accompanying definitions of the groups they aimed to identify. We were able to abstract these into three broad types: 1) *source-based* campaigns which are a group sent from a single source as part of a single attack; 2) *scam-based* campaigns which are disseminated with a specific purpose and share common narrative; and 3) *response-based* campaigns which can be effectively responded to as a collective incident rather than individual incidents. Furthermore, we summarize the proposed use cases in order to understand the range of applications for such efforts, and recognize four prevalent applications: 1) identifying the common spammer; 2) identifying botnets; 3) profiling attackers; and 4) reducing the load of manual analysis. The identified definitions and use cases also determined which features and algorithms were the most effective. Our analysis also provides a set of recommendations for future work in the field. First, it is crucial for publications to define terms like 'campaign' or 'group' clearly, as understanding these definitions is key to interpreting the implications of the

work. Approaches that are effective in one domain may not necessarily translate to another. Additionally, given recent advancements in context modeling, emerging technologies such as large language models (LLMs) should be explored to better capture the contextual nuances of emails. Establishing a common benchmarking or comparison framework, along with improving access to public datasets, would significantly advance research efforts. Finally, there is a pressing need for standardized data labeling methods to ensure consistency and reliability. Our work makes several key contributions to research on malicious email campaigns:

1) **Systematization of Knowledge:** We present the *first-ever systematization of knowledge (SoK)* on grouping malicious emails to identify coordinated campaigns or attack groups. Given the increasing scale and frequency of large-scale attacks, our work provides a structured understanding of this emerging threat landscape.
2) **Formalization of Key Definitions for Consistency in Research:** We propose *three formal definitions* to establish a common terminology. These definitions help ensure clarity, consistency, and precision in future research on campaign identification.
3) **Comprehensive Analysis of Features and Methodologies:** We conduct an *in-depth analysis* of the existing literature, examining the methodologies and features used for campaign identification. This synthesis highlights effective approaches, identifies critical limitations, and uncovers research gaps that must be addressed to advance the field. Additionally, we analyze the features used in prior studies and their effectiveness in different use cases, offering insights into how various clustering and campaign identification methods align with specific applications.
4) **Guidance for Future Research and Practical Applications:** We provide *actionable recommendations* for future researchers by outlining key areas that need further investigation.

II. MOTIVATION

Existing research has primarily focused on detecting individual spam and phishing emails as they are usually filtered one at a time by a mail server. However, grouping spam and phishing emails also holds significant potential, enabling researchers and practitioners to analyze recurring patterns, uncover tactics used in malicious campaigns, and identify broader trends in email-based threats. Although many organizations and email security providers implement clustering techniques, little is known about the underlying mechanisms driving these processes. Prior literature reviews do not focus on grouping or clustering emails based on shared characteristics. Therefore, we conducted a review of research articles to gather insights into this emerging area.

The field of email security, particularly in spam and phishing mitigation, has a robust foundation of research, with numerous literature reviews and surveys examining various defense mechanisms and threat landscapes. Studies have focused on identifying and filtering malicious or unsolicited emails to protect users from scams, malware, and phishing attacks [10], [11], [12], [14], [15], [16], [17], [18]. For example, Khonji et al. [10] reviewed the literature on phishing detection and categorized four primary detection solutions: blacklists, rule-based heuristics, visual similarity, and machine-learning classifiers. Their findings highlight that machine learning techniques achieve high accuracy, particularly when analyzing similar features used by rule-based methods. Salloum et al. [15] specifically examined the use of Natural Language Processing (NLP) techniques in phishing detection to further the understanding of state-of-the-art methods in this area. Researchers have also conducted extensive reviews on the human component of spam and phishing, analyzing how users perceive and respond to potential threats [2], [19], [20], [21]. Additional surveys have assessed various attack vectors and techniques used in email-based attacks [9], [12], [22], [23]. For instance, Lee et al. [9] classified phishing attacks into categories like spoofed emails, vulnerabilities in email content, and file spoofing, and conducted detailed analyses to profile a phishing group targeting individuals in defense and security sectors, noting a rapid escalation of these attacks in recent years.

A. CAMPAIGN IDENTIFICATION

In recent years, threat analysts have faced growing challenges in analyzing new threats and identifying emerging trends due to the overwhelming volume of security data. Incident grouping or clustering is a popular approach to managing this volume [24], [25], [26], [27], [28], [29]. For instance, Tang et al. [26] clustered *SMS spam messages* reported on Twitter into groups based on their URLs and subsequently profiled the spam campaigns. Their findings reveal the presence of cross-language campaigns, suggesting that organized spam operations may target different regions. Additionally, they observed that SMS spam messages often share templates across various target services, indicating that either a single spam campaign targets multiple services or that different campaigns utilize a common text generator. Phillips and Wilder [27] analyzed *public online and blockchain data* to investigate cryptocurrency scams, applying the DBSCAN clustering technique to identify patterns in advance-fee and phishing scams. Their study reveals that the same entities operate multiple scams, manipulating online infrastructure and blockchain activity to appear legitimate. Additionally, funds analysis shows victims often send money from fiat-accepting exchanges, while scam operators launder proceeds through exchanges, gambling sites, and mixers. Cova et al. [28] conducted a longitudinal analysis of the rogue *antivirus threat ecosystem* and applied attack attribution techniques to correlate campaigns likely to be associated with the same individuals or groups, identifying 127 rogue software campaigns across 4,549 domains. Thonnard and Dacier [29] conducted a strategic analysis of *spam botnet operations*, examining inter-relationships among botnets through their

spam campaigns to identify operational patterns. Their contributions included a long-term analysis of botnet behavior across over one million spam records, demonstrating the value of identifying campaigns.

B. INDUSTRY PRACTICES

Many email security providers now employ advanced AI-driven techniques to mitigate phishing threats, including using clustering to remediate mass attacks. For instance, Microsoft Defender for Office 365 [30] uses automated investigations to analyze email threats by clustering emails based on attributes like sender and content, detecting malicious patterns, and assessing threat levels. It examines URLs and attachments to identify malware and phishing attempts. This enables security analysts to refine and remediate threats by creating clusters and providing detailed query results, ensuring comprehensive threat detection and response for email-based attacks. Similarly, Ironscale's AI-powered platform (MSOAR) [31] uses machine learning to identify threats and cluster-related incidents, automatically triggering remediation processes. Employee-reported suspicious emails are quickly analyzed, and if further investigation is needed, manual review processes are streamlined, often concluding within one minute. PhishER Plus by KnowBe4 [32] enhances email security by automating the response to threats and using clustering to group and categorize suspicious emails based on patterns, rules, tags, and actions. This clustering capability is meant to identify widespread phishing attacks within an organization and enables security teams to respond more quickly and effectively to these threats by recognizing patterns across multiple incidents. While they use terms like AI and NLP, little information is provided on the specific methods and features involved, creating a gap in transparency regarding how these sophisticated mechanisms truly operate. This highlights that email grouping is increasingly utilized by organizations, yet there remains a limited, comprehensive understanding of the approaches being explored by researchers and the challenges within this field.

III. METHODOLOGY

Section II shows that clustering incidents and analyzing them as groups provided critical insights into campaign patterns and emerging attack techniques, which may not captured by analyzing them in isolation. While clustering methods are commonly used by many email security providers, the academic community has limited access to the methodologies and inner workings of these techniques. Our work bridges that gap, offering smaller organizations valuable knowledge to implement or understand these methods without relying solely on proprietary solutions. A consolidated review of email grouping will establish a baseline for future research in clustering and campaign identification, offering a framework for evaluating and advancing machine learning techniques in email security.

In this section, we provide a detailed account of the methodology used in conducting the review. First, we describe the construction of the search query, we then explain the search process, outlining the databases utilized, and the inclusion and exclusion criteria applied.

A. CONSTRUCTING THE SEARCH QUERY

In order to construct our search query, we focus on the following three aspects: (i) *Malicious aspect:* These are terms meant to capture the malicious or problematic aspect of emails and include all papers that talk about either 'spam', 'phishing', and 'malicious' emails. (ii) *Email data:* These terms are used to make sure we only include articles that use email data. Other common forms of phishing or spam include phone calls, text messages, social media messages, and website data. (iii) *Groups:* These terms are meant to capture studies that perform some kind of grouping of emails. We include the terms 'group' and 'cluster', and 'campaign'. The latter is a commonly used term for groups of related spam or phishing emails. The final boolean query to identify articles for analysis was as follows:

```
"(malicious* OR phish* OR spam*) AND
(email*) AND (group* OR campaign* OR
cluster*)"
```

B. REVIEWING THE PAPERS

We used the ACM Digital Library, IEEE Xplore, and Web of Science databases to capture publications across a range of research venues. The search was limited to peer-reviewed studies in English that were available as of September 2023. The search query was identical across databases and was searched against the titles and abstracts of all included articles. Furthermore, we manually searched through the titles of journals and conferences that had not been covered by the previous three databases. This included the USENIX Network and Distributed System Security Symposium (NDSS) and the USENIX Security Symposium, which added 2 articles. Our systematic literature review follows the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)* guideline, developed to help authors in improving the reporting of systematic review results [33].

The inclusion criteria for papers were that they had to: (1) be a full paper in a peer-reviewed publication, (2) focus on phishing or spam emails, (3) attempt to cluster or group the emails, (4) identify meaningful groups of emails, and (5) contain methodological details about the clustering or grouping approach used.

We exported the search results of each database (ACM Digital Library: 73; IEEE Xplore: 140; Web of Science: 248; USENIX: 2) resulting in an initial set of 463 papers, which reduced to 349 after removing 114 duplicates. The lead researcher manually screened the titles and abstracts of these papers against the inclusion criteria resulting in 43 papers. These were further analyzed in detail by reviewing the full texts resulting in 20 papers (Table 1 details exclusion reasons). The references of the 20 papers were then reviewed

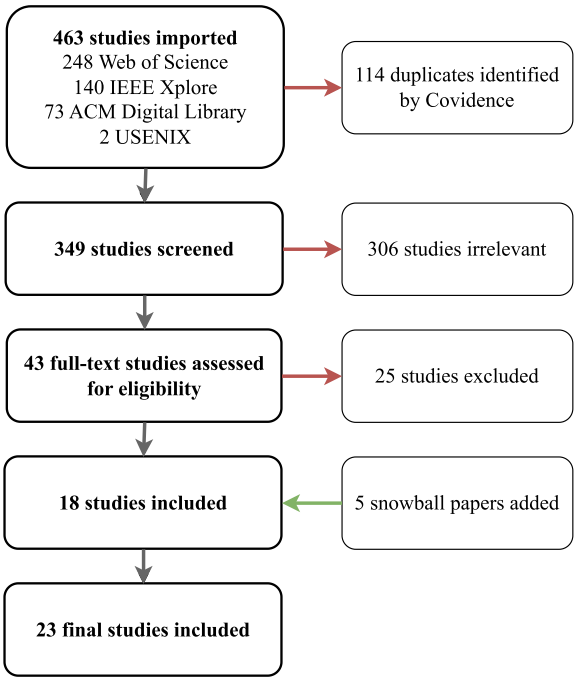


FIGURE 1. PRISMA diagram of literature screening process.

TABLE 1. Exclusion reasons encountered during the full-text review and the corresponding count of excluded papers.

Exclusion Criteria	Number of papers
Poster, Demo or Theoretical paper	2
Not available online	1
Unclear features or methodology	3
No email dataset	6
No grouping of emails	5
Not about identifying groups	8

for relevant-sounding titles followed by a review of their full texts, resulting in an additional 3 papers. Throughout the process, the lead researcher identified and discussed any paper where the inclusion was unclear with the other authors. The final dataset consisted of 23 papers. Figure 1 shows the flowchart that details the number of screened, excluded, and included articles following the PRISMA statement.

IV. LITERATURE ANALYSIS AND RESULTS

To systematize our analysis, we propose six criteria that we applied to the identified literature and used the Covidence software[1] to conduct the analysis. The identified six criteria are as follows: (C1) definition of groups, (C2) use case or motivation, (C3) dataset, (C4) feature set, (C5) algorithms, and (C6) evaluation. These criteria were chosen to better understand the problems these publications aimed to solve (C1-C2) and the approaches employed (C3-C6) to answer RQ3. We now present a detailed analysis of the studies with respect to these criteria and identify interesting trends and limitations. Table 2 presents an overview of the methodological criteria (C3-C6) for the whole literature sample.

[1]https://www.covidence.org/

A. C1: DEFINITION OF GROUPS

The terms 'campaign', 'cluster', and 'group' frequently appear in security literature, yet there is no single accepted definition for them. This criterion (C1) answers the following question: *How do the selected papers, either explicitly or implicitly, define 'campaign', 'cluster', or 'group'?* Among these terms, 'campaign' is commonly used in security, whereas 'cluster' and 'group' are commonly used terms in AI/ML literature. Hence, we start our analysis by looking at papers that use the term 'campaign', and analyse how these papers define the term, focusing on the associated characteristics and attributes. Through this section, we address in detail our first research question (RQ1).

1) EMAIL CAMPAIGNS

The term 'campaign' was used by 14 out of the 23 papers to refer to the set of emails they were aiming to group. We further identified three aspects that were commonly used in their definition of an email campaign: *Common source, Similarities, and Variations.* We present an overview of our findings in Table 3. Definitions of campaigns contained one or more of the following:

1) **Common source (11 papers).** Campaign emails share a common source. Some researchers define common source based on the sender such as the same attacker or spammer [37], [41], [47], [48], [49] or the same group of attackers [35]. The source could also be a single spamming network [42], the same botnet [55], or the same generator tool [40]. Wei et al. [53] imply common origin by using the subject line as a feature based on the presumption that two emails sharing the same subject are more likely to have originated from the same source.
2) **Similarities (11 papers).** These definitions are based on the assumption that attackers are employing a single email template as the foundation for each attack which naturally results in similarities. Haider and Scheffer [40] stated that spam dissemination tools produce emails according to probabilistic templates. Althobaiti et al. [35] state that attackers create one sophisticated email and then base an attack on it. The similarities within campaign emails can manifest in various aspects, including email content [34], [57], writing style [41], message format [49], or structural similarity [47], [48], [49]. In the case of spam email campaigns, several researchers assert that emails within the same campaign share the same goal or purpose and often focus on promoting a single product or service [37], [39], [55].
3) **Variations (7 papers).** The possibility of variation within emails was also included in campaign definitions. To evade blacklists and filters, attackers sometimes introduce subtle variations to emails within a campaign,

TABLE 2. Overview of the C3-C6 criteria. The research paper, dataset used (Phish/Spam), feature classes used (Header, URL, Subject, Attachment, Content), algorithms used for campaign identification, and types of evaluation strategies (Internal, External, Manual).

Paper	Dataset		Feature Classes					Clustering/Grouping Algorithms	Evaluation		
	P	S	H	U	S	A	C		I	E	M
Alazab *et al.* [34]		•					•	NUANCE Authorship Attribution			•
Althobaiti *et al.* [35]	•		•	•	•	•	•	Meanshift and DBSCAN	•	•	
Calais *et al.* [36]		•		•	•		•	Frequent-Pattern Tree			
Chen *et al.* [37]		•	•				•	Fuzzy Hashing [38]	•		
Dinh *et al.* [39]		•	•	•	•	•	•	Frequent-Pattern Tree			
Haider *et al.* [40]		•					•	Bayesian Hierarchical Clustering		•	
Halder *et al.* [41]		•		•			•	K-Means and Expectation Maximization		•	
Halder *et al.* [42]		•	•	•	•	•	•	Nearest Neighbour algorithm K-Means and Expectation Maximization	•		•
Han *et al.* [43]	•		•		•	•	•	Nearest Neighbors Affinity Graph		•	
Husna *et al.* [44]		•	•				•	Hierarchical and K-means clustering		•	
Saka *et al.* [45]	•		•	•	•	•	•	K-Means, DBSCAN, and Agglomerative Hierarchical Clustering	•	•	
Seifollahi *et al.* [46]	•						•	K-Means, DCClust, MS-MGKM and INCA		•	
Sheikhalishahi *et al.* [47]		•	•	•	•	•	•	CCTree algorithm	•	•	
Sheikhalishahi *et al.* [48]		•	•	•	•	•	•	CCTree algorithm	•	•	
Sheikhalishahi *et al.* [49]		•	•	•	•	•	•	CCTree algorithm	•	•	
Shen and Thonnard [50]	•	•	•		•	•		MR-TRIAGE (Multi-Criteria Clustering)		•	
Song *et al.* [51]		•		•				Text-Shingling Optimized K-Means Clustering		•	
Song *et al.* [52]		•		•				Text-Shingling and IP Clusterer		•	
Wei *et al.* [53]		•		•	•			Agglomerative Hierarchical Clustering Weighted Connected Components			•
Woo *et al.* [54]		•	•					K-Means Clustering			•
Xie *et al.* [55]		•		•				AutoRE Framework			•
Yearwood *et al.* [56]	•			•			•	K-Means, Modified Global k-means		•	
Zhuang *et al.* [57]		•	•	•			•	Text Shingling and Connected Components			

altering elements such as the sender ID, email body, or email subject [35]. The most frequently observed variations occur in the email text [35], [37], [49], [57]. Attackers usually obfuscate the email content such that each email has slightly different text from the others to evade detection. This often involves altering frequently filtered words or inserting obfuscated terms [37]. Another strategy includes the insertion of random text or links [48], [49]. Additionally, attackers may leverage spam dissemination tools that generate emails based on probabilistic templates [40]. According to Halder et al. [42], spammers register multiple domains to minimize the risk of domain blacklisting.

2) NON-CAMPAIGN GROUPS

Of the identified studies, eight did not explicitly use the term 'campaign'. Some defined groups based on email content such as similar scams [45], URLs [51], and resolved IP addresses [52]. Others had the goal of identifying common senders, so they defined groups based on their own analysis such as authorship analysis to find common authors [46], clustering content features to identify groups exhibiting the same modus operandi [56], or simply clustering with the expectation that clusters likely aligned with malicious groups [44], [54]. Finally, one paper contained no explicit definition of what was meant to be in a 'cluster' [50].

Takeaway 1: No matter how much the content (text, sender address, URL domain) of an email is obfuscated, all emails belonging to a single campaign have to share a set of ''common identifiers'' to achieve their initial purpose. This implies that the fundamental ruse or scam remains consistent throughout. For example, emails within a spam campaign selling a product or phishing campaigns regarding parcel deliveries typically share common descriptive elements.

TABLE 3. An overview of papers utilizing the term 'Campaign' to refer to their targeted clusters. We dissect their campaign definitions into three aspects: common source or origin, email similarity, and email variation.

Lead author/Year	Study	Definition		
		Common Source/Origin	Email Similarity	Email Variation
Calais (2008)	[36]	•	•	•
Althobaiti (2023)	[35]	•	•	•
Dinh (2015)	[39]	•	•	
Zhuang (2008)	[57]		•	•
Han (2016)	[43]	•		
Chen (2014)	[37]	•	•	•
Haider (2009)	[40]	•	•	•
Alazab (2013)	[34]		•	
Wei (2008)	[53]	•		
Sheikhalishahi (2020)	[49]	•	•	•
Sheikhalishahi (2016)	[48]	•	•	•
Sheikhalishahi (2015)	[47]	•	•	
Xie (2008)	[55]	•	•	
Halder (2011)	[41]	•	•	
Halder (2012)	[42]	•		•

In other words, the context of the email remains consistent, and an accurate representation of context can play a crucial role in identifying campaigns.

B. C2: USE CASE OR MOTIVATION

Most of the papers we analyzed motivate the work by presenting a gap in the literature or a use case that the proposed approach would address. To understand this criterion better, we present the most common high-level use cases and motivations for grouping malicious emails as follows:

1) IDENTIFY BOTNETS

A botnet is a group of Internet-connected devices running automated programs called bots that can be used to perform Distributed Denial-of-Service attacks, steal data, and send spam and phishing messages. Security teams prioritize detecting botnets, often by identifying multiple bots in shared spam campaigns. Identifying botnets was the primary motivation in six of the papers [37], [44], [52], [54], [55], [57].

Woo et al. [54] state that responding to every incoming spam is an impractical task and suggest that by effectively clustering spam emails and discerning those originating from botnets, organizations can design more adaptable incident response plans. According to Zhuang et al. [57], there exists anecdotal evidence that spam is a driving force in the economics of botnets. They state that a common strategy for monetizing botnets is to send spam emails, where spam is defined liberally to include traditional advertisement emails, phishing emails, emails with viruses, and other unwanted emails. Hence, they aimed to map botnet membership and other botnet characteristics using spam traces. Similarly, Chen et al. [37] stated that an effective approach to inferring spamming botnets is to first identify spam emails belonging to the same campaigns. Song et al. [52] state that to cope with spam-based attacks, it is important to characterize their infrastructure and how they are grouped. They assert that the majority of spam emails sent by bots contain URLs that direct recipients to malicious web servers, and hence present a new spam clustering method that relies on IP addresses extracted from URLs. Husna et al. [44] asserted that botnet detection can significantly improve the control of unwanted traffic. In their work, they investigated the behavioral patterns of spammers based on the underlying similarities in spamming. Xie et al. [55] focused on characterizing spamming botnets by leveraging both spam payload and spam server traffic properties.

2) IDENTIFY COMMON SOURCE OR SPAMMER

One effective strategy for addressing the growing problem of spam and phishing is to trace the source of the attacks and hold the responsible parties accountable. Five studies [41], [46], [47], [48], [49] had this motivation for their research. According to Sheikhalishahi et al. [47], [48], [49], finding spammers is important not only to tackle the source of the problem but also to legally prosecute those responsible. They assert that early analysis of correlated spam emails is crucial to identifying spammers and propose an algorithm to automate this challenging and time-consuming task. This process is also called relative attribution or attack attribution. Seifollahi et al. [46] used authorship analysis techniques to

cluster phishing emails for identifying attacks performed by the same person to enable the tracking of perpetrators and increase evidence-gathering opportunities. Halder et al. [41] do the same using the premise that spam originating from the same spammer will have similarities in writing style, irrespective of the bot used.

3) REDUCE HUMAN-LOAD AND MANUAL ANALYSIS

Humans play a crucial role in the mitigation of large-scale email attacks. These roles include support staff tasked with managing reported emails within organizations, researchers engaged in the study of email and spammer behavior, and security professionals who analyze emails to extract valuable insights. Effectively addressing the growing threat of malicious emails requires the development of advanced methodologies and tools to alleviate the burden on human resources engaged in manual analysis, and formed the motivation in five of the identified papers [35], [39], [43], [45], [46], [53].

Dinh et al. [39] emphasized the impracticality of manually analyzing the overwhelming volume of spam emails and campaigns for cyber-crime investigations and argued the need for automatic techniques and tools to facilitate efficient analysis. They proposed a software framework for on-the-fly spam campaign detection, analysis, and investigation. Similarly, Han et al. [43] argued that attributing spear-phishing emails to known campaigns requires considerable time and manual effort and hence only a small number of suspicious emails are being attributed with the majority not being investigated. They developed an automated spear-phishing campaign attribution model to expedite the process and automatically attribute unlabeled suspicious emails to known campaigns or identify newly emerged spear-phishing campaigns. Wei et al. [53] aimed to provide support to law enforcement personnel engaged in the analysis of spam emails through applied data mining techniques that aid humans in finding clues among spam emails.

Another effective way to support human staff involves grouping reported phishing emails into campaigns and assisting IT teams in the efficient analysis of these emails. Such a defense strategy can help organizations take swift and targeted actions and efficiently mitigate large-scale attacks. This concept was investigated by Althobaiti et al. [35] and Saka et al. [45]. Althobaiti et al. [35] attempted to group potential phishing emails into meaningful clusters that can help IT teams analyze potential phishing emails in terms of campaigns and reduce the effort spent on manually checking each email individually. Saka et al. [45] aimed to give IT teams a way to address threats collectively by employing unsupervised clustering techniques to group similar phishing emails based on underlying scams.

4) PROFILING ATTACKERS

Attack profiling involves the systematic analysis of various elements associated with a cyberattack, such as the content of the emails, sender information, malicious payloads, attack infrastructure, victim targeting, and recurring attack patterns. This process aims to create detailed profiles or characteristics of attack attempts, enabling cybersecurity professionals to identify patterns, understand attacker methodologies, and develop effective strategies for detection, prevention, and response. An effective method for constructing attacker profiles involves the aggregation of emails originating from the same attacker. This approach reveals behaviors and characteristics that cannot be noticed when looking at individual emails but become discernible through the patterns formed by the whole set of emails. Five studies in the dataset had this motivation [34], [36], [46], [54], [56].

One common way to profile cyber attacks is through authorship analysis. Alazab et al. [34] employed authorship attribution techniques to profile various attack behaviors that match a group and used these profiles to recognize future attacks by the same groups of spammers, improve spam detection tools, and develop better user education. Similarly, Seifollahi et al. [46] used authorship analysis in criminal profiling to establish characteristics such as age, gender, and country of origin through clues in writing styles. Both of these studies used email text to perform authorship analysis and hence encountered the inherent limitations associated with text-based methods. In contrast, Woo et al. [54] used user behavior-based features to detect malicious spamming groups and provided systematic behavior profiles. They argued that such a behavior profiling strategy enhances detection methods, making it more challenging for spammers to evade identification than text-based methods. Yearwood et al. [56] clustered phishing emails using structural characteristics, the content of emails, and information about their likely origins, and assumed that these correspond to different phishing groups with certain modus operandi. They proposed using the information provided by these clustering results to construct comprehensive group profiles. Calais et al. [36] took a two-stage approach, first identifying spam campaigns and then using the result to characterize how each campaign has exploited the network resources and how their contents have been obfuscated.

***Takeaway** 2:* Grouping malicious emails may serve different goals. For example, individual organizations care about reducing the load on IT workers; while organizations with wider network impact care about identifying botnets and common sources. Law enforcement and those tracking trends care about profiling attackers and collecting evidence. While similar, each goal implies a slightly different definition of what emails should be grouped together.

C. C3: DATASET

The research papers used a wide range of datasets, varying in terms of availability, size, and source. Seventeen papers used spam email datasets, six worked with phishing emails, and one used both, as summarized in the dataset column in Table 2. Only two studies focused on directed or personalized phishing attacks, commonly referred to as spear phishing [43], [50]. Datasets were predominantly private

with only two groups of authors using publicly available datasets [45], [47], [48], [49]. Sheikhalishahi et al. [47], [48], [49] used emails from a publicly available spam archive collected since early 1998 called the Untroubled Archive.[2] Saka et al. [45] used the Nazario phishing corpus, a publicly available dataset of phishing emails collected by a single researcher.[3] Datasets ranged in size from 1277 emails [46] to 97 million emails [36]. The data collection periods also varied greatly, spanning from as short as one day [54] to as long as seven years [45]. Dataset sourced ranged from being collected by email service providers [35], [43], [50], [55], [57], personal email accounts [37], [45], honeypots [36], [47], [48], [49], [54], catch-all email accounts [41], [42], [53], and mail servers [51], [52], [55]. Two studies got data from financial organizations [46], [56].

We observed several aspects of the datasets that affect the generalizability of the findings in these papers. Firstly, the majority of publications relied on private datasets, which inherently limits the reproducibility and verification of the research. While this limitation is somewhat expected due to the sensitive nature of email data, the absence of standardized benchmarking datasets in this field remains a notable issue. Furthermore, semi-public datasets such as the Cambridge Cyber Crime Center dataset of phishing emails[4] which are available under a Non-Disclosure Agreement (NDA) were also notable in their absence. Secondly, the methods used to classify emails as spam or phishing varied, including automation, manual labeling, and the presumption that emails sent to non-existent addresses are indicative of spam or phishing attempts. Such automatic classification of emails introduces a potential bias towards features that are more amenable to algorithmic analysis, thereby potentially skewing the results.

***Takeaway** 3:* In the current research landscape, there is significant variation in the email datasets used, including the source of the data, size of the data, collection period, and year of collection. This poses a significant challenge to the generalizability and comparability of the studies.

D. C4: FEATURE SET

As summarized in the 'Feature Classes' column in Table 2, we define five classes of features that researchers consider while working with phishing/spam emails. 1) *Content (C)*: Features extracted from the body of the email, such as the email text, images, and HTML content. 2) *URLs (U)*: Features extracted from URLs such as URL count, domains, and paths. 3) *Header (H)*: Features extracted from the email header, such as From address and authentication fields. 4) *Subject (S)*: Features from the subject line, such as the number of words. And 5) *Attachment (A)*: Features regarding attachments in emails such as number, type, and size of attachments. We will now discuss each of these classes in detail.

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

[3]https://monkey.org/ jose/phishing/

[4]https://www.cambridgecybercrime.uk/datasets.html

1) CONTENT-BASED FEATURES

The most commonly used feature class was observed to be 'Content'. Seventeen papers used content-based features in their work. Some works used very limited content-based features, such as email layout [36], [39], email length [44], email text language [36], [47], [48], [49], character encoding [39], [43], and presence and types of email elements (HTML content, URLs, images) [35], [36], [44], [49], [56]. As discussed in Section IV-A, campaign emails are defined by similar content and therefore a comprehensive description of the email body can be very useful in achieving good accuracy. Yearwood et al. [56] defined a set of 13 orthographic features, including the size of the text, greeting line, signature, presence of HTML content, and images. The orthographic features mainly consisted of style characteristics that are used to convey the role of words, sentences or sections that describe the email content. Althobaiti et al. [35] defined an extensive set of Body-Based Features derived from the plain text part and the HTML part of the email object to effectively describe the email body. This included types and numbers of email elements, presence and number of images, URLs, HTML tags, scripts, CSS specifications, number of lines, number of words, average word length, and greeting type. Halder, Tiwari, and Sprague [41], [42] defined 57 stylistic features, including total word count, number of punctuations, contractions, obfuscated words, email IDs, and number of new lines. They argued that stylistic features based on the writing style of the authors or attackers can be very effective in achieving good accuracy. Finally, Sheikhalishahi et al. [47], [48], [49] defined a set of 21 categorical features representative of email structure, including the presence of HTML tags, email size, email language, and the number of images in the email text, for both email clustering and classification.

A very important aspect of the email content is the *email text*. This is the part of the email that contains the most information and context about the scam and hence provides a stronger representation of the email. There are many ways to represent the text. Haider et al. [40] represented every email in their dataset by a binary vector of almost 1.92 million attributes that indicate the presence or absence of a word. While this approach captures a significant amount of information, such a large dimensionality can also lead to noise and potential inaccuracies. Similarly, Althobaiti et al. [35] converted the email text into a bag of words and then used Latent Semantic Analysis (LSA)[5] to extract the top ten terms describing the email's content. Zhuang et al. [57] explored the use of text shingling to generate a set of fingerprints that represent the email text. They argue that this method is robust to the different obfuscation methods of different spammers. Chen et al. [37] make a similar

[5]LSA is a technique in natural language processing that analyzes relationships between terms and documents, identifying hidden patterns to improve information retrieval and document similarity assessment by capturing the underlying semantic structure of the text through mathematical transformations.

claim about using fuzzy hashing to compare the content of spam emails. Their approach is based on the assumption that spam emails from one campaign should have a high similarity score among each other and a low score with other emails. Alazab et al. [34] use local n-gram methods to create a document profile for a given email by taking the most distinctive n-grams from that email text. A similar representation was also used by Halder et al. [41], [42], where they used the count of the top bigrams in the email text. Some studies have used more advanced text-embedding techniques to capture semantic and contextual information from email body text. The most common of these was found to be Term Frequency–Inverse Document Frequency (TF-IDF)[6] [41], [42], [46], [56]. However, Seifollahi et al. [46] also incorporated word similarity measures to reduce sparsity and improve accuracy. Han and Shen [43] use Latent Semantic Indexing, an information retrieval technique commonly used in NLP, to find topics in the email subject and body text. Another study that used advanced NLP techniques is the work of Saka, Vaniea, and Kokciyan [45]. They used the BERT transformer to represent the context of the email body text and to create a topic model for their subject line. This study is the only one to use large language models to represent email body text.

These observations collectively highlight the diverse set of features and techniques used in spam clustering, ranging from basic text analysis to advanced natural language processing and semantic similarity measures.

2) URL-BASED FEATURES

This was the next most common feature class, with 15 studies using URL-based features. URLs present in emails are an important criterion for detecting or classifying attacks. The URL destinations are aspects that attackers cannot fully conceal or manipulate and hence provide vital information regarding the attack. The most common URL feature is the number of URLs present in an email [41], [42], [45], [47], [48], [49], [56]. Another common technique is to tokenize URLs into various parts, such as the domain, hostname, path, and parameters, and use each token as a feature [35], [36], [39], [56]. Wei et al. [53] only used the URL domain to create clusters. Other URL-based features include the presence of non-ASCII characters, IP address links, short URLs, and blacklisted links. Some studies used secondary, derived features derived from URLs as input, such as the IP addresses the URL resolves to [52], webpages linked to by URLs [51], domain registration information [35], and spam URL signatures generated by an algorithm [55]. URLs are an excellent source of features, either directly or derived.

[6]TF-IDF is a numerical statistic that reflects the importance of a word in a document relative to a collection of documents. It combines the frequency of a term (word) in a document (TF) with the inverse of its frequency across all documents (IDF), providing a measure that highlights unique and significant terms in a specific document. Higher TF-IDF scores indicate terms that are both frequent in a document and rare across the document collection.

3) HEADER-BASED FEATURES

The header of an email contains metadata and routing information about the email message. It provides vital information about the source, recipient, time, and authentication of an email. Thirteen studies used header-based features in their experiments. The most common type of features extracted from the headers are origin or source information, including From/Sender [35], [42], [43], [45], [50], sender IP address [35], [37], [43], [50], [54], [57], domain registration information [35], and origin country [43]. Recipient information includes recipient addresses and counts [35], [47], [48], [49], [50] and time-related information includes date and time the email was sent [35], [43], [44], [50]. Husna et al. [44] show that spammers demonstrate highly clustering structures based on time-based features such as time of arrival, frequency of email, active time, and inter-arrival time.

4) SUBJECT-BASED FEATURES

The email subject line is the first single-line text that recipients see when they receive an email and is usually a summary of the email's content. Cybercriminals design subject lines in a way that creates urgency, personalization, and pressure to trick victims into clicking on malicious links, downloading malware, and so on. It is generally assumed that emails that are part of the same attack or campaign share similar subject lines [36], [39]. Eleven studies used subject-based features in their research.

The most common representation of the subject line is the number of characters or the length of the subject [35], [43], [47], [48], [49]. Wei et al. [53] used 'subject' in the first iteration of their proposed approach, and merged two clusters if they share a common subject. Similarly, Dinh et al. [39] considered the entire subject line as a feature. Although this approach provides high accuracy, it might be too fine-grained as subject lines within a campaign or group are occasionally altered subtly to evade filters. To overcome this, other researchers have used TF-IDF [35], Latent Semantic Indexing [43], n-gram similarity [50], or even advanced methods like BERT-based topic models [45]. Halder et al. [42] used Part-of-Speech (POS) tagging to assign a grammatical category or part-of-speech label (such as noun, verb, adjective, etc.) to each word in a sentence. They then used the exact similarity of the subject POS tags to cluster emails. Sheikhalishahi et al. [47], [48], [49] used some other relevant features such as the language of the subject line, number of words, presence of common words like ''Re'' or ''Fwd'', and number of non-ASCII characters.

5) ATTACHMENT-BASED FEATURES

Some spam or phishing emails contain attachments, and email recipients are lured into believing that they contain vital information, either about their health, wealth, or career or about important business procedures. Some attachments contain malware or other malicious elements, and opening such an infected attachment can have serious consequences.

Nine studies used attachment-based features for this purpose. The most common representation is the number of attachments present [35], [42], [45], [47], [48], [49], size and type of attachment [35], [42], [43], [47], [48], [49], and the name of the attachment [39], [50].

E. C5: ALGORITHMS

In this section, we discuss the various grouping algorithms employed by researchers to identify meaningful clusters of emails. The choice of an algorithm is a critical decision that involves considering the nature of the task, the characteristics of the data, computational requirements, and the desired interpretability of the model. It requires thoughtful analysis and often involves experimentation to identify the algorithm that performs optimally for a given problem. An ideal algorithm should be able to: (i) efficiently group together emails that share common attributes, (ii) be robust to noise like benign emails and singleton emails, and (iii) be interpretable, as real-world applications need to be explainable to ensure the system works as expected and that justifications could be provided if needed.

1) CLUSTERING APPROACHES

As summarized in Table 2, the most commonly used method for identifying email campaigns and groups is *clustering*. Clustering is an unsupervised technique in machine learning and data analysis that involves grouping together similar data points based on certain characteristics or features. The goal of clustering is to identify patterns, similarities, and structures within a dataset without any predefined labels [58]. Clustering algorithms can be Centroid-based, Density-based, Connectivity-based, or Distribution-based. The most popular of the clustering techniques is *K-Means clustering*. Eight studies employed the *K-Means* algorithm to cluster emails [34], [41], [42], [44], [45], [46], [51], [54], [56]. K-means [59] is a popular centroid-based clustering algorithm, renowned for its simplicity and speed. It uses the Euclidean distance between points in space to identify which points belong together. The main limitation of this algorithm is that it requires the number of clusters as an input parameter, which in most cases is unknown. This parameter can be computed using the Elbow Method [60] and the Silhouette Method [61]. Alazab et al. [34] used the NUANCE authorship attribution algorithm, which utilizes K-Means as an initial clustering algorithm to generate a co-association matrix for the emails. Various modifications of K-Means have also been used in this context. For instance, Song et al. [51] utilized O-means, an optimized spam clustering method based on K-means, Seifollahi et al. [46] used the multi-start modified global k-means (MS-MGKM) [62], and Yearwood et al. [56] used the Modified Global k-means algorithm [63] to cluster emails based on a set of 13 orthographic features.

Another commonly used clustering algorithm is Hierarchical Clustering which seeks to build a hierarchy of clusters [64]. Agglomerative Hierarchical Clustering is the most commonly used hierarchical algorithm and has been used in four studies [34], [44], [45], [53]. This is a bottom-up approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Alazab et al. [34] used agglomerative clustering to cluster the co-association matrix generated in the initial step of NUANCE. Other clustering algorithms used in the dataset are DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [35], [45] and Meanshift algorithm [35].

2) GRAPH-BASED APPROACHES

Graph-based clustering methods involve the use of graph theory concepts and algorithms to partition the data into clusters. In these methods, data points are represented as nodes in a graph, and the edges between nodes encode the relationships or similarities between the corresponding data points. The objective is to identify groups of nodes (clusters) that are more tightly connected to each other than to the rest of the graph. Six studies used graph-based approaches to cluster emails [37], [42], [43], [50], [53], [57]. In these studies, each spam or phishing email is represented as a node. However, a key challenge lies in determining the appropriate method to construct edges, which signify similarity between nodes, and defining a threshold for clusters. For example, Chen et al. [37] calculated similarity using spam hash values, while Halder et al. [42] relied on exact matches of POS-tagged subject lines and sender names with over 70% similarity. Zhuang et al. [57] connected two nodes if the corresponding messages were connected in content or shared the same embedded links.

Many works used the Connected Components algorithm to identify campaigns [50], [53], [57]. In graph theory, a connected component refers to a subgraph where every pair of vertices is connected by a path, with no connections to vertices outside the subgraph. An important note here is that graph-based methods can be used in the semi-supervised setting, which involves using a small set of labeled emails. Han and Shen [43] constructed a K-Nearest Neighbours (k-NN) attribute graph, and then used a semi-supervised technique called label propagation, along with a labeled set, to assign labels to all emails.

3) TREE-BASED APPROACHES

Tree-based methods are a class of machine learning algorithms that make decisions in a hierarchical, tree-like structure. They use a series of if-then rules to generate predictions from one or more decision trees. Five studies use tree-based approaches to cluster emails [36], [39], [47], [48], [49]. Two studies [36], [39] used the Frequent-Pattern Tree algorithm to identify spam campaigns. Sheikhalishahi et al. [47], [48], [49] developed and presented a new framework called the Categorical Clustering Tree (CCTree) that analyzes and clusters large numbers of raw spam emails into spam campaigns. The root of the CCTree contains all the elements to be clustered and each element is described through a set of categorical attributes with finite discrete values. The

CCTree is constructed iteratively through a decision tree-like structure, where the leaves of the tree are the desired clusters.

F. C6: EVALUATION

Evaluation or validation of results is a crucial step for assessing the performance of a model and determining how well it generalizes to new, unseen data. It allows researchers to measure the accuracy of their methods and enables objective comparison of models based on quantitative metrics, aiding in model selection, hyperparameter tuning, and benchmarking against baselines. In this application scenario, the lack of ground-truth labels makes evaluation difficult. In general, we see two types of validation methods used to evaluate the performance of an algorithm: internal and external. *Internal measures* evaluate how well clusters are formed with respect to their compactness and separation. These measures do not require prior labels or ground truths. On the other hand, *external evaluation* measures gauge the degree to which the cluster labels match the class labels supplied externally. Furthermore, several studies conducted a manual analysis to validate their results. Hence, we define three evaluation strategies Internal, External, and Manual. Some studies have also used a combination of these techniques, as shown in Table 2.

1) INTERNAL EVALUATION

Six studies employed internal measures to evaluate performance. The most commonly used internal evaluation metric is the *Silhouette Score* [61], as reported in [35], [42], [45], [48], and [49]. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to $+1$, where a high value indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters. A clustering with an average silhouette score of over 0.7 is considered to be ''strong'', a value over 0.5 ''reasonable'', and over 0.25 ''weak''. The silhouette score is specialized for measuring cluster quality when the clusters are convex-shaped, and may not perform well if the data clusters have irregular shapes or are of varying sizes. Several other internal measures are used in the literature. Sheikhalishahi et al. [47], [48], [49] used the Dunn Index, and Saka et al. [45] used the Davies–Bouldin Index.

Due to the popularity of the Silhouette score (SS) method, we conducted further analysis of the reported results to assess its reliability and determine what values of SS are generally considered indicative of good clustering performance. However, our analysis revealed variable results, indicating inconsistencies in the interpretation of the SS across different studies and datasets. The challenge in comparing these studies is the use of a diverse range of input features, which has a significant impact on the scores obtained. Halder et al. [42] assumed that a silhouette score surpassing 0.3 suggests the potential presence of highly homogeneous clusters, and observed the highest average silhouette value to be 0.41. On the other hand, Sheikhalishahi et al. [48], [49] leveraged tree-based approaches and observed remarkably high silhouette values, such as 0.99, with their proposed CCTree approach, which can be attributed to their hierarchical structure and the way they partition the data. Similarly, Althobaiti et al. [35] achieve scores ranging between 0.30 and 0.6, and state that the silhouette and homogeneity scores indicate well-formed clusters resulting from their clustering analysis. Conversely, the investigation of Saka et al. [45] yielded notably low silhouette scores, particularly for DBSCAN, indicating substantial cluster overlap. This suggests the challenge of effectively segregating closely related entities into distinct clusters.

2) EXTERNAL EVALUATION

External evaluation methods refer to techniques that assess the performance of a model by comparing its predictions to external, ground truth information. These methods are applicable when there are known or labeled outcomes for the data, allowing for a quantitative assessment of the model's accuracy and effectiveness. Fourteen studies used external evaluation measures to evaluate performance, making it the most popular choice.

In order to perform external evaluation, most researchers manually labeled a part of their dataset. There were no common techniques for doing so with each author group using a different approach. Althobaiti et al. [35] identified 10 campaigns by manually comparing emails based on the sender names, sender emails, email subjects or topic features. They used these labels to calculate the homogeneity score of their clusters. In [45], the authors used a key phrase matching technique to identify and label emails. They identified common and important phrases occurring in the email text, ran a comparison against their whole dataset, and tagged 493 emails with 14 unique numbers. They measured accuracy using the Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), and V-Measure. Sheikhalishahi et al. [47], [48], [49] collected 276 spam emails from different mailboxes and manually analyzed and classified them into 29 groups according to the structural similarity of the raw email message. The external evaluation set does not intersect with that used for the internal evaluation. Song et al. [51], [52] created labeled data by considering two emails as members of the same cluster if the HTML content obtained from Web sites linked to their respective URLs was similar to each other. Halder et al. [41] evaluated the purity of the clusters using ground truth data that was manually collected, although not much information is provided on how the data was labeled.

Two studies used external evaluation measures to compare the results of different clustering algorithms, rather than comparing them to the ground truth. Yearwood et al. [56] attempted to find the ''most accurate clustering method among the three by determining the intersection between the final consensus of the three clustering results and each individual clustering result. They employed three measures: Normalized Mutual Information (NMI), purity, and the

number of edge cuts. Seifollahi et al. [46] used the Rand Index to compute the similarity between four clustering techniques. Rand index compares two cluster distributions by considering all pairs of samples and counting pairs assigned to the same or different clusters in the predicted and true cluster distributions.

Two studies performed external evaluations on downstream tasks. Haider et al. [40] evaluated their proposed Bayesian clustering method by comparing it to a Support Vector Machine (SVM) and an Expectation Maximization (EM) algorithm. They posit that a Mail Transfer Agent (MTA) must be extremely confident when deciding to refuse a message for delivery from a contacting agent. Hence, they measured the rate of true positives (spam messages identified as such) at a false positive rate of zero. They found that their Bayesian clustering method outperformed both methods. Husna et al. [44] hand-labeled a spam dataset by manually labeling each email as either spam or legitimate. They then measured the accuracy of their technique by comparing the classification results of the clustering technique with those of the hand-labeled dataset. The general idea behind such an evaluation is that a good clustering algorithm should result in largely homogeneous clusters with either only benign emails or malicious emails. Hence, the downstream task of identifying phishing/spam versus benign should have good accuracy.

Two studies obtained ground-truth labels from external sources [43], [50]. Han and Shen [43] had each email manually examined by security experts and attributed to a specific campaign with strong confidence. Shen and Thonnard [50] evaluated the effectiveness of the MR-TRIAGE algorithm by comparing its results to previously validated results. They also used the Adjusted Rand Index and Adjusted Mutual Information.

Takeaway 4: External evaluation using ground truth labels is a challenging task in this domain. First, it is difficult to obtain true labels. Second, the absence of a universal definition of clusters exacerbates this issue as researchers often label data according to specific use cases or perspectives, making cross-study comparisons difficult.

3) MANUAL EVALUATION

Five studies [34], [42], [53], [54], [55] manually analyzed the resulting clusters for accuracy. Alazab et al. [34] analyzed the five largest clusters using a set of linguistic, structural, and syntactic features. Wei et al. [53] evaluated the resulting clusters using visual inspection of graphical images of websites fetched from the URLs found in the emails and the *WHOIS* data of the IP address of the computer hosting the advertised sites. Similarly, Xie et al. [55] examine the similarity between the text in the destination Web pages to verify whether each spam campaign is correctly grouped. Halder et al. [42] evaluated the purity of the resulting clusters by visual inspection and evaluated the performance of different features (Subject and Sender Domain) by analyzing the number of resultant clusters, number of emails, and number of unique domains. Woo et al. [54] evaluated the maliciousness of suspected clusters by calculating the ratio of malicious IP addresses in each cluster, for classifying the clusters into botnets, worms, or individual spammers.

V. DISCUSSION

In this section, we present the key findings from our literature survey related to our research questions. We will begin by introducing three definitions of a phishing campaign based on our analysis (RQ1.1). Next, we will discuss various features or attributes typically used by researchers to characterize email clusters, with a particular emphasis on the importance of email text. Following this, we will examine the input features utilized in our proposed use cases, reiterating the significant role of email content and text. We will also address the challenges that researchers face regarding email body text, as documented in the relevant literature. Finally, we will identify the gaps and limitations in the current state of research that require further investigation (RQ3). Additionally, we will share our recommendations for security researchers to consider when developing future cybersecurity defenses. Please note that our second research question (RQ2), which addresses the methodological approaches to email grouping, is answered and discussed in detail using criteria C3-C6 in Section IV.

A. THREE DEFINITIONS OF A PHISHING CAMPAIGN

In Section IV-A, we discuss in detail the definition of a "group OR campaign OR cluster". In particular, we dissect the definition of a campaign into three commonly used aspects: *common source*, *similarities*, and *variations*. One key observation we note is the absence of a consistent definition in this domain, which is not surprising given that defenders have a range of goals when trying to track attacks, such as blocking the attack, responding to attacks, collecting evidence, and trying to understand the tools used. Given this range, it is more logical to consider a collection of definitions rather than a single, all-encompassing one. Such a collection of definitions could also assist the research community in better articulating and signposting their work, making communication more effective and facilitating easier access to relevant research. Based on our analysis, we propose the following three definitions.

1) SOURCE-BASED

A group of spam or phishing emails sent from a single source (i.e. attacker group, bot, botnet) as part of a single attack. *Goals* of this definition include profiling attackers, identifying botnets or spammers to block or prosecute them, informing incident response policy, and building evidence. *Features* tend to be source-based like URLs and sender addresses, while IP addresses would be expected to show some similarity and are likely the most important feature.

2) SCAM-BASED

A group of spam or phishing emails disseminated with a specific purpose and share a common narrative (such as advertising a product or impersonating an organization). *Goals* of this definition include tracking scams and is particularly important for impersonated organizations to keep track of and craft user advice. This is particularly crucial for organizations that have been impersonated, as it enables them to keep track of scams and offer advice. For example, WeTransfer's page on ''Phishing attempts and other unusual WeTransfer imitations'' is vital for ensuring a good user experience.[7] *Features* of these campaigns typically focus on consistencies across the email subject and body text, Logos, and Named Entities.

3) RESPONSE-BASED

A group of spam or phishing emails that can be effectively responded to as a collective incident rather than individual incidents. *Goals* of this definition focus on efficient response to a specific attack. They include: improving the processing speed of reports, accurately blocking an attack, identifying potential victims, responding to victims, and generally reducing manual analysis load born by humans. *Features* exhibit commonalities in terms of scams, content, and the actions they aim to elicit.

The proposed definitions aim to encompass the variety of definitions and motivations observed in our set of papers. They highlight different areas of focus at various temporal points in the phishing or spam response process. For instance, response-based groupings concentrate on short-term mitigation strategies. This approach prioritizes quick responses with minimal human effort and is less concerned with advanced persistent threats (APTs). In this context, even minutes can make a significant difference. In contrast, source-based groupings focus on identifying the origins of threats, allowing for larger mitigation strategies such as taking down websites, deactivating botnets, modeling threats, or pursuing legal action. While this approach needs to be timely, achieving some of its goals may take months. A more deliberate yet precise technical approach can often be more effective for this grouping. Finally, scam-based groupings examine the content of emails to identify issues, such as brand impersonation, product advertisements, or the types of scams being used. These groupings are stable over longer time periods. For example, package delivery scams are quite common and will likely remain so for years. These models may be used as part of the short-term response, but they are likely to be more stable over time.

B. CHARACTERIZATION OF CLUSTERS BASED ON ATTRIBUTES

Our first research question (RQ1) focuses on the definition of spam or phishing clusters, as elaborated in Section IV-A. An important component of the definition was found to be the similarities shared among emails within a cluster. In this study, we refer to a *cluster attribute* as such a shared characteristic or quality among member emails, which provides valuable insights into the nature of the cluster. In this section, we discuss the various attributes that researchers found consistent within the clusters in their studies with the most significant attributes being: *URLs*, *time-based attributes*, and *text-based attributes*.

1) URL-BASED ATTRIBUTES

are considered reliable indicators of a common source by many researchers. Domains are usually owned by a single individual or organization, so emails that all point to the same domain likely have a common source either in terms of sender or the software platform being used. Several studies used URLs, their domains, and additional derived information like IP addresses, WHOIS data, and DNS (Domain Name System) data to create ground truth labels for validation purposes. Dinh et al. [39] utilized this information along with the URL geo-location to assist investigators in tracing spammers. Moreover, they employed visualization tools to illustrate the relationships between spam emails, spam campaigns, domains, and IP addresses. Song et al. [51] use web content from URLs to define clusters and if the IP addresses resolved from URLs are completely the same [52]. Woo et al. [54] claim that maliciousness of URLs can be used to classify the source of a spam email as a botnet, worm, or individual spammer. Wei et al. [53] perform cluster validation using WHOIS data, the IP address of the advertised sites, and graphical images of website fetches. They claim that this validation technique was able to identify relationships between spam campaigns that were not identified by human researchers. Halder et al. [42] observed that small groups of IP addresses are used to host a large number of spam domains, and hence state that IP blacklisting is more efficient to stop spammers.

However, some researchers also found that URLs are commonly obfuscated by attackers to avoid such detection methods. Calais et al. [36] identified three types of campaigns in terms of obfuscation of URLs: static campaigns, campaigns with sub-campaigns, and random-obfuscated campaigns. *Static campaigns* are the ones in which the attacker inserts the same URL in all the messages of the campaign. In *campaigns with sub-campaigns*, the attacker generates a set of URLs in which each URL corresponds to a different product from the same website. For example, dvd1.htm, dvd2.htm, and dvd3.htm are different products associated with the same campaign. In such cases, purely URL-based would fall short. Finally, in *random-obfuscated campaigns*, attackers constantly obfuscate their URLs inserting random fragments which are different for each message. In their IP-based clustering algorithm, Song et al. [52] observed that each IP cluster had many different URLs and domain names. They thus propose that spammers or attackers frequently

[7] https://help.wetransfer.com/hc/en-us/articles/208554176-Phishing-attempts-and-other-weird-WeTransfer-imitations

change their URLs and domain names even if they are connected to a single server.

URL attributes are also an example of a feature that ages poorly since attack domains are taken down once identified, metadata about the domain may change over time, and features like search results change quite quickly. So researchers using public datasets that are even days old may find different results about the URL than were available when the email was first sent. The importance of such features suggests that public dataset creation should also consider including common features that are known to change over time as the IP address links resolve.

2) TIME-BASED ATTRIBUTES

are another important aspect when identifying phishing/spam campaigns [34], [37], [54]. In [34], the authors found time-based features to be consistent within clusters, such as most emails in each cluster appearing within the same month. However, they speculated that this could be because of the low recall rate of the algorithm. Chen et al. [37] analyze different characteristics of spam campaigns, including activity time period, to identify the behavior of spammers and evaluate the probability of a spam message with new content belonging to previously detected campaigns. Zhuang et al. [57] analyze the behavior of botnets using the duration of spam campaigns as a metric. They observed that over 50% of spam campaigns finish within 12 hours, and only about 20% of campaigns persisted for more than 8 days. Xie et al. [55] compute the standard deviation (std) of spam email sending time for each campaign and observed that 50% of campaigns have std less than 1.81 hours. The rest of the campaigns had a larger variation. Overall, 90% of campaigns have stds less than 24 hours. Similarly, Chen et al. [37] investigated the temporal behavior of spam campaigns by plotting active periods of the top 35 campaigns in their dataset. However, they found that many campaigns last a few months.

It is hard to tell if the observation that campaigns last a short amount of time is an artifact of how campaigns are defined and labeled or if it is related to the sender changing tactics regularly. When a campaign is being defined using approaches like containing the same link, it makes sense that they would be short because emails from the campaign are detected and the attack domain is taken down. Naturally creating a short campaign. If a campaign is defined as a combination of a topic and a single sender, such as DHL package scams all sent from one source, then it becomes harder to tell how long a campaign might last. Without a consistent set of language around words like ''campaign'', interpreting numbers like those above is challenging. For instance, Zhuang et al. [57] who observed that more than half of spam campaigns last less than 12 hours, considered a ''campaign'' to be a set of email messages with the same or almost the same content. As opposed to Xie et al. [55] who observed a quite different 1.81 hours and considered ''campaign'' to involve bots promoting a single product or service combined with server features.

3) TEXT-BASED ATTRIBUTES

are often considered important for characterizing clusters because attackers often use templates or kits to send bulk emails with similar text and wording [34], [39], [45], [49], [52]. Researchers seem to be using a combination of text-based features followed by a manual review to create labels. For instance, Sheikhalishahi et al. [48], [49] used the structural and semantic similarity of emails to manually label 276 spam emails into 29 groups for external evaluation of the resulting campaigns. Similarly, Saka et al. [45] used a key-phrase matching technique and manual inspection to tag a subset of their data into 14 groups. During their hand labeling process, Husna et al. [44] observed that most spammers send large numbers of emails and that they send the same content repeatedly with different IDs.

Several researchers have used text-based similarity to analyze their resultant clusters. Dinh et al. [39] calculated text similarity scores for each campaign using three measures and found up to 78.95% average similarity in clusters. Alazab et al. [34] analyzed the five largest clusters in their results using a set of 29 linguistic, structural, and syntactic features to gain insight into the different attack strategies used by different spam groups or individuals. However, they found that linguistic features were not very informative in their study, and features like payload length had a very high standard deviation within a cluster, making it of little use. They suggest that further in-depth analysis is needed to discover significant traits within spam campaigns, as low-level features alone are insufficient. However, Xie et al. [55] observed that even though a spam campaign directs email users to the same/similar set of destination Web pages, the text content of the emails in a campaign can vary significantly. Hence, they suggest that detecting botnets solely based on the email text is not reliable.

Through the analysis of cluster attributes, researchers have identified common characteristics within their resultant clusters. Nevertheless, the contrasting findings from different studies make it hard to formulate definitive conclusions. It becomes evident that no single attribute or feature will be universally consistent within a campaign, presenting a challenge for researchers. Therefore, experts must consider these attributes as a whole to develop more effective methodologies for identifying and preventing spam and phishing threats.

Recommendation 1: **Adopting source-based and context-based features.** Researchers need to use multiple attributes in combination to develop more effective methodologies for identifying and preventing spam and phishing threats. In particular, considering the definitions, motivations, and variations observed, source-based and context-based features can be very useful.

C. SELECTION OF FEATURES BASED ON USE CASES

One of the biggest challenges in any machine learning research is identifying the right set of input features. The

TABLE 4. Features used: common origin (spammer).

Author/Year	Study	Feature Classes				
		H	U	S	A	C
Sheikhalishahi (2015)	[47]	•	•	•	•	•
Sheikhalishahi (2016)	[48]	•	•	•	•	•
Sheikhalishahi (2020)	[49]	•	•	•	•	•
Seifollahi (2017)	[46]					•
Halder (2011)	[41]		•			•

feature set and representation are crucial in the model's performance. In this review, we identified features that have been used to classify malicious email campaigns or groups. It is important to note that there is no universally accepted set of features, researchers typically select features based on their specific use case or objectives. Below, we break down the observed feature sets by use case, looking at the use of *header (H), URL (U), subject (S), attachment (A), and content (C).*

1) **Identifying the common source (spammer):** The most common feature class is *Content-based features*, as shown in Table 4. Two studies [41], [46] approached this problem from the authorship attribution angle, which involves identifying the author of a text through computational techniques. These techniques typically rely on textual analysis, with common approaches including Term Frequency-Inverse Document Frequency (TF-IDF), n-gram counts, and semantic similarity measures. Additional content-based features include the presence of HTML tags, email size, email language, and the number of images within the email text.
2) **Identifying botnets:** The most common feature class is *Header-based features*, as shown in Table 5. Email headers play a crucial role in identifying the source of an email by providing detailed information about the path the email took from the sender to the recipient, including timestamps and IP addresses. Additionally, the "From" field, "Return-Path" and "Reply-To" fields can offer further insights into the sender's identity. Hence, email headers are important in identifying botnets by revealing patterns consistent with botnet activity, such as large volumes of identical or similar emails originating from multiple IP addresses or domains associated with compromised systems. All relevant studies employed timestamps and IP addresses extracted from email headers to infer the presence of botnets.
3) **Profiling attackers**: The most common feature class is *Content-based features*, as shown in Table 6. Two studies used authorship analysis to approach this problem, which involves identifying the author of a text through computational techniques. Other content features mainly consist of style characteristics that are used to convey the role of words, images, and HTML content. Content-based features encompass various attributes such as language patterns, keywords, and stylistic elements that may reveal the attacker's intent or affiliation.
4) **Reducing the load of manual human analysis:** The most common feature class is *Content-based features and Subject-based features*, as shown in Table 7. In the given use case, where humans are tasked with managing reported emails, studying email and spammer behavior, and analyzing emails to extract insights, content-based features offer a robust approach to automate and streamline these processes. By leveraging attributes such as language patterns, keywords, and stylistic elements, content-based features enable automated systems to efficiently categorize and prioritize emails based on their content, thereby reducing the need for manual scrutiny.

TABLE 5. Features used: identify botnets.

Author/Year	Study	Feature Classes				
		H	U	S	A	C
Chen (2014)	[37]	•				•
Husna (2008)	[44]	•				•
Song (2010)	[52]		•			
Woo (2014)	[54]	•				
Xie (2008)	[55]		•			
Zhuang (2008)	[57]	•	•			•

TABLE 6. Features used: profiling attackers.

Author/Year	Study	Feature Classes				
		H	U	S	A	C
Alazab (2013)	[34]					•
Calais (2008)	[36]		•	•		•
Seifollahi (2017)	[46]					•
Woo (2014)	[54]	•				
Yearwood (2009)	[56]			•		•

TABLE 7. Features used: reduce manual analysis load.

Author/Year	Study	Feature Classes				
		H	U	S	A	C
Althobaiti (2023)	[35]	•	•	•	•	•
Dinh (2015)	[39]	•	•	•	•	•
Han (2016)	[43]	•		•	•	•
Saka (2022)	[45]	•	•	•	•	•
Seifollahi (2017)	[46]					•
Wei (2008)	[53]		•	•		

Based on the analysis, we observe that content-based features hold significant importance in categorizing malicious emails across various use-cases. This is not surprising since attackers typically create emails based on a pre-defined template, making it a relatively stable feature class. The

content of an email provides a wealth of information that can be analyzed to uncover valuable insights. First, the email text contains all the essential details about the underlying scam. Second, the content of an email may also contain clues about the authorship of the message. For instance, the use of specific language, tone, or style. Moreover, the inclusion of signatures or disclaimers can provide additional information. Thirdly, the layout and formatting of an email can also offer valuable insights into its nature. For example, the use of indentation, HTML tags, images, or other visual elements can help categorize emails based on their content. Furthermore, content-based features can be adapted to evolving threats and changing email behaviors, because even with various obfuscation techniques used, the email content must exist.

***Recommendation** 2:* The definition of 'campaign' or 'group' should be made explicit in work as it can have a large impact on the feature sets selected, the labels produced, and ultimately the type of algorithm recommended.

D. VARIATION IN THE STRUCTURE OF EMAILS

The unstructured nature of emails makes it difficult to efficiently process them. While the header part of an email has a structure with information stored and presented as Key:Value pairs, the email body has no predefined structure. It entirely depends on the sender and can contain salutations, signatures, images, links, text, disclaimers, and so on. The email text itself can be highly variable with respect to length, language, grammar, indentation, and context. This variation can be challenging for the algorithms to process. Phishing emails in particular may also purposely make text harder for a computer to read by putting it in images or by including white-on-white text that is invisible to humans.

According to Halder et al. [41], the length of an email has a significant impact on the performance of clustering algorithms. They performed clustering analysis using a set of stylistic features, semantic features, and combined features. They found that stylistic clustering gave good results when the email length was short, and semantic clusters gave good results when the semantic body was rich in content. This makes sense because semantic features require actual context in the emails, and datasets with less text usually have no context. Furthermore, they observed that when multiple groups share similar writing styles, they get clustered together. Alazab et al. [34] also state that features such as payload length that have a very high standard deviation are of little use in developing profiles for larger spam groups. Their dataset had a significant variation in the content of the payloads (i.e. content of the email) with a mean length of 1069 characters/letters excluding spaces/separators, and a substantial standard deviation of 3023 characters/letters. Similarly, Han and Shen [43] state that when emails are extremely short, such as a single sentence or a few phrases, the text features become much less stable and less informative for classification. According to experts at Egress Defend, the accuracy of most detection tools increases with longer sample sizes, often requiring a minimum of 250 characters to work. Their statistics show that 44.9% of phishing emails do not meet the 250-character limit and a further 26.5% fall below 500, and currently AI detectors either won't work reliably or at all on 71.4% of attacks.

Wei et al. [53] show that a weakness of their algorithm is that coincidence, common phrases, and sheer luck can introduce untrustworthy relationships. According to them, spammers commonly use simple phrases for their likelihood of intriguing a reader to open the message. Phrases like ''Was this from you?'' or ''Alert!'' or ''Thank you'' may be chosen by unrelated spammers, and hence create an overlap between different campaigns. Han and Shen [43] shared a similar finding, where one campaign (named 'layork') was often misclassified as another campaign ('krast'), which caused deterioration in their campaign attribution performance. Further analysis revealed that emails from the two campaigns shared considerable similarities in body text and readability indices. Hence, they posit that when two spear-phishing emails from two different campaigns share similar features in text and readability, the algorithms are confused by them. Saka et al. [45] made another important observation in their experiment: Most phishing emails misclassified as benign emails were short text emails with random words and junk text. They also observed low Silhouette Scores in their clusters, which they attribute to a generally high overlap between phishing email structures and text. This type of variation in email structure makes it difficult for machine learning algorithms to reliably cluster malicious emails.

***Recommendation** 3:* **Increasing focus on context.** Regardless of the length or size of an email, each email typically contains some form of information or context that communicates its purpose. Even when an email lacks explicit context, the absence itself can be indicative of its nature. Moreover, this context tends to remain consistent, to some extent, within emails of the same campaign or cluster, as discussed in Section IV-A.

***Recommendation** 4:* **Talking about Context.** Researchers should develop reliable methods for modeling the context of emails. For instance, a standardized template should be employed to extract the most critical information from malicious emails and present it in a clear and concise manner. Standardizing the representation of highly variable emails would enhance the efficiency of spam and phishing email processing tools, thereby improving their effectiveness in detecting and mitigating malicious content.

E. RESEARCH GAPS AND LIMITATIONS

In this section, we discuss the various research gaps and limitations identified through this systematic review, as addressed in the third research question (RQ3). We delve into the effects of these limitations in detail and provide practical recommendations for future researchers to address them.

1) LIMITATION IN THE TYPE OF TECHNOLOGY

Remarkably, only three studies in the set were published in the past five years (since 2018) [35], [45], [49], indicating a gap in the type of technology and algorithms that have been tried. Technology has evolved significantly during this period, particularly in the domains of natural language processing (NLP) and machine learning (ML) by the introduction of large-language models (LLMs). These advancements hold immense potential for effectively identifying and mitigating phishing and spam attacks and have been employed in other areas of security research such as spam detection [65], malware detection [66], and phishing email detection [67], [68]. Such advancements could have tremendous potential in the identification of malicious groups and campaigns. For instance, Saka et al. [45] demonstrated the importance of context-based features when clustering emails into similar scams. In order to capture the contextual information inherent in the text, they used a transformer-based language model, BERT (Bidirectional Encoder Representations from Transformers) [69], to represent text as semantic vectors that are known to adapt to new unseen situations. An approach that was not readily available to researchers of earlier works.

Furthermore, it is essential to acknowledge that these technological advancements have also provided perpetrators with tools to create more sophisticated attacks. For instance, Roy et al. [70] identified malicious prompts capable of exploiting ChatGPT's[8] abilities to generate phishing websites mimicking renowned brands and employing various evasive techniques. This growing trend of using LLMs is particularly troublesome because any individual, without any technical knowledge, could use LLMs to generate highly deceiving phishing websites that look genuine, hence increasing the difficulty and sophistication of these attacks. According to the 2023 Phishing Threat Trends Report, 71% of email attacks created through AI go undetected [71]. Therefore, it becomes necessary for researchers to experiment with similar tools to combat attacks generated by the use of such tools. Given the potential benefits of integrating NLP and ML into cybersecurity, it is imperative to bridge the gap between research and practice.

***Recommendation** 5:* **Staying current with technology trends.** Researchers need to explore newer technologies, such as LLMs, to represent emails. Integrating such models into email security research is a promising future direction, as shown in [45] for clustering emails into scams.

2) DATASET LIMITATIONS

The datasets used by the reviewed papers (Section IV-C) represented quite a range of sizes and sources but they also had some important limitations to consider. Firstly, it is important to highlight that out of all the datasets used, only two are publicly available – one for phishing and one for spam. Public datasets are very crucial for researchers because they allow for comparing and evaluating different methods on the same data. However, all studies have used different datasets, mostly private, making comparison and generalization challenging. Secondly, only one study from the identified set utilized a dataset from the past five years. It is crucial to emphasize that phishing and spam attacks evolve continuously, incorporating new obfuscation techniques and tricks on a daily basis [72]. According to new research from Egress Software Technologies, hackers are using increasingly sophisticated tactics to get their phishing emails past companies' cybersecurity defenses [71]. They observed that the percentage of phishing emails utilizing obfuscation techniques surged by 24.4% in 2023, accounting for 55.2% of all malicious emails. Unfortunately, the limitations of old datasets become apparent in such scenarios.

***Recommendation** 6:* **Sharing Datasets.** Researchers should consider sharing more datasets for research purposes. The advances in current technology could help researchers to generate synthetic data from a sample set.

3) LACK OF ACCESSIBLE LABELED DATA

A very important limitation identified in existing literature is the lack of: (i) publicly available labeled datasets and (ii) widely accepted labeling techniques in malicious email clustering research. The task of spam/phishing filtering, which separates the "bad" email from the "good" email, has some labeled data in the form of independent spam datasets, phishing datasets, and benign email datasets. Unfortunately, for the task of identifying malicious groups or email campaigns, there are no publicly available labeled datasets or a standardized labeling technique. In Section IV-A, we presented how the various papers define the concept of a campaign or group. We identified that there is a lack of a single agreed-upon definition for campaigns or a definite set of features that campaign emails share. This observation makes it challenging to build labeled datasets that are useful for all the various research aimed at grouping campaigns.

The lack of readily accessible labeled data poses two major limitations to researchers. Firstly, it makes it difficult to train efficient models. Most machine learning algorithms that achieve good performance are supervised learning algorithms, where the algorithm is trained on a labeled dataset, generalizes from the training data, and makes predictions or classifications on new, unseen data. Good research practice suggests that researchers should share their code together with datasets they train/test the models on, for reproducibility purposes. Such training cannot be done unless public datasets become available.

Secondly, the lack of ground truth labels makes the evaluation and comparison of models challenging. While some studies used labeled datasets obtained from security service providers [43], others manually labeled the data themselves for evaluation. For instance, two recent studies [45], [49]

[8]ChatGPT is a large-language model which interacts with users in a conversational way. The model can be accessed online: https://openai.com/blog/chatgpt

employed a manual approach to analyze emails and assign labels based on different textual attributes. However, the absence of a standardized method for this process presents a significant challenge in terms of reproducibility and comparability.

While making labeled datasets public would be best for researchers, it is also important to acknowledge the sensitive nature of the data. Both spam and phishing emails often contain the name and email address of the person they were sent to. If the full source is included, it may also include the mail servers sent through and various checks run by those servers. Datasets may also contain false positives where a real legitimate email is included as phishing or spam. While a small number of such false positives will cause minimal impact on model building, they can have large privacy issues for email subjects and possibly intellectual property issues for a company. There is a need for more exploration of how to create datasets that can be released safely without endangering email recipients. Possible directions include synthetic data or automated scripts that change email addresses, names, and unique links throughout a corpus to limit potential harm.

***Recommendation** 7:* **Speaking a common language.** Researchers need to develop and adopt a common framework for labeling and characterizing malicious emails. This would enhance the reliability and validity of research findings and facilitate the comparability of different algorithms.

VI. LIMITATIONS

In this study, we used multiple publication databases and additional security-specific databases to include a variety of publications and avoid overlooking any relevant work. Our search query was devised based on our prior research and knowledge of the field. However, the list of publications analyzed may not be exhaustive. Although it is unlikely that relevant studies were missed, it is a possibility we must acknowledge. Furthermore, the criteria used to evaluate the papers and the categories established to organize the information were carefully considered and deliberated. However, it is possible that some aspects were categorized differently by us than they would be by the authors themselves or other researchers. Despite these limitations, our systematization of knowledge should still serve as a basis for identifying the current state of research, recognizing potential gaps, and guiding future researchers in their work.

VII. CONCLUSION

In this study, we examined the existing research on detecting groups of malicious emails and their implications on email security. Given the range of use cases for grouping spam and phishing emails highlighted by our review, this type of research is critical for enhancing several aspects of cybersecurity defense. The alarming surge in the frequency and scale of email-based threats necessitates the development of efficient grouping algorithms to identify meaningful clusters of emails. While grouping techniques are actively used in practical security operations, our study addresses the gap in a comprehensive, holistic understanding of the research landscape in this field. Understanding what, why, and how such algorithms can be used is a critical research question. This paper presents a systematic analysis of 23 research articles on this subject, focusing on two foundational aspects (definition of a cluster and use-case) and four methodological aspects (dataset, input features, clustering or grouping algorithm, and evaluation strategies). We hope this article provides researchers with an overview of the current state of research and useful directions for future works on spam and phishing interventions. One significant finding of our review is the inconsistency in how research defines and applies ''campaign'' groupings. To address this, we propose three alternative definitions for campaigns: *source-based, scam-based, and response-based.* Using consistent language will not only streamline research comparisons but also enhance collaboration between cybersecurity teams by creating a shared understanding of campaign categorization. Based on our findings, in Section V, we provide practical recommendations to researchers and promising future directions, which can be summarized as follows:

Utilize advancements in Machine Learning (ML) and Natural Language Processing (NLP). *What are some new text-representation techniques? What clustering algorithms are being used in other domains, and could they be tailored for email analysis? What levels of explainability and generalizability do these algorithms offer?* To advance email grouping for spam and phishing detection, we recommend leveraging recent ML and NLP innovations like transformer-based embeddings (e.g., BERT, RoBERTa) for richer text representations [73], [74], which could capture subtle patterns in email language. Clustering algorithms from other domains, such as deep clustering models [75] and graph-based clustering [76], could be adapted to enhance group discovery based on both content and network features. Incorporating explainability tools, like LIME [77], would improve transparency. These approaches can help build a robust, adaptable, and interpretable framework for email threat grouping.

Create a standardized approach to label data. *How can we better document the labeling process? What should be taken into consideration while labeling? Can we define a systematized way to label based on the proposed use-case?* To create a standardized approach for labeling data in spam and phishing email research, we recommend establishing clear, well-documented guidelines that outline labeling criteria aligned with specific use cases. Such frameworks should incorporate consistency, and contextual factors (e.g., sender information, language cues), and should be adaptable to evolving threats. This can be done by evaluating the labeling approach over multiple datasets from various sources and times. Researchers can also explore the use of generative AI to create synthetic datasets of email campaigns by simulating realistic variations in content, structure, and sender profiles [78], [79].

Find innovative techniques to model the context of an email. *How can we normalize the variation seen in malicious*

emails? Can we extract the most important information from emails? What contextual features exist in most emails that can be extracted? Contextual embeddings, such as those from models like BERT fine-tuned on phishing data [69], can help identify commonalities despite superficial differences. Furthermore, to extract key contextual information from emails, researchers should consider information extraction techniques [80], [81] or keyword extraction [82] to identify common phrases or structures. Additionally, image and link fingerprinting can detect similarities in altered media or URLs, and tokenization methods like stemming and lemmatization can reduce word variation [83]. Together, these approaches enhance consistency in comparing malicious emails.

ACKNOWLEDGMENT

This work has been partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under award number RGPIN-2024-06737. This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the Cluster of Excellence Cyber Security in the Age of Large-Scale Adversaries, CASA (EXC 2092 - 390781972). We would like to express our gratitude for this support, which has been instrumental in this research.

REFERENCES

[1] Government Canada. (2019). *Understanding Canadas Anti-spam Legislation*. Accessed: Jan. 27, 2025. [Online]. Available: https://ised-isde.canada.ca/site/canada-anti-spam-legislation/en

[2] A. Franz, V. Zimmermann, G. Albrecht, K. Hartwig, C. Reuter, A. Benlian, and J. Vogt, "SoK: Still plenty of phish in the sea—A taxonomy of user-oriented phishing interventions and avenues for future research," in *Proc. 17th Symp. Usable Privacy Secur. (SOUPS)*, Aug. 2021, pp. 339–358. [Online]. Available: https://www.usenix.org/conference/soups2021/presentation/franz

[3] R. B. C. Ellis. (2024). *Spam Statistics (2024): New Data on Junk Email, AI Scams & Phishing*. Accessed: Jan. 27, 2025. [Online]. Available: https://www.emailtooltester.com/en/blog/spam-statistics/

[4] APWG. (2023). *Phishing Activity Trends Report, 4th Quarter*. Accessed: Jan. 27, 2025. [Online]. Available: https://apwg.org/trendsreports/

[5] Proofpoint Inc. (2023). *State of the Phish*. Accessed: Jan. 27, 2025. [Online]. Available: https://www.proofpoint.com/sites/default/files/threat-reports/pfpt-us-tr-state-of-the-phish-2023.pdf

[6] IBM Secur. (2023). *IBM Security X-Force Threat Intelligence Index 2023*. Accessed: Jan. 27, 2025. [Online]. Available: https://www.ibm.com/reports/threat-intelligence

[7] IBM Secur. (2024). *Cost of a Data Breach Report 2024*. Accessed: Jan. 27, 2025. [Online]. Available: https://www.ibm.com/security/data-breach

[8] A. Chrysanthou, Y. Pantis, and C. Patsakis, "The anatomy of deception: Measuring technical and human factors of a large-scale phishing campaign," *Comput. Secur.*, vol. 140, May 2024, Art. no. 103780.

[9] J. Lee, Y. Lee, D. Lee, H. Kwon, and D. Shin, "Classification of attack types and analysis of attack methods for profiling phishing mail attack groups," *IEEE Access*, vol. 9, pp. 80866–80872, 2021.

[10] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 2091–2121, 4th Quart., 2013.

[11] A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani, "A survey of phishing email filtering techniques," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 2070–2090, 4th Quart., 2013.

[12] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, "A review of spam email detection: Analysis of spammer strategies and the dataset shift problem," *Artif. Intell. Rev.*, vol. 56, no. 2, pp. 1145–1173, Feb. 2023.

[13] B. Naqvi, K. Perova, A. Farooq, I. Makhdoom, S. Oyedeji, and J. Porras, "Mitigation strategies against the phishing attacks: A systematic literature review," *Comput. Secur.*, vol. 132, Sep. 2023, Art. no. 103387.

[14] T. Gangavarapu, C. D. Jaidhar, and B. Chanduka, "Applicability of machine learning in spam and phishing email filtering: Review and approaches," *Artif. Intell. Rev.*, vol. 53, no. 7, pp. 5019–5081, Oct. 2020.

[15] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "Phishing email detection using natural language processing techniques: A literature survey," *Proc. Comput. Sci.*, vol. 189, pp. 19–28, Jan. 2021.

[16] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A systematic literature review on phishing email detection using natural language processing techniques," *IEEE Access*, vol. 10, pp. 65703–65727, 2022.

[17] A. Aleroud and L. Zhou, "Phishing environments, techniques, and countermeasures: A survey," *Comput. Secur.*, vol. 68, pp. 160–196, Jul. 2017.

[18] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," *Telecommun. Syst.*, vol. 76, no. 1, pp. 139–154, Jan. 2021.

[19] L. Gallo, D. Gentile, S. Ruggiero, A. Botta, and G. Ventre, "The human factor in phishing: Collecting and analyzing user behavior when reading emails," *Comput. Secur.*, vol. 139, Apr. 2024, Art. no. 103671.

[20] S. Zhuo, R. Biddle, Y. S. Koh, D. Lottridge, and G. Russello, "SoK: Human-centered phishing susceptibility," *ACM Trans. Privacy Secur.*, vol. 26, no. 3, pp. 1–27, Aug. 2023.

[21] G. Desolda, L. S. Ferro, A. Marrella, T. Catarci, and M. F. Costabile, "Human factors in phishing attacks: A systematic literature review," *ACM Comput. Surveys*, vol. 54, no. 8, pp. 1–35, Nov. 2022.

[22] K. L. Chiew, K. S. C. Yong, and C. L. Tan, "A survey of phishing attacks: Their types, vectors and technical approaches," *Expert Syst. Appl.*, vol. 106, pp. 1–20, Sep. 2018.

[23] R. Alabdan, "Phishing attacks survey: Types, vectors, and technical approaches," *Future Internet*, vol. 12, no. 10, p. 168, Sep. 2020.

[24] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "Spam: The underground on 140 characters or less," in *Proc. 17th ACM Conf. Comput. Commun. Secur.*, Oct. 2010, pp. 27–37.

[25] A. Ramachandran and N. Feamster, "Understanding the network-level behavior of spammers," in *Proc. Conf. Appl., Technol., Archit., Protocols Comput. Commun.*, Aug. 2006, pp. 291–302.

[26] S. Tang, X. Mi, Y. Li, X. Wang, and K. Chen, "Clues in tweets: Twitter-guided discovery and analysis of SMS spam," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2022, pp. 2751–2764.

[27] R. Phillips and H. Wilder, "Tracing cryptocurrency scams: Clustering replicated advance-fee and phishing websites," in *Proc. IEEE Int. Conf. Blockchain Cryptocurrency (ICBC)*, May 2020, pp. 1–8.

[28] M. Cova, C. Leita, O. Thonnard, A. D. Keromytis, and M. Daciér, "An analysis of rogue AV campaigns," in *Proc. Int. Workshop Recent Adv. Intrusion Detection*, Jan. 2010, pp. 442–463.

[29] O. Thonnard and M. Dacier, "A strategic analysis of spam botnets operations," in *Proc. 8th Annu. Collaboration, Electron. Messaging, Anti-Abuse Spam Conf.*, Sep. 2011, pp. 162–171.

[30] Microsoft Corp. (2024). *Email Analysis in Investigations for Microsoft Defender for Office 365*. Accessed: Sep. 9, 2024. [Online]. Available: https://learn.microsoft.com/en-us/defender-office-365/email-analysis-investigations?view=o365-worldwide%5C

[31] Ironscales. (2024). *AI-Driven Security Platform for Email Threat Protection*. Accessed: Sep. 27, 2024. [Online]. Available: https://ironscales.com/platform/ai

[32] KnowBe4, Inc. (2024). *PhishER Plus—AI-Powered Anti-Phishing De Bfense*. Accessed: Sep. 9, 2024. [Online]. Available: https://www.knowbe4.com/products/phisher-plus

[33] M. D. F. McInnes, D. Moher, B. D. Thombs, and T. A. McGrath, "Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: The PRISMA-DTA statement," *Jama*, vol. 319, no. 4, pp. 388–396, Jan. 2018.

[34] M. Alazab, R. Layton, R. Broadhurst, and B. Bouhours, "Malicious spam emails developments and authorship attribution," in *Proc. 4th Cybercrime Trustworthy Comput. Workshop*, Nov. 2013, pp. 58–68.

[35] K. Althobaiti, M. K. Wolters, N. Alsufyani, and K. Vaniea, "Using clustering algorithms to automatically identify phishing campaigns," *IEEE Access*, vol. 11, pp. 96502–96513, 2023.

[36] P. H. Calais, D. E. Pires, D. O. Guedes, W. Meira Jr., C. Hoepers, and K. S. Jessen, "A campaign-based characterization of spamming strategies," in *Proc. 5th Conf. Email Anti-Spam (CEAS)*, Mountain View, CA, USA, Aug. 2008.

[37] J. Chen, R. Fontugne, A. Kato, and K. Fukuda, "Clustering spam campaigns with fuzzy hashing," in *Proc. Asian Internet Eng. Conf. (AINTEC)*, 2014, pp. 66–73.

[38] J. Kornblum, "Identifying almost identical files using context triggered piecewise hashing," *Digit. Invest.*, vol. 3, pp. 91–97, Sep. 2006.

[39] S. Dinh, T. Azeb, F. Fortin, D. Mouheb, and M. Debbabi, "Spam campaign detection, analysis, and investigation," *Digit. Invest.*, vol. 12, pp. S12–S21, Mar. 2015.

[40] P. Haider and T. Scheffer, "Bayesian clustering for email campaign detection," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, Jun. 2009, pp. 385–392.

[41] S. Halder, R. Tiwari, and A. Sprague, "Information extraction from spam emails using stylistic and semantic features to identify spammers," in *Proc. IEEE Int. Conf. Inf. Reuse Integr.*, Aug. 2011, pp. 104–107.

[42] S. Halder, R. Tiwari, and A. Sprague, "Identifying features to improve real time clustering and domain blacklisting," in *Proc. 50th Annu. Southeast Regional Conf.*, Mar. 2012, pp. 238–243.

[43] Y. Han and Y. Shen, "Accurate spear phishing campaign attribution and early detection," in *Proc. 31st Annu. ACM Symp. Appl. Comput.*, Apr. 2016, pp. 2079–2086.

[44] H. Husna, S. Phithakkitnukoon, S. Palla, and R. Dantu, "Behavior analysis of spam botnets," in *Proc. 3rd Int. Conf. Commun. Syst. Softw. Middleware Workshops (COMSWARE)*, Jan. 2008, pp. 246–253.

[45] T. Saka, K. Vaniea, and N. Kökciyan, "Context-based clustering to mitigate phishing attacks," in *Proc. 15th ACM Workshop Artif. Intell. Secur.*, Nov. 2022, pp. 115–126.

[46] S. Seifollahi, A. Bagirov, R. Layton, and I. Gondal, "Optimization based clustering algorithms for authorship analysis of phishing emails," *Neural Process. Lett.*, vol. 46, no. 2, pp. 411–425, Oct. 2017.

[47] M. Sheikhalishahi, A. Saracino, M. Mejri, N. Tawbi, and F. Martinelli, "Digital waste sorting: A goal-based, self-learning approach to label spam email campaigns," in *Proc. Int. Workshop Secur. Trust Manage.*, Jan. 2015, pp. 3–19.

[48] M. Sheikhalishahi, A. Saracino, M. Mejri, N. Tawbi, and F. Martinelli, "Fast and effective clustering of spam emails based on structural similarity," in *Proc. 8th Int. Symp. Found. Pract. Secur.*, Clermont-Ferrand, France. Cham, Switzerland: Springer, Oct. 2016, pp. 195–211.

[49] M. Sheikhalishahi, A. Saracino, F. Martinelli, A. La Marra, M. Mejri, and N. Tawbi, "Digital waste disposal: An automated framework for analysis of spam emails," *Int. J. Inf. Secur.*, vol. 19, no. 5, pp. 499–522, Oct. 2020.

[50] Y. Shen and O. Thonnard, "MR-TRIAGE: Scalable multi-criteria clustering for big data security intelligence applications," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2014, pp. 627–635.

[51] J. Song, M. Eto, H. C. Kim, D. Inoue, and K. Nakao, "A heuristic-based feature selection method for clustering spam emails," in *Proc. 17th Int. Conf. Neural Inf. Process. Theory Algorithms (ICONIP)*, Sydney, NSW, Australia. Cham, Switzerland: Springer, Nov. 2010, pp. 290–297.

[52] J. Song, D. Inque, M. Eto, H. C. Kim, and K. Nakao, "An empirical study of spam: Analyzing spam sending systems and malicious Web servers," in *Proc. 10th IEEE/IPSJ Int. Symp. Appl. Internet*, Jul. 2010, pp. 257–260.

[53] C. Wei, A. Sprague, G. Warner, and A. Skjellum, "Mining spam email to identify common origins for forensic application," in *Proc. ACM Symp. Appl. Comput.*, Mar. 2008, pp. 1433–1437.

[54] J. Woo, H. J. Kang, A. R. Kang, H. Kwon, and H. K. Kim, "Who is sending a spam email: Clustering and characterizing spamming hosts," in *Proc. 16th Int. Conf. Inf. Secur. Cryptol. (ICISC)*, Seoul, South Korea. Cham, Switzerland: Springer, Nov. 2014, pp. 469–482.

[55] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov, "Spamming botnets: Signatures and characteristics," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, pp. 171–182, 2008.

[56] J. Yearwood, D. Webb, L. Ma, P. Vamplew, B. Ofoghi, and A. Kelarev, "Applying clustering and ensemble clustering approaches to phishing profiling," in *Proc. 8th Australas. Data Mining Conf.*, vol. 101, Dec. 2009, pp. 25–34.

[57] Z. Li, J. Dunagan, D. Simón, H. J. Wang, and J. D. Tygar, "Characterizing botnets from email spam records," *Leet*, vol. 8, no. 1, pp. 1–9, Apr. 2008.

[58] N. Grira, M. Crucianu, and N. Boujemaa, "Unsupervised and semi-supervised clustering: A brief survey," *Rev. Mach. Learn. Techn. Process. Multimedia Content*, vol. 1, no. 2004, pp. 9–16, 2004.

[59] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math., Stat., Prob*, vol. 1, Jan. 1967, pp. 281–297.

[60] K. D. Joshi and P. S. Nalwade, "Modified K-means for better initial cluster centres," *Int. J. Comput. Sci. Mobile Comput.*, vol. 2, no. 7, pp. 219–223, 2013.

[61] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[62] B. Ordin and A. M. Bagirov, "A heuristic algorithm for solving the minimum sum-of-squares clustering problems," *J. Global Optim.*, vol. 61, no. 2, pp. 341–361, Feb. 2015.

[63] A. M. Bagirov, "Modified global -means algorithm for minimum sum-of-squares clustering problems," *Pattern Recognit.*, vol. 41, no. 10, pp. 3192–3199, Oct. 2008.

[64] F. Nielsen, "Hierarchical clustering," in *Introduction to HPC with MPI for Data Science*. Cham, Switzerland: Springer, 2016, pp. 195–211. [Online]. Available: https://doi.org/10.1007/978-3-319-21903-5_8

[65] I. AbdulNabi and Q. Yaseen, "Spam email detection using deep learning techniques," *Proc. Comput. Sci.*, vol. 184, pp. 853–858, Jan. 2021.

[66] J. Qiu, J. Zhang, W. Luo, L. Pan, S. Nepal, and Y. Xiang, "A survey of Android malware detection with deep neural models," *ACM Comput. Surveys*, vol. 53, no. 6, pp. 1–36, Nov. 2021.

[67] C. Catal, G. Giray, B. Tekinerdogan, S. Kumar, and S. Shukla, "Applications of deep learning for phishing detection: A systematic literature review," *Knowl. Inf. Syst.*, vol. 64, no. 6, pp. 1457–1500, Jun. 2022.

[68] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Syst. Appl.*, vol. 117, pp. 345–357, Mar. 2019.

[69] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[70] S. Saha Roy, K. Vamsi Naragam, and S. Nilizadeh, "Generating phishing attacks using ChatGPT," 2023, *arXiv:2305.05133*.

[71] Egress Softw. Technol. (2023). *Phishing Threat Trends Report*. Accessed: Feb. 14, 2024. [Online]. Available: https://pages.egress.com/Whitepaper-PhishingThreatTrendsReport-10-232023-Landing-Page.html

[72] F. Carroll, J. A. Adejobi, and R. Montasari, "How good are we at detecting a phishing attack? Investigating the evolving phishing attack email and why it continues to successfully deceive society," *Social Netw. Comput. Sci.*, vol. 3, no. 2, p. 170, Mar. 2022.

[73] W. Xin Zhao et al., "A survey of large language models," 2023, *arXiv:2303.18223*.

[74] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Comput. Surveys*, vol. 56, no. 2, pp. 1–40, Feb. 2024.

[75] Y. Ren, J. Pu, Z. Yang, J. Xu, G. Li, X. Pu, P. S. Yu, and L. He, "Deep clustering: A comprehensive survey," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 4, 2024, doi: 10.1109/TNNLS.2024.3403155.

[76] A. Tsitsulin, J. Palowitch, B. Perozzi, and E. Müller, "Graph clustering with graph neural networks," *J. Mach. Learn. Res.*, vol. 24, no. 127, pp. 1–21, Jan. 2023.

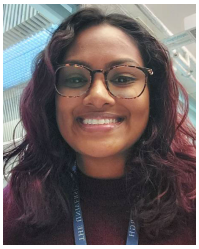
[77] A. Bhattacharya, *Applied Machine Learning Explainability Techniques: Make ML Models Explainable and Trustworthy for Practical Applications Using LIME, SHAP, and More*. Birmingham, U.K.: Packt Publishing Ltd, 2022.

[78] M. Goyal and Q. H. Mahmoud, "A systematic review of synthetic data generation techniques using generative AI," *Electronics*, vol. 13, no. 17, p. 3509, Sep. 2024.

[79] J. Singh, "The rise of synthetic data: Enhancing AI and machine learning model training to address data scarcity and mitigate privacy risks," *J. Artif. Intell. Res. Appl.*, vol. 1, no. 2, pp. 292–332, 2021.

[80] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, H. Chen, Z. Liu, Z. Dou, and J.-R. Wen, "Large language models for information retrieval: A survey," 2023, *arXiv:2308.07107*.

[81] J. Cowie and W. Lehnert, "Information extraction," *Commun. ACM*, vol. 39, no. 1, pp. 80–91, 1996.
[82] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," in *Text Mining: Applications and Theory*. Hoboken, NJ, USA: Wiley, 2010, ch. 1, pp. 1–20. [Online]. Available: https://doi.org/10.1002/9780470689646.ch1
[83] S. Vijayarani and R. Janani, "Text mining: Open source tokenization tools—An analysis," *Adv. Comput. Intell., Int. J. (ACII)*, vol. 3, no. 1, pp. 37–47, Jan. 2016.



TARINI SAKA received the Ph.D. degree in AI-assisted usable security from the Artificial Intelligence and its Applications Institute, The University of Edinburgh, in 2024. She is currently a Postdoctoral Researcher with Ruhr-Universität Bochum, where she is also part of the Chair for Security and Privacy of Ubiquitous Systems and the Cluster of Excellence CASA. Her research interests include applied artificial intelligence, user security, organizational phishing mitigation, and user guidance.

KAMI VANIEA (Member, IEEE) received the Ph.D. degree in computer science from Carnegie Mellon University, in 2012. She was a Postdoctoral Researcher with Michigan State University, from 2012 to 2014, and an Assistant Professor with Indiana University Bloomington, from 2014 to 2015. She was a Reader with the School of Informatics, The University of Edinburgh (2015–2023). She is currently an Associate Professor in electrical and computer engineering with the University of Waterloo. Her research interests include usable security and privacy, along with related fields, such as human–computer interaction, cybersecurity, and privacy. Her work aims to enhance understanding and support the protection needs of diverse users.

NADIN KÖKCIYAN received the Ph.D. degree from Boğaziçi University, in 2017. She held a postdoctoral research position with the King's College London, until 2019. She is currently a Lecturer in artificial intelligence with the School of Informatics, The University of Edinburgh; and a Senior Research Affiliate with the Centre for Technomoral Futures, Edinburgh Futures Institute. Her research interests include human-centered AI, privacy, argument mining, responsible AI, and AI ethics. She regularly serves on the program committees for leading AI conferences, such as AAMAS, IJCAI, AAAI, and ECAI. In 2021 and 2025, she served as a Guest Editor for Special Issues on "Sociotechnical Perspectives of AI Ethics and Accountability" and "Humans Meets AI" in IEEE Internet Computing.