Phish and Chips: Language-agnostic classification of unsolicited emails

Carlos H. Gañán, Siôn Lloyd, Samaneh Tajalizadehkhoob
Office of the CTO
ICANN
Email: {carlos.ganan, sion.lloyd, samaneh.tajali}@icann.org

***Abstract*—Email remains a popular communication tool despite the emergence of new messaging systems, however, this popularity also attracts individuals with malicious intentions. Despite the efforts of current email filtering to keep up with the email-based threat vectors, unsolicited emails still keep reaching millions of targets. The current solutions are mainly focused on distinguishing ham from spam/phishing, leaving a gap in the identification and analysis of other unsolicited emails such as scams and adult content. In this paper, we present a study on the development of a more granular approach for sanitizing and categorizing unsolicited emails, specifically focusing on spam, phishing, scam and adult content. We design and evaluate a method for classifying unsolicited emails that can aid incident response teams in extracting contextual potential Threat Indicators (TIs). We train a machine learning language-agnostic classifier that achieves high accuracy with a novel set features such as attachments and TIs characteristics. Our results show that spam continues to drive a great portion of unsolicited emails together with phishing. Our analysis of URLs extracted from unsolicited emails revealed a surprising finding - over 80% of these TIs were not flagged as malicious by other threat feeds. This highlights the need for more effective methods of sharing malicious emails and their associated TIs.**

I. Introduction

Email remains a widely used method of communication for both business and personal use [1], but its popularity also attracts individuals with malicious intentions. A large portion of the population has had to deal with spam, so much so that it is estimated that 70% of all business email is spam and 40% of social media accounts are used to distribute spam [2]. Unsolicited emails are also used very commonly as a constant source of various scams and online threats including deception and distribution of phishing or malware attacks and is therefore recognized as one of the most important threat types when it comes to DNS abuse mitigation and identification [3]. However, since the spectrum of threats that spam can deliver could be broad, specially for incident response and threat mitigation, it is important to be able to identify the specific type of threat an unsolicited email attempts to deliver [4].

Unsolicited emails can be detected and reported via various means. Commercial email clients (e.g., Gmail and Outlook) have a mail filter that determines whether an email is unsolicited or not [5], [6]. Typically, within these clients there is a also a feedback loop where the user can tag emails as spam. This process is complex; due to the constant evolution of the techniques used to camouflage content and deceive spam filters, it is necessary to read the subject of the email and analyze the body of the email. Alternatively, there are open-source solutions for identifying unsolicited emails such as SpamAssassin [7]. Similarly, the Anti Phishing Working group (APWG) [8] provides an unsolicited email feed that contains millions of emails as captured by its members.

Despite the availability of solutions for identifying unsolicited emails (e.g., [9], [10]), there is a gap in the ability to distinguish between email-delivery scams, spam, and phishing emails. Distinguishing between different types of unsolicited emails is not merely an academic pursuit; it is a critical necessity in the battle against cyber threats. Each type of unsolicited email, whether it be spam, phishing, scams, or adult content, presents distinct risks and consequences for individuals and organizations [11]. Failure to accurately categorize and understand these threats can lead to ineffective mitigation strategies and increased vulnerability [12]. This paper aims at unraveling the complexities of unsolicited emails and accurately classifying them. This empowers incident responders to prioritize their actions, deploy targeted countermeasures, and protect against psychological, financial, and data-related harm. This leads us to investigate the following research questions:

(**RQ1**): What are the different types of emails are reported as unsolicited emails?

(**RQ2**): What threats that can be identified from these emails? How do the extracted URLs from unsolicited emails compare to those on high-reputable reputation block lists and AV vendors in terms of coverage?

(**RQ3**): How can we categorize the different types of emails to contextualize TIs? How does the language of the emails impact the categorization?

To answer these questions, we leverage a total of 10.8 million emails collected during 4.5 years (May 2018–Dec 2022) which were reported as phishing to the APWG exchange [13]. Using this dataset, we create an email processing pipeline to sanitize the emails and extract features that can be used to split these emails intro four categories: spam, scam, phishing and adult content. The main contributions of this paper are:

- We develop a method for classifying unsolicited emails, specifically for spam, phishing, scam, and adult content.
- We evaluate the adequacy of different models with various features such as an analysis of links within the email, sender

domains, and attachments.

- We evaluate the effectiveness of the models on classifying non-English languages (i.e., Spanish, French, Russian, Japanese, Portuguese and German).
- We extract TIs that can aid incident response teams to take appropriate actions. We assess the coverage of these TIs by comparing the extracted URLs against both reputation block lists and Virustotal [14].

II. Related work

A. Email classification

Several surveys have already captured the different techniques (e.g., [15]), mostly leveraging machine learning. Youn and McLeod [16] classified email data using four different classifiers (Neural Network, SVM classifier, Naive Bayesian Classifier, and J48 classifier) and showed that a simple J48 classifier could be efficient for datasets of emails that could be classified as binary tree. Dredze et al. [17] presented several algorithms for automatically recognizing emails as part of an ongoing activity, including SimSubset and SimOverlap algorithms that compare the people involved in an activity against the recipients of each incoming message, and a SimContent algorithm that uses IRR to classify emails into activities using similarity based on message contents. There have also been some studies focusing on the specifics of a particular language to classify emails. Alsmadi and Alhami [18] used clustering and classification algorithms to perform folder and subject classifications on a large set of personal emails. They showed that classification based on NGram is effective for such large text collections with Bi-language content (English and Arabic). Our work builds on the knowledge of automatic email classification, but differs from these studies as we aim to build a language-agnostic classifier specific for unsolicited emails, i.e., our aim is not to identify ham.

B. Ham, Spam, Phishing detection

Machine learning and artificial intelligence techniques have been widely used in the detection of email-based attacks such as phishing and spam [19], [20]. Existing studies have proposed various approaches to improve the accuracy of detection and have identified future research directions in this area [21]. Fette et al. [22] proposed a method for detecting phishing emails using machine learning. They evaluated their approach on a dataset of approximately 860 phishing emails and 6950 non-phishing emails, achieving a detection rate of over 96% and a mis-classification rate of only 0.1% of legitimate emails. Yasin and Abuhasan [23] proposed an intelligent classification model for detecting phishing emails using knowledge discovery, data mining, and text processing techniques. Hayati and Potdar [24] analyzed existing works in two different categories of spam domains, including email spam and image spam. They presented an evaluation of spam detection frameworks and identified future research directions.

In this paper, we select machine learning models that have been previously identified as having good performance [25] to use them in a new context, i.e., multiclass classification of unsolicited emails. We build on the features already proven to be useful [26] for detection to create a novel set of features for classification of exclusively unsolicited emails. By extracting non only content-specific features, we train different algorithms using the threat indicators within the emails subject, body and attachments. This novel expansion of features increases the performance of the classifier achieving an accuracy above 93.1% (TF-IDF).

III. Methodology

A. Data collection

The unsolicited email dataset used in our study was obtained from the APWG [8] archive of emails reported as phishing, spanning from May 2018 to December 2022. The APWG is a collaboration of security experts, businesses, law enforcement, and other stakeholders impacted by, or combating, phishing attacks. In our effort to better understand the current landscape of unsolicited emails, we collected 10,849,051 (10.85M) million emails reported as phishing. Although our manual process of generating ground truth reveals that there are no legitimate emails (ham) in this dataset, it is evident that not all emails in the dataset are related to phishing. Table I provides a summary of the data fields in the dataset.

TABLE I: Dataset fields

Field	Description
ID	Identifier of the reported email
Email subject	Content of the email subject in the reported email
Sender email	Email address of the reporter who sent the email
Recipient email	Email addresses to where the reported email was sent
Email body	Raw content of the email's body
Email headers	Raw content of the email's headers
Date Reported	Email's reported date and time
Attachments	Files attached to the email

B. Data processing

Figure 1 shows the different phases of the email processing pipeline. In the next sections, we delve into each stage of this pipeline in more detail, exploring the different techniques and tools that we used to accomplish each task.

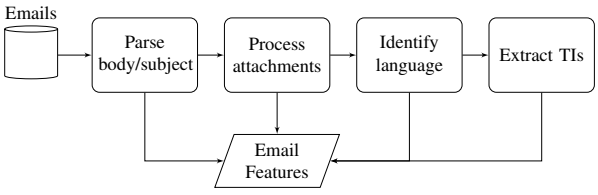


Fig. 1: Pipeline to extract email features

1) Parse email body/subject: The initial step in processing the email body and subject is to remove HTML tags, alert messages generated by email servers, empty lines, non-alphanumeric characters, and characters added to the beginning of lines in certain email clients when forwarding messages. The result is only the text of the subject and the body

of the messages, which provides the necessary information to facilitate the task of classification. The data process as described in Figure 2 is an iterative process that involves several steps to ensure that the final text extracted from each email is free of extraneous characters and noise. The first step involves the decoding of base64 emails, which are then passed through a parser to remove any HTML code in the email, preventing the generation of noise or unwanted text. The next step involves the removal of any text obfuscation in the email and the decoding of emojis. Emojis are converted to text, decoded, and then converted back to the original emojis to ensure they are retained in the final text. During the email cleaning process, we found that it is common to have text obfuscation in both the subject and body of the email. The obfuscation consists of replacing Latin alphabet letters with visually similar letters that have a different ASCII code.

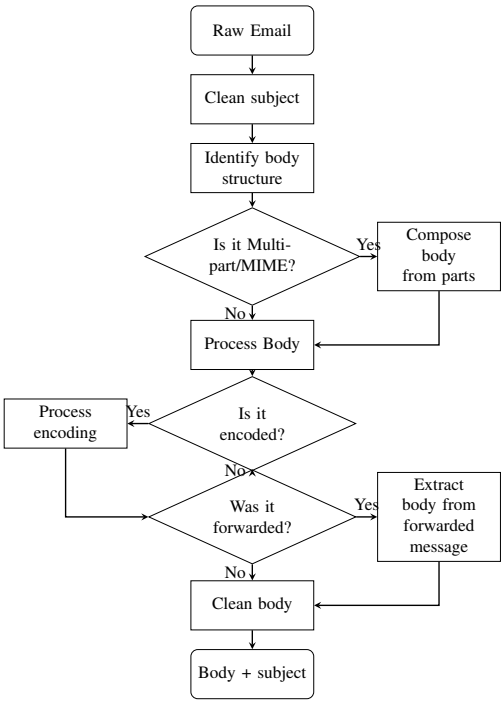


Fig. 2: Email cleaning pipeline.

2) Processing attachments: To process email attachments, we first extract the text embedded in any file attached to the emails. This requires the use of OCR techniques (e.g., [27]), which can be computationally expensive when handling a large volume of emails. Some cybercriminals use content obscuring techniques, such as distorting images, to the extent that OCR techniques cannot extract any meaningful text. In our dataset, we did not encounter content obscuring techniques for text embedded into attached images.

3) Language identification: Detecting the language of an email can affect the vectorization and extraction of features because different languages have unique characteristics in terms of grammar, syntax, and vocabulary. We start by combining the text from the subject and the body to then detect the language of the merged text. To do this, we use Freeling [28]. For each query, Freeling returns the language of the text and the confidence level of the detected language.

4) Extraction of Threat Indicators (TI): We extracted potential TIs (i.e., bitcoin addresses, hashes, URLs, email addresses, IP addresses and domain names) that may be linked to the type of attack. We utilized regular expressions to identify URLs and domain names. We extract of URLs and domain names derived from both the text of emails and text derived from images. When we found defanged strings [29], we refanged them and extracted the URL and domain name that matched the regular expression. We developed a regular expression to match Bitcoin addresses. Bitcoin addresses are base58check encoded integers with a checksum. Our regular expression to match Bitcoin addresses is: `\b[13][a-km-zA-HJ-NP-Z1-9]{25,34}\b`. We validate the checksum of the identified Bitcoin addresses.

C. Email classification

We manually inspected a large set of emails (a random sample of 2,500) and identified common categories of unsolicited emails. Then, two independent researchers developed a coding scheme based on these categories and used an in-house content analysis tool to analyze a larger set of emails. This process resulted in the identification of four distinct categories of unsolicited emails:

- *Phishing*: emails designed to deceive and trick the recipient into revealing sensitive or confidential information, such as login credentials or financial details. Phishing emails often appear to come from a legitimate source, such as a bank or a trusted organization, but are actually sent by criminals seeking to obtain personal information. The content of a phishing email may include a request to click on a link, download an attachment, or provide personal information through a form or reply message.
- *Scam*: emails that are designed to trick the recipient into taking a particular action that benefits the scammer, such as sending money, or providing personal information. Unlike phishing emails, which typically seek to obtain personal or financial information, scam emails may involve a wide range of deceptive tactics, such as offering fake job opportunities, lottery winnings, or other fraudulent schemes.
- *Spam*: emails that are unsolicited and usually sent in bulk to a large number of recipients. Spam emails often contain advertising or promotional text, and may be sent for commercial purposes, or to drive traffic to a website.
- *Adult content*: emails that contains adult-related content. While adult content emails can be an annoyance and may be considered inappropriate or offensive by some recipients, they generally do not pose a direct security threat or attempt to extract sensitive information from the recipient.

While these categories are not always exclusive (e.g., a phishing email can also use adult content), we use a simple rule of thumb that always favored labeling an email with the most harmful category. To facilitate the agreement between the two coders, we used the following preference classification rule: Phishing $\succ$ Scam $\succ$ Spam $\succ$ Adult content.

IV. Descriptive Results

A. Reported Emails

Over the span of 56 months, more than 10.8 million emails were reported to the APWG. The number of emails per month exhibited a clear increasing trend (see Figure 3). The average monthly number of reported emails in 2019 was around 85,000, while in 2022 it was 364,000, representing a four-fold increase. Given the nature of this inbox, it was common for the same phishing email to be reported multiple times by different senders. Out of the 10.8 million emails, only 7.5 million contained unique combinations of subject and body, meaning that 30.5% were duplicates.

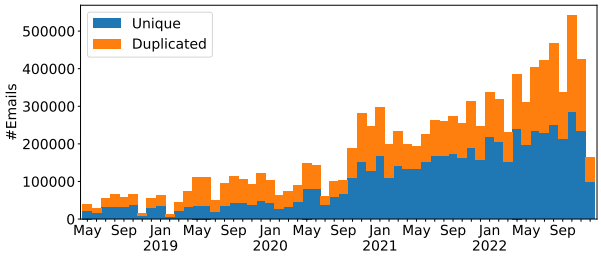
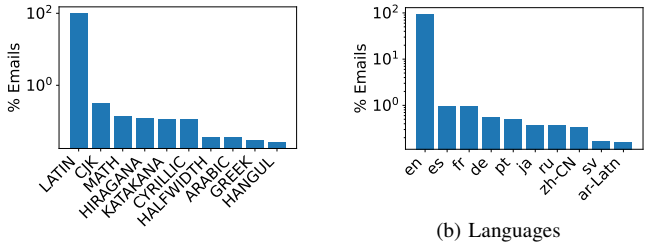


Fig. 3: Number of reported unsolicited emails per month.

B. Email languages and alphabets

The alphabets found in the body of text and subject lines of unsolicited emails are varied, but the majority of the text is composed of Latin characters. As shown in Figure 4a, Latin characters make up 99.96% of the text, which is not surprising since Latin is the most commonly used alphabet in the world. The other alphabets found in these emails are CJK (Chinese, Japanese, and Korean), Mathematical, Hiragana, Katakana, Cyrillic, Katakana-Hiragana, Arabic, Halfwidth, and Greek. These other alphabets are likely used by spammers and scammers to try to bypass spam filters or to appear as if the email is coming from a legitimate source.



(a) Alphabets

(b) Languages

Fig. 4: Percentage of emails per alphabet and language

Figure 4b shows the top 10 detected languages in the dataset. The great majority –94.1% of the emails– are in English, followed by Spanish (0.96%), French (0.94%), and German (0.55%), with the remaining 3.43% distributed among 79 other languages. This distribution of languages is not surprising as spammers and scammers aim at creating a sense of familiarity or authenticity that would not always be possible with a message written in English. For example, an email written in Spanish may be more likely to fool a Spanish-speaking individual into thinking that the message is legitimate. Similar to the use of different alphabets, using languages other than English is also an attempt to evade spam filters that are designed to flag emails with certain keywords or phrases in English [30].

C. Threat Indicators (TIs)

Table II provides insights into the types and frequency of TIs extracted from the sanitized email dataset. These TIs will be later used to classify the type of threat that was reported. The table shows that email addresses[1] are the most common type of TI extracted, with 7.5 million instances and being present in 69.46% of the total number of emails. This indicates that attackers frequently use email addresses in their attacks, in this case as means of communication. Note that some contributors do not forward the whole unsolicited email but a summary, and hence, not all emails contain the original email address from which the email was sent. The second most common TI type is domain names, accounting for 5.9 million instances (54.65%). Attackers often use domain names to host malicious content. Emails also contain potentially malicious URLs, but they are less prevalent (26.64%) IPv4 addresses are the third most common TI type, reported in 2.5 million instances (23%). IPv6 addresses are rare (0.79%). Other TI types like MD5 hash and SHA1 hash have lower frequencies, indicating they are less commonly used by attackers. However, even a small number of occurrences of these TI types can help identify potential threat types.

TABLE II: Extracted TIs

TI Type	Count	Ratio (%)
BTC address	6,890	0.06
Domain name	5,901,806	54.65
Email address	7,501,933	69.46
IPv4 address	2,484,354	23.00
IPv6 address	85,550	0.79
MD5 hash	277,900	2.57
SHA1 hash	23,492	0.22
SHA256 hash	2,1289	0.20
URL	2,877,114	26.64

Figure 5 shows the distribution of TIs extracted from unsolicited emails over time. The number of domains and emails remained high throughout the monitored period, with the number of emails increasing to a peak in January 2021. The use of IP addresses as TIs also followed a similar pattern, with the number of IPv4 and IPv6 addresses used as TIs peaking in January 2021 and April 2022, respectively. In contrast, the number of Bitcoin addresses used as TIs varied widely over the monitored period. The number of URLs used as TIs was relatively low. The number of hashes used as TIs was also relatively low and did not follow a specific pattern over the monitored period.

[1]Note that these addresses do not represent the reporter's email address, but rather the address of the sender of the unsolicited email which is not always present.

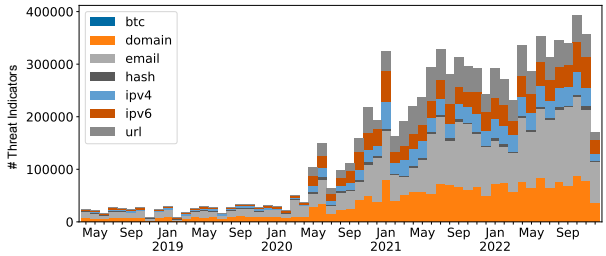


Fig. 5: Number of extracted threat indicators over time

V. Building a classifier

A. Feature Engineering

In order to classify the type of unsolicited email, we extract features that are relevant to this task. We use four different types of features: reported features, content features, TI features, and attachment features.

a) Reporter Features: Understanding the differences in the reporting behavior of different users can be crucial in identifying patterns in unsolicited emails. By extracting reporter features, we can identify the type of emails that different users are more likely to report. For example, some users may be more likely to report phishing emails, while others may be more likely to report spam emails. By identifying these differences, we can more accurately classify the type of unsolicited emails. We consider the number of emails reported by the sender, the domain name of the email address, and the sender's activity period (i.e., the number of days from the first to the the last reported email).

b) Content Features: The content of an email can reveal important information about its type. By extracting content features, we can identify patterns that are indicative of different types of unsolicited emails, such as phishing or scam emails. We include eight types of features such as the number of characters, number of words, number of URLs or domain names, presence of URL in the subject, obfuscation of content, and non-Latin characters. For example, the use of obfuscation techniques or non-Latin characters may be indicative of a spam email.

c) TI Features: TIs are pieces of information that can indicate the presence of a security threat. By extracting TI features, we can identify patterns that are indicative of different types of unsolicited emails. We first capture the type of TI as a feature in itself, and then design the following five features specific to URLs, domains, and email/IP addresses: number of characters in the domain name, number of characters in path, number of digits, and top-level domain. For example, certain types of URLs or domains may be associated with phishing attempts, while others may be associated with spam emails.

d) Attachment Features: Attachments can contain important information that can help identify the type of unsolicited email. For example, the use of certain symbols or characters in an attached image may be indicative of a phishing attempt. The five types of attachment features we designed include the number of characters, number of words, number of symbols, number of digits, and number of URLs or domain names. These features help us assess if the images in the emails are related to potential attacks by capturing differences in the extracted strings.

B. Feature selection

Feature selection is a crucial step in unsolicited email classification to ensure that only the most informative features are used for model training and inference. By selecting relevant features, the model can achieve higher accuracy while reducing overfitting. One widely used method for feature selection is Boruta [31], which compares and statistically tests the importance of features by training a random forest with shadow features that do not contribute to classification. In the context of unsolicited email classification, the feature selection process involves selecting the most informative features from the 869 dimensional feature space. Using the Boruta method, 24 out of the 869 features were considered important, and the remaining 845 features were deemed unimportant. In fact, only content features and TI features were considered important.

C. Balancing classes

Based on the ground truth created manually and upon checking the class distribution, it is observed that they are imbalanced, with some categories having more samples than others. The spam category has 47% of all samples, followed by phishing with 20%, scam with 18%, and adult content with 15%. Such imbalanced datasets can cause bias during training or deteriorate performance. Therefore, three techniques have been considered to balance the datasets: class augmentation, downsampling and upsampling. After analyzing the different methods used to balance the datasets (see Table III), we determined that, while data downsampling and data augmentation were also considered, data upsampling provided the best performance. To mitigate the risk of overfitting when using upsampling, we employed cross-validation.

TABLE III: Classifier performance results for balanced classes

	Accuracy	Fscore	Precision	Recall
Unbalanced	77.07	76.87	78.22	77.07
Balanced				
Augmented	85.71	85.57	86.03	85.71
Downsampled	77.81	77.57	78.83	77.81
Upsampled	89.90	89.89	90.39	89.90

D. Training and evaluation

For this phase, we selected four models to verify their effectiveness in creating a classifier for unsolicited emails: SVC [32], NB [33], LSTM [34], and TF-IDF [35]. Additionally, the variations LinearSVC [36] and BILSTM [37] have been used, making a total of six models. All six models are supervised learning models that require a properly labeled dataset.

For the testing and training phase, the combination of email fields is a crucial component to be tested with the different models. The fields subject, body, and text from the attachments

are combined together in tuples. The newly formed text is then vectorized to generate an integer value matrix, based on the vectorization technique used in each model.

In the training phase, the selected models will be tested with a combination of email fields and different features. These experiments will test the models using the following email field combinations: body, subject+body, subject+body+attachment, features+body, features+subject+body, and features+ subject+body+imagetext. For evaluating the models, we employed Stratified K-fold Cross-Validation (SKCV) [38] with $K = 10$.

The results showed that the combination of fields (subject, body, and attachments) in the email impacted the models' performance (see Table IV). Our results unveil noteworthy disparities in performance across the examined models, with BILSTM and LSTM standing out as the formidable frontrunners. Notably, BILSTM exhibited commendable precision (91.8%) and F-score (91.6%), when using subject, body and attachment data. LSTM, on the other hand, showcased exceptional precision (93.5%) when incorporating subject and body information, underscoring its efficacy in distinguishing unsolicited emails. Conversely, LinearSVC encountered considerable challenges, consistently yielding lower precision scores across all feature combinations. Meanwhile, the resolute NB and SVC models displayed robust performance, yielding precision rates ranging from 82% to 87.4%. The incorporation of TIs did not improve the performance of the machine learning models for unsolicited email classification, as evidenced by the marginal or no significant changes observed in accuracy, F-score, precision, and recall metrics across different models.

TABLE IV: Models performance results

Model	Features	Accuracy	F-score	Precision	Recall
BILSTM	TIs, body	88.7	89.1	90.3	88.7
	TIs, subject, body	87.3	87.5	87.9	87.3
	TIs, subject, body, attch	90.7	90.7	90.8	90.7
	body	90.0	90.1	90.3	90.0
	subject, body	91.1	91.3	91.9	91.1
	subject, body, attch	91.6	91.6	91.8	91.6
LSTM	TIs, body	90.9	91.1	91.6	90.9
	TIs, subject, body	92.4	92.6	92.9	92.4
	TIs, subject, body, attch	90.4	90.6	90.9	90.4
	body	90.9	91.0	91.4	90.9
	subject, body	92.2	92.5	93.5	92.2
	subject, body, attch	91.6	91.7	92.2	91.6
LinearSVC	TIs, body	85.9	85.7	85.8	85.9
	TIs, subject, body	86.5	86.3	86.4	86.5
	TIs, subject, body, attch	88.3	88.2	88.2	88.3
	body	85.0	84.7	85.0	85.0
	subject, body	86.1	85.9	86.0	86.1
	subject, body, attch	87.4	87.2	87.2	87.4
NB	TIs, body	82.0	81.8	82.2	82.0
	TIs, subject, body	83.7	83.6	84.0	83.7
	TIs, subject, body, attch	84.1	84.0	84.4	84.1
	body	82.1	81.9	82.0	82.1
	subject, body	83.6	83.6	83.9	83.6
	subject, body, attch	84.2	84.1	84.4	84.2
SVC	TIs, body	81.5	81.8	84.0	81.5
	TIs, subject, body	84.2	84.4	86.0	84.2
	TIs, subject, body, attch	85.2	85.4	86.8	85.2
	body	81.5	81.8	84.1	81.5
	subject, body	83.7	84.0	85.8	83.7
	subject, body, attch	84.8	85.0	86.5	84.8
TFIDF	TIs, body	92.8	92.8	92.8	92.8
	TIs, subject, body	91.1	91.0	91.2	91.1
	TIs, subject, body, attch	91.6	91.6	91.9	91.6
	body	92.3	92.3	92.4	92.3
	subject, body	93.1	93.1	93.1	93.1
	subject, body, attch	92.0	91.9	92.2	92.0

1) Language impact: Using the LSTM classifier with all features, we evaluate its performance against emails which were predominantly written in a specific language. We limited this analysis to the top seven common language in our dataset, i.e., English, Spanish, French, German, Portuguese, Japanese, and Russian. The results are shown in Figure 6.

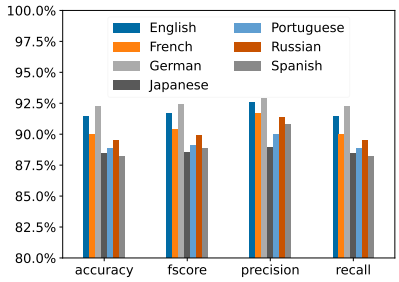


Fig. 6: Classifier performance per email language

Overall, the classifier performed well in classifying unsolicited emails in all languages, with precision, recall, F-score, and accuracy scores ranging from 87.4% to 92.9%. The highest-performing language was German, followed closely by English and French, while the lowest-performing language was Japanese in terms of precision, recall, and F-score.

There could be several reasons why the performance of the classifier varies depending on the language of the unsolicited emails. One reason could be the structural differences between languages. Some languages, such as Japanese, have more complex grammatical structures, while others, such as English, have simpler structures. The classifier may perform better on languages with simpler structures because it can more easily learn patterns in the text.

VI. Threat Evolution

Using the classifier with the highest F-score as trained in section V, we classified all emails in our dataset. The results show that unsolicited emails have evolved over the years in terms of their type and volume. Figure 7 shows the number of reported cases of different categories of unsolicited emails from 2018 to 2022.

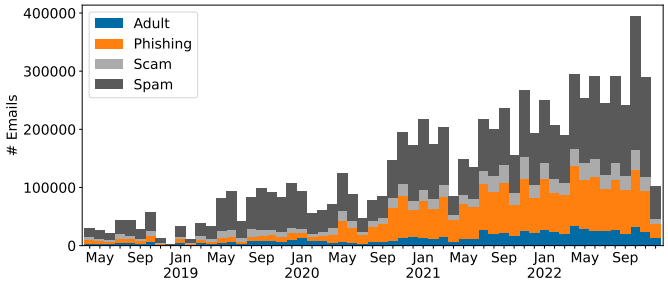


Fig. 7: Number of unsolicited emails per type over time.

The number of reported cases of unsolicited emails has increased steadily over the years, with the most significant increase in spam emails in 2022. The type of unsolicited emails has also evolved, with adult and phishing-related emails increasing significantly. Between 2018 and 2023, the number

of reported cases of adult-related unsolicited emails increased steadily, with the highest number of reports in 2022. Phishing-related emails also increased during this period, with a sharp increase in 2022. Scam emails peaked in 2019 and then experienced a slight decrease in 2020 and 2021, with a slight increase in 2022. Spam emails, on the other hand, have consistently remained the most prevalent category.

A. TIs per email type

Table V shows the percentage of emails containing a particular TI categorized by the type of unsolicited email. The distribution of TIs varies depending on the type of unsolicited email, with domain-based TIs being the most common type for phishing and spam, email-based TIs being the most common for scams. Looking at the distribution of TIs, the most common type of TI for all four categories of unsolicited email are domain names, with a significantly higher number of TIs for phishing and spam than the other two categories. The number of domain-based TIs for phishing and spam is in the millions, which is several times higher than the domain-based TIs for the other two categories.

TABLE V: Threat indicator concentration per email category

TI	Adult	Phishing	Scam	Spam
BTC	0.04	0.02	0.02	0.05
domain	36.98	39.32	32.29	43.74
email	49.39	21.99	50.00	29.75
hash	0.73	1.38	1.25	1.33
IP address	12.16	16.67	15.86	22.46
URL	0.70	20.62	0.57	2.67

VII. Extracted URLs coverage

To assess the uniqueness and overlap of the extracted URLs, we conducted a comparative evaluation against VirusTotal (VT) [14] and multiple reputation blocklists (RBLs) that publish malicious URLs.

A. Comparison with VirusTotal

We used VT to analyze whether the antivirus engines flag malicious URLs collected by our method. VT evaluates the maliciousness as reported by around 90 different types of antivirus engines. We selected a random sample of 500,000 URLs that we extracted from the unsolicited emails and ran them through VT. This sample size was chosen based on the scanning quota that was available to us.

Only 49.65% of the scanned URLs had ever been scanned by VT. This implies that more than 50% of the URLs that we extracted from the unsolicited emails were never scanned by VT publicly. We retrieved the scanning results for the remaining set of URLs. VT offers four categories of outcomes for scanned URLs: harmless, malicious, suspicious, and undetected. On average, 74.94 (83.34%) AV engines flagged these URLs as harmless at the very first scan, while only 5.2 (5.78%) AV engines labeled them as malicious. In fact, 24.98% of all the URLs that were scanned were never labeled by any engine as malicious. Even though a portion of these could be deemed as falsely reported, given the high level of confidence of these reporters (e.g., national CERTs, CSIRTs, etc.), this highlights the increased coverage that the extracted URLs provide.

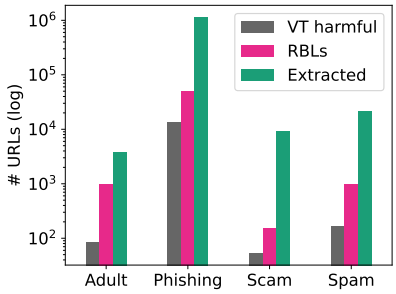


Fig. 8: Number of URLs flagged as malicious

Figure 8 shows that there is a low rate of overlap between the extracted URLs and those flagged as harmful by VT, ranging from 0.6% to 2.2%. For all categories, the ratio of overlap is less than 3%, indicating that the majority of the URLs are not harmful according to VT. However, there are some differences between the categories. The adult category has the highest ratio of overlap, with 2.17% of the extracted URLs flagged as harmful by VT. In contrast, the scam category has the lowest ratio of overlap, with just 0.6% of the extracted URLs flagged as harmful by VT. The results also suggest that phishing and spam URLs have a low ratio of overlap with VT. Both categories have a ratio of overlap between 0.77-1.11%.

B. Comparison with RBLs

We collected six datasets for comparative evaluation from different RBL providers: OpenPhish [39], PhishTank [40], Phishstats [41], APWG [8], WMC PhishFeed [42], and URLHaus [43]. These blocklists contain URLs that are identified as being associated with attacks, and are used by organizations and individuals to protect against such attacks [44], [45], [46]. In this context, it is important to evaluate the overlap between the URLs extracted from unsolicited emails, and those included in RBLs. These RBLs were collected from 2020 to 2023 so for a fair comparison we also restricted the URLs extracted from unsolicited emails to this period.

TABLE VI: Comparison of common URLs between the extracted URLs and RBLs.

	Common URLs	
	Number	Percentage (%)
Phishstats	25,946	3.93
APWG	19,870	3.01
WMC PhishFeed	16,764	2.54
Phishtank	16,033	2.43
Openphish	11,945	1.81
URLhaus	2,275	0.34

When comparing the URLs from RBLs to those extracted from unsolicited emails, we discover that there is only a 7.71% overlap, which amounts to 51,163 URLs in common. The extent of this overlap varies depending on the specific RBL (see Table VI). We find that Phishstats has the highest

number of common URLs, with 25,946 URLs. APWG has 19,870 unique common URLs, while WMC PhishFeed and PhishTank have 16,764 and 16,033 unique common URLs, respectively. OpenPhish has 11,945 common phishing URLs, while URLHaus has only 2,275 common phishing URLs.

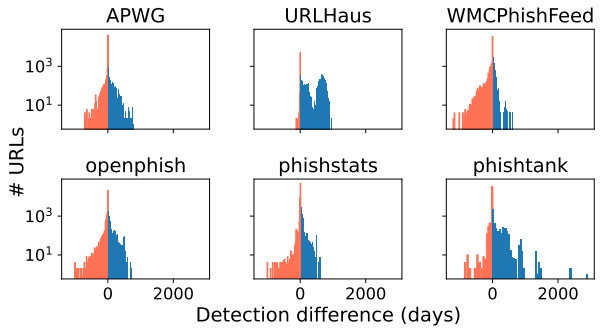


Fig. 9: Histogram of the detection latency per RBL.(red for negative latency, blue for positive latency)

Besides the number of common URLs, we also analyzed the difference in days between the moment a URL is detected in an unsolicited email and the moment it is included in an RBL. 19,956 URLs were reported faster as seen in unsolicited emails that they were included in any RBLs. As shown in Figure 9, only a small portion of common URLs (22.34%) appeared before in one of the six RBLs than they were reported as unsolicited email. On average, it took 3.5 days for a URL to be added to an RBL after being having appeared already in an unsolicited email. In most cases, the URL is added to the block list on the same day or soon after being included in the unsolicited email.

VIII. Limitations

While the use of machine learning to classify spam, phishing, and scam emails demonstrated promising results, our research has some limitations. Firstly, although the LSTM and TF-IDF classifiers showed high performance, none of the techniques reached 100% classification accuracy. This highlights the complexity of distinguishing between the different types of unsolicited emails and underscores the need for further research in this field. Another limitation is that we only trained the classifiers with one dataset, and, even though it is currently the largest collection of unsolicited emails, this dataset may not encompass all the characteristics of future unsolicited emails. Hence, the classifiers' performance may differ when applied to different types of emails. Moreover, we assumed that all unsolicited emails reported by the members of APWG contain some kind of threat. Finally, the study did not assess the classifiers' vulnerability to adversarial machine learning attacks. Adversarial attacks aim to manipulate the classifiers' behavior by modifying the input data, and they are becoming increasingly sophisticated. Therefore, it is essential to evaluate the classifiers' resilience against such attacks and develop countermeasures to prevent them.

IX. Discussion

Our analysis of one of the largest datasets of unsolicited emails reveals a crucial finding: intrinsic heterogeneity exists in the various types of unsolicited emails. Identifying and classifying these emails accurately is essential to take effective countermeasures. By doing so, researchers and security professionals can gain valuable insights into the tactics, techniques, and procedures used by attackers. This understanding of the evolving threat landscape can lead to the development of more effective defense mechanisms. For example, unsolicited emails containing phishing content can serve as an ideal training source for machine learning models to enhance their detection capabilities. By identifying trends in unsolicited emails with malicious attachments or links, we can design targeted awareness campaigns about the risks and best practices for safe email usage.

One of the dominant types of unsolicited emails is spam. These emails can contain a wide range of content, including unwanted commercial offers and misinformation. It is crucial to identify these emails as spam because the threat indicator in these cases is rooted in the email address. Identifying and blocking these emails can help prevent further disruption. Managing spam emails focuses on reducing inbox clutter and minimizing productivity impact through the use of filters and user guidelines. Moreover, unsolicited email classification can also help in identifying emerging threats. For example, the rise of cryptocurrency scams and the use of social engineering techniques such as sextortion have become prevalent in recent years. By tracking the volume and content of unsolicited emails related to these emerging threats, security professionals can identify patterns and anticipate new tactics used by attackers. Our analysis has also shown that a significant number of unsolicited emails contain adult content. Although these emails could also be seen as spam, given their volume and content, they indicate a different type of threat. In such cases, the TIs extracted from those emails deserve a different consideration than TIs from spam, phishing, and scam emails. For example, the presence of BTC addresses in unsolicited emails can serve as an early indicator of the type of threat.

It is also important to note that the volume of unsolicited emails has been increasing over time. This increase is not necessarily driven by an increasing threat landscape, but also by the number of active reporters. Being able to set apart the different types of unsolicited emails also helps to gain a deeper understanding of the current threat landscape.

X. Conclusions

In this paper, we have analyzed a staggering 10.8 million unsolicited emails and found that four categories dominate the space: spam, phishing, scam, and adult content. Our longitudinal analysis reveals a consistent increase in the number of reported unsolicited emails over the past five years; however, not all types of unsolicited emails have seen the same rate of growth. This analysis also served to characterize the activity of reporters as well as the characteristics of the reported emails. Despite fluctuations in the overall volume of unsolicited emails

and reporters, the prevalence of phishing and spam as the primary categories of such emails has remained constant.

We leveraged this dataset containing over 50 languages to train a neural network, achieving high classification accuracy of over 87.4% independently of the language used. This is a significant improvement over publicly available datasets, which are typically limited to English only, making them challenging to use for email classification in other languages.

The analysis of the extracted TIs from the emails allowed us to contextualize the type of threat for each TI. Our analysis of the extracted URLs revealed that more than 95% of these URLs were not previously flagged as malicious by any of the 90 AV engines used by VirusTotal, highlighting the increased coverage that our method provides. Furthermore, our method revealed a strikingly low overlap with RBLs, with some having as few as 0.34% in common. Finally, our findings demonstrate that our method can detect malicious URLs faster than they are added to any RBL, with almost 20,000 URLs reported within unsolicited emails before being added to any blocklist.

References

[1] R. Group, "Email Statistics Report, 2021-2025," https://www.radicati.com/wp/wp-content/uploads/2020/12/Email-Statistics-Report-2021-2025-Executive-Summary.pdf, 2020.

[2] O. Kharif, "'Likejacking': Spammers Hit Social Media," https://www.bloomberg.com/news/articles/2012-05-24/likejacking-spammers-hit-social-media, 2012.

[3] J. Jung and E. Sit, "An empirical study of spam traffic and the use of DNS black lists," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, 2004, pp. 370–375.

[4] M. Alazab and R. Broadhurst, "Spam and criminal activity," *Trends and issues in crime and criminal justice*, no. 526, pp. 1–20, 2016.

[5] K. Wiggers, "Gmail is now blocking 100 million more spam emails a day, thanks to Tensorflow," https://www.computerworld.com/article/2498784/has-the-spam-problem-been-solved-.html, 2019.

[6] C. Metz, "Google says its ai catches 99.9 percent of Gmail spam," https://www.wired.com/2015/07/google-says-ai-catches-99-9-percent-gmail-spam/, 2015.

[7] A. Schwartz, *SpamAssassin*. O'Reilly Germany, 2005.

[8] APWG, "Anti-phishing working group," https://apwg.org/, 2023.

[9] E. Marková, T. Bajtoš, P. Sokol, and T. Mézešová, "Classification of malicious emails," in *2019 IEEE 15th International Scientific Conference on Informatics*. IEEE, 2019, pp. 000 279–000 284.

[10] M. D. Grilli, K. S. McVeigh, Z. M. Hakim, A. A. Wank, S. J. Getz, B. E. Levin, N. C. Ebner, and R. C. Wilson, "Is this phishing? older age is associated with greater difficulty discriminating between safe and malicious emails," *The Journals of Gerontology: Series B*, vol. 76, no. 9, pp. 1711–1715, 2021.

[11] J. A. Teixeira da Silva, A. Al-Khatib, and P. Tsigaris, "Spam emails in academia: issues and costs," *Scientometrics*, vol. 122, pp. 1171–1188, 2020.

[12] M. Jartelius, "The 2020 data breach investigations report–a cso's perspective," *Network Security*, vol. 2020, no. 7, pp. 9–12, 2020.

[13] APWG, "The APWG eCrime Exchange," https://apwg.org/ecx/, 2023.

[14] VirusTotal, https://virustotal.com/, 2023.

[15] Xiao-Lin Wang and Cloete, "Learning to classify email: a survey," in *2005 International Conference on Machine Learning and Cybernetics*. IEEE, 2005.

[16] S. Youn and D. McLeod, "A comparative study for email classification," in *Advances and innovations in systems, computing sciences and software engineering*. Springer, 2007, pp. 387–391.

[17] M. Dredze, T. Lau, and N. Kushmerick, "Automatically classifying emails into activities," in *Proceedings of the 11th international conference on Intelligent user interfaces*. ACM, jan 29 2006.

[18] I. Alsmadi and I. Alhami, "Clustering and classification of email contents," *Journal of King Saud University - Computer and Information Sciences*, vol. 27, no. 1, pp. 46–57, 1 2015.

[19] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab, "A Comprehensive Survey for Intelligent Spam Email Detection," *IEEE Access*, vol. 7, pp. 168 261–168 295, 2019.

[20] S. Smadi, N. Aslam, and L. Zhang, "Detection of online phishing email using dynamic evolving neural network based on reinforcement learning," *Decision Support Systems*, vol. 107, pp. 88–102, 2018.

[21] A. El Aassal, S. Baki, A. Das, and R. M. Verma, "An in-depth benchmarking and evaluation of phishing detection research for security needs," *IEEE Access*, vol. 8, pp. 22 170–22 192, 2020.

[22] I. Fette, N. Sadeh, and A. Tomasic, "Learning to detect phishing emails," in *16th international conference on WWW*. ACM, may 8 2007.

[23] A. Yasin and A. Abuhasan, "An Intelligent Classification Model for Phishing Email Detection," *International Journal of Network Security & Its Applications*, vol. 8, no. 4, pp. 55–72, jul 30 2016.

[24] P. Hayati and V. Potdar, "Evaluation of spam detection and prevention frameworks for email and image spam," in *Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services*. ACM, nov 24 2008.

[25] A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani, "A survey of phishing email filtering techniques," *IEEE communications surveys & tutorials*, vol. 15, no. 4, pp. 2070–2090, 2013.

[26] F. Toolan and J. Carthy, "Feature selection for spam and phishing detection," in *2010 eCrime Researchers Summit*. IEEE, 2010, pp. 1–12.

[27] R. Smith, "An overview of the Tesseract OCR engine," in *Ninth international conference on document analysis and recognition (ICDAR 2007)*, vol. 2. IEEE, 2007, pp. 629–633.

[28] L. Padró and E. Stanilovsky, "FreeLing 3.0: Towards wider multilinguality," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012, pp. 2473–2479.

[29] X. Mertens, "Defang all the things!" https://isc.sans.edu/diary/Defang+all+the+things/22744, 20317.

[30] A. Gendre, "The rise of non-English language spear phishing emails," https://www.helpnetsecurity.com/2021/02/26/non-english-spear-phishing-emails/, 2022.

[31] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, "Boruta–a system for feature selection," *Fundamenta Informaticae*, vol. 101, no. 4, pp. 271–285, 2010.

[32] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.

[33] D. J. Hand and K. Yu, "Idiot's Bayes—not so stupid after all?" *International statistical review*, vol. 69, no. 3, pp. 385–398, 2001.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[35] A. Aizawa, "An information-theoretic perspective of tf–idf measures," *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, 2003.

[36] S. S. Keerthi, D. DeCoste, and T. Joachims, "A modified finite newton method for fast solution of large scale linear svms." *Journal of Machine Learning Research*, vol. 6, no. 3, 2005.

[37] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[38] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2. Montreal, Canada, 1995, pp. 1137–1145.

[39] OpenPhish, https://openphish.com/, 2023.

[40] PhishTank, https://phishtank.org/, 2023.

[41] PhishStats, https://phishstats.info/, 2023.

[42] WMC Global, "PhishFeed," https://www.wmcglobal.com/phishfeed, 2023.

[43] Abuse.ch, "URLhaus — Malware URL exchange," https://urlhaus.abuse.ch/, 2023.

[44] S. Lloyd, C. Gañán, and S. Tajalizadehkhoob, "Block and Roll: A Metric-based Evaluation of Reputation Block Lists," in *APWG.EU Tech*, 2023, pp. 1–8.

[45] E. Rodríguez, R. Anghel, S. Parkin, M. van Eeten, and C. Gañán, "Two sides of the shield: Understanding protective DNS adoption factors," in *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Aug. 2023, pp. 3135–3152.

[46] S. Tajalizadehkhoob, R. Böhme, C. Ganán, M. Korczyński, and M. V. Eeten, "Rotten apples or bad harvest? what we are measuring when we are measuring abuse," *ACM Transactions on Internet Technology (TOIT)*, vol. 18, no. 4, pp. 1–25, 2018.