Methodology to Classify Unsolicited Email Threats

ICANN Office of the Chief Technology Officer

Carlos Hernández-Gañán
Siôn Lloyd
Samaneh Tajalizadehkhoob

OCTO-038
17 January 2024

TABLE OF CONTENTS

This document is part of ICANN's Office of the Chief Technical Officer (OCTO) document series. Please see the OCTO publication page for a list of documents in the series. If you have questions or suggestions on any of these documents, please send them to octo@icann.org.

This document supports ICANN's strategic goal to improve the shared responsibility for upholding the security and stability of the Domain Name System (DNS) by strengthening DNS coordination in partnership with relevant stakeholders. It is part of ICANN's strategic objective to strengthen the security of the DNS and the DNS root server system.

Executive Summary

Email, a fundamental form of communication, faces increasing threats from unsolicited messages. Differentiating these types of threats is essential to take appropriate mitigation measures and deploy effective security controls. It is also an important part of ICANN's mission to monitor, understand, and report email-based DNS threats. Being able to correctly classify a report as being a genuine threat or not means that ICANN, and others, can have greater confidence in our conclusions.

Spam is typically the largest portion of DNS abuse in the data that is listed by reputation feed providers. Therefore, being able to confidently separate spam as a delivery mechanism of malicious content from other types of content is essential for our work.

This research delves into the complexities of this issue, examining the diverse categories, inherent threats, and the role of language in classifying unsolicited emails. To build a dataset of 10.8 million unsolicited emails (spam), which cover a period of four and a half years, this study constructed a robust email processing pipeline and methodology for categorizing unsolicited emails into spam, scam, phishing, and adult content.

The dataset reveals a significant surge in reported unsolicited emails. The predominant language is English, but the emails include a diverse set of languages, potentially employed to deceive recipients. Threat indicators, including email addresses and domain names, play a crucial role in identifying and understanding threats, providing a nuanced view of the tactics used by malicious actors.

The study uses machine learning models, including the Long Short-Term Memory (LSTM) neural network and the frequency-inverse document frequency (TF-IDF) statistical measure, which, together, excel in classifying unsolicited emails across various languages. This process extends beyond English, achieving high classification accuracy across 80+ languages, and demonstrating the adaptability of the models.

A longitudinal analysis of reported cases shows an evolution in the dissemination of unsolicited emails, with a surge in spam in 2022 and a continuous increase in adult and phishing-related emails. Phishing and spam maintain their prevalence, while scam emails fluctuate, highlighting the dynamic nature of email threats over time. The methodology proves robust in classifying unsolicited emails, offering valuable insights into threat types and their evolution.

1 Introduction

Email remains a prevalent form of communication for both private and public interactions. However, its popularity also exposes it to be exploited by individuals with malicious intent. A significant portion of Internet users encounter unsolicited emails, with statistics suggesting that approximately 70% of all business email traffic consists of spam. However, spam, defined as any electronic message sent to a large number of recipients without their consent or permission,[1] is just one type of unsolicited emails, among many others. While the term

[1] Spamhaus Project. The Spamhaus Project – The Definition of Spam. Retrieved from https://www.spamhaus.org/consumer/

"unsolicited emails" is generally understood to refer to unwanted or irrelevant messages, it is worth considering the nuances of this categorization. For the scope of this research, unsolicited emails are considered to be unwelcome and/or malicious.

Unsolicited emails can be detected and reported through various mechanisms. Commercial email clients (e.g., Gmail and Outlook) use mail filters to determine whether emails are solicited or unsolicited. Within these clients, users typically have the option to label emails as unsolicited through a feedback loop. This process entails intricate analyses to keep up with the constant evolution of techniques aimed at concealing content and evading spam filters. This requires examining both email subject lines and email content. Alternatively, open-source solutions like SpamAssassin are available for unsolicited email identification. The Anti-Phishing Working Group (APWG)[2] offers an unsolicited email feed containing millions of emails captured by its members. Nevertheless, given the diverse spectrum of threats that unsolicited emails can potentially deliver, especially concerning incident response and threat mitigation, it is crucial to delineate the type of threats found in unsolicited emails.

This study delves into the complexities surrounding unsolicited emails, including the diverse categories, inherent threats, and the role of language in their classification. It also explores the intricate ecosystem of unsolicited emails, encompassing the various attack vectors employed by threat actors to exploit unsuspecting users.

To address this research, a dataset comprising 10.8 million emails, reported as phishing to the APWG exchange, and collected over a span of four and half years (May 2018–Dec 2022) is used. Employing this dataset, an email processing pipeline is constructed to sanitize email content and extract features that facilitate categorization into four distinct categories: spam, scam, phishing, and adult content.

In this study we describe a methodology for classifying unsolicited emails. Our goal is not to distinguish between solicited and unsolicited emails, but rather to identify various types of unsolicited emails. We also show the effectiveness of various machine learning models that use a range of features, such as the analysis of email links and attachments. We share our assessment of the models' performance in classifying non-English language emails. Finally, we cover the extraction of Threat Indicators (TIs) designed to aid incident response teams in taking appropriate actions.

2 Methodology

2.1 Data Collection

The unsolicited email dataset used in this study was obtained from the APWG archive of phishing emails, spanning from May 2018 to December 2022. In this effort to better understand the current landscape of unsolicited emails, 10.8 million (10,849,051) emails were reported as "phishing." We checked the dataset manually and found that there are no real, legitimate solicitated emails. We also noted the dataset includes emails beyond phishing – some involve other threat vectors (deceptive, harmful emails).

[2] Anti-Phishing Working Group (APWG). (2023). https://apwg.org/

2.2 Data Processing

Data processing is an iterative procedure that encompasses several steps to ensure that the final text extracted from each email remains free of extraneous characters and noise. Figure 1 shows the different phases of the data processing pipeline. Each stage of this pipeline is explained in more detail, exploring the different techniques and tools that are used to accomplish each task.

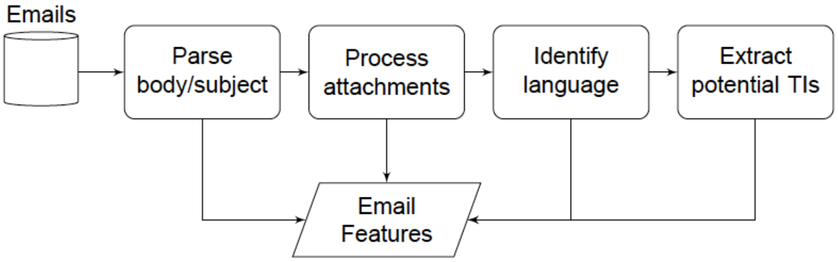


Figure 1. Pipeline to extract email features

Parse email body/subject: The initial step in processing the email body and subject line involves the removal of 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. This leaves only the text of the subject and the body of the messages, which is necessary for facilitating the classification task.

In this initial step, Base64 emails, which are emails containing binary data encoded using the Base64 encoding scheme, are decoded to extract the original binary information. This binary data is then fed into a parser, a software program designed to analyze and extract meaningful information from data. The parser removes any HTML tags or other formatting elements that could introduce noise or unwanted text, leaving behind the essential content of the email. This process of decoding and parsing prepares the email data for subsequent analysis or processing ensures that the extracted information is clean and relevant.

Subsequently, any text obfuscation within the email is removed; obfuscation typically involves the substitution of Latin alphabet letters with visually similar characters that possess different ASCII codes. The obfuscation removal process entails a character-by-character examination of the text, checking whether it corresponds to entries in a language-specific dictionary of visually similar characters. The dictionary maps each character to its corresponding Latin alphabet character.

Finally, any emojis are converted to text, decoded, and reconverted back into their original emojis to ensure their retention in the final text. Throughout the email cleaning process, it was observed that text obfuscation is commonly present in both the subject and body of the email.

Processing attachments: For the processing of email attachments, the text embedded within any files attached to the emails is extracted. This entails the use of Optical Character Recognition (OCR) techniques. Previous research[3] has determined that content obscuring

[3] Arshad, A., Rehman, A. U., Javaid, S., Ali, T. M., Sheikh, J. A., & Azeem, M. (2021). A systematic literature review on phishing and anti-phishing techniques. arXiv preprint arXiv:2104.01255.

techniques, such as image distortion, are often used by cybercriminals to an extent that renders the extraction of meaningful text unfeasible through OCR techniques. In this dataset, no instances of content obscuring techniques being applied to text embedded within attached images were encountered.

Language identification: The detection of an email's language can be important. Unique characteristics in terms of grammar, syntax, and vocabulary are exhibited by different languages, which can impact the email's structure and composition. The initiation of this process involves the combination of text from both the subject and the body, followed by language detection on the merged text, which is accomplished through the use of FreeLing.[4] FreeLing is an open-source library that provides language analysis services, including language detection. It can identify the language of a given text with a high degree of accuracy. For each query, the language of the text and the associated confidence level in the detected language are returned. Subsequently, a filter was applied to discard all emails for which the language detection yielded a confidence level of 30% or lower. This resulted in the discarding of fewer than 2% of the records.

Extraction of TIs: The objective of this step is to extract potential TIs, encompassing bitcoin addresses, file hashes, URLs, email addresses, IP addresses, and domain names. In the extraction of URLs and domain names from emails, regular expressions were used to identify and locate them. Emphasis was placed on identifying matches of URLs and domain names within both the text content of emails related to reports and text extracted from images. In cases where defanged[5] strings were present in the text, a refanging[6] process was executed to restore them to their original form, subsequently allowing the extraction of the matching URL and domain name based on the regular expression.

Additionally, a dedicated regular expression was developed to identify Bitcoin addresses. Each address is a base58check encoded integer, which means it is a string of characters generated from a specific encoding scheme that ensures the integrity of the address. To identify valid Bitcoin addresses in a text, a regular expression pattern is used. This pattern, "\b[13][a-km-zA-HJ-NP-Z1-9]{25,34}\b," matches strings that start with either "1" or "3" (indicating the Bitcoin network prefix) and are followed by 25 to 34 alphanumeric characters (a-km-zA-HJ-NP-Z1-9). Once potential Bitcoin addresses are identified using the regular expression, a validation check is performed to verify the checksum integrity of the address. This checksum is a mathematical calculation that ensures the address is valid and has not been corrupted. If the validation check passes, the identified string is confirmed to be a valid Bitcoin address.

2.3 Ground Truth Generation

A manual inspection of a substantial subset of emails, specifically a random sample of 2,500, was conducted to identify prevalent categories of unsolicited emails. Using an in-house content

[4] FreeLing, The Center for Language and Speech Technologies and Applications (TALP) Research Center – Universitat Politècnica de Catalunya, https://github.com/TALP-UPC/FreeLing and https://nlp.lsi.upc.edu/freeling/

[5] Defanging involves replacing certain characters or parts of a URL with placeholder characters to make it less recognizable or to prevent it from being executed directly.

[6] Refanging is the process of reversing the defanging operation, restoring the original URL or sensitive information to its original form.

analysis tool, this manual set was extended automatically. This systematic process enabled the identification and categorization of diverse types of unsolicited emails within the dataset. As a result, four distinct categories of unsolicited emails emerged:

- **Phishing:** These emails are crafted with the intent to deceive recipients into divulging sensitive or confidential information, such as login credentials or financial details. Often, phishing emails masquerade as messages from legitimate sources like banks or trusted organizations. In reality, they originate from malicious actors aiming to illicitly acquire personal information for fraudulent purposes. The content of a phishing email might include requests to click on links, download attachments, or provide personal details via forms or reply messages.

- **Scam:** Scam emails are designed to manipulate recipients into taking actions that benefit the scammer, which may involve sending money or divulging personal information. Unlike phishing emails, which primarily seek personal or financial information, scam emails use a broad array of deceptive tactics. These tactics include offering fictitious job opportunities, lottery winnings, or other fraudulent schemes.

- **Spam:** These unsolicited emails are typically dispatched in bulk to a large number of recipients. Spam emails frequently contain advertising or promotional content and are often sent with the objective of promoting products or services or driving traffic to a website.

- **Adult Content:** Emails categorized under this label contain content of an adult nature. While adult content emails may be considered bothersome and potentially offensive to some recipients, they generally do not pose direct security threats or try to extract sensitive information from the recipient.

It is noteworthy that these categories may not always be mutually exclusive; for instance, a phishing email may also incorporate adult content. However, a straightforward rule of prioritization was adopted to consistently label an email with the most severe category. To streamline agreement between the two coders, a preference classification rule was established as follows: Phishing ≻ Scam ≻ Spam ≻ Adult content. This prioritization rule provides a structured approach to categorization. However, it is crucial to recognize that the potential harmfulness of an email may not always align with the assigned category. A well-crafted phishing email, for instance, could pose a more significant threat than a less sophisticated scam email, even though both fall under the same priority level. Therefore, it is essential to be cautious and consider the specific context and content of each email when evaluating its potential harmfulness.

3 Case Study: APWG Email Feed

3.1 Number of Unsolicited Emails

Over a period spanning 56 months, the APWG received reports of more than 10.8 million emails. As illustrated in Figure 2, the monthly volume of received emails increased steadily. In 2019, the average monthly count of reported emails was about 85,000, while by 2022, it had surged to about 364,000, a four-fold increase. Given the nature of this email repository, often identical phishing emails were reported on multiple occasions by different senders. Among the

10.8 million emails, only 7.5 million featured unique combinations of subject lines and email bodies. This implies that 30.5% of the total email count consisted of duplicates. Note that the lower volume in the final month can be attributed to the fact that we had only partial data available for that period.

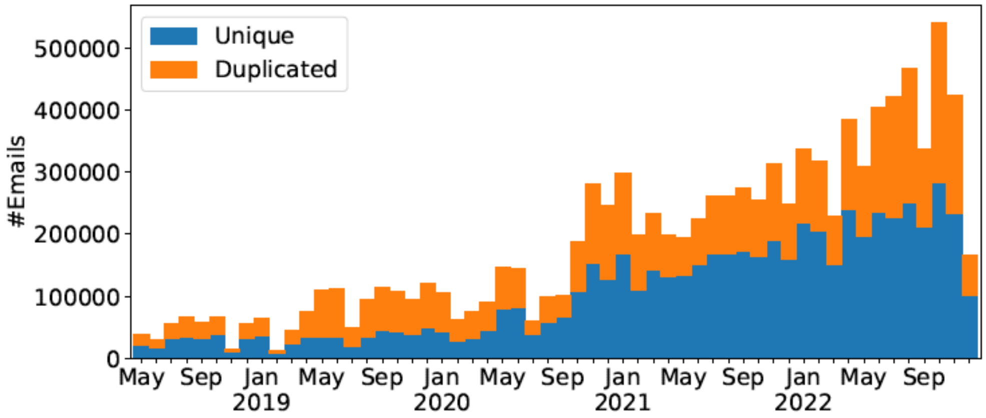
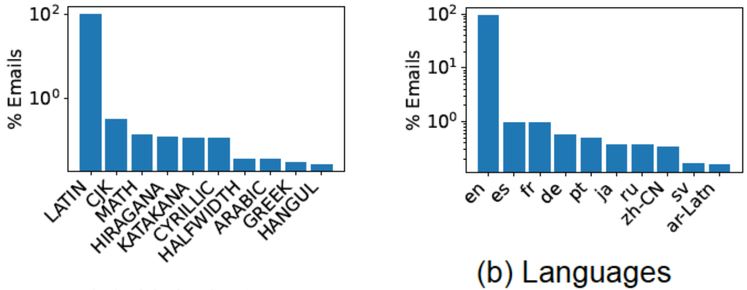


Figure 2. Number of reported unsolicited emails per month

3.2 Email Languages and Alphabets

The text and subject lines of unsolicited emails show a diverse set of alphabets, with Latin characters dominating the composition, as depicted in Figure 3a. Notably, 99.96% of the emails contained text content that includes at least one Latin character, which aligns with expectations considering Latin's global prevalence as the most widely used alphabet. Other alphabets are likely harnessed by spammers and scammers in attempts to circumvent spam filters or create the illusion of legitimate emails.



(a) Alphabets

(b) Languages

Figure 3. Percentage of emails per alphabet and language (note the log scale on the y-axes)

Figure 3b provides insights into the top 10 detected languages within the dataset. A substantial majority, accounting for 94.1% of the emails, are composed in English, followed by Spanish (0.96%), French (0.94%), and German (0.55%). The remaining 3.43% are distributed across 79 other languages. This language distribution is consistent with tactics employed by malicious actors who try to establish a sense of familiarity or authenticity, a feat not always achievable with messages exclusively in English. For instance, an email composed in Spanish might be more effective in deceiving a Spanish-speaking recipient into perceiving a message as genuine.

The use of languages other than English may also serve as an evasion strategy against spam filters designed to flag emails containing specific English keywords or phrases.[7]

3.3 Threat Indicators

TIs are observable pieces of evidence that can be used to detect, identify, and understand cyber threats. They can include specific email addresses, URLs, file names, IP addresses, or other digital artifacts that are associated with malicious activity. In Table 1, an overview of the types and frequencies of TIs extracted from the sanitized email dataset is presented. These TIs play a crucial role in subsequently classifying the reported threat types. The table highlights that email addresses are the most prevalent type of TI, appearing in 7.5 million instances (69.46%) of the total email count. This underscores the frequency with which attackers use email addresses as a means of communication in their attacks. It is important to note that some contributors forward summaries rather than the entire unsolicited email, which can result in not all emails containing the original sender's email address.

TI Type	Count	Ratio (Percent)
Email address	7,501,933	69.46
Domain name	5,901,806	54.65
URL	2,877,114	26.64
IPv4 address	2,484,354	23.00
MD5 hash	277,900	2.57
IPv6 address	85,550	0.79
SHA1 hash	23,492	0.22
SHA256 hash	21,289	0.20
BTC address	6,890	0.06

Table 1. Extracted TIs

The second most common TI type is domain names, accounting for 5.9 million instances, or 54.65% of TIs. Attackers frequently use domain names as hosts for malicious content. While emails also feature potentially malicious URLs, they are less widespread, constituting 26.64% of the TIs. IPv4 addresses rank as the third most common type of TI, with 2.5 million instances (23%), whereas IPv6 addresses are relatively rare at 0.79%. Other TI types, such as MD5 hash and SHA1 hash, are less frequent. Nonetheless, even a few occurrences of these TI types can prove valuable in identifying potential threat types.

Figure 4 shows the distribution of TIs extracted from unsolicited emails over time. Throughout the monitored period, the counts of domains and emails remained consistently high, with the number of emails peaking in January 2021. The use of IP addresses as TIs exhibited a similar trend, with the counts of both IPv4 and IPv6 addresses used as TIs reaching their peaks in January 2021 and April 2022, respectively. In contrast, the number of Bitcoin addresses employed as TIs fluctuated significantly over the monitored period. The use of URLs as TIs remained relatively low compared to domain names. Additionally, the number of hashes used as TIs was relatively low and did not follow a specific pattern during the monitoring period.

[7] Liu, C., & Stamm, S. (2007, October). Fighting unicode-obfuscated spam. In Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit (pp. 45-59).

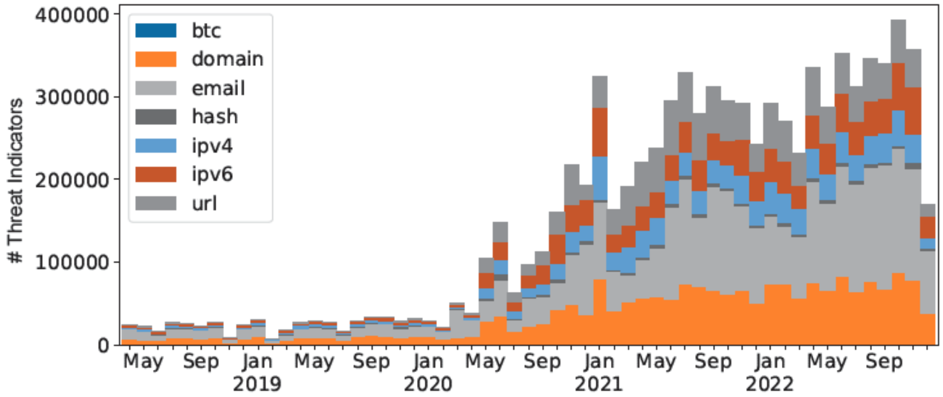


Figure 4. Number of extracted threat indicators over time

3.4 Building a Classifier

3.4.1 Feature Engineering

To classify the type of unsolicited emails accurately, different properties of the individual emails within our sample were examined. These properties are known as features. Four distinct categories of features are created: Reporter Features, Content Features, TI Features, and Attachment Features.

Reporter Features: Understanding variations in user reporting behavior is important to discern patterns within unsolicited emails. Extracting reporter features allows us to identify the specific types of emails that users are more inclined to report. For instance, certain users may be more prone to reporting phishing emails, while others may favor reporting spam emails. These differences aid in refining the classification of unsolicited emails. Three key features are considered within this category: the volume of emails reported by the sender, the domain name associated with the sender's email address, and the sender's activity period, which measures the number of days between the first and last reported email.

Content Features: The content of an email can unveil essential information regarding its type. This category comprises seven distinct features, i.e., the number of characters, words, URLs, domain names, the presence of URLs in the subject, content obfuscation, and the presence of non-Latin characters.

TI Features: Threat indicators can signify potential security threats. Extracting TI features assists in detecting patterns is indicative of different unsolicited email types. Initially, the type of TI is captured as a feature itself and then five additional features related to URLs, domains, and email/IP addresses are extracted. These features include the number of characters in the domain name, the number of characters in the path, the count of digits, and the identification of top-level domains. For example, particular URL or domain types may be associated with phishing attempts, while others may be linked to spam emails.

Attachment Features: Attachments often contain crucial information that helps to identify the type of unsolicited email. For instance, certain symbols or characters within an attached image may signal a phishing attempt. Five types of attachment features are extracted: the number of characters, words, symbols, digits, and URLs of domain names. These features enable us to assess if the images within the emails are related to potential attacks by capturing disparities in the extracted strings.

3.4.2 Feature Selection

Feature selection stands as a pivotal phase in the classification of unsolicited emails, ensuring that only the most informative features are incorporated into model training and inference.

Through the careful selection of relevant features, the model can become more accurate while mitigating the risk of overfitting. Overfitting occurs when a model becomes overly focused on the specific details of the training data, resulting in its inability to generalize effectively to new, unseen data. By selecting only the most relevant and informative features, the model can avoid overfitting and maintain its ability to accurately classify new emails.

One widely adopted technique for feature selection is Boruta,[8] which systematically assesses and statistically tests the significance of features. It achieves this by training a random forest classifier with shadow features, which are artificial features that do not contribute to the classification process. By comparing the importance scores of original features to those of shadow features, Boruta can identify truly relevant features and eliminate redundant or irrelevant ones, leading to a more accurate and robust machine learning model.

3.4.3 Balancing Classes

The ground truth, created through manual annotation, has revealed an imbalance in the class distribution of the dataset, with certain categories containing more samples than others. Specifically, the "spam" category constitutes 47% of all samples, followed by "phishing" with 20%, "scam" with 18%, and "adult content" with 15%. Imbalanced datasets of this nature can introduce bias during model training and potentially deteriorate performance. Consequently, three techniques were considered to address this issue: class augmentation, downsampling, and upsampling.

	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

Table 2. Classifier performance metrics[9] for balanced vs. unbalanced classes

After a comprehensive analysis of these balancing methods (as outlined in Table 2), it was determined that data upsampling yielded the most favorable performance outcomes. To mitigate the risk of overfitting associated with upsampling, a cross-validation approach was employed, ensuring robust and reliable model performance.

[8] Boruta-Shap, Ekeany, https://github.com/Ekeany/Boruta-Shap

[9] Fundamentals and Methods of Machine and Deep Learning: Algorithms, Tools, and Applications. United Kingdom, Wiley, 2022.

3.4.4 Training and Evaluation

In this phase, four models were selected to assess their effectiveness in constructing a classifier for unsolicited emails: Support Vector Classifier (SVC), Naive Bayes (NB), Long Short-Term Memory (LSTM), and Term Frequency-Inverse Document Frequency (TF-IDF). Additionally, two variations, Linear Support Vector Classifier (LinearSVC) and Bidirectional LSTM (BILSTM), were considered, bringing the total to six models. All six models are supervised learning models that rely on a properly labeled dataset.

In the testing and training phase, the combination of email fields played a pivotal role and was systematically assessed with different models. The fields "subject," "body," and "attachment" were combined in tuples. This newly constructed text was then subjected to vectorization, resulting in an integer-valued matrix based on the specific vectorization technique used by each model.

During the training phase, various combinations of email fields and different features were used to evaluate the models. These experiments assessed the models with the following email field combinations: "body," "subject+body," "subject+body+attachment," "features+body," "features+subject+body," and "features+subject+body+attachment." To evaluate model performance, Stratified K-fold Cross-Validation (SKCV) with K = 10 was employed.

The results revealed that the performance of the models was significantly impacted by the combination of email fields (subject, body, and attachments). Notable variations in performance were observed among the models, with BILSTM and LSTM emerging as the top performers. Particularly, remarkable precision (91.8%) and F-score (91.6%) were achieved by BILSTM when incorporating subject, body, and attachment data. LSTM demonstrated outstanding precision (93.5%) when using subject and body information, highlighting its effectiveness in distinguishing unsolicited emails.

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

Table 3. Results of the models' performance

Conversely, considerable challenges were encountered by LinearSVC, consistently resulting in lower precision scores across all feature combinations. However, robust performance was demonstrated by the NB and SVC models, with precision rates ranging from 82% to 87.4%. Interestingly, the inclusion of TIs did not improve the performance of machine learning models for unsolicited email classification. This was evident from the marginal or insignificant changes observed in accuracy, F-score, precision, and recall metrics across different models.

3.4.5 Assessing Language Impact

The impact of language was assessed by using the trained LSTM classifier, incorporating all features, to evaluate its performance on a subset of emails predominantly composed in specific languages. This analysis was limited to the top seven commonly encountered languages within our dataset, namely English, Spanish, French, German, Portuguese, Japanese, and Russian. The results, depicted in Figure 5, reveal the classifier's performance across various languages.

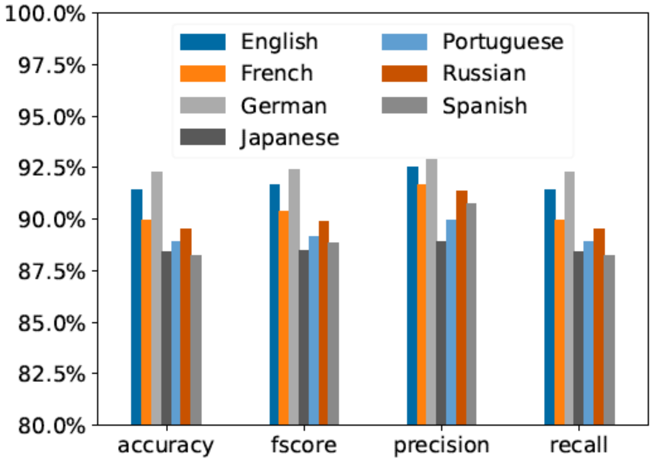


Figure 5. Classifier performance vs. email language

Overall, the classifier demonstrated strong performance in classifying unsolicited emails across all languages, achieving precision, recall, F-score, and accuracy scores ranging from 87.4% to 92.9%. German exhibited the highest performance, closely followed by English and French, while Japanese displayed the lowest performance in terms of precision, recall, and F-score.

Several factors may contribute to variations in the classifier's performance based on the language of unsolicited emails. Structural differences between languages could be one influencing factor. Some languages, such as Japanese, possess more intricate grammatical structures, whereas others, like English, have simpler ones. The classifier may perform exceptionally well in languages with simpler structures because it can easily spot text patterns. Another factor to consider is that more formal languages tend to offer fewer correct ways to convey the same information, resulting in reduced "grammatical complexity," which could further enhance the classifier's accuracy. Another factor could be the quality and volume of training data accessible for each language. Insufficient training data for a particular language may result in diminished classifier performance for that language.

3.5 Evolution of Threats Delivered by Unsolicited Emails

Using the classifier with the highest F-score, as trained in the previous section, all emails within our dataset were classified. The results unveiled an evolution in unsolicited emails over the years, both in terms of their types and volume. Figure 6 provides insight into the number of reported cases for different categories of unsolicited emails spanning from 2018 to 2023.

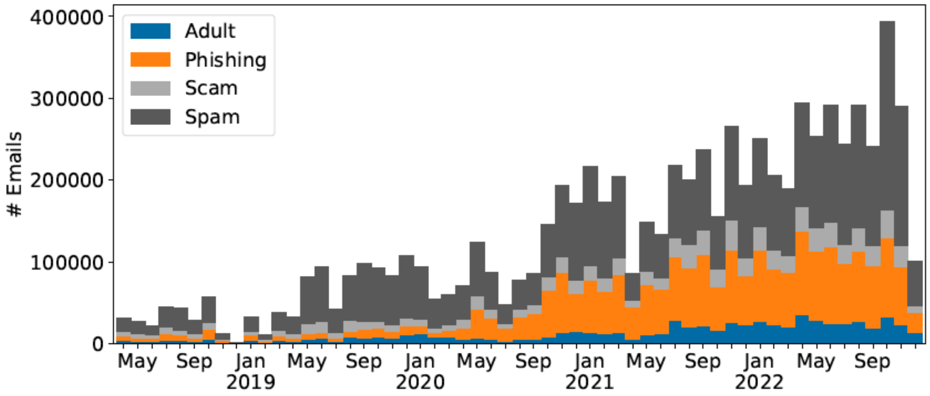


Figure 6. Number of unsolicited email per type over time

The count of reported unsolicited email cases shows a steady increase over the years, with a substantial surge in spam emails witnessed in 2022. Furthermore, the nature of unsolicited emails has evolved, notably with a significant rise in adult and phishing-related emails. Between 2018 and 2023, the prevalence of adult-related unsolicited emails continuously increased, peaking in 2022. Phishing-related emails also surged during this period, particularly in 2022. Scam emails reached their apex in 2019 and subsequently declined slightly in 2020 and 2021, with a modest resurgence in 2022. Conversely, spam emails have consistently maintained their position as the most prevalent category.

3.5.1 Threat Indicators by Email Type

Table 4 presents the percentage of emails containing specific TIs, categorized by the type of unsolicited email. Note that the TIs themselves are not one of the features being used by the classifier. The distribution of TIs varies depending on the type of unsolicited email, with domain-based TIs being the most prevalent for phishing and spam, while email-based TIs dominate for scam and adult emails. The number of domain-based TIs for phishing and spam reaches the millions, surpassing the count of domain-based TIs in the other two categories by several orders of magnitude.

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

Table 4. Threat indicator concentration per email category

4 Limitations

While promising results were demonstrated in using machine learning to classify spam, phishing, adult, and scam emails, several limitations need to be acknowledged in our research. Firstly, despite the high performance achieved by LSTM and TF-IDF classifiers, none of the techniques reached complete classification accuracy. This highlights the complexity of distinguishing between different types of unsolicited emails and underscores the need for further research in this field.

While the classifiers were trained on a substantial dataset of unsolicited emails, it is important to acknowledge that this dataset may not fully capture the evolving nature of unsolicited emails. As new forms of unsolicited emails emerge, the classifiers' performance may vary when applied to these novel types of communications. This underscores the need for continuous monitoring and refinement of the classifiers to ensure they remain effective in detecting a wide range of unsolicited emails.

Moreover, we assume that all unsolicited emails reported by the members of APWG contain some kind of threat. While our manual validation of these emails showed no legitimate emails, this assumption may not hold true for every reported case.

Lastly, an assessment of the classifiers' vulnerability to adversarial machine learning attacks was not conducted in our study. Adversarial attacks aim to manipulate the behavior of classifiers by modifying input data and are becoming increasingly sophisticated. Therefore, it is essential for the classifiers' resilience against such attacks be evaluated with the goal of developing countermeasures to mitigate potential vulnerabilities.

5 Conclusions

In this study, we thoroughly explored the development and application of a methodology to analyze unsolicited emails. Our case study involved examining an extensive dataset comprising 10.8 million unsolicited emails, culminating in the discovery of four dominant categories: spam, phishing, scam, and adult content. The longitudinal analysis demonstrated a consistent increase in the number of reported unsolicited emails over the past five years. This analysis also helps elucidate the characteristics of the reported emails. Despite fluctuations in the overall volume of unsolicited emails, the prevalence of phishing and spam, as the primary categories of such emails, has remained stable.

The methodology presented in this study for classifying unsolicited emails into distinct categories based on the threats they pose has proven to be robust and effective. This methodology used machine learning, particularly LSTM and TF-IDF classifiers, which exhibited commendable performance in distinguishing between these unsolicited email categories. Additionally, our analysis of threat indicators provided valuable insights into contextualizing threat types. Finally, our approach extended beyond the conventional focus on English-only datasets, achieving high classification accuracy across more than 80 languages.