From Reports to Insights: Challenges and Opportunities in Citizen-Driven Malicious Website Datasets

Daan Vansteenhuyse
DistriNet, KU Leuven
3001 Leuven, Belgium
daan.vansteenhuyse@kuleuven.be

Hadji Musaev
DistriNet, KU Leuven
3001 Leuven, Belgium
hadji.musaev@kuleuven.be

Lieven Desmet
DistriNet, KU Leuven
3001 Leuven, Belgium
lieven.desmet@kuleuven.be

***Abstract*—Cybercriminals increasingly exploit the web, targeting millions of users and causing substantial financial losses. To combat these online scams, industry and academia have created databases consisting of malicious websites. By analyzing its properties, various detection mechanisms have been proposed to automatically identify fraudulent activity on the web. Although proven useful, these databases are curated, focus on the global perspective and lack insights of benign websites perceived as malicious by users. In this paper, we analyze user-reported scams from an anti-scam initiative, deployed in a European country, using topic modeling to uncover regional trends and user perceptions. Our findings inform the design of localized anti-cybercrime datasets and detection strategies.**

Based on an initial manual analysis, we find most reported malicious activity takes place in the form of dating scams while a main portion of the dataset contains benign newsletters indicating the varying accuracy of user reports. Using BERTopic to extend the manual analysis, we show how it can be used to study the evolution of campaigns over time. We combine our insights into advice that can be used by anti-cybercrime organizations to set up similar datasets and describe how tools, such as topic modeling, can further aid both industry partners, to harden their anti-phishing defenses, and research institutions, to better study regional and psychological aspects associated with online fraud.

I. Introduction

Since its inception, mail has been a medium prone to misuse by so-called spam messages. Users receive a plethora of unwanted messages ranging from advertisements to malicious phishing campaigns. As long as people will click on spam email, attackers will continue to send out unsolicited messages [37]. Through the years the spam has evolved into full-fledged underground economies of online fraud, such as phishing, targeting thousands worldwide and hurting the confidence of regular users in electronic communication [8].

Luckily, organizations such as the Anti-Phishing Working Group (APWG) [7] and PhishTank [54] maintain databases of malicious websites which are highly valued by the community for cybercrime research. These datasets are composed by collecting reports from a global user base and consist of URLs which lead to fraudulent websites. While we acknowledge the value these organizations and their datasets bring to the anti-fraud community, they can be complemented on two factors. First, previous research has shown that malicious campaigns can be tailored to the victim's location [13], [33], [59], which necessitates a more regional approach for data collection. Second, while useful for training binary classification models, well-known phishing databases filter their reports to only retain malicious content leaving out key insights into what users initially report and thus perceive as malicious.

Thanks to, among others, these type of datasets, various techniques emerged to filter benign from malicious content, based on email content [6], [9], [58] or based on hyperlinks [5], [15], [23], [38], [39], [49] with various accuracies and use cases. Most techniques, however, focus on binary classification leaving out insights about targeted companies or type of scams (e.g. investment, gambling ...).

Where previous work often focused on finding new features or classifiers, we introduce topic modeling as a new analysis tool capable of clustering similar campaigns and identifying how targeted sectors change over time in scam datasets. Topic modeling is an ML technique that sorts related documents into clusters, introducing contextual structure. Applying such analysis to a scam dataset can aid companies in determining if and how malicious actors are targeting their sector of business. By filtering the dataset on topics of interests, cybersecurity analysts can more quickly gather relevant insights and study the evolution of their interests over time.

Regional Dataset. We apply our analysis to a long-standing national anti-fraud initiative in Belgium that is composed of unfiltered user reports of suspicious links. Users can report links by sending them to a reporting email address. By collecting all links that the day-to-day internet user determines to be malicious, this anti-fraud initiative is able to quickly gather insights into regional cybercrime but can also provide insights to companies if certain benign campaigns are being reported as malicious. Contrary to existing datasets of online fraud, this dataset is unfiltered and gives raw insights into what users perceive as suspicious making it a great additional

Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb) 2026
27 February 2026, San Diego, CA, USA
ISBN 978-1-970672-06-0
https://dx.doi.org/10.14722/madweb.2026.23033
www.ndss-symposium.org

source of cybercrime information that can be leveraged in the regional ecosystem. Previous research has found that 44% of citizens residing in our targeted country have reported at least once to this dataset, accumulating to over 9 million reports in 2024, it shows that the dataset encompasses a broad view of regional online fraud [57]. We explore the characteristics of this user-reported dataset and investigate the most important challenges in order to provide guidance on setting up similar services and on automating the data collection and processing pipeline. Using our findings, we are able to identify more specific hurdles that one needs to tackle before a community-driven phishing reporting dataset can be reliably used to study malicious websites. We also identify specific areas of cybercrime research that can especially benefit from this type of dataset that might otherwise find lacking results in conventional databases.

Topic Modeling. We establish base insights into the dataset by manually clustering all reports made in a single day, finding an especially high prevalence of scams with pornographic content but also find that the majority of reports made by users actually consist of benign marketing campaigns from known companies. By employing our topic modeling analysis on a subset of this dataset based on TLD, we confirm our initial findings and measure how reports change over time. We train the model on a limited set of domains and later apply this model to a larger set to show the transferability of the model Additionally, we find that the amount of reports that is sorted into a cluster, remains stable over time.

To summarize, our main contributions are the following:

- We present an in-depth study of suspicious website links in user-reported emails with a regional focus, showing the unique insights such a dataset can provide.
- We show how topic modeling can help cybersecurity analysts to gather insights into malicious datasets.

II. Regional anti-fraud initiative

While datasets such as PhishTank [54] or APWG [7] are highly valued by the community, they provide a global image on cybercrime. Malicious online campaigns, and phishing in particular, however, are increasingly focusing on regional brands [43], [59] necessitating a more narrow dataset tailored to the region.

This motivates why the Centre of Cybersecurity in Belgium (CCB) has set up an initiative where citizens can report suspicious websites to complement existing datasets. Reports can easily be done by forwarding a suspicious email or sending suspicious links to a well-known email address: `suspicious@safeonweb.be`. Although primarily focused on collecting phishing reports, reports are welcome for any link a user finds suspicious. Therefore, we expect to also find benign websites in such a dataset, yet these can provide valuable information for website developers to change the website content to prevent it from being flagged as suspicious by users. Thanks to its perceived trust as a governmental organization (in combination with broad and continued media coverage), the service has been used by over 44% of citizens amassing 9 million yearly reports [57].

In this paper we further investigate how to leverage upon these user-reported messages with an automated data collection pipeline (Section III), provide insights into the user-reported websites (Section IV), and apply topic modeling to measure its evolution over time (Section V).

III. Automated data collection pipeline

Our research leverages upon the reports made to the regional initiative. Based on these reports, an automated data collection pipeline is triggered as shown in Figure 1. In this section, we summarize some insights and challenges in setting up such a pipeline.

The data collection consists of the following coarse-grained steps:

1) Extracting the URLs (i.e. links) from the reports.
2) Visiting the extracted URLs.
3) Storing the website content, crawling metadata and screenshots for further processing.
4) Enrich the data with external sources (e.g. DNS records, blocklist services ...).

A. Challenges

Before such a pipeline can be applied, certain issues need to be taken into account. We identify and discuss four challenges inherent to processing the user-reported links and how we addressed them:

1) Coping with cloaking.
2) Privacy sensitivity of the links.
3) Limited availability of the sites.
4) Varying suspiciousness of the links.

a) Coping with cloaking behavior: A recurring challenge in representatively crawling malicious websites is the possible cloaking behavior of these websites. Cloaking refers to the techniques by which websites try to disguise themselves as being benign or by preventing automated tools to access them. A variety of cloaking mechanisms exist, simple solutions such as hard coded IP or user-agent blocklists but also more advanced techniques such as fingerprinting or Captcha's [2], [62], [71].

In order to prevent our crawler from being detected as one, we employed a couple of anti-cloaking techniques on a best effort basis. We used Chromium `new-headless` mode which removes some bot detection issues previously found in the headless implementations of the browser [18]. Additionally, we manually set a user-agent to match a real Chrome user. Finally, we mimic internal browser aspects such as the Permissions API and override WebGL parameters which can be used to identify webcrawlers. It is, however, likely that malicious websites will still be able to detect our webcrawler and cloak themselves despite our employed anti-cloaking techniques.

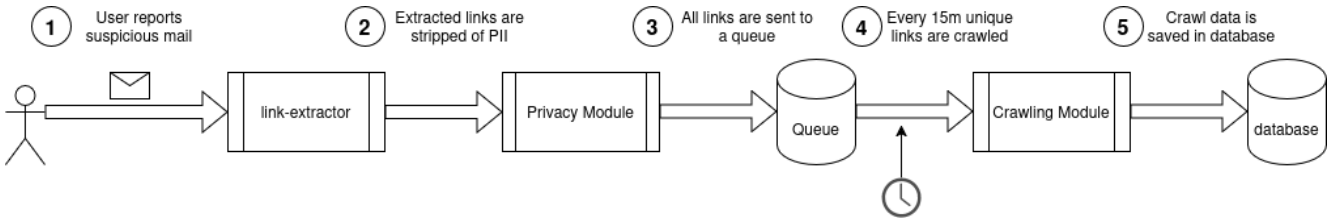

Fig. 1. Overview of the data collection pipeline to crawl user-reported links.

b) Privacy sensitivity of the links: When handling data originating from internet users, there's always a chance Personally Identifiable Information (PII) is exposed. Since reports can only happen via email in our case, the email address from the reporter and possibly the body of the email can link a report back to the original user. Therefore, received emails are automatically parsed and only the hyperlinks found are stored.

Additionally, the reported links themselves can contain personal information such as email addresses in the query parameters [68]. To prevent that, we employ pattern matching to mask all email addresses as part of the user-reported link. This can, though, initiate additional cloaking behavior as the links are modified leading to a trade-off between privacy and accuracy. [1]

c) Limited availability of the sites: Phishing websites are notorious for having a short lifetime [51]. This is likely also the case for other types of malicious websites, such as fake webshops. The availability of the dataset, therefore, poses a concern. A link reported yesterday, has a high chance of being unavailable today. It is thus crucial that malicious websites get investigated as soon as possible.

To limit the time between reporting and data gathering, we set up a pipeline that crawls reported links as soon as possible. Every 15 minutes a webcrawler takes the latest unique reported links and gathers data about the website resulting in a maximum of 30 minutes between initial reporting of the URL and the automated visit of the website.

d) Varying suspiciousness of the links: By relying on users to report suspicious websites, we also rely on their ability to correctly judge what is malicious. Previous work has shown varying results in the correctness of individual users correctly classifying malicious behavior [16], [45]. This type of dataset should therefore not be considered as ground truth but be subject to extensive evaluation and filtering before using it to e.g. train machine learning models. However, it does provide insights into the thought patterns of the day-to-day internet user showing what they perceive to be malicious.

B. End-to-end data collection pipeline

We set up a pipeline as shown in Figure 1 to account for the challenges discussed in this section. The pipeline starts with a user reporting a link or forwarding a suspicious email. To limit the privacy impact for the reporting user, only the links, stripped of possible email addresses in its query parameters, are kept. The resulting links are stored in a temporary queue for at most 15 minutes, before a distributed crawler, with necessary anti-cloaking adjustments, visits the website and stores the gathered data in the final database.

Afterwards, it becomes possible to store DNS data and the status of the website in Google Safe Browsing (GSB), a popular blocklisting service [27].

[1]Note that also the website content itself can embed sensitive information, therefore a best practice pattern matching approach is applied to the website content as well (as will be discussed in Section VI-C).

TABLE I
THE FOUR CATEGORIES OF WEBSITES FOUND IN THE DATASET.

Category	Amount of sites
Undetermined	3,932 (19.0%)
Benign	11,619 (56.1%)
Malicious	954 (4.6%)
Adult Content (likely malicious)	4,238 (20.5%)

IV. Manual analysis baseline

To obtain an accurate picture of the dataset content and of reporting behavior, we perform an in-depth analysis of all links reported by citizens on January 17th, 2025 as a case study. After filtering out non-reachable websites, we are left with a dataset of 20,724 websites reported as being suspicious. This initial analysis will serve as a baseline reference for the later topic modeling experiments.

To avoid possible bias introduced by AI models, we conduct a manual analysis of the content by classifying the reports in three categories (Undetermined, Benign and Malicious) and provide insights into the type of malicious content found by leveraging a fraud taxonomy. By using the insights gained from this in-depth case study, we identify both unique issues as possibilities this dataset has to offer.

First, we manually labeled all websites with one of the three categories based on visual features and the website URL. We marked a website as malicious when it hosted fraudulent content aiming to deceive users into providing the attacker personal information or monetary gain. Benign, when a website showed no malicious intent and undetermined when there was not enough information to correctly classify. Table I shows the distribution in the dataset across the three categories.

Secondly, we investigated to what extent the locality of the leveraged anti-phishing campaign influences the dataset followed by an analysis of the lifetime of the websites.

A. Undetermined content

Although we performed a thorough manual inspection of all dataset entries and asked for second opinions when unsure, 19.0% could not be sorted in a category due to a lack of information. Screenshots taken of these websites were either completely blank, leading to a captcha challenge or did not reveal any information about its content. Although the majority of the sites returns a successful HTTP status code (see Table II), 11.5% has a non-successful status code. 1,600 sites showed a CAPTCHA challenge our crawler was unable to solve, while 1,175 displayed a blank page, showing no text.

TABLE II
STATUS CODES OF WEBSITES WHO LACKED INFORMATION TO BE CORRECTLY CLASSIFIED.

Status Code	Count	Status Code	Count
200	3478	502	8
404	252	521	3
403	140	405	3
400	24	302	1
500	12	429	1
401	9	204	1

Although part of the sites are likely to be sites that were already taken down, the high amount of CAPTCHA challenges suggest that a big portion is also employing various cloaking techniques. We acknowledge that our crawler is not perfect in bypassing cloaking techniques, yet it raises a question on how existing, curated, datasets handle such websites as it is possible even more advanced crawlers are subject to cloaking. Removing cloaked websites from a curated dataset, since they can't be accurately classified as malicious, can lead to a dataset biased towards less complex malicious campaigns that don't use cloaking.

B. Benign content

We consider benign in the broad sense, encompassing both links leading to known legitimate sites (established companies) and any website with no fraudulent or misleading intent. They mostly stem from newsletters or marketing campaigns such as promotions on certain products or coupons. Although they don't have a malicious intent, those newsletters can be annoying and arrive unexpectedly in one's mailbox as users frequently forget they signed up [22], [25]. The large volume of benign reports (56.1%) clearly shows users are receiving more emails than they deem necessary and wanted. We suspect the cause for this is two-fold.

On the one hand, users could be unable to determine the difference between benign and malicious. The dataset contains legitimate authentication portals to often mimicked services such as banking and governmental applications which hints towards the inability of the reporter to accurately determine the benignness of a site. An example of this is shown in Figure 2 where on the left the benign log-in portal is shown, whereas the right is a phishing campaign. Both screenshots stem from our dataset.

On the other hand, we see users forwarding all mails they don't want rather than only the suspicious ones. This is notable

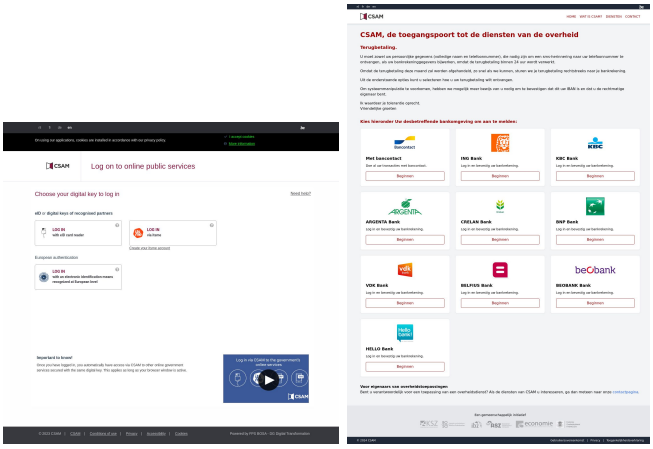

Fig. 2. Screenshots of reported sites showing a benign respectively a malicious log-in portal.

as we see lots of newsletters from benign companies recurring advertising products. It is possible these emails were received without the user ever subscribing to them, yet we have no way of confirming this.

C. Malicious content

In order to map what kinds of malicious activity is found in our dataset, we leverage the taxonomy of fraud by Beals et al. [12], which uses a systematic numbering to sort fraudulent activities into categories. Appendix A provides an overview of the used categories and their description together with two extra types we added to the taxonomy related to crypto fraud and a visualization. This section highlights some key findings.

1) Adult content: A large part of the dataset contained adult content such as nudity or pornographic content. To limit our exposure to this content, we used an open-source vision transformer which was able to detect 4,118 screenshots containing inappropriate content [20]. During the manual analysis we further identified 120 sites with adult content incorrectly classified by the model.

A closer look into these entries reveal they mostly contain fake dating websites where users can buy coins to arrange a date. Other types of sites advertise to sell erotic content in exchange for payment often using AI-generated or highly edited images. Due to ethical concerns we did not further investigate this part of the dataset and did not include it in the taxonomy, yet we believe most sites will not deliver their promised content and play with sexual desire to elicit payment or financial information, and we therefore mark them as malicious.

2) Fraudulent e-commerce websites: We classified a website as being a fraudulent e-commerce website (FCW) or fake webshop, if they advertised services or goods with no intent of delivering them. To distinguish FCW from benign shops, we leveraged features such as social media links, discounts and domain names [14], [46], [65]. Out of 289 FCW reported, 60 are hosted on `github.com`, likely due to its ease of freely deploying basic websites. Cybercriminals

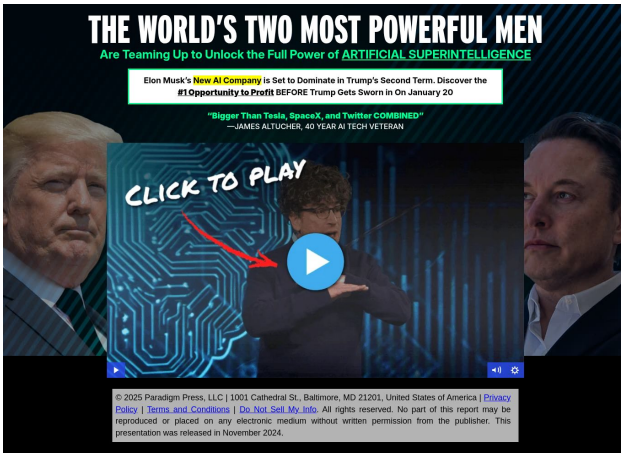

Fig. 3. Example of a website hosting fraudulent investment advice taking advantage of current events.

also leverage existing E-commerce platforms such as Shopify (23.5%) and WooCommerce (22.8%) which are also the most used platforms for benign shops [52].

3) Investment advice: 292 of reported sites are marked as Consumer Investment Fraud which entails forged information about investment opportunities leading to high financial returns. In that category, we mostly see websites advertising stocks one should invest in to make profit, often accompanied by urgent and flashing text. Malicious actors can use these techniques to artificially raise the price of shares they also own to later sell it as part of a pump-and-dump scheme. Additionally, we see recent events being used to influence victims, such as the inauguration of the U.S. President which happened 4 days after this data was collected or the hype of AI stocks (Figure 3).

D. Regional aspect

Although the campaign is rolled out on a national level, Table III shows that only 8.8% of reported URLs have the national TLD (`be`) and the majority uses the `.com` TLD. Additionally, we see that ccTLDs from neighboring countries (`nl`, `fr`, `de`, `lu`) make up 7% of the dataset.

More interesting are the results filtered on only the malicious websites, where we see similar trends with `.com` being the most popular. The second and third most popular domain in our malicious dataset are `.shop` and `.io`. All malicious `.shop` domains are used to host FCW, likely due to the cheap registration fee of this TLD and the semantic link between the TLD and the hosted content [63].

These trends are also present when looking at the primary languages used in the webpage detected via LibreTranslate [44]. The majority is using English whilst 28.9% are shown in a local language (Dutch, French or German) as noted in Table IV.

E. Lifetime

Previous research has shown that phishing sites have a limited lifetime, often only a couple of days before being taken down [51]. We checked the status of the reported websites 5 months after initial reporting and found that 52.3% of malicious website were still online. Table V shows per malicious category, how many websites were reachable and how many sites were operational, meaning they still contained their original content.

TABLE III
DISTRIBUTION OF THE TLDS USED BY REPORTED WEBSITES.

	All	Malicious
com	15,921 (61.8%)	557 (58.4%)
shop	123 (0.5%)	107 (11.2%)
io	105 (0.4%)	62 (6.5%)
be	2,277 (8.8%)	11 (1.2%)
net	1,441 (5.6%)	4 (0.4%)
de	662 (2.6%)	2 (0.2%)
ru	575 (2.2%)	1 (0.1%)
nl	501 (1.9%)	9 (0.9%)
fr	372 (1.4%)	1 (0.1%)
lu	277 (1.1%)	20 (2.1%)
other	3,511 (13.6%)	180 (18.9%)

TABLE IV
USED LANGUAGES OF THE REPORTED WEBSITES IN THE DATASET.

	All	Malicious
English	13,874 (67.7%)	752 (78.8%)
Dutch	4,175 (20.4%)	101 (10.6%)
French	1,425 (6.9%)	27 (2.8%)
German	325 (1.6%)	22 (2.3%)
Other	684 (3.3%)	52 (5.5%)

TABLE V
AMOUNT OF MALICIOUS WEBSITES STILL REACHABLE AND OPERATIONAL AFTER 5 MONTHS PER CATEGORY.

Category	Reachable	Operational
Equity Investment Fraud	171 (68.4%)	171 (68.4%)
Cryptocurrency Fraud	17 (77.3%)	17 (77.3%)
Rare Object Scam	3 (15.0%)	2 (10.0%)
Worthless Products	10 (41.6%)	10 (41.6%)
Paid never received (FCW)	127 (43.9%)	105 (36.3%)
Travel Booking Scam	2 (100.0%)	1 (50.0%)
Fake Credit Lines and Loans	7 (100.0%)	7 (71.4%)
Fortune Telling Fraud	8 (100.0%)	8 (100.0%)
Phishing Websites	44 (23.7%)	3 (1.6%)
Free Product	9 (90.0%)	9 (90.0%)
Cash Prize	72 (97.3%)	11 (14.9%)
Free Crypto	4 (14.3%)	3 (10.7%)
Loan Debt Scam	3 (13.0%)	3 (13.0%)
Other Bogus Charity	11 (100.0%)	11 (100.0%)

We see that 23.7% of phishing websites are still reachable, yet closer inspection reveals that their sites have been seized by their hosting provider or law enforcement showing a phishing warning. Only 3 sites are still online and functional.

Other types of malicious websites are largely online. Notably fraud related to investing and the stock market remain operational.

Half of the FCW, rather, have been taken down protecting end-users from its harm. Anti-phishing solutions and protections against FCW are well studied and thus those types of malicious activities can be taken down rather quickly. In practice, however, the situation is not perfect as 36.3% of fake

webshops are still operational. We see that research into other types of online fraud is more limited, giving more leeway to attackers exploiting these types of campaigns and likely granting those campaigns a longer uptime [14], [24].

V. Topic Modeling

While manual inspection should deliver results as precise as possible, it is not scalable. Therefore, we propose topic modeling as a helpful tool for cybersecurity analysts. Topic modeling is an unsupervised ML technique from the domain of Natural Language Processing (NLP) used for automatic grouping of documents, based on their textual content, in topics identified by keywords. These keywords enable analysts to identify topics of interests and recurring themes. This technique is widely applied in similar scenarios both in the context of phishing [29] and for other text-based tasks [28], [30].

In this section we discuss how topic modeling can be leveraged to classify what sectors (e.g. Banking, Social media) the suspicious reports are coming from. The insights gained from clustering reports together in topics can be used by analysts in these sectors to determine what malicious activity is happening and what benign activities users misinterpret as malicious. Such a dual perspective can support both threat intelligence and usability improvements for benign services. In contrast to static categories, topic modeling allows for easy retraining to capture the latest trends.

To show the possibilities of this method and check the findings from the manual analysis, we use all reported websites with the `eu` TLD reported between May and September 2025. Of these reports, we use the first month as training data to establish the initial topics. By measuring the decay of topics over time, we gain insights in the evolution of the online fraud landscape in Belgium. Finally, we take a larger subset of our dataset (reports with the national TLD `be`) to confirm the transferability of the initial topics.

A. Methodology

As we are interested in the content of the reported websites, we apply the topic modeling to HTML content. Our preprocessing consists of three steps: 1) Language filtering 2) Removal of boilerplate text 3) Extraction of the text.

First, since classical ML models that work with texts are sensitive to noise, it is necessary to select the languages our model should support, and strive to filter out tokens that could hamper the performance of the model. In this study we keep the pages in the most prevalent languages: English, Dutch, French and German.

Second, we apply a popular filter, used by ad blockers [2], to remove cookie banners from the HTML as they are not representative of the content and will likely skew the topic modeling results.

Finally, each of these pages is stripped of the HTML structure and only the raw textual data is extracted and stored as documents for the algorithm to use. The documents are also filtered with sets of stop-words, i.e. high-frequency terms which bring little semantic value (a/an, then, the ...) based on an open-source list [3].

After preprocessing, we use BERTopic [28] with the default model (`all-MiniLM-L6-v2`) for embeddings and HDBSCAN with a minimal cluster size of 25 documents. We limit our initial experiment to reports of suspicious websites hosted on the `.eu` domain name since using all reports (around $40,000$ daily) is technically unfeasible. The used dataset, therefore, consists of reports made from May 2025 until September 2025, totaling $44,601$ documents. By using the reports from May 2025 to identify the set of topics and training our model, we study how these evolve over time. Later, using the same timeframe, we sort reports with the national TLD ($177,629$ documents) into the same topics to check the transferability of our model.

1) Topic expansion: Normally, BERTopic assigns a single topic to each document, yet, in the context of suspicious reports, a document could fit in multiple topics. For example, a website of a benign newsletter from a bank can be classified in topics containing other newsletters but also in topics related to the financial world. This way, analysts investigating what newsletters are perceived as malicious by users, as well as analysts investigating what financial institutions are being reported can find the relevant documents.

To achieve this, we leverage the confidence scores assigned by the model to each document. For every topic, BERTopic calculates the probability a document belongs to said topic where the total sum is 1, called the confidence score. For each document, we sort it into said topic if the confidence score is above 5%.

B. Initial topics

To gather the initial set of topics, we train the model on all reports made in May 2025 ending with the `eu` TLD. This leads to 69 topics wherein 82.9% of reports can be sorted, other reports are classified as outliers for not fitting into the found topics. The full list of topics can be found in the appendix (Table VI).

The found topics can be linked back to the findings of the manual analysis. For example topic 19 (`apple`, `ipad`, `shop`) clearly is related to Apple products and webshops (where we make no classification on the benignness), where topic 42 (`sex`, `women`, `age`) and 61 (`18 ans`, `gratuit confirme`, `granny`) shows dating scams or related NSFW content. It also reveals that, although our best effort of filtering these out, some topics are represented by cookie banners (e.g. Topic 17 and 56). Manual inspection reveals they all contain cookie banners of the same company, in this case Meta and LinkedIn, as they have custom cookie banners not found on other websites.

Topic 45 (`javascript`, `supported`, `browser`) shows that, although our best efforts, some websites still recognize

[2]https://secure.fanboy.co.nz/fanboy-cookiemonster.txt

[3]https://github.com/stopwords-iso/stopwords-iso

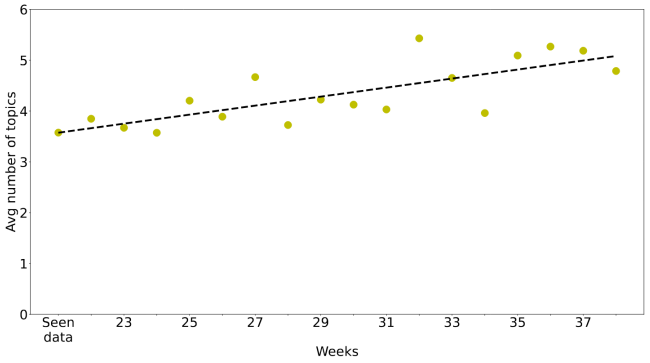

Fig. 4. Over time, the average number of topics a document is classified in, increases showing the model becomes less accurate over time.

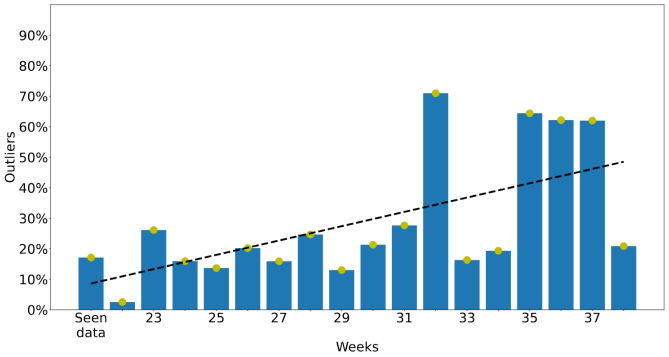

Fig. 5. In the dataset of `eu` reports, the amount of outliers, documents not sorted into a single topic, increases over time. This evolution can be attributed to a single widespread campaign appearing around week 32.

our crawler as an automated tool by clustering pages together that show a warning related to the lack of JavaScript execution.

The manual analysis found a large amount of newsletters, users reported as suspicious. This is further confirmed by topics 1, 21 and 44 whose representations all mention `unsubscribe` and `data`, indicating these as websites to unsubscribe the user from mailing lists.

C. Evolution over time

As time evolves, we expect the content of reported messages to change and thus the confidence of which documents are assigned to a topic will decrease. This can be seen in Figure 4 where the average number of topics assigned to a document increases, shown using linear regression as a dotted line. Starting from the seen data the model was trained on, for every week following, numbered according to the calendar year, the average number of topics is plotted. Since the number of topics increases, it depicts pages become less and less fit for a single topic which shows the dataset becomes more diverse from the original topics over time.

Secondly, Figure 5 shows how many reports are being classified as outliers, i.e. not belonging to a topic. Similar to the average topics per report, the linear regression model shows an increase in outliers, however this is largely due to 4 weeks showing a large percentage of outliers.

Week 32, 35, 36 and 37 stand out with around 70% of reports not belonging to a topic. A manual look inside this data shows it largely consists of a campaign advertising cheap electricity providers, not previously seen in our training data. Respectively 67%, 58%, 48% and 47% of reports made in those weeks, belong to that campaign. In general, this shows that a large spike in the outliers is likely to indicate a new, previously unclustered, campaign. For the timeframe used in this analysis, we see that the initial topics remain a good fit, but the evolution shows that this will diminish over time and that after a certain period of time, retraining the model is likely needed to avoid topic drift.

D. Transferability

By applying the identified topics from our model to a broader dataset, we check how transferable the model is. We

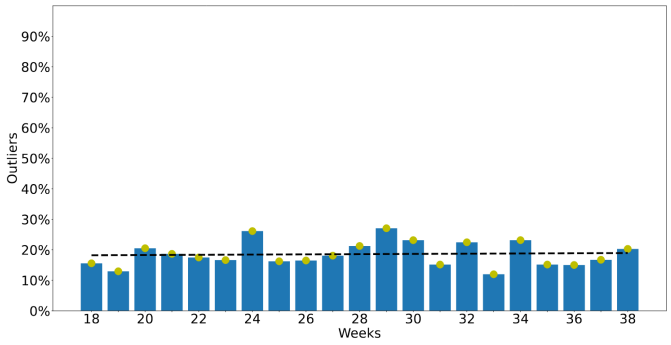

Fig. 6. The amount of outliers for reports with the national TLD shows no evolution over time.

cluster all reports hosted on the national TLD (`be`), a dataset of 177,629 documents, in their topics and check, in general, how well the reports fit in the dataset.

1) Outliers: In contrast to the dataset of `eu` reports, we see that reports with the national TLD continue to fit into the initial topics, even after 20 weeks. Although the reports in this larger dataset were not used to train the model upon, we see a relatively similar percentage of outliers when compared to the `.eu` dataset. The most notable difference is the absence of the large spikes, related to the electricity provider campaign, in the later weeks. This shows topic modeling is able to capture a broad sense of campaigns in its topics that can be reused over time and in similar datasets.

VI. Issues and Opportunities

Collecting reports from users in order to study the malicious activity on the web has its advantages, but does not come without challenges. It opens up valuable insights into the trust users place in the mailing ecosystem by knowing what they find suspicious. Using the insights collected in this paper, we identify key difficulties while handling a dataset of this nature but also show the opportunities this can offer for both industry and the research community.

A. Adult Content

One of the major categories of websites reported on a daily basis is dating sites or other sites specifically aimed at adults due to their pornographic nature. That nature makes it hard to investigate them without running into ethical issues. Therefore, we propose to filter those entries out of the dataset by using ML models to limit the exposure of the analyst to this type of content. Unfortunately, the non-perfect accuracy of such models can still lead to some exposure which will be unavoidable when working with user-reported websites.

Additionally, removing such entries from a dataset loses valuable insights as they are likely malicious. Users who report scams based on adult content, do so since they find these messages suspicious or at least unwanted.

The anti-scam ecosystem should be hardened for these types of threats by conducting more research into adult content scams for which a user-reported dataset can be of help.

B. Benign filtering

By manually analyzing the dataset, we found numerous benign sites as opposed to a limited set of malicious websites. Existing user-reported datasets, such as PhishTank, rely on volunteers verifying reported websites which assures accurate datasets, but is not scalable. Numerous AI models exist claiming various accuracies and properties, yet they will fail to capture newer trends which are likely to occur in a constantly updated dataset or malicious sites targeted towards local companies [43].

One possible solution therefore, is to rely on whitelists to identify the gross of the benign websites. Researchers have used popularity lists for this purpose, although they can include malicious or unregistered domains [11], [41], [55]. Bayer et al. proposed a methodology for constructing a whitelist starting from a set of reference domains which we partially implemented to check the feasibility of this technique for this type of dataset [11].

Starting from a set of domains, we leveraged the Public Suffix List (PSL) to generate a list of possible whitelisted domains [48]. The reference domains were derived from Wikidata [64] and governmental data. We queried Wikidata for companies and businesses located in the studied European country or its neighboring countries giving us $1,962$ domains. Secondly, we looked up domain names from companies registered in a governmental database giving us additional $42,740$ domains. From there, we combined these domains with the public suffix list and kept all domains that shared the same `NS` records as the original domain it derived from. This gave a final whitelist of $47,409$ entries.

Employing this whitelist on our dataset only filters out 7.23% of websites. The low effectiveness is due to the variety of benign campaigns reported by users. They make use of temporary domains specifically created for marketing purposes hosted by other services redirecting to the effective brand (e.g. `lnxbn1b.r.eu-west-1.awstrack.me` → `netflix.com`). The mailing domain is likely not whitelisted due to its uniqueness whilst the final domain is not a reliable source to be checked against a whitelist. Most cloaking mechanisms redirect bots to a benign website which would lead to incorrectly flagging cloaked websites as benign. Additionally, it takes resources to fetch the final domain as a crawler needs to be instructed, whilst whitelisting would be most effective if this leads to a reduced crawlable dataset.

Whitelisting on regional domains can thus be an aid in applying a first filter, though, due to the nature of the dataset it does not prove to be a sufficient solution. More elaborate whitelists or other techniques will be required to filter out the gross of benign websites.

C. Privacy

Users who report suspicious websites to an organization or government, can do so in order to help harden the ecosystem and prevent harm to other victims. Unfortunately, we see that this also exposes these users to privacy risks. Naturally, they expose their email address and all metadata incorporated in that email which can be geo-specific, such as time zones. An initial step to protect the privacy of user reports, is to only keep the links incorporated in the reported emails. This effectively protects the user from unwanted parties accessing their email, but also throws away the context of malicious emails. A second step, also employed in this research, is to replace email addresses still found in the reported links by a dummy address using regular expressions.

Even when employing the steps above, a lot of PII is still exposed to the end-user of the dataset. Most marketing campaigns are specifically tailored towards the client, meaning they include mail and possibly name in the final website. We see that the URL often contains identifiers that are used by a site to personalize the content [68]. Removing these identifiers from the URL, however, can render an error page.

Additionally, we also find some peculiar reports especially prone to privacy risks, for example invoices. Those invoices, originating from benign domains, contain personal and legal information about the payer, including address and national identity number as shown in Figure 7.

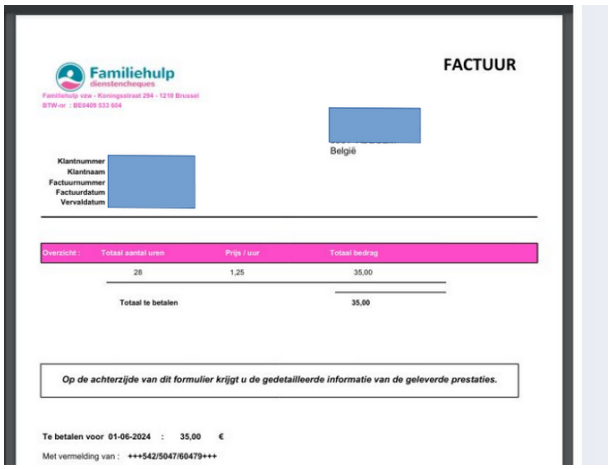

Fig. 7. Screenshots of a reported site showing an invoice that includes physical address and identity numbers.

Preventing PII from leaking in the dataset is not straightforward. Although it's easy to filter out emails from the URL, most links contain unique identifiers from which it is impossible to determine what information is stored. When leaving out the identifier to prevent access to personal information, it either gives an error page or redirects to a completely different website. This shows that only a preprocessing step does not suffice to filter out all PII. Therefore, we propose to also employ a post-processing module that can detect and anonymize PII in HTML pages and screenshots.

D. Classification tools

By leveraging topic modeling as an additional tool, analysts can gain insights into what campaigns are happening aside from what reports are benign or malicious. Our analysis showed that the relevance of topics remains stable over time and the topics are transferable to a different dataset. We do, however, see some new campaigns (e.g. the electricity provider campaign) pop up, showing that, over time, it can be helpful to recalibrate the model to capture the newest trends.

Aside from cybersecurity analysts, tasked with studying malicious campaigns, such a classification tool can also be used by e.g. marketing specialists who see their benign campaigns reported. Filtering on topics relevant to their company, they could check if their campaigns are found suspicious by users and identify what changes they can make.

VII. Limitations

The results discussed in this paper are limited by certain aspects. Due to privacy concerns, there is no metadata about emails or the user who made the report present in the dataset. Therefore, we have no data on how many unique users made daily reports, possibly introducing a bias towards a handful of very active reporters.

To get a first grasp on the dataset, we limited ourselves to reports made in one day which made us able to investigate each data entry manually and avoid bias from AI models. Although, the data was sorted manually, it is possible human mistakes happened leading to a small number of misclassified sites. Additionally, we only looked at screenshots of the website's landing page, whereas it could have improved accuracy if user interaction was simulated [61].

Due to technical limitations in the crawler, it is possible that malicious webpages were successful in cloaking themselves as benign webpages. Although we classified webpages with no visual content as possible cloaking, some webpages redirected to known benign sites such as `google.com` which makes it hard to accurately classify them as either benign or malicious [72]. This highlights the issue of cloaking in studying crime on the web and the need for more advanced crawlers specialized in bypassing cloaking techniques.

In our topic modeling analysis, certain parameters (such as the minimal cluster size) were decided and fixed to, arbitrary, values. These values can be easily adapted for specific use cases and datasets, yet further work is needed to establish ideal fine-tuning methods.

VIII. Future Work

We thoroughly described and studied the dataset showing its unique composition. After filtering out unwanted entries such as benign pages or adult content websites, this dataset can provide an insight into regional fraudulent activity taking place on the web. We showed the dataset is highly regionalized, meaning it can provide long term insights into cybercrime for local companies not offered in conventional, more international, datasets. This locality also opens up possibilities for more specialized AI models focused on identifying phishing using brand detection as those lack knowledge of regional context [43].

Thanks to user reports, this type of dataset provides a direct insight into the thought process of internet users otherwise missing from curated datasets. Future work could leverage this to perform a user study on why certain websites are reported gaining psychological insights.

We found a majority of benign newsletter and advertisement campaigns meaning something in the mail or website triggered users to report it as suspicious. This can be of value for marketing departments who wish to reach as many potential clients as possible. By studying what type of emails are reported as suspicious, they can adapt their communication strategies.

A large chunk of the reports contained adult content and fake dating sites. By employing these reports, advancements in mitigating and detecting NSFW scams can be made.

IX. Related work

Our work focuses on applying topic modeling to an unfiltered dataset of suspicious websites as identified by the day-to-day internet user to measure changes in malicious campaigns. This section discusses previous work to highlight its similarities and differences with our work.

A. Regional cybercrime

Previous research has identified various methods to identify online cybercrime. Early research was mostly focussed on spam [5], [6], [9], [23], [58] where the focus later shifted more towards phishing by analyzing their properties [42], [50], [51] that can be used to construct various detection models [3], [24], [43], [56], [70]. Other approaches tried to construct whitelists instead of blocklists in order to filter in datasets [1], [11], [17].

As seen in our dataset, fraudulent websites are not limited to phishing. Other popular examples include gambling scams [34], which gives users free spins or money in an (online) casino, crypto scams [49], where users are tricked into giving their payment details in exchange for (worthless) crypto coins, and dating scams [60], [69] where attackers incite romantic feelings ultimately leading to victims sending over money.

One notable recent evolution is the study of e-Commerce scams where malicious actors mimic web shops. Kotzias et al. tracked real customers showing the impact of online e-Commerce scams leveraging those insights to create a detection model [40]. In addition, Bitaab et al. also proposed a novel

detection framework for detecting fraudulent e-Commerce Websites [14]. Similar to our work, they collected reports from users and uncovered the low accuracy of GSB in this type of scam.

However, since the web exhibits strong ties between regions [31], cybercrime should be studied on a regional level for optimal mitigation. Houtti et al. conducted a user survey in 12 countries to map the type of scams experienced and the collective actions taken [35]. They found that Belgium was one of the countries were people reported scams the most, proving the relevance and comprehensiveness of our studied dataset. Complementary, Moura et al. studied phishing campaigns in three European countries, including Belgium, finding the most targeted companies and listing advice for blocking malicious campaigns at the DNS level which was later implemented by Daniels et al. [21], [47]. Both Park et al. and Bijmans et al. focused on phishing in their respective countries where they also took note of the low effectiveness of blacklists [13], [53]. Böttger et al. investigated the regional aspect further, by comparing filter lists for ad blockers on a country-level showing their usefulness to handle local languages [19].

B. Topic modeling

Topic modeling has proven to be a useful tool to cluster similar entries in a dataset. In electronic fraud, this technique has been mostly used to identify spam emails [4], [26], [32], [67]. More similar to our work, Jáñez-Martino et al [36] and Wang et al. [66] leveraged topic modeling to find clusters of similar spam campaigns, yet focused on email content where the contribution of this work lies in applying it to website content.

X. Conclusion

In this paper we explored the properties of a dataset generated directly from user reports, consisting of what they perceive as suspicious. By asking citizens of a European country, Belgium, to report suspicious emails and links they encounter, this type of dataset encompasses a regional view on cybercrime. Since the dataset is unfiltered, contrary to other popular available datasets, it can be used to gain psychological insights and conduct user studies.

We incepted an automated data collection pipeline to timely harvest the suspicious websites, as reported by the end-users, and discussed various challenges in setting up such a pipeline.

By performing an initial manual analysis, later complemented by topic modeling, we identified and quantified the different types of reported websites. Less than half of the reported links can be attributed to malicious websites, such as investment or crypto scams, bogus webshops, phishing pages or (fake) dating websites. Surprisingly, the later one (broadly categorized as adult content) is the biggest malicious category.

We show how topic modeling can help analysts to gain insights into malicious datasets by clustering reports in categories they are interested in together allowing for easy measurement how topics evolve over time.

XI. Ethical considerations

By only collecting the links which were stripped of any potential email addresses we tried to limit our exposure to PII. Since we found that some PII was still present despite our best efforts, all data was stored in local environments and will not be made public.

To limit network traffic, every reported site was only crawled once adhering to the Menlo report [10].

By using a vision transformer for adult content classification, we tried to limit our exposure to this type of content.

XII. LLM usage considerations

We used Generative AI (Claude Sonnet 4[4] and Claude Sonnet 4.5[5]) to partially generate code used for the visualizations of our data. All outputs were manually verified for correctness.

Acknowledgment

This research is partially funded by the Internal Funds KU Leuven, and by the Cybersecurity Research Program Flanders. We'd like to thank the Centre for Cybersecurity Belgium for their support.

References

[1] Y. Abyaa, O. Hureau, A. Duda, and M. Korczyński, "LogoTrust: Leveraging BIMI to Build a Validated Dataset of Brands, Domain Names, and Logos," in *2025 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, Jun. 2025, pp. 132–137, iSSN: 2768-0657. [Online]. Available: https://ieeexplore.ieee.org/document/11129546/

[2] B. Acharya and P. Vadrevu, "PhishPrint: Evading phishing detection crawlers by prior profiling," in *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Aug. 2021, pp. 3775–3792. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/acharya

[3] M. AlSabah, M. Nabeel, Y. Boshmaf, and E. Choo, "Content-Agnostic Detection of Phishing Domains using Certificate Transparency and Passive DNS," in *Proceedings of the 25th International Symposium on Research in Attacks, Intrusions and Defenses*, ser. RAID '22. New York, NY, USA: Association for Computing Machinery, Oct. 2022, pp. 446–459. [Online]. Available: https://doi.org/10.1145/3545948.3545958

[4] K. Althobaiti, M. K. Wolters, N. Alsufyani, and K. Vaniea, "Using Clustering Algorithms to Automatically Identify Phishing Campaigns," *IEEE Access*, vol. 11, pp. 96 502–96 513, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10235951/

[5] D. S. Anderson, C. Fleizach, S. Savage, and G. M. Voelker, "Spamscatter: Characterizing internet scam hosting infrastructure," in *16th USENIX Security Symposium (USENIX Security 07)*. Boston, MA: USENIX Association, Aug. 2007. [Online]. Available: https://www.usenix.org/conference/16th-usenix-security-symposium/spamscatter-characterizing-internet-scam-hosting

[6] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C. D. Spyropoulos, and P. Stamatopoulos, "Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach," vol. cs.CL/0009009, 2000. [Online]. Available: https://arxiv.org/abs/cs/0009009

[7] APWG, "The APWG eCrime Exchange (eCX)." [Online]. Available: https://s29837.pcdn.co/ecx/

[8] ——, "APWG Trends report Q4 2024," Tech. Rep., 2025. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q4_2024.pdf

[9] W. Awad, "Machine Learning Methods for Spam E-Mail Classification," *IJCSIT*, vol. 3, no. 1, pp. 173–184, Feb. 2011. [Online]. Available: http://www.airccse.org/journal/jcsit/0211ijcsit12.pdf

[4] https://www.anthropic.com/news/claude-4

[5] https://www.anthropic.com/news/claude-sonnet-4-5

[10] M. Bailey, D. Dittrich, E. Kenneally, and D. Maughan, "The Menlo Report," *IEEE Security And Privacy*, vol. 10, no. 2, pp. 71–75, 2012.

[11] J. Bayer, S. Maroofi, O. Hureau, A. Duda, and M. Korczynski, "Building a Resilient Domain Whitelist to Enhance Phishing Blacklist Accuracy," in *2023 APWG Symposium on Electronic Crime Research (eCrime)*, Nov. 2023, pp. 1–14, iSSN: 2159-1245. [Online]. Available: https://ieeexplore.ieee.org/document/10485549

[12] M. Beals, M. DeLiema, and M. Deevy, "Framework for a taxonomy of fraud," *Stanford Center on Longevity*, Jul. 2015.

[13] H. Bijmans, T. Booij, A. Schwedersky, A. Nedgabat, and R. van Wegberg, "Catching phishers by their bait: Investigating the dutch phishing landscape through phishing kit detection," in *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Aug. 2021, pp. 3757–3774. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/bijmans

[14] M. Bitaab, H. Cho, A. Oest, Z. Lyu, W. Wang, J. Abraham, R. Wang, T. Bao, Y. Shoshitaishvili, and A. Doupé, "Beyond Phish: Toward Detecting Fraudulent e-Commerce Websites at Scale," in *2023 IEEE Symposium on Security and Privacy (SP)*, May 2023, pp. 2566–2583, iSSN: 2375-1207. [Online]. Available: https://ieeexplore.ieee.org/document/10179461/

[15] M. Bitaab, H. Cho, A. Oest, P. Zhang, Z. Sun, R. Pourmohamad, D. Kim, T. Bao, R. Wang, Y. Shoshitaishvili, A. Doupé, and G.-J. Ahn, "Scam Pandemic: How Attackers Exploit Public Fear through Phishing," in *2020 APWG Symposium on Electronic Crime Research (eCrime)*, Nov. 2020, pp. 1–10, iSSN: 2159-1245. [Online]. Available: https://ieeexplore.ieee.org/document/9493260/

[16] X. Bouwman, V. L. Pochat, P. Foremski, T. V. Goethem, C. H. Ganan, G. C. M. Moura, S. Tajalizadehkhoob, W. Joosen, and M. van Eeten, "Helping hands: Measuring the impact of a large threat intelligence sharing community," in *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 1149–1165. [Online]. Available: https://www.usenix.org/conference/usenixsecurity22/presentation/bouwman

[17] R. Burton and L. Rocha, "Whitelists that Work: Creating Defensible Dynamic Whitelists with Statistical Learning," in *2019 APWG Symposium on Electronic Crime Research (eCrime)*. Pittsburgh, PA, USA: IEEE, Nov. 2019, pp. 1–10. [Online]. Available: https://ieeexplore.ieee.org/document/9037505/

[18] M. Bynens and P. Kvitek, "Chrome Headless mode," 2024. [Online]. Available: https://developer.chrome.com/docs/chromium/headless

[19] C. Böttger, N. Demir, J. Hörnemann, B. Acharya, N. Pohlmann, T. Holz, M. Grosse-Kampmann, and T. Urban, "Understanding Regional Filter Lists: Efficacy and Impact," in *Proceedings on Privacy Enhancing Technologies*, 2025. [Online]. Available: https://petsymposium.org/popets/2025/popets-2025-0063.php

[20] A. Codd, "vit-base-nsfw-detector," Jan. 2025. [Online]. Available: https://huggingface.co/AdamCodd/vit-base-nsfw-detector

[21] T. Daniels, M. Bosteels, P. Robberechts, and J. Davis, "RegCheck: A Real-Time Approach for Flagging Potentially Malicious Domain Name Registrations," *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, May 2025.

[22] J. Dev, E. Rader, and S. Patil, "Why Johnny Can't Unsubscribe: Barriers to Stopping Unwanted Email," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, Apr. 2020, pp. 1–12. [Online]. Available: https://dl.acm.org/doi/10.1145/3313831.3376165

[23] D. Fetterly, M. Manasse, and M. Najork, "Spam, damn spam, and statistics: using statistical analysis to locate spam web pages," in *Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004*. Paris France: ACM, Jun. 2004, pp. 1–6. [Online]. Available: https://dl.acm.org/doi/10.1145/1017074.1017077

[24] 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 *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. USENIX Association, Aug. 2021, pp. 339–358. [Online]. Available: https://www.usenix.org/conference/soups2021/presentation/franz

[25] I. Gamzu, L. Lewin-Eytan, and N. Silberstein, "Unsubscription: A Simple Way to Ease Overload in Email," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. Marina Del Rey CA USA: ACM, Feb. 2018, pp. 189–197. [Online]. Available: https://dl.acm.org/doi/10.1145/3159652.3159698

[26] T. Gokcimen and B. Das, "Topic Modelling Using BERTopic for Robust Spam Detection," in *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, Apr. 2024, pp. 1–5, iSSN: 2768-1831. [Online]. Available: https://ieeexplore.ieee.org/document/10527342/

[27] Google, "Google Safe Browsing," 2025. [Online]. Available: https://safebrowsing.google.com

[28] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 2022, arXiv:2203.05794 [cs]. [Online]. Available: http://arxiv.org/abs/2203.05794

[29] E. S. Gualberto, R. T. De Sousa, T. P. De B. Vieira, J. P. C. L. Da Costa, and C. G. Duque, "From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection," *IEEE Access*, vol. 8, pp. 76 368–76 385, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9075252

[30] R. Guan, H. Zhang, Y. Liang, F. Giunchiglia, L. Huang, and X. Feng, "Deep Feature-Based Text Clustering and its Explanation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3669–3680, Aug. 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9215004

[31] R. Habib, K. Ruth, G. Akiwate, and Z. Durumeric, "Formalizing dependence of web infrastructure," in *Proceedings of the ACM SIGCOMM 2025 Conference*, ser. SIGCOMM '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 1132–1153. [Online]. Available: https://doi.org/10.1145/3718958.3750492

[32] I. R. A. Hamid and J. H. Abawajy, "Profiling Phishing Email Based on Clustering Approach," in *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, Jul. 2013, pp. 628–635, iSSN: 2324-9013. [Online]. Available: https://ieeexplore.ieee.org/document/6680895/

[33] X. Han, N. Kheir, and D. Balzarotti, "PhishEye: Live Monitoring of Sandboxed Phishing Kits," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. Vienna Austria: ACM, Oct. 2016, pp. 1402–1413. [Online]. Available: https://dl.acm.org/doi/10.1145/2976749.2978330

[34] G. Hong, Z. Yang, S. Yang, X. Liaoy, X. Du, M. Yang, and H. Duan, "Analyzing Ground-Truth Data of Mobile Gambling Scams," in *2022 IEEE Symposium on Security and Privacy (SP)*, May 2022, pp. 2176–2193, iSSN: 2375-1207. [Online]. Available: https://ieeexplore.ieee.org/document/9833665/

[35] M. Houtti, A. Roy, V. N. R. Gangula, and A. Walker, "A Survey of Scam Exposure, Victimization, Types, Vectors, and Reporting in 12 Countries," *Journal of Online Trust and Safety*, vol. 2, no. 4, Sep. 2024, number: 4. [Online]. Available: https://tsjournal.org/index.php/jots/article/view/204

[36] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre, "Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach," *Applied Soft Computing*, vol. 139, p. 110226, May 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494623002442

[37] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage, "Spamalytics: an empirical analysis of spam marketing conversion," *Commun. ACM*, vol. 52, no. 9, pp. 99–107, Sep. 2009. [Online]. Available: https://dl.acm.org/doi/10.1145/1562164.1562190

[38] A. Kharraz, W. Robertson, and E. Kirda, "Surveylance: Automatically Detecting Online Survey Scams," in *2018 IEEE Symposium on Security and Privacy (SP)*. San Francisco, CA: IEEE, May 2018, pp. 70–86. [Online]. Available: https://ieeexplore.ieee.org/document/8418597/

[39] P. Kotzias, M. Pachilakis, J. A. Iuit, J. Caballero, I. Sanchez-Rola, and L. Bilge, "Ctrl+Alt+Deceive: Quantifying User Exposure to Online Scams," in *Proceedings 2025 Network and Distributed System Security Symposium*. San Diego, CA, USA: Internet Society, 2025. [Online]. Available: https://www.ndss-symposium.org/wp-content/uploads/2025-1816-paper.pdf

[40] P. Kotzias, K. Roundy, M. Pachilakis, I. Sanchez-Rola, and L. Bilge, "Scamdog Millionaire: Detecting E-commerce Scams in the Wild," in *Annual Computer Security Applications Conference*. Austin TX USA: ACM, Dec. 2023, pp. 29–43. [Online]. Available: https://dl.acm.org/doi/10.1145/3627106.3627184

[41] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczynski, and W. Joosen, "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation," in *Proceedings 2019 Network and Distributed System Security Symposium*. San Diego, CA: Internet Society, 2019.

[42] K. Lee, K. Lim, H. Kim, Y. Kwon, and D. Kim, "7 Days Later: Analyzing Phishing-Site Lifespan After Detected," in *Proceedings of the ACM on Web Conference 2025*. Sydney NSW Australia: ACM, Apr. 2025, pp. 945–956. [Online]. Available: https://dl.acm.org/doi/10.1145/3696410.3714678

[43] Y. Li, C. Huang, S. Deng, M. L. Lock, T. Cao, N. Oo, H. W. Lim, and B. Hooi, "KnowPhish: Large language models meet multimodal knowledge graphs for enhancing Reference-Based phishing detection," in *33rd USENIX Security Symposium (USENIX Security 24)*. Philadelphia, PA: USENIX Association, Aug. 2024, pp. 793–810. [Online]. Available: https://www.usenix.org/conference/usenixsecurity24/presentation/li-yuexin

[44] LibreTranslate, "LibreTranslate - Free and Open Source Machine Translation API," 2025. [Online]. Available: https://libretranslate.com

[45] T. Moore and R. Clayton, "Evaluating the Wisdom of Crowds in Assessing Phishing Websites," in *Financial Cryptography and Data Security*, G. Tsudik, Ed. Berlin, Heidelberg: Springer, 2008, pp. 16–30.

[46] W. Mostard, B. Zijlema, and M. Wiering, "Combining Visual and Contextual Information for Fraudulent Online Store Classification," in *IEEE/WIC/ACM International Conference on Web Intelligence*. Thessaloniki Greece: ACM, Oct. 2019, pp. 84–90. [Online]. Available: https://dl.acm.org/doi/10.1145/3350546.3352504

[47] G. C. M. Moura, T. Daniels, M. Bosteels, S. Castro, M. Müller, T. Wabeke, T. Van Den Hout, M. Korczyński, and G. Smaragdakis, "Characterizing and Mitigating Phishing Attacks at ccTLD Scale," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. Salt Lake City UT USA: ACM, Dec. 2024, pp. 2147–2161. [Online]. Available: https://dl.acm.org/doi/10.1145/3658644.3690192

[48] Mozilla, "Public Suffix List," 2025. [Online]. Available: https://publicsuffix.org/

[49] M. Muzammil, A. Pitumpe, X. Li, A. Rahmati, and N. Nikiforakis, "The Poorest Man in Babylon: A Longitudinal Study of Cryptocurrency Investment Scams," in *Proceedings of the ACM on Web Conference 2025*. Sydney NSW Australia: ACM, Apr. 2025, pp. 1034–1045. [Online]. Available: https://dl.acm.org/doi/10.1145/3696410.3714588

[50] A. Oest, Y. Safei, A. Doupe, G.-J. Ahn, B. Wardman, and G. Warner, "Inside a phisher's mind: Understanding the anti-phishing ecosystem through phishing kit analysis," in *2018 APWG Symposium on Electronic Crime Research (eCrime)*. San Diego, CA: IEEE, May 2018, pp. 1–12. [Online]. Available: https://ieeexplore.ieee.org/document/8376206/

[51] A. Oest, P. Zhang, B. Wardman, E. Nunes, J. Burgis, A. Zand, K. Thomas, A. Doupé, and G.-J. Ahn, "Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale," in *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, Aug. 2020, pp. 361–377. [Online]. Available: https://www.usenix.org/conference/usenixsecurity20/presentation/oest-sunrise

[52] J. Pagel, N. Demir, and N. Kaleev, "Ecommerce," in *The 2024 Web Almanac*. HTTP Archive, 2024, section: 13. [Online]. Available: https://almanac.httparchive.org/en/2024/ecommerce

[53] H. Park, K. Lim, D. Kim, D. Yu, and H. Koo, "Demystifying the Regional Phishing Landscape in South Korea," *IEEE Access*, vol. 11, pp. 130 131–130 143, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10320327/

[54] PhishTank, "PhishTank," 2025. [Online]. Available: https://phishtank.org/

[55] V. L. Pochat, T. V. Goethem, and W. Joosen, "Evaluating the long-term effects of parameters on the characteristics of the tranco top sites ranking," in *12th USENIX Workshop on Cyber Security Experimentation and Test (CSET 19)*. Santa Clara, CA: USENIX Association, Aug. 2019. [Online]. Available: https://www.usenix.org/conference/cset19/presentation/lepochat

[56] S. S. Roy and S. Nilizadeh, "PhishLang: A Real-Time, Fully Client-Side Phishing Detection Framework Using MobileBERT," Apr. 2025, arXiv:2408.05667 [cs]. [Online]. Available: http://arxiv.org/abs/2408.05667

[57] Safeonweb, "44% van de Belgen stuurt phishingberichten door naar verdacht@safeonweb.be," Jan. 2025. [Online]. Available: https://safeonweb.be/nl/actueel/44-van-de-belgen-stuurt-phishingberichten-door-naar-verdachtsafeonwebbe

[58] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail," in *Learning for Text Categorization: Papers from the 1998 workshop*, vol. 62. Madison, Wisconsin, 1998, pp. 98–105.

[59] C. Simoiu, A. Zand, K. Thomas, and E. Bursztein, "Who is targeted by email-based phishing and malware? Measuring factors that differentiate risk," in *Proceedings of the ACM Internet Measurement Conference*, ser. IMC '20. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 567–576. [Online]. Available: https://dl.acm.org/doi/10.1145/3419394.3423617

[60] G. Suarez-Tangil, M. Edwards, C. Peersman, G. Stringhini, A. Rashid, and M. Whitty, "Automatically Dismantling Online Dating Fraud," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1128–1137, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8768406/

[61] K. Subramani, W. Melicher, O. Starov, P. Vadrevu, and R. Perdisci, "PhishInPatterns: measuring elicited user interactions at scale on phishing websites," in *Proceedings of the 22nd ACM Internet Measurement Conference*. Nice France: ACM, Oct. 2022, pp. 589–604. [Online]. Available: https://dl.acm.org/doi/10.1145/3517745.3561467

[62] X. Teoh, Y. Lin, R. Liu, Z. Huang, and J. S. Dong, "PhishDecloaker: Detecting CAPTCHA-cloaked phishing websites via hybrid vision-based interactive models," in *33rd USENIX Security Symposium (USENIX Security 24)*. Philadelphia, PA: USENIX Association, Aug. 2024, pp. 505–522. [Online]. Available: https://www.usenix.org/conference/usenixsecurity24/presentation/teoh

[63] TLD-LIST, ".shop Domain Information," 2025. [Online]. Available: https://tld-list.com/tld/shop

[64] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, Sep. 2014. [Online]. Available: https://dl.acm.org/doi/10.1145/2629489

[65] J. Wadleigh, J. Drew, and T. Moore, "The E-Commerce Market for "Lemons": Identification and Analysis of Websites Selling Counterfeit Goods," in *Proceedings of the 24th International Conference on World Wide Web*. Florence Italy: International World Wide Web Conferences Steering Committee, May 2015, pp. 1188–1197. [Online]. Available: https://dl.acm.org/doi/10.1145/2736277.2741658

[66] D. Wang, D. Irani, and C. Pu, "A Study on Evolution of Email Spam Over Fifteen Years," in *Proceedings of the 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing*. Austin, United States: ICST, 2013. [Online]. Available: http://eudl.eu/doi/10.4108/icst.collaboratecom.2013.254082

[67] S. Wen, Z. Zhao, and H. Yan, "Detecting Malicious Websites in Depth through Analyzing Topics and Web-pages," in *Proceedings of the 2nd International Conference on Cryptography, Security and Privacy*. Guiyang China: ACM, Mar. 2018, pp. 128–133. [Online]. Available: https://dl.acm.org/doi/10.1145/3199478.3199500

[68] A. G. West and A. J. Aviv, "Measuring Privacy Disclosures in URL Query Strings," *IEEE Internet Computing*, vol. 18, no. 6, pp. 52–59, Nov. 2014, conference Name: IEEE Internet Computing. [Online]. Available: https://ieeexplore.ieee.org/document/6879052/?arnumber=6879052

[69] M. T. Whitty, "Who can spot an online romance scam?" *JFC*, vol. 26, no. 2, pp. 623–633, Apr. 2019. [Online]. Available: https://www.emerald.com/insight/content/doi/10.1108/JFC-06-2018-0053/full/html

[70] C. Yue and H. Wang, "BogusBiter: A transparent protection against phishing attacks," *ACM Trans. Internet Technol.*, vol. 10, no. 2, pp. 1–31, May 2010. [Online]. Available: https://dl.acm.org/doi/10.1145/1754393.1754395

[71] P. Zhang, A. Oest, H. Cho, Z. Sun, R. Johnson, B. Wardman, S. Sarker, A. Kapravelos, T. Bao, R. Wang, Y. Shoshitaishvili, A. Doupé, and G.-J. Ahn, "CrawlPhish: Large-scale Analysis of Client-side Cloaking Techniques in Phishing," in *2021 IEEE Symposium on Security and Privacy (SP)*, May 2021, pp. 1109–1124, iSSN: 2375-1207. [Online]. Available: https://ieeexplore.ieee.org/document/9519414/?arnumber=9519414

[72] P. Zhang, Z. Sun, S. Kyung, H. W. Behrens, Z. L. Basque, H. Cho, A. Oest, R. Wang, T. Bao, Y. Shoshitaishvili, G.-J. Ahn, and A. Doupé, "I'm SPARTACUS, No, I'm SPARTACUS: Proactively Protecting Users from Phishing by Intentionally Triggering Cloaking Behavior," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '22. New York, NY, USA: Association for Computing Machinery, Nov. 2022, pp. 3165–3179. [Online]. Available: https://dl.acm.org/doi/10.1145/3548606.3559334

APPENDIX

TABLE VI
ALL TOPICS BY TRAINING BERTOPIC ON THE INITIAL DATASET OF .EU WEBSITES.

Topic	Count	Representation
0	561	['privacy', 'https', 'data', 'policy', 'privacy policy', 'marketing', 'personal', 'email', 'services', 'opt']
1	262	['confirm', 'unsubscribe', 'button confirm', 'receive emails', 'chosen', 'survey', 'button', 'correct', 'emails', 'receive']
2	167	['cleaning', '2025', 'ld cleaning', 'ld', 'sociale', 'kbc', 'partijen', 'mei', 'mei 2025', 'samenleving']
3	115	['hotel', 'reis', 'vertrek', 'betaling', 'boeken', 'vakantie', '099', 'tarief', 'flex', 'aanpassen']
4	114	['fietsknoop', 'string', 'route', 'données', 'array', 'cookieduur', 'privacybeleid toestemming', 'toestemming', 'privacybeleid', 'routes']
5	99	['bewaartermijn', 'type', 'cookie', 'type cookie', 'gebruikt', 'informatie', 'informatie aanbieder', 'aanbieder', 'type lokale', 'lokale']
6	99	['vpn', 'beveiliging', 'nordvpn', 'chat', 'digitale', 'svenska türkçe', 'norsk polski', 'chat online', 'bescherming versterken', 'jaarabonnement']
7	98	['bewaartermijn', 'lokale', 'lokale opslag', 'opslag', 'cookie', 'permanent', 'type', 'gebruikt', 'gebruiker', 'type lokale']
8	96	['español', 'français', 'ads', 'english united', 'personalized', 'english', 'italiano', 'português', 'deutsch']
9	90	['telenet', 'center', 'backup', 'sint', 'denijs', 'sint denijs', 'denijs westrem', 'westrem', 'internet', '00']
10	76	['lichaam', 'helpt', 'ondersteunt', 'natuurlijke', 'vitaminen', 'krachtige', 'ingrediënten', 'kruiden', 'huid', 'balans']
11	74	['kaarten', 'moederdag', 'geboorte', 'gefilterd', 'kaart', 'verjaardag', '99', 'fotokaarten', 'kleur', 'gefilterd kleur']
12	70	['yorgo', 'linkedin', 'vennootschap', 'entreprise', 'bedrijfswinkel yorgo', 'bedrijfswinkel', 'achat', 'entreprises', 'agree', 'vente']
13	65	['jobs', 'report', 'followers', 'report post', 'linkedin', 'premed', 'play', 'post', 'checkpoint', 'reprobel']
14	62	['kaarten', 'geboorte', 'teksten', 'moederdag', 'fotokaarten', 'ballonnen', 'verjaardag', 'liefde', 'felicitatie', 'chocolade']
15	62	['jobs', 'europe', 'wemove', 'european', 'parliament', 'error', 'wemove europe', 'download', 'european parliament', '404']
16	59	['enter', 'bespaar simpel', 'actief gemeente', 'bespaar', 'gemeente', 'simpel', 'enter enter', 'postcode', 'enter characters', 'actief']
17	59	['products', 'cookies', 'close', 'close products', 'meta products', 'meta', 'glasses', 'optional', 'cookies cookies', 'learn']
18	56	['enter', 'characters enter', 'enter valid', 'enter enter', 'postcode', 'enter characters', 'valid', 'characters', 'gebruiken', 'checken moment']
19	55	['auva', 'apple', 'ipad', 'iphone', 'mac', 'watch', 'shop', 'airpods', 'support', 'apple watch']
20	55	['informatie', 'dak', 'kijzer', 'project', 'kijker', 'zonnepanelen', 'gevel', 'woning', 'isolatie', 'realisaties']
21	53	['data unsubscribed', 'unsubscribed removed', 'powered', 'unsubscribed', 'removed', 'data', '', '', '', '']
22	52	['schoonmaken', 'huis', 'schoon', 'vuil', 'chemicaliën', 'center', 'sint', '09', 'geverifieerde klant', 'oppervlakken']
23	51	['vie', 'guide', 'enfants', 'leçons', 'moments', 'questions', 'transmettre', '500', 'valeurs', 'histoire']
24	50	['kaarten', 'geboorte', '99', 'fotokaarten', 'moederdag', 'felicitatie', 'hallmark', 'cadeaus', 'cadeau', 'teksten']
25	48	['brico', 'persoonsgegevens', 'gegevens', 'gebruiken', 'weigeren energie', 'verwerking', 'cookies gebruiken', 'gepersonaliseerde', 'verzameld', 'toepassing gebruiken']
26	47	['btw', 'incl btw', 'excl btw', 'excl', 'incl', 'verkocht', 'stuk', 'stuks', 'btw verkocht', 'stuk verkocht']
27	45	['spaans', 'consumenten', 'panel', 'lessen', 'gebruiken', 'spaanse', 'leren', 'cursus', 'toepassing gebruiken', 'bezoekersgegevens verbeteren']
28	44	['organization', 'unsubscribe', 'email preferences', 'receiving emails', 'emails', 'change', 'receiving', 'preferences', 'email', 'update']
29	44	['opleiding', 'coga', 'gebruiken', 'persoonsgegevens', 'bvba', 'dalende', 'hrd academy', 'hrd', 'coach training', 'academy']
30	44	['marketing bureau', 'bureau', 'views', 'inhouse', 'playing', 'marketing', 'bedrijf', 'online marketing', 'krachtig', 'online']
31	43	['snoep', 'chips', 'winkelmand', 'frisdranken', 'gram', 'terug', 'drop', 'candy', 'chocolade', 'lay']
32	43	['english', 'bolt', 'paris', 'lou', '37 38', '36 37', 'ontdek', 'winkel beschikbaarheid', 'beschikbaarheid']
33	42	['device', 'advertising', 'profiles', 'select', 'days', 'consent', 'storage', 'content', 'identifiers', 'retention period']
34	42	['musea', 'brugge', 'musea brugge', 'tiktok', 'log', 'bezoek', 'collectie', 'brusk', 'plan bezoek', 'webshop']
35	42	['cookies', 'meta products', 'meta', 'optional', 'products', 'cookies cookies', 'learn', 'learn cookies', 'cookies companies', 'products provide']
36	39	['afmelden', 'mails', 'uitschrijven', 'ontvangen', 'uitschrijving', 'weet wilt', 'wilt', 'bevestigen', 'adreslijst', 'afmeldt']
37	38	['map', 'jump', 'opleiding', '75 jump', 'jump 75', 'map data', '75', '2025 google', 'data 2025', 'zoom']
38	38	['coach', 'opleiding', 'burn', 'formation', 'leidinggeven', 'opleiding leidinggeven', 'humor wijsheid', 'mensen beweging', 'coachtechnieken', 'wijsheid provocatieve']
39	38	['audi', 'diffuser', 'oem', 'cars', 'background music', 'class', 'background', 'detailing', 'count', 'music']
40	37	['telenet', 'eneco', 'center', 'internet', 'afspraak', 'checkpoint', 'advies', 'tijdstip', 'telenet center', 'sint']
41	37	['00', 'center', 'lokeren', '18 00', '00 18', '18', 'telenet', '00 00', 'checkpoint lokeren', 'eneco']
42	37	['sex', 'women', 'agree', 'girls', 'question', 'online', 'quick', 'photos', 'girl', 'dating']
43	37	['stedentrip', 'portugal', 'bestemmingen', 'gepersonaliseerde', 'español', 'français', 'europa', 'verenigde', 'deal', 'gepersonaliseerde content']
44	36	['data unsubscribed', 'unsubscribed removed', 'unsubscribed', 'removed', 'data', '', '', '', '', '']
45	36	['javascript', 'supported', 'bolt', 'service privacy', 'sign', 'terms service', 'corp', 'browser', 'imprint', 'disabled browser']
46	35	['ez', 'mercedes', '00 excl', '00', '00 uur', 'btw', '000 00', 'info', 'uur', 'excl btw']
47	34	['opleiding', 'trainer', 'train trainer', 'train', 'trainer opleiding', 'leidinggevenden', 'coaching opleiding', 'opleiding trainer', 'opleiding leidinggevenden', 'opleiding coaching']
48	34	['avontuurlijke', 'nieuws', 'uitgelicht', 'inloggen', 'opzijnplek', 'producten gevonden', 'totaal', 'gevonden', 'sluiten', 'ontvang']
49	33	['verbinding', 'voeten', '70', 'apparaten', '70 korting', 'blokkeert', 'ontvang 70', 'korting', 'werkt', 'snelle']

TABLE VII
ALL TOPICS BY TRAINING BERTOPIC ON THE INITIAL DATASET OF .EU WEBSITES.

Topic	Count	Representation
50	33	['toon', 'oorspronkelijke', 'taal', 'versie', 'geverifieerde klant', 'geverifieerde', 'klant', '2023', 'instant', 'bestelling']
51	32	['javascript', 'graag leren', 'kbc graag', 'kbc', 'survey', 'graag', 'campaign 50', 'emails campaign', 'header kbc', 'leren detected']
52	32	['interesses', 'verwijdert bepaalde', 'bijwerken user', 'content mails', 'content onderwerp', 'interesses verwijdert', 'krijg aangepaste', 'interesses basis', 'interesses hieronder', 'basis interesses']
53	32	['kbc', 'vragenlijst', 'javascript', 'graag leren', 'kbc graag', 'graag', 'survey', 'bedrijf', 'minuten beslag', 'employee']
54	31	['enter', 'characters enter', 'enter valid', 'enter enter', 'postcode', 'enter characters', 'valid', 'characters', 'huisnummer controleer', 'toev besparing']
55	31	['waterval', 'stress', 'slaap', 'echt', 'geur', 'oliën', 'essentiële', 'natuurlijke', 'kijken', 'hele']
56	30	['linkedin', 'sign', 'essential', 'password', 'tagalog', 'chinese', 'cookie policy', 'user agreement', 'agreement privacy', 'essential cookies']
57	30	['pins', 'symposia', 'hallmark nederland', 'explore', 'sections', 'touch', 'fisher', 'fisher investments', 'investments', 'pinterest']
58	30	['producten', 'sluiten', 'sluiten producten', 'sale', 'bekijk bericht', 'bericht', 'bekijk', 'cookinglife', 'pannen', '99']
59	30	['hartjes', '00', 'oudenaarde', 'maandag gesloten', 'dinsdag zaterdag', 'zondag maandag', '10u 18u', '10u', 'dames', '18u']
60	29	['pom', 'pep', 'exposed', 'login', 'javascript app', 'persons', 'person', 'certified', 'betalen', 'pay']
61	29	['18 ans', 'ans', 'gratuit confirme', 'granny ronde', 'ans genre', 'ans 18', 'besoin payer', 'adolescente granny', 'confirme', 'confirme âgé']
62	29	['cleaning', 'interesses', 'ld cleaning', 'ld', 'interesses verwijdert', 'interesses basis', 'aangepaste content', 'content mails', 'content onderwerp', 'hieronder krijg']
63	28	['payconiq', 'handelaarsportaal', 'bancontact', 'betalingen', 'cookies', 'bancontact payconiq', 'payconiq company', 'bedrijfsinformatie', 'oplossingen', 'periodieke']
64	27	['formation', 'coaching', 'formation coaching', 'coach', 'begeleiden faciliteren', 'faciliteren', 'authentiek', 'coaching formation', 'carrière', 'begeleiden']
65	27	['spaans', 'spaanse', 'lessen', 'leren', 'cursus', 'praten', 'spaans leren', 'makkelijk', 'spaans praten', 'toegang']
66	27	['primadonna', 'cup', 'volle', '90', 'twist', 'voorgevormde', 'shop', 'lingerie', 'maat', 'body']
67	27	['spaans', 'spaanse', 'lessen', 'leren', 'cursus', 'spanje', 'makkelijk', 'weken', 'spaans leren', 'cintha']
68	26	['checkpoint', 'winkels', 'iphone', 'laptop', 'backup', 'desktop', 'laptop smartwatch', 'goede backup', 'actie', 'waardebon 20']

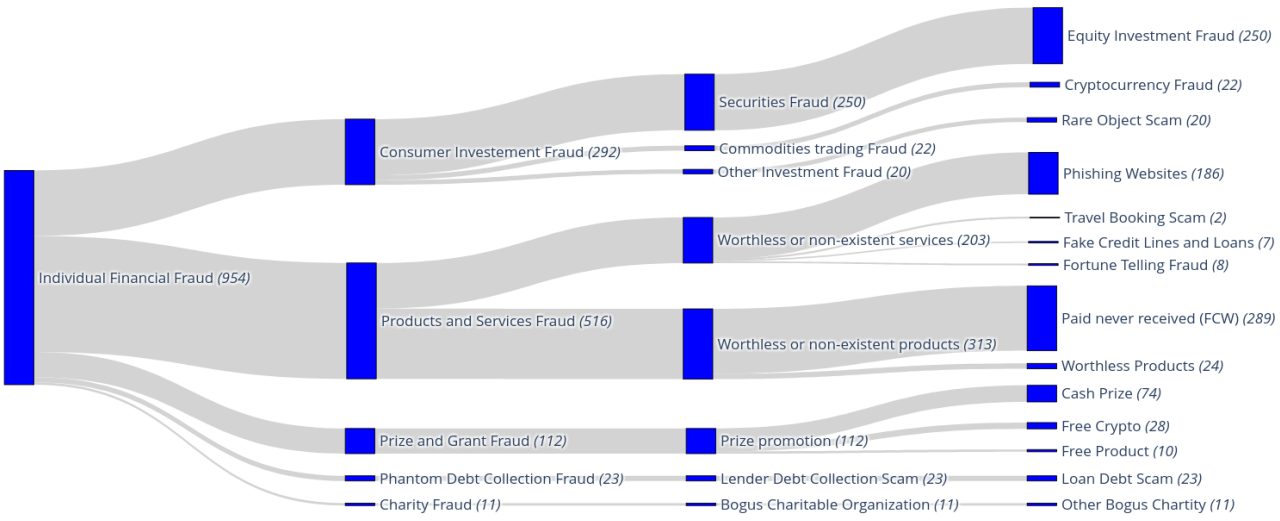

Fig. 8. Visualization of malicious domains sorted using the taxonomy of fraud [12].

TABLE VIII
OVERVIEW OF THE TAXONOMY OF FRAUD BY BEALS ET AL. [12] HIGHLIGHTED ARE THE NEWLY ADDED TYPES OF FRAUD.

Code	Amount	Title	Description
1	954	Individual Financial Fraud	This refers to intentionally and knowingly deceiving the victim by misrepresenting, concealing, or omitting facts about promised goods, services, or other benefits
1.1	292	Consumer Investment Fraud	Investors gain and lose money in financial markets for a variety of legitimate reasons, yet the following definitions refer to investment fraud, where someone knowingly misleads an investor on the basis of false information.
1.2	516	Consumer Products and Services Fraud	This broad category covers all fraud related to the purchase of tangible goods and services.
1.4	112	Prize and Grant Fraud	The hallmark of this category of fraud is that victims are led to believe they will receive winnings in the form of a prize, lottery, grant, or windfall of money, provided that they first purchase certain products or make advance payments to cover fictitious fees and taxes.
1.5	23	Phantom debt collection fraud	This category of fraud refers to fake debt collectors who deceive and possibly threaten individuals to convince them to pay debts they don't owe.
1.6	11	Charity Fraud	This category of fraud involves scam artists collecting money by posing as a genuine charity.
1.1.1	250	Securities Fraud	Investment fraud dealing with securities, which are tradable financial assets in the form of an ownership position in a publicly-traded corporation (stock/equity), or a creditor relationship with governmental body or corporation (bond/debt).
1.1.2	22	Commodities trading fraud	Fraud in connection with commodities transactions (such as transactions in foreign currency, agricultural products, energy products, and precious metals).
1.1.3	20	Other investment opportunities fraud	Other investment opportunities that are not securities or commodities such as rare objects scam.
1.2.2	203	Worthless, unnecessary, or non-existent services	Schemes in which victims intentionally enter an agreement for a service that turns out to involve false or misleading promotions or is not provided at all.
1.2.1	313	Worthless or non-existent products	This category includes schemes that involve false or misleading information about products, goods, housing, or experiences (including vacations and concerts) that over exaggerate claims, turn out to be worth much less than anticipated, or are non-existent.
1.4.1	112	Prize promotion/sweepstakes scam	Fraudsters send out letters, use email, or call potential victims telling them they have won a prize or are entitled to a financial reward, but they need to pay a small 'administrative' or shipping and handling fee to obtain their winnings.
1.5.2	23	Lender debt collection scams	A sub-type of "Phantom Debt Collection Fraud" in which fraudsters impersonate well-known lending institutions to convince individuals to pay back debt they don't owe.
1.6.1	11	Bogus charitable organization	A sub-type of charity fraud in which the fraudsters collect donations by pretending to be genuine non-profit organizations representing a variety of causes, from natural disaster relief to children's issues.
1.1.1.1	250	Equity Investment Fraud	A type of securities fraud in which the fraudulent investment vehicle is equity or stock in a publically-traded corporation.
1.1.2.4	22	**Crypto Investment Fraud**	**A type of commodity trading fraud in which individuals are promised high profits by investing in new or non-existing cryptocurrencies.**
1.1.3.3	20	Rare objects scam	Investment fraud dealing in rare objects, like coins, artwork, or stamps.
1.2.2.9	186	Phishing websites/emails/calls	Phishing is a fraudulent attempt to acquire sensitive information by masquerading as a trustworthy entity.
1.2.2.15	2	Travel Booking Scam	In this worthless services scam, a person claiming to be a travel agent or representative of a travel booking company deceives the victim into paying for a vacation or a vacation rental that either does not exist or is grossly misrepresented.
1.2.2.7	7	Fake credit lines and loans	These scams are often targeted at people with bad credit or in need of a loan. The scam may start as a legitimate appearing email or website offering online lending services, or a victim is informed they qualify for a loan over the telephone but first must pay money up front.
1.2.2.8	8	Fortune Telling Fraud	A type of scam in which a purported psychic or fortune teller convinces victims that a curse has been placed on them and their family.
1.2.1.2	289	Paid never received	These are schemes in which victims intentionally enter an agreement and pay for something that they never receive. A common example are fraudulent e-commerce shops.
1.2.1.1	24	Worthless Products	These are schemes that involve false or misleading information about products or goods that over exaggerate claims or turn out to be worth much less than anticipated.
1.4.1.3	74	Cash prize	A type of prize promotion scam in which the supposed prize is cash.
1.4.1.5	28	**Free Crypto**	**A type of prize promotion scam in which the supposed prize is a cryptocurrency.**
1.4.1.1	10	Free product	A type of prize promotion scam in which the supposed prize is a product, ranging from small trinkets to brand new cars.
1.5.2.2	23	Loan debt scam	In this scam, victims are contacted by scam artists claiming that they have unpaid debt associated with a personal loan that was provided sometime in the past.
1.6.1.99	11	Other bogus charity	Other types of bogus charities.