Beneath the Cream: Unveiling Relevant Information Points from CrimeBB with Its Ground Truth Labels

Felipe Moreno-Vera[1], Daniel Sadoc Menasché[1], and Cabral Lima[1]

Federal University of Rio de Janeiro (UFRJ), Brazil
felipe.moreno@ppgi.ufrj.br,sadoc@dcc.ufrj.br, cabrallima@ufrj.br

Abstract. In the realm of cybersecurity, identifying and mitigating the exploitation of vulnerabilities is crucial. Building on prior research that analyzes underground hacking forums, this study refines methodologies for detecting vulnerability exploitation within underground discussion forums. Using the CrimeBB dataset, previous works employed machine learning approaches to extract insights, label textual information, build predictive models, and classify forum posts discussing Common Vulnerabilities and Exposures (CVE). Recently, the PostCog framework was released to facilitate navigation of the CrimeBB data. Building on this, the current study integrates the PostCog extension with ChatGPT, enhancing the labeling of posts by type, intent, and crime category into new classifications such as Proof-of-Concept (PoC), Weaponization, and Exploitation. Additionally, using the SHAP explanation method, we uncover insights into the keywords frequently found in the text—such as "fud", "sell", "buy", and "pm"—which have emerged as significant indicators of exploitation.

Keywords: Cybersecurity · CrimeBB · underground forums · NLP · PostCog · explainability · SHAP

1 Introduction

In recent years, monitoring underground hacking forums has become increasingly important as these platforms serve as valuable sources of threat intelligence (TI) on vulnerabilities and exploits [4, 8]. These forums enable cybercriminals to exchange knowledge, discuss newly discovered vulnerabilities, and share techniques. Analyzing these discussions provides insights into emerging trends, exploit methods, and potential targets [3]. Consequently, early detection and proactive responses informed by real-world intelligence from these forums can significantly enhance vulnerability management and incident response strategies.

In [15] we have proposed a machine learning approach, leveraging the CrimeBB dataset, to classify forum posts discussing Common Vulnerabilities and Exposures (CVE) into categories like PoC and Exploitation, achieving F1 scores above 90%. Ground truth information was not available at CrimeBB, and a major effort in previous works involved the manual labeling of posts and threads. More

recently, a new extension to CrimeBB, namely PostCog, provides (noisy) ground truth labels for a subset of CrimeBB posts.

In this work, building upon [15],[1] we integrate the PostCog extension into the dataset, allowing for enhanced labeling for each post type, intent and crime type. Notably, terms like "fud" (fully undetectable) and "pm" (private message) emerged as significant indicators of exploitation. Those terms already appeared as relevant in [15], and our new results leveraging the PostCog extension serve to reinforce some of our previous findings.[2]

Utilizing the refined labels, the study advances methods for detecting exploitation, reaffirming previous conclusions while gaining a deeper understanding of hacking community dynamics. The presentation of these findings highlights the synergy between initial methodologies and new data, stressing the importance of continuous data refinement for cybersecurity threat detection and understanding.

Contributions. In summary, our key contributions are threefold. First, we provide an analysis of the exploitation of vulnerabilities in the wild using the PostCog dataset and GPT labeling. Second, we present a classifier to assess imminent threats based on underground forums. Third, we perform an explanation of our classifier to understand the most relevant words related to vulnerabilities in underground forums.[3]

Outline. The remainder of this paper is organized as follows. In Section 2 we report related work. Section 3 presents the methodology, followed by results in Section 4. Finally, Section 5 concludes.

2 Related works and background

In what follows, we discuss related work and background related to the main themes of our work.

Exploitation of vulnerabilities in the wild. The analysis of the exploitation of vulnerabilities in the wild poses significant challenges. While attackers actively target vulnerabilities to compromise systems, often sharing techniques and information on online forums, researchers must understand these exploits to build effective defense strategies. Manual methods, such as reverse engineering and fuzzing, are particularly useful in uncovering new vulnerabilities and weaknesses, helping to strengthen defenses against threats, whether discovered in controlled environments or exposed on blackhat forums [12].

Collaboration and information sharing through platforms such as the National Vulnerability Database (NVD), along with responsible disclosure and bug bounty programs, enhance cybersecurity by enabling efficient reporting and resolution of vulnerabilities, ensuring timely responses to emerging threats [9, 11]. In this paper, in particular, we focus on posts shared at CrimeBB and PostCog

[1] The title of this paper is a reference to [15], indicating that it is a follow-up.

[2] A preliminary version of this work was presented at the Cambridge Cybercrime Conference `https://www.cambridgecybercrime.uk/conference2024.html`.

[3] The full version of this paper is available at `https://tinyurl.com/cscmlmoreno`

that explicitly cite CVEs. CVE identifiers allow us to enrich, augment and cross-relate information from those forums against information available at NVD.

NLP and threat intelligence (TI). The use of Natural Language Processing (NLP) for the analysis of hacker forums has been considered in [15–17,19]. In this work, we complement such a body of literature by focusing on discussions about software vulnerabilities within CrimeBB forums, which have previously been considered for the analysis of eWhoring and other cybercrimes [10,17].

In our previous work [15] we focused on the detection of CVE identifiers within forums and on the comparison of market prices of vulnerabilities across different underground forums [2]. In [14] we leveraged topic modeling to infer and track the themes discussed within the forums. In this work, we extend [14, 15] focusing on the analysis of threads within CrimeBB that can be linked to known CVEs leveraging PostCog for that matter.

3 Methodology

In this section, we present the dataset, explain our methodology, and detail preprocessing steps.

3.1 Datasets

Cambridge Cybercrime Center (CCC) makes available 38 underground forums collected in the CrimeBB dataset [17]. The PostCog framework [18] extends CrimeBB providing a web application to navigate through CrimeBB data, to find keywords and to label posts following a crowd sourcing approach.

CrimeBB. Most posts in online forums consist primarily of plain text, though they may also incorporate additional elements such as images, videos, or attachments. By using the CrimeBot, CCC identifies and annotates various elements within posts, including images, video (via iframes), snippets of source code, external website links, internal thread links, and attachments [17].

The CrimeBB dataset comprises hierarchical data based on websites, containing around 45Gb of textual information. The structure of underground forums is composed of forums or "websites", followed by boards related to "topics", threads corresponding to "questions", and posts corresponding to "answers". As of August 29, 2024, CrimeBB has 6,739,073 users interacting on 38 websites, 10,600,580 discussion threads, and 117,365,492 posts.

PostCog. PostCog [18] is a web application providing curated data from CrimeBB underground forums. It is now in its second phase of development. The initial prototype was constructed using NodeJS, ExpressJS, and PostgreSQL. Following insights derived from tests with users, the technology stack was expanded to integrate ReactJS and ElasticSearch. PostCog users can filter search results by forum and subforum, date, and NLP tags. Results can be exported as CSV files for analysis by external applications.

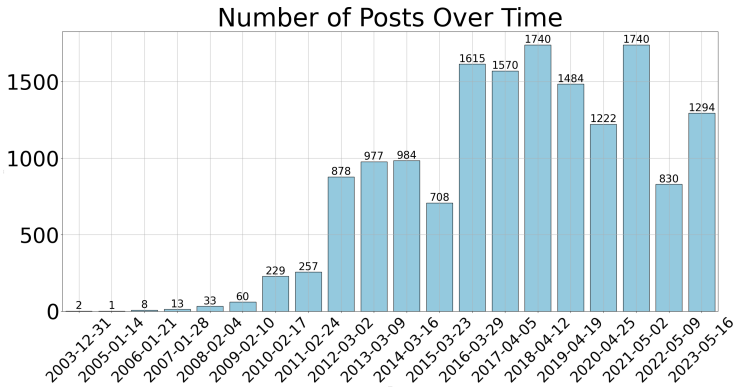

Fig. 1: Longitudinal report of posts that cite CVEs. We use the case-insensitive regular expression `cve-[0-9]{4}-[0-9]{4,}` to filter such posts. The posts were created between January 8, 2004, 23:13:00 and May 23, 2023, 22:08:23. We have observed a growing trend in the number of posts.

3.2 Labeling Targets Classes

We focus on threads that mention Common Vulnerabilities and Exposures (CVE) identifiers. We search for CVE ids using the case-insensitive regular expression `cve-[0-9]{4}-[0-9]{4,}`. In Figure 1, we show the number of posts that cite CVEs, as a function of the post creation dates. After analyzing our dataset, we found about 650 null values in thread titles, and 7,203 posts without content. Those posts were removed from our analysis.

PostCog labels We obtain three sets of labels from PostCog:

1. **post type**, which refers to the content of the post, and is retrieved from [7];
2. **intention**, which refers to user intention (e.g., negative, positive, or neutral) and is retrieved from [7];
3. **crime type**, which refers to the criminal activity identified within the thread, and is obtained from [22].

These labels are currently available only for the HackForum site; therefore, most of our analysis focus on it.

Expert annotations in HackForum We leverage the manual labeling by experts reported in [15].[4] We asked experts to label a subset of threads in the CrimeBB dataset using the HackForum website. The experts used the following code book to manually label the threads:[5]

[4] In [15] we already had all expert labels considered in this work, but we used only three of them (exploitation, PoC and weaponization).

[5] Accounting for slang and abbreviations that are typical in those communities is left as a subject for future work.

Crime type labels	
Labels	Samples
Not criminal	2,307
Bots/Malware	604
Sql Injection	208
Credentials	41
VPN/proxy	34
DDoS/booting	12
Spam/marketing	7
CurrencyXchange	4
Identity fraud	2
eWhoring	1

(a)

Intention labels	
Labels	Samples
Neutral	2,184
Other	494
Positive	197
Gratitude	170
Aggression	53
Negative	37
PrivateMessage	30
Moderate	28
Vouch	27

(b)

Post type labels	
Labels	Samples
InfoRequest	912
Comment	909
Other	494
OfferX	490
Exchange	137
RequestX	137
Tutorial	76
Social	65

(c)

Expert annotations labels	
Labels	Samples
Weaponization	397
PoC	242
Others	190
Exploitation	102
Warning	55
Help	41
Scam	10

(d)

Table 1: We present the 3,220 threads and set of labels obtained from the PostCog framework: (a) Crime type labels correspond to cybercrime types discussed in a post, (b) Intention labels refer to the intention expressed within a post, and (c) Post type labels correspond to the content of the post. In addition, in (d) we present the 1,037 HackForum threads annotated by experts.

- **Exploitation:** (1) mention a well-known hacker group; (2) reference cryptocurrencies and keywords like bitcoin, exploitation, and attack (in the context of attacks in the wild); (3) discuss approaches to make exploits fully undetectable; (4) involve markets of exploits.
- **Proof-of-Concept (PoC):** (1) contain keywords such as PoC, tutorial, or guide (in the context of producing tools in a lab or controlled environment); (2) provide a tutorial on building a PoC; or (3) discuss vulnerabilities without using exploits in the wild.
- **Weaponization:** (1) contain keywords like vulnerability and exploit (in the context of weaponization); (2) discuss the availability of fully functional or highly mature exploits, providing references or source code.
- **Warning:** contains advice or warning about the detection (e.g., by some company) of exploitation of some vulnerability.
- **Help:** contains keywords such as help or coding referring to users asking help to exploit some vulnerability or execute some code of general purpose.
- **Scam:** contain information about transactions, or selling an exploit or vulnerability that does not work or has already been published.
- **Others:** Other discussion not related to any above.

In Table 1, we summarize the number of samples for each class for each set of labels analyzed in this work. Note the imbalance of samples for each set of labels: "not criminal", "infoRequest" and "neutral" are the most representative classes for crime type, post type, and intention, respectively.

ChatGPT labeling We ask ChatGPT [6] to label our threads using prompts given the context and content of each thread. We ask each set of labels to summarize: (1) post type, (2) crime type, (3) intention, and (4) expert annotations in HackForums. One of our aims was to reduce the number of classes and to have a representative set of up to four labels for each of the above four dimensions. We use ChatGPT 3.5-turbo and perform a few-shot learning [1], by giving five samples of input and output for each class of each set of labels (crime type, post type, intention, and expert annotations). Below we give the prompt template example used for crime type labeling:

```
I have a list of possible labels {not criminal, bots/malware, ...} related to cyber criminal
↪  activites.

I want to perform two tasks: the first one is to group the list of possible labels into a
↪  smaller list of labels: e.g,  {not  criminal,  criminal activity A, criminal activity B,
↪  ... }, this  new list of labels should be the most accurate to group all of them.

The second task is to set a new label using the new smaller list of labels given an input raw
↪  {text} and their corresponding {label}.

Based on the following samples { (text 1, label 1, new label 1), \dots, (text 5, label 5, new
↪  label 5)}, I would like to label the  following texts {text 1, text 2, text 3, ...}}

Please return in a list of tuples: [ ( "input", {text},   "new_label", {new label}), ...]
```

Consequently, we execute this prompt to obtain a new group of labels from each set of labels and the new labels for each sample. In Table 2, we present the new set of labels, classes, and the number of samples assigned by the ChatGPT prompt. In all cases, we significantly reduce the number of classes, while preserving a high imbalance in crime type and intention labels.

From this, ChatGPT learns to classify threads using the following criteria:

- (1) **crime type:** recategorized into "not criminal" defined as activities that are not considered criminal, "cybercrime activities" defined as illegal activities related to cybercrime, and "cybercrime support services" defined as services that facilitate cybercrime activities.
- (2) **post type:** recategorized into "communication/interaction" defined as forms of interaction or content type, "requests" defined as seeking information or services, "offer/exchanges" defined as providing services or trading, and "others".
- (3) **intention:** recategorized into "sentiment" defined as emotions or attitudes, "expression" defined as ways of communicating or expressing oneself, "intensity" defined as levels of strength or forcefulness, and "others".

[6] OpenAI ChatGPT: `https://www.openai.com/chatgpt`

Crime type labels	
Labels	Samples
Not criminal	2,307
Cybercrime activities	875
Cybercrime support services	38

(a)

Intention labels	
Labels	Samples
Sentiment	2,418
Other	494
Expression	280
Intensity	28

(b)

Post type labels	
Labels	Samples
Communication/Interaction	1050
Requests	1049
Other	494
Offer/exchanges	627

(c)

Expert labels	
Labels	Samples
Malicious activity	509
Informal	297
Others	190
Support and assistance	41

(d)

Table 2: We present the new labels assigned by ChatGPT, (a) crime type labels were grouped into three new labels, (b) intention labels were grouped into four new labels, (c) post type labels were grouped into four new labels, and (d) expert annotations in HackForums were grouped into four new labels.

– (4) **expert annotations in HackForum:** recategorized into "malicious activity" defined as content including about weaponization, exploitation, or scam, "informational" defined as content including PoC, alerts, tutorials, or warnings, "support and assistance" defined as content asking help, and "others".

3.3 Text Preprocessing

Our preprocessing is divided into three steps: (i) NLP preprocessing, (ii) language evaluation, and (iii) feature extraction. Our preprocssing was conducted using the NLPToolKit.[7]

1. **Stop words.** It is key to select which words to filter. Our first step is to filter the following: stop words, punctuations, special characters, and emojis.
2. **English posts.** Following this, we conducted a language evaluation to keep the posts in English.
3. **Feautre extraction.** Finally, we perform a feature extraction process using text encoding-based methods, such as Term Frequency, Inverse Document Frequency (TF-IDF) [21].

4 Results and Discussion

In this section, we will discuss experiments and results. To evaluate the performance of our models, we evaluate our dataset using the following labeling:

[7] Library to preprocess NLP raw texts `https://github.com/fmorenovr/nlpToolkit`

Table 3: Random Forest (RF) performance summary for all experiments.

	Target classes	Accuracy	Precision	Recall	F1
Crime type	PostCog labels	**0.97**	0.97	0.99	0.98
	ChatGPT labels	0.95	0.98	0.94	0.96
	Previous work [22]	0.89	0.9	0.89	0.89
Intention	PostCog labels	0.98	0.95	0.97	0.95
	ChatGPT labels	**0.99**	0.97	0.99	0.98
	Previous work [7]	–	0.78	0.49	0.61
Post type	PostCog labels	**0.81**	0.79	0.89	0.82
	ChatGPT labels	0.74	0.75	0.76	0.75
	Previous work [7]	–	0.91	0.78	0.84
Expert annotations	Expert labels	**0.96**	0.97	0.98	0.97
	ChatGPT labels	0.91	0.92	0.93	0.92
	Previous work [15]	0.86	0.87	0.86	0.86

- **PostCog Labels**: These labels were provided by PostCog domain experts and served as the baseline for our analysis (see Table 1(a), 1(b), and 1(c)).
- **Expert Labels**: These labels were provided by domain experts using a code book reported in [15] (see Table 1(d)). We compare our new results against those ones, focusing on three classes: "poc", "weaponization", and "exploitation".
- **ChatGPT Labels**: We used ChatGPT to generate new labels, to explore the potential of AI-driven annotations and contrast the classification results using the new labels (see Table 2) against the PostCog labels.

HackForum contains 4 million threads, of which only approximately 3200 are labeled with their crime type by PostCog. In the case of post type and intent labels, there are even fewer labels at PostCog (approximately 2700), while the remainder were cataloged as "other" (corresponding to no label). Therefore, one of our attempts while utilizing ChatGPT was to label more threads, to fill up missing data.

Model Training and Evaluation. We trained multiple classification models, including SVM [5], Logistic Regression, Ridge Classifier [24], Decision Tree [23], and random forest [6], and performed GridSearch with five-fold cross-validation to optimize hyperparameters. We used random over-sampling to balance the dataset. Random forest outperformed other models, achieving accuracies between 74% and 99%, while other methods stayed below 60%. Notably, using ChatGPT labeling did not consistently improve results, particularly for predicting post and crime types.

Model Explanation. In order to explain the outcomes produced by random forests, we use the SHAP (SHapley Additive exPlanations), which enables the interpretation of text classifiers, offering insights into model predictions [13, 20].

Predictive Power of Random Forests for PostCog and ChatGPT labels. Table 3 reports the performance of random forests to infer labels provided by PostCog, ChatGPT and by our own expert annotations. We have a separate random forest to infer post type, crime type, intention, and the expert annotations. We note that for crime type and intention, models trained using ChatGPT labels and PostCog labels produce similar performance. However, for

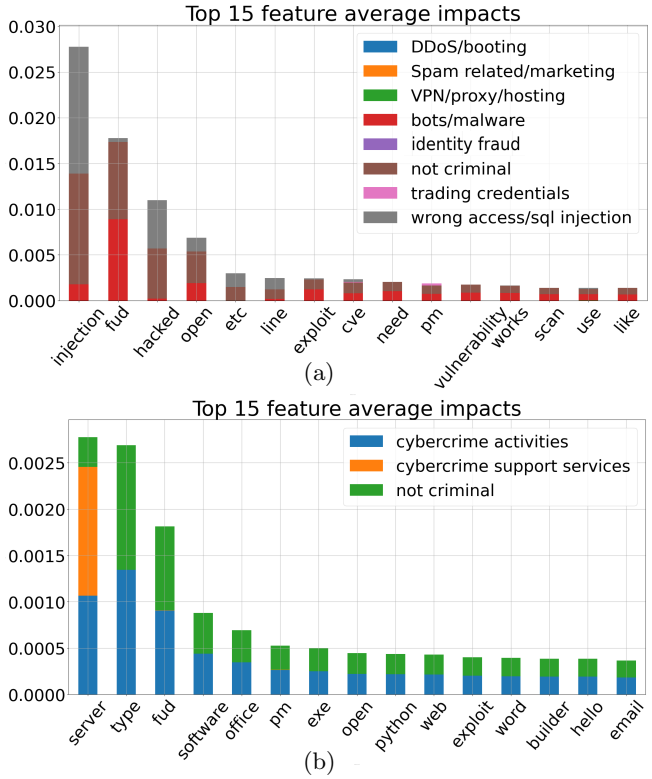

Fig. 2: The 15 most relevant features based on SHAP explanation values were determined from models trained with (a) the **PostCog crime type** labels and (b) the **ChatGPT crime type** labels.

post type and expert annotations, ChatGPT labels were harder to predict. In addition, we take results from our previous work [7, 15, 22], and compare them against our new results, indicating that our new results, irrespectively of the use of ChatGPT labels, correspond to better performance (last line of each experiment in Table 3).

Crime Type. Next, we focus on crime types. Figure 2 reports the classification results and SHAP-generated explanations for the PostCog crime type labels, which achieved a model accuracy of 97%, and the ChatGPT crime type labels, which achieved a model accuracy of 95%.

In Figure 2(a) we focus on PostCog labels. Note that we have eight classes, but classes with low number of samples have minimal relevance for the explanations generated by SHAP. The words "injection", "fud" (fully undetectable),[8]

[8] A fully undetectable exploit is an exploit that is not detected by any of the known antivirus tools.

"hacked", and "open" present a high relevance for the classes "bots/malware", "not criminal", and "wrong access/sql injection".

The word "pm" is relevant when assessing the class "trading credential". "pm" means 'private message', and indicates that a private message will be used to share details about the post, due to privacy issues. Words such as "btc" (for Bitcoin) and "selling" are also relevant for "spam-related/marketing".

In Figure 2(b), the explanations using ChatGPT labels focus on two classes: "not criminal" and "cybercrime activities". The words "server" and "prices" shows a relevance to the class "cybercrime support services". We also find that "php", "malicious", "scan", and "free" are words related to "cybercrime support services".

5 Conclusion

In this work we reported results on the classification of posts from underground forums based on various cybercriminal activities, textual intentions and content type labels. In particular, we leveraged TF-IDF and a random forest classifier, together with the SHAP explainer. We determined the importance of specific keywords in our classification model, revealing that the relevance of keywords such as "fud" (fully undetectable) and "pm" (private message) is closely related to the labels assigned. These findings underscore the effectiveness of our approach in identifying key terms that are crucial to distinguish between different types of cybercriminal activity, textual intentions, or post content type, ultimately contributing to more accurate and insightful cybersecurity analysis.

Acknowledgment

This project was sponsored by CAPES, CNPq, and FAPERJ (315110/2020-1, E-26/211.144/2019, and E-26/201.376/2021).

References

1. Ahmed, T., Devanbu, P.: Few-shot training llms for project-specific code-summarization. In: Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering. pp. 1–5 (2022)
2. Allodi, L.: Economic factors of vulnerability trade and exploitation. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. pp. 1483–1499 (2017)
3. Anderson, R., et al.: Measuring the changing cost of cybercrime. The 2019 Workshop on the Economics of Information Security (2019)
4. Basheer, R., Alkhatib, B.: Threats from the dark: a review over dark web investigation research for cyber threat intelligence. Journal of Computer Networks and Communications **2021**, 1–21 (2021)
5. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. pp. 144–152. ACM (1992)

6. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)
7. Caines, A., Pastrana, S., Hutchings, A., Buttery, P.: Automatically identifying the function and intent of posts in underground forums. Crime Science **7** (2018)
8. Campobasso, M., Allodi, L.: Threat/crawl: a trainable, highly-reusable, and extensible automated method and tool to crawl criminal underground forums. In: APWG eCrime 2022 (2022), arXiv:2212.03641
9. Chen, D.D., Woo, M., Brumley, D., Egele, M.: Towards automated dynamic analysis for linux-based embedded firmware. In: Network and Distributed System Security Symposium (2016)
10. Deguara, N., et al.: Threat miner: A text analysis engine for threat identification using dark web data. In: Big Data. pp. 3043–3052 (2022)
11. Edkrantz, M., Truvé, S., Said, A.: Predicting vulnerability exploits in the wild. 2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing pp. 513–514 (2015)
12. Liang, H., Pei, X., Jia, X., Shen, W., Zhang, J.: Fuzzing: State of the art. IEEE Transactions on Reliability **67**, 1199–1218 (2018)
13. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Neural Information Processing Systems (2017)
14. Moreno-Vera, F.: Inferring discussion topics about exploitation of vulnerabilities from underground hacking forums. In: ICTC. pp. 816–821 (2023)
15. Moreno-Vera, F., et al.: Cream skimming the underground: Identifying relevant information points from online forums. In: 2023 IEEE International Conference on Cyber Security and Resilience (CSR). pp. 66–71 (2023)
16. Pastrana, S., Hutchings, A., et al.: Measuring ewhoring. In: Proceedings of the Internet Measurement Conference. pp. 463–477 (2019)
17. Pastrana, S., Thomas, D.R., et al.: CrimeBB: Enabling cybercrime research on underground forums at scale. In: Proceedings of the 2018 World Wide Web Conference. pp. 1845–1854 (2018)
18. Pete, I., Hughes, J., Caines, A., Vu, A.V., Gupta, H., Hutchings, A., Anderson, R.J., Buttery, P.: Postcog: A tool for interdisciplinary research into underground forums at scale. 2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW) pp. 93–104 (2022)
19. Rahman, M.R., et al.: What are the attackers doing now? automating cyberthreat intelligence extraction from text on pace with the changing threat landscape: A survey. ACM Computing Surveys (2021)
20. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. SIGKDD (2016)
21. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**, 513–523 (1988)
22. Siu, G.A., Collier, B., Hutchings, A.: Follow the money: The relationship between currency exchange and illicit behaviour in an underground forum. EuroS&PW pp. 191–201 (2021)
23. Speybroeck, N.: Classification and regression trees. International Journal of Public Health **57**, 243–246 (2012)
24. Tikhonov, A.N.: On the stability of inverse problems. In: Dokl. Akad. Nauk SSSR. vol. 39, pp. 195–198 (1943)