UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

PPGI PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

UNIVERSITY OF CAMBRIDGE

Automated Detection of Vulnerability Exploitation in Underground Hacking Forums

Felipe A. Moreno-Vera

Advisor: Prof. Daniel Sadoc Menasché

CNPq

Conselho Nacional de Desenvolvimento Científico e Tecnológico

FAPERJ

Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro

CAPES

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

About me

B.Sc. Felipe A. Moreno

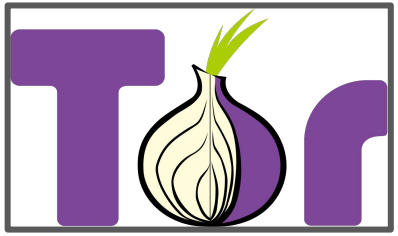
www.fmorenovr.com

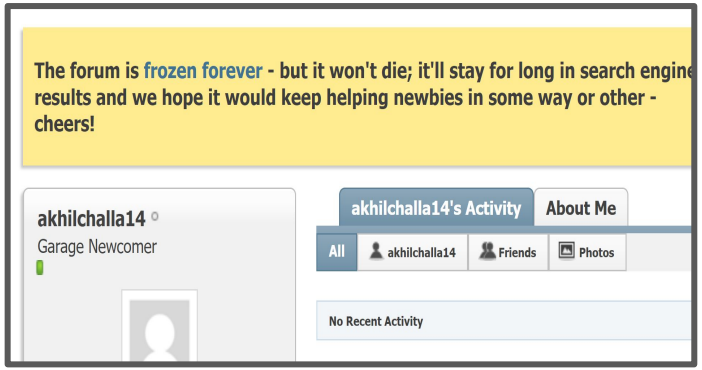
Content

- Context & Motivation
- Datasets
  - CrimeBB
  - PostCog
  - National Vulnerability Database (NVD)
- Methodology
  - Dataset Preparation
  - Exploratory Dataset Analysis
- Cyber-Crimes Forums Analysis
  - Data Pre-processing
  - Models Configurations
  - Experiments and Results
- Conclusions

Context & Motivation

Forbidden/Non Accessible Content





Context

Underground forums are frequently used by malicious actors to discuss vulnerabilities and exploitation strategies. Exploitation of vulnerabilities in the wild poses a significant threat to the Internet ecosystem.

There is a lack of effective methods to process discussions about threats and identify potential exploitation in underground forums.

Motivation

Developing methods to analyze these forums can help predict and prevent cyber attacks, safeguarding critical systems and data.

Analyzing these forums allows for tracking

- Analyse keywords, their usage
- Exploit prices, demand, and targets
- Classify vulnerability level of treats
- Classify discussions threads

Key contributions

- We **propose a methodology** to explore and analyze the **CrimeBB** dataset, which contains 117,365,492 posts from 37 underground-forum websites.
- We **implement a text-based model** that can efficiently **classify the discussion about cybercrime** in threads-posts content while considering the behavior and distribution of the analyzed data.
- We **apply explanation methods** to understand and analyze our model's predictions, aiming to **identify relevant keywords** that provide insights into the discussions within a forum.
- We **use Generative Pre-trained Transformers (GPT)** to analyze and obtain new information from posts and their content. We aim to **evaluate the performance of GPT models when classifying** unstructured information such as forum content.

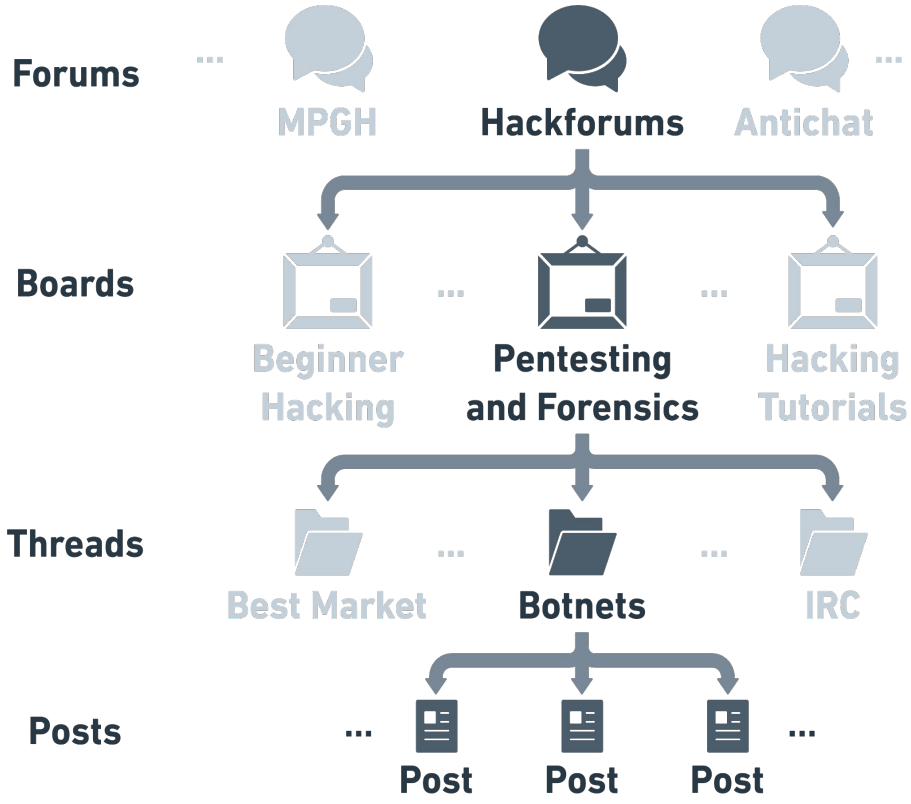
Datasets

CrimeBB

Made available by Cambridge Cybercrime Centre.

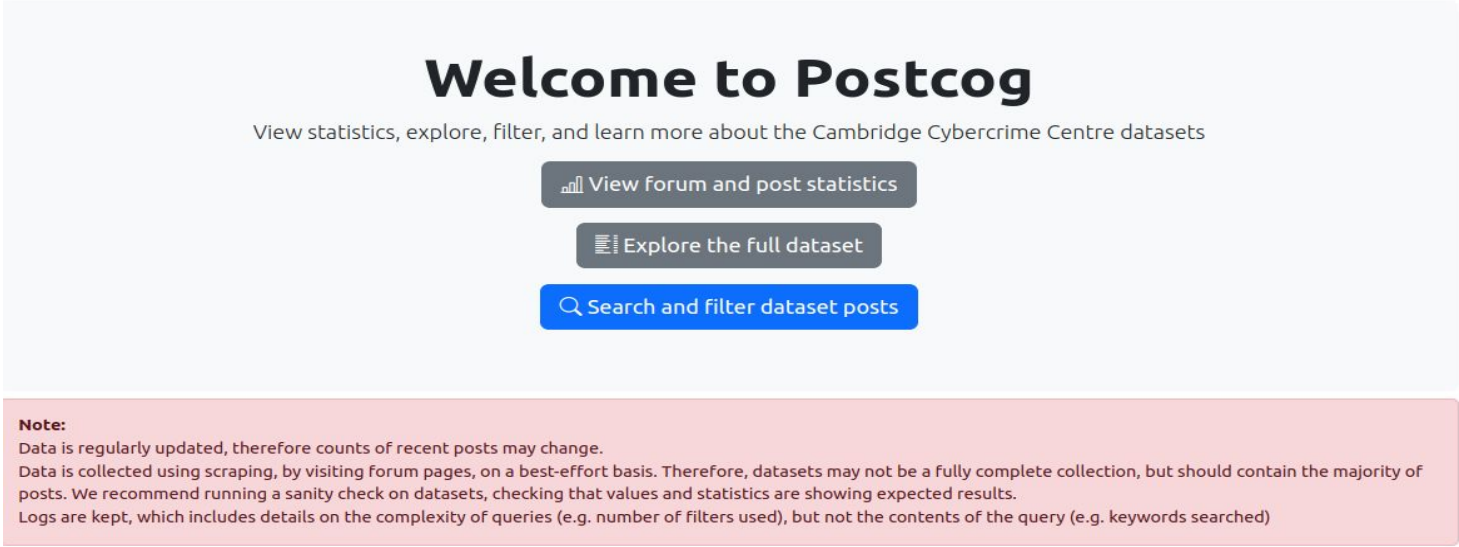
Contains data scraped from multiple underground forums.

Organized in forums, boards, threads and posts.



PostCog

PostCog, a framework to navigate through CrimeBB data.



Obtained from: https://postcog.cambridgecybercrime.uk/

National Vulnerability Database (NVD)

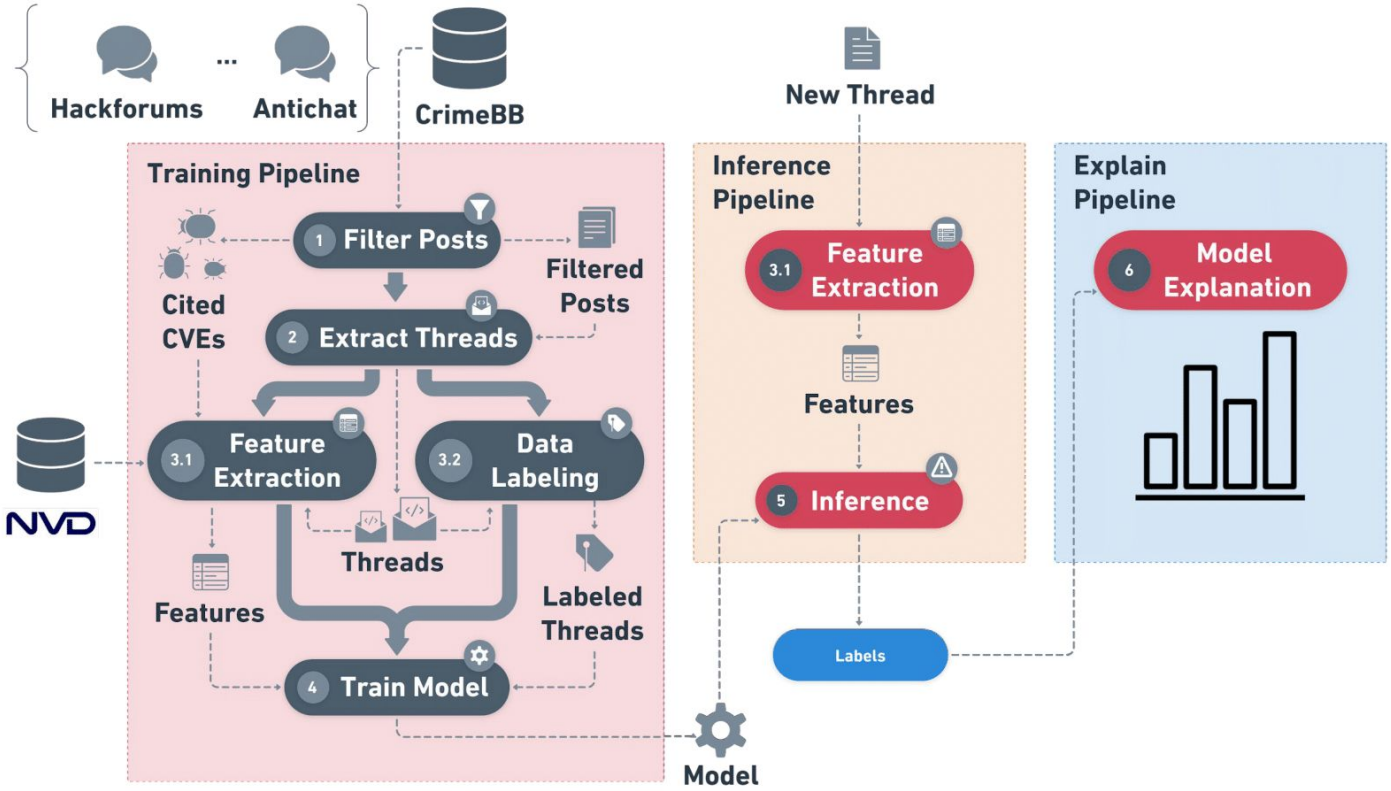
A comprehensive database of reported known vulnerabilities which are assigned CVEs.

Frequency of words of vulnerabilities summary description.

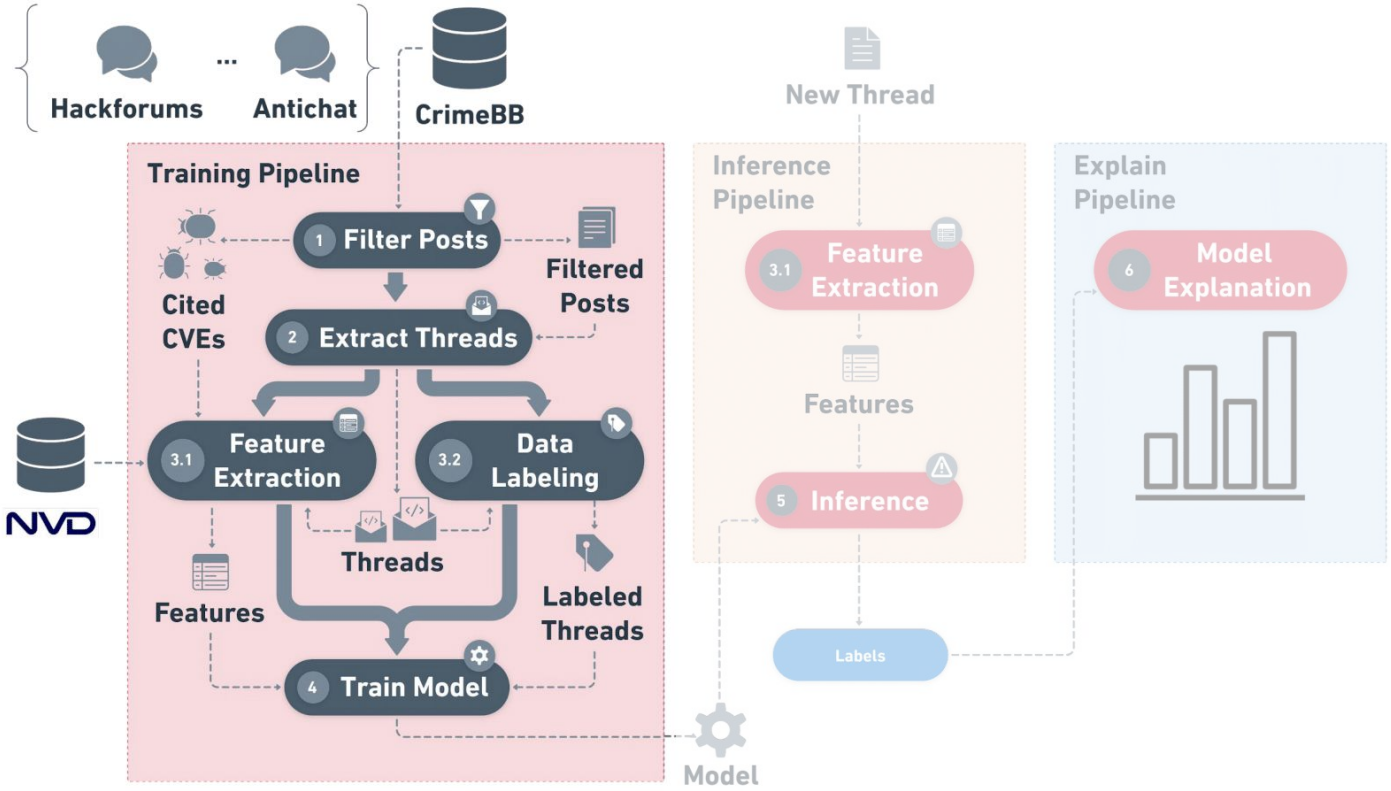
Obtained from: https://nvd.nist.gov/general/visualizations

Methodology

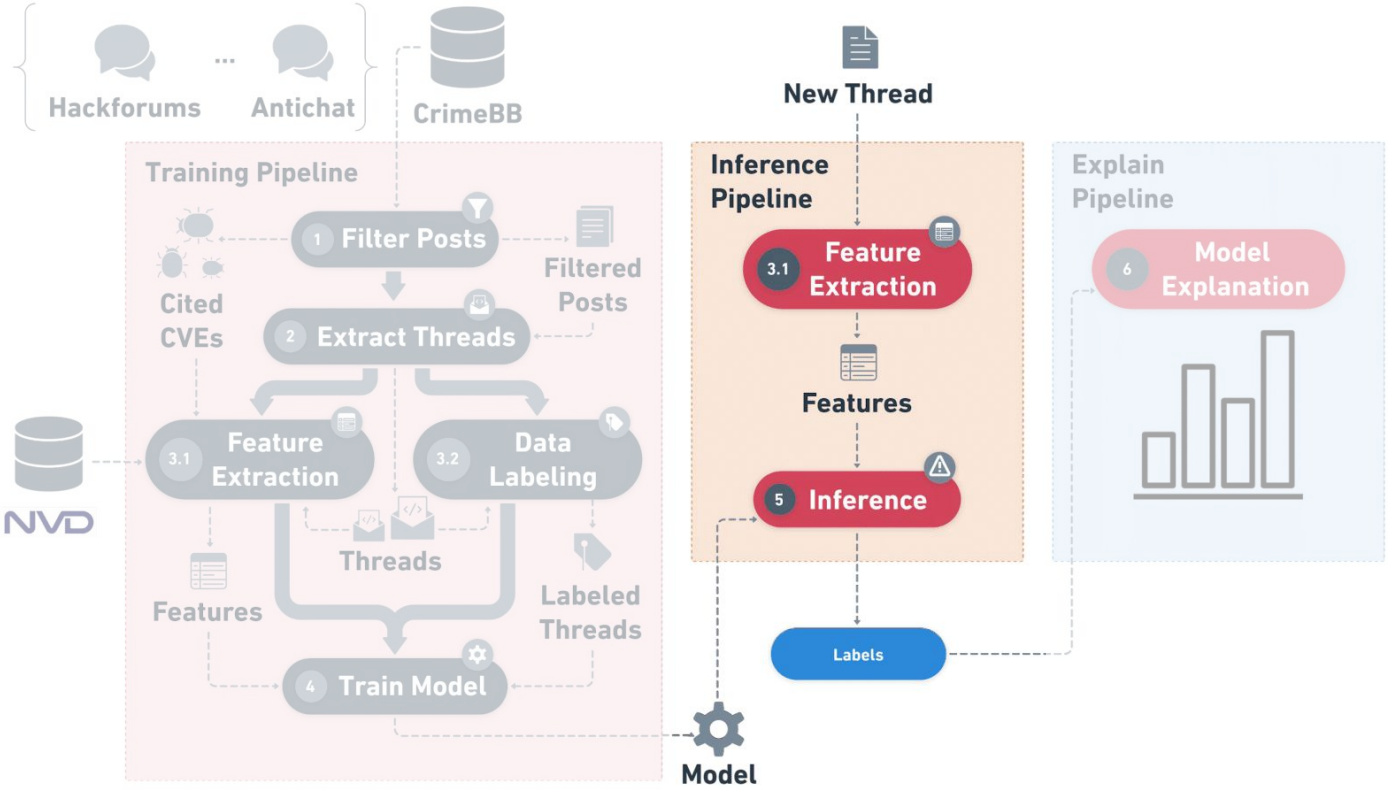
Pipeline


Hackforums
...
Antichat
CrimeBB
New Thread
Training Pipeline
Inference Pipeline
Explain Pipeline
1 Filter Posts
Filtered Posts
Cited CVEs
2 Extract Threads
NVD
3.1 Feature Extraction
3.2 Data Labeling
Threads
Features
Labeled Threads
4 Train Model
Model
3.1 Feature Extraction
Features
5 Inference
Labels
6 Model Explanation

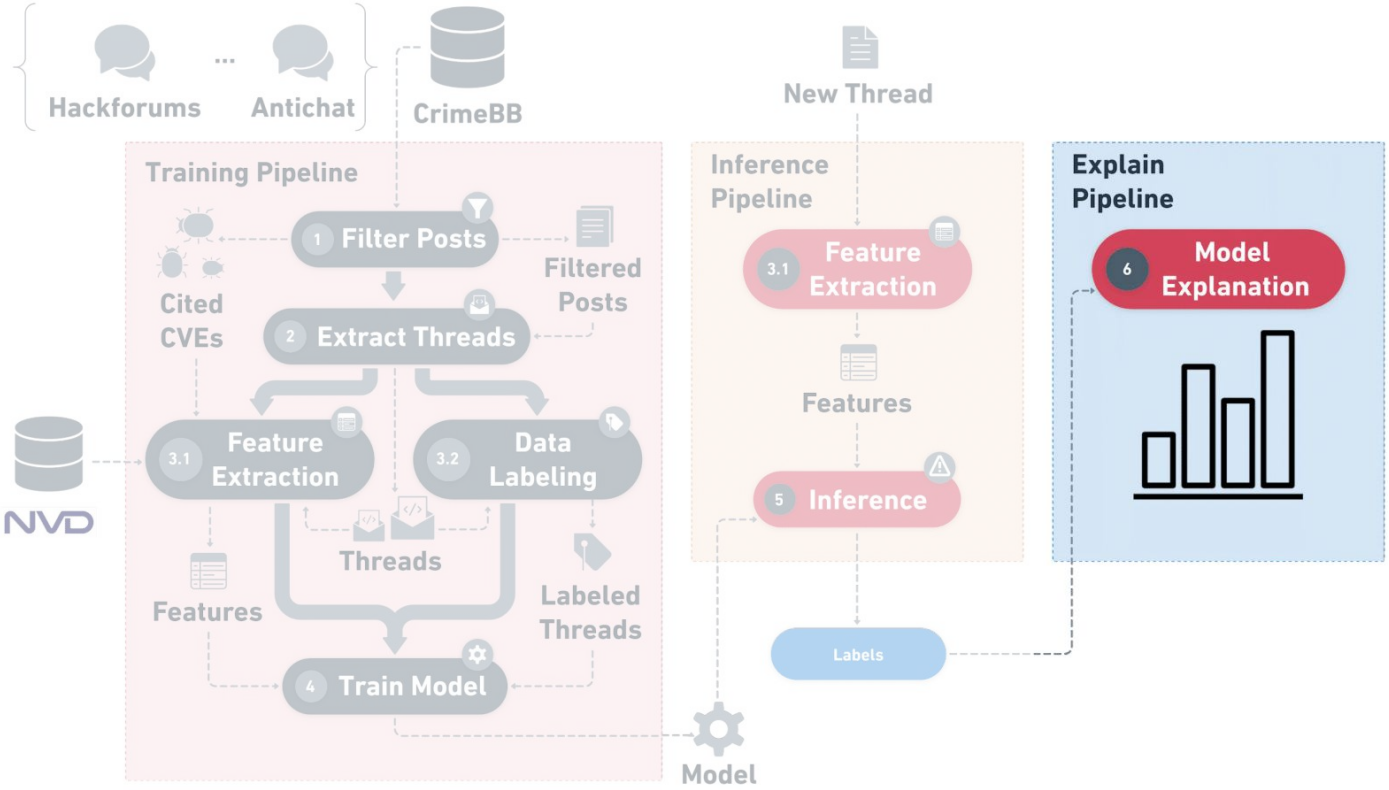
Pipeline


Hackforums
...
Antichat
CrimeBB
New Thread
Training Pipeline
Inference Pipeline
Explain Pipeline
1 Filter Posts
Filtered Posts
Cited CVEs
2 Extract Threads
NVD
3.1 Feature Extraction
3.2 Data Labeling
Threads
Features
Labeled Threads
4 Train Model
Model
3.1 Feature Extraction
Features
5 Inference
Labels
6 Model Explanation

Pipeline


Hackforums
...
Antichat
CrimeBB
New Thread
Training Pipeline
Inference Pipeline
Explain Pipeline
1 Filter Posts
Filtered Posts
Cited CVEs
2 Extract Threads
NVD
3.1 Feature Extraction
3.2 Data Labeling
Threads
Features
Labeled Threads
4 Train Model
Model
3.1 Feature Extraction
Features
5 Inference
Labels
6 Model Explanation

Pipeline


Hackforums
...
Antichat
CrimeBB
New Thread
Training Pipeline
Inference Pipeline
Explain Pipeline
1 Filter Posts
Filtered Posts
Cited CVEs
2 Extract Threads
NVD
3.1 Feature Extraction
3.2 Data Labeling
Threads
Features
Labeled Threads
4 Train Model
Model
3.1 Feature Extraction
Features
5 Inference
Labels
6 Model Explanation

Data Preparation

CrimeBB

As of August 28, 2024, CrimeBB have:

- **6,739,073** users interacting on **37** websites.
- **4,339** boards
- **10,600,580** discussion threads
- **117,365,492** posts.
- More than **~45Gb** of information.

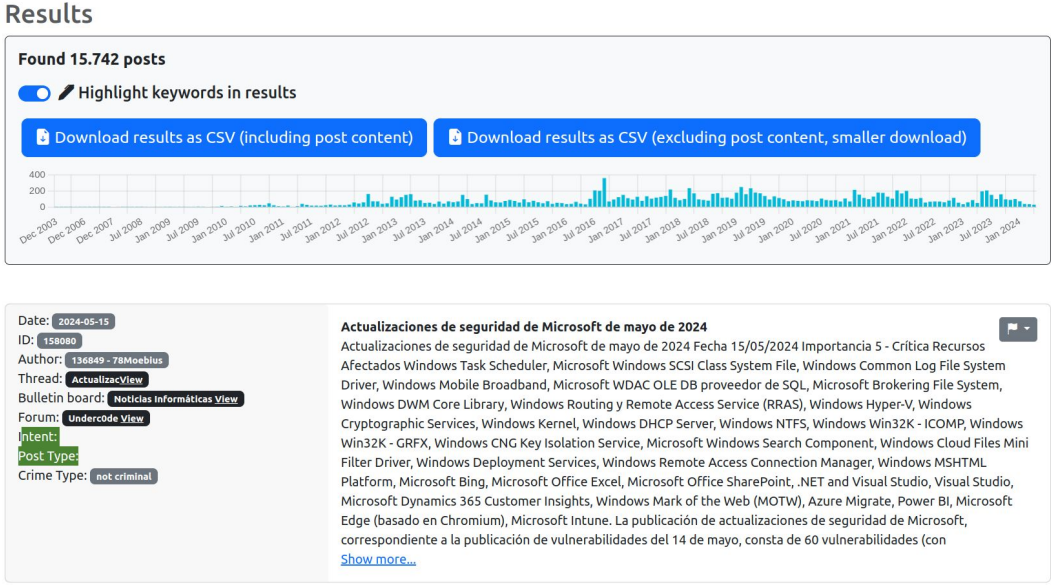
Forum	#Boards	#Threads	#Posts	First post	Recent post
Hack Forums	212	4,301,893	42,686,891	2007-01-27	2024-05-24
Zismo	39	546,832	12,194,525	2010-05-26	2024-05-04
MPGH	770	918,439	12,193,797	2005-12-26	2024-05-26
Blackhatworld	112	1,017,226	12,132,290	2005-10-31	2024-05-20
Nulled	169	687,522	9,546,230	2013-04-02	2024-05-16
lolzteam	292	577,642	6,196,005	2013-03-10	2019-09-01
Cracked	163	419,517	3,911,032	2018-04-03	2024-04-17
OGUsers	58	244,766	3,608,306	1990-01-01	2019-04-09
UnKnoWnCheaTs	248	182,667	2,837,509	2002-11-02	2024-05-24
Antichat	80	254,810	2,642,161	2002-05-29	2024-03-15
V3rmillion	40	456,262	2,459,519	2016-02-02	2019-11-11
Raidforums	88	114,450	1,231,126	2015-03-20	2022-02-20
Elhacker	53	212,081	987,039	2002-08-21	2024-05-28
Probiv	168	123,023	909,007	2014-11-05	2024-04-25
Breached	72	34,412	737,922	2022-03-16	2023-03-19
Hackers Armies	53	42,548	468,880	2009-06-01	2024-04-01
Forum Team	201	44,404	433,901	2017-10-31	2024-03-26
BreachForums	76	28,800	331,357	2023-05-12	2024-05-14
Indetectables	72	32,274	328,539	2006-02-20	2024-05-19
XSS Forum	49	48,718	310,796	2004-11-13	2023-04-27
Dread	446	75,122	294,596	2018-02-15	2020-01-09
Runion	19	16,867	240,632	2012-01-11	2020-01-05
Offensive Community	71	119,251	161,492	2012-06-30	2018-12-11
Undercode	73	27,054	95,723	2010-02-10	2024-05-26
The Hub	62	11,286	88,753	2014-01-09	2019-08-09
Ifud	65	11,827	72,851	2012-05-10	2022-12-19
PirateBay Forum	33	11,526	60,678	2013-10-23	2020-12-03
OnniForums	27	3,542	45,094	2023-02-08	2024-05-24
Torum	11	4,346	28,485	2017-05-25	2019-08-07
Safe Sky Hacks	50	12,963	27,018	2013-03-28	2019-01-23
Kernelmode	11	3,606	26,815	2010-03-11	2019-11-29
Freehacks	228	5,106	26,471	2013-07-27	2023-04-23
Deutschland im Deep Web	43	4,075	20,185	2018-11-22	2020-06-04
GreySec	28	2,232	11,925	2015-06-10	2022-01-04
Garage for Hackers	47	2,329	8,710	2010-07-06	2018-10-13
Stresser Forums	17	708	7,069	2017-04-09	2018-04-09
Envoy Forum	93	454	2,163	2019-07-06	2019-08-09
Total	4,339	10,600,580	117,365,492		

CrimeBB: Manual annotations

- HackForums: **3,037 posts** (in **1,162 threads**) cite a CVE.
- **Manually labeled threads** by the posts content: **1,067**
- Hackforums: **2,666 posts** (in **1,042 threads**) were labeled
- A total of **8,915** (**969 unique**) CVE codes were found

Label	Threads labeled	Threads citing CVE	Posts citing CVE
Weaponization	410	397	891
PoC	247	244	861
Others	195	192	520
Exploitation	107	102	232
Warning	55	55	67
Help	43	42	60
Scam	10	10	35
Total	1,067	1,042	2,666

PostCog



- From PostCog, at the date of **2024-08-28**, we search the word “CVE” and found about **15,742** posts since **2004-01-08** until **2024-05-26**.
- We identify that only post scrapped from **HackForums** has **crime type**, **post type**, and **intent** tags included.

PostCog labels

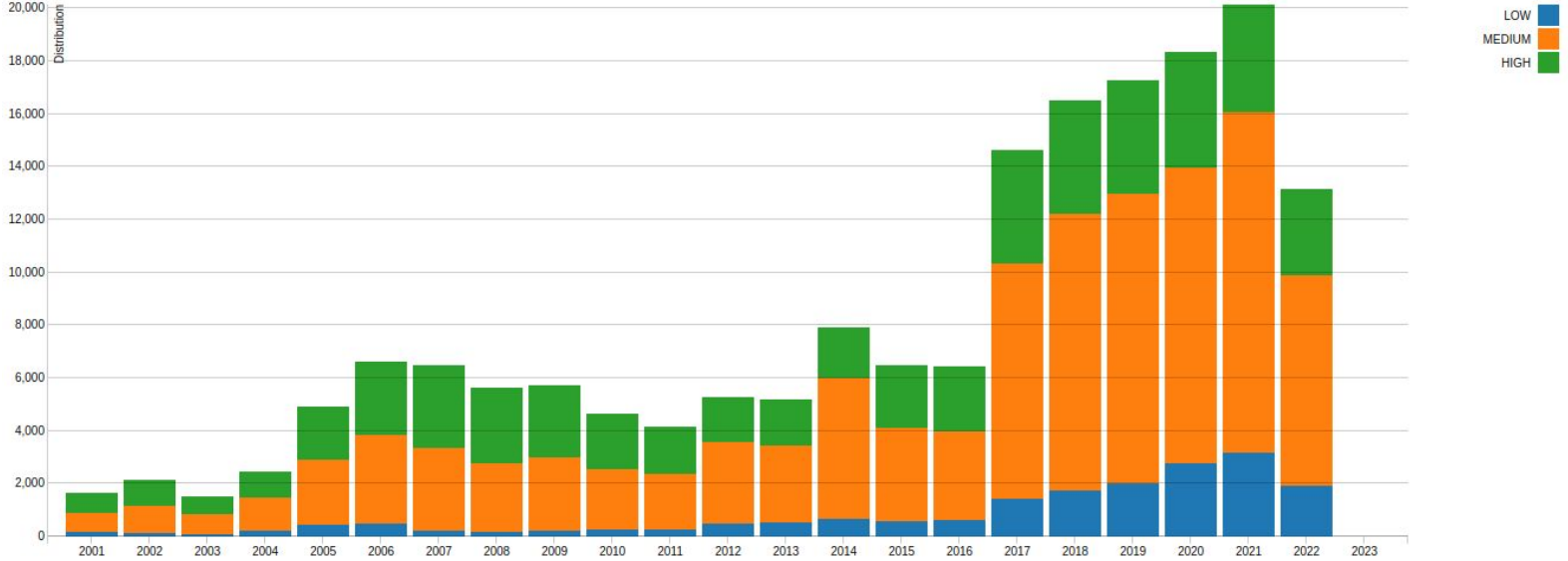
Crime type labels	
Labels	Samples
Not criminal	2,307
Bots/Malware	604
Sql Injection	208
Credentials	41
VPN/proxy	34
DDoS/booting	12
Spam/marketing	7
CurrencyXchange	4
Identity fraud	2
eWhoring	1

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

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

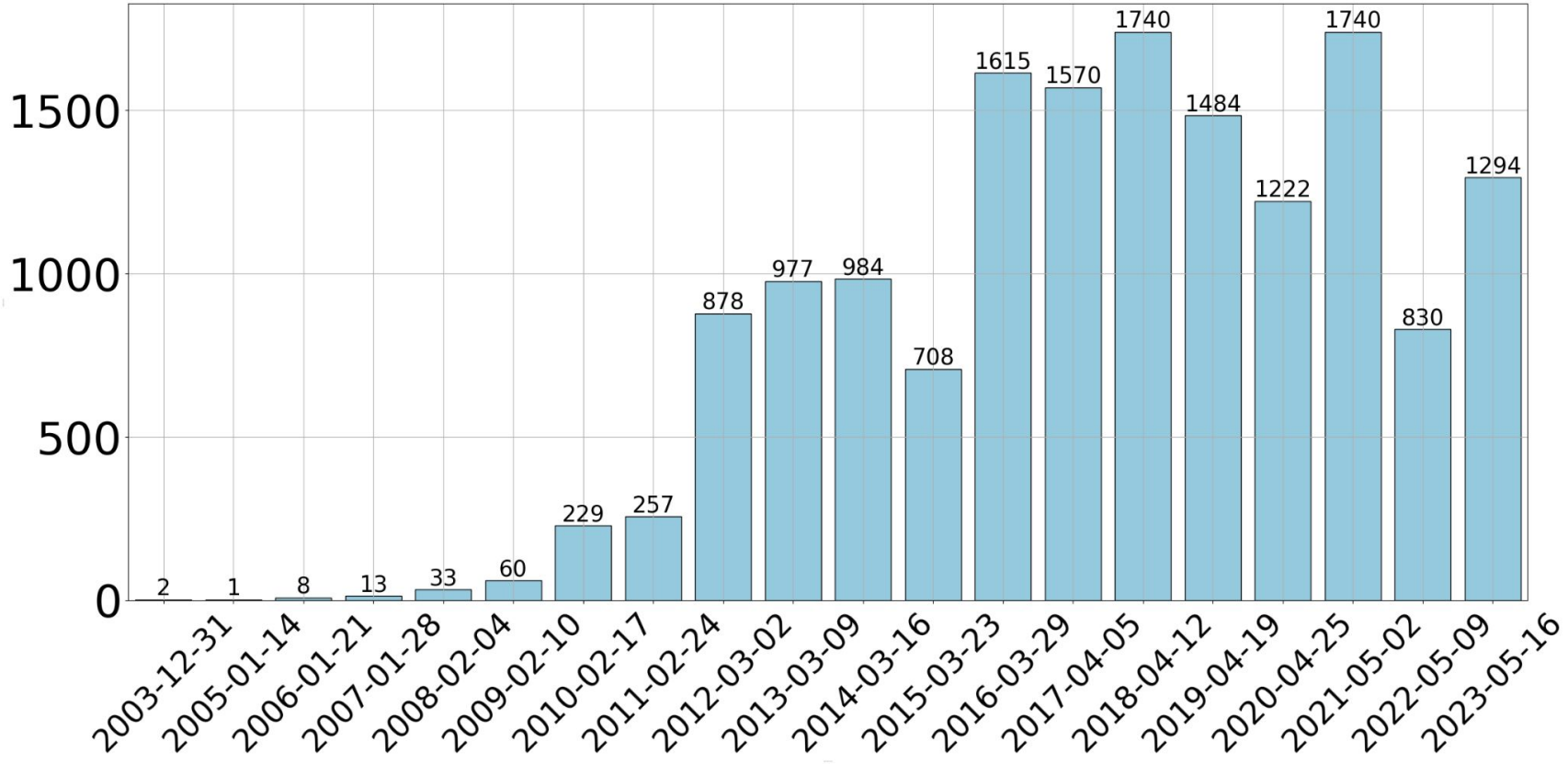
National Vulnerability Database (NVD)

- From NVD, at the date of **2024-08-28**, we present the **CVSS Severity Distribution Over Time**.
- The choice of LOW, MEDIUM and HIGH is based upon the CVSS V2 Base score.



Exploratory Data Analysis

Number of post over time

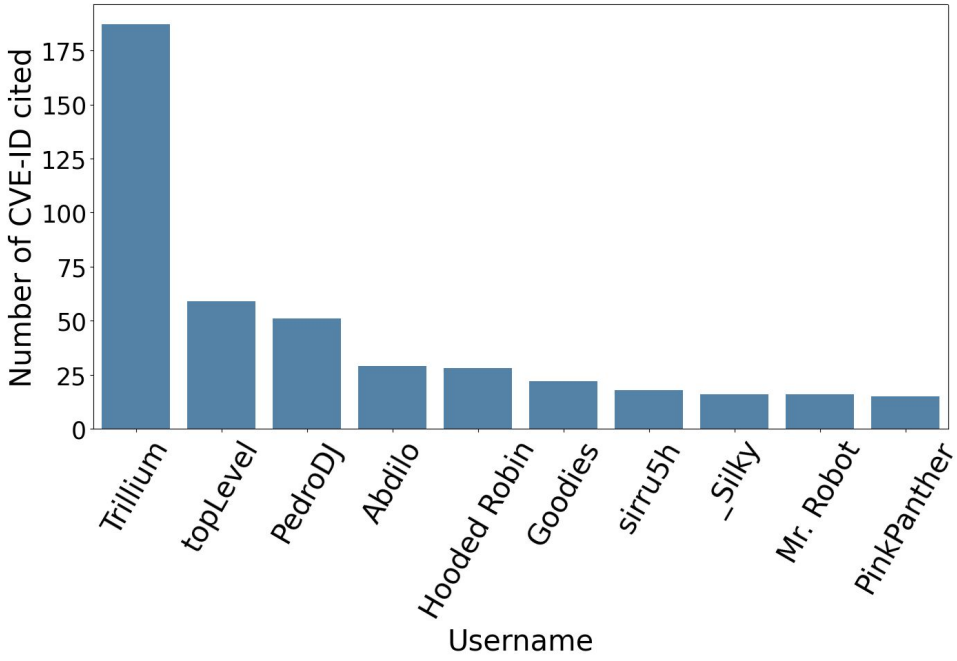


Top 10 users who cites CVE codes

*User **Trillium** mentioned the largest number of CVEs across various posts*

Seems to be advertising exploits

We found some names repeated across forums, but hard to match

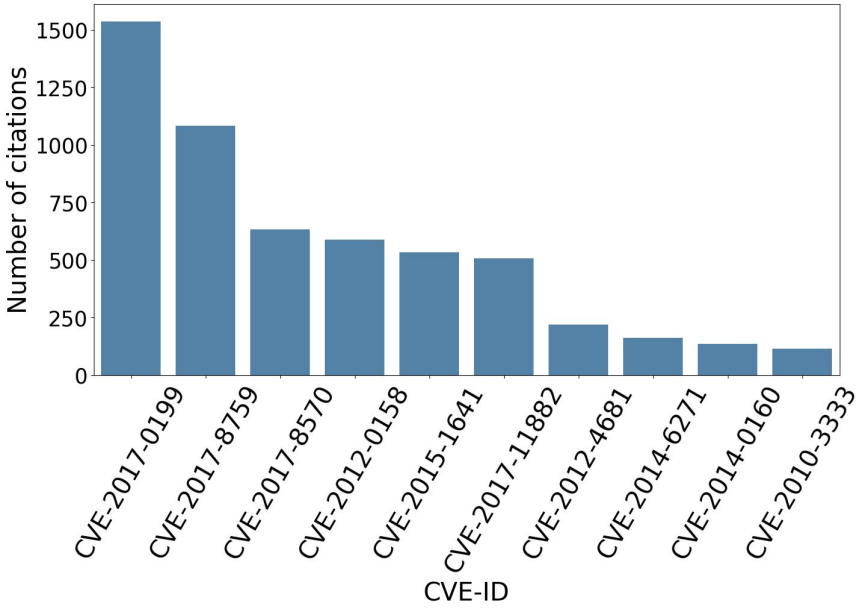


Top 10 CVE codes cited in posts

CVE-2017-0199 affects Microsoft Office: remote execution of arbitrary code.

Top CVEs cover a wide time horizon: from 2010 to 2017

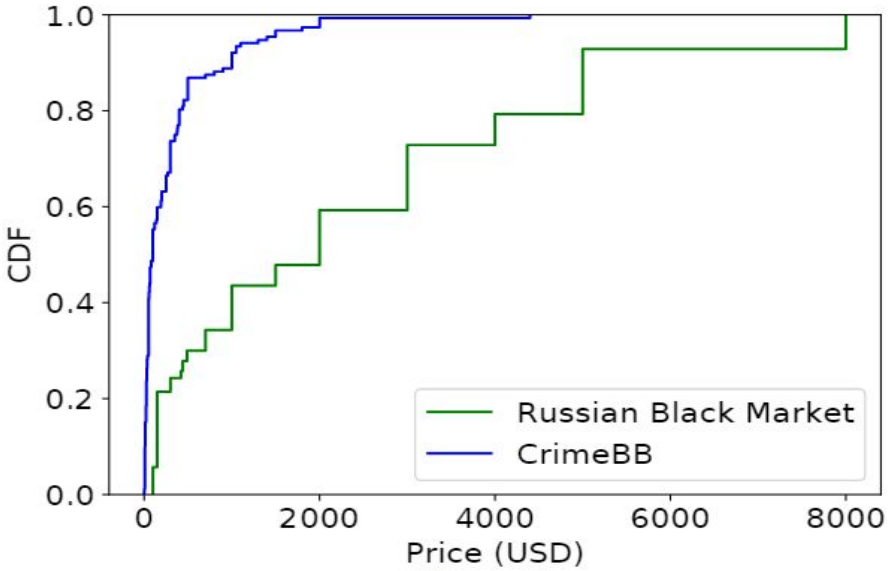
Top 3 most cited CVEs are most recent



CrimeBB vs Russian Market: Prices

How do prices of artifacts sold at CrimeBB compare against Russian Market?

- *Russian Market (ACM CCS, Luca Allodi)*
- ***Prices at Russian Market are larger***
  - *Median value at CrimeBB < 100 USD*
    and > 2000 USD at Russian Market
- ***Why?***
  - *Russian Market is closed market*
  - *Admission control to enter the market*
  - ***Artifacts sold at Russian market are more valuable***



CDF of hacking tools prices

CrimeBB vs Russian Market: Publication delays

How do delays at CrimeBB compare against Russian Market?

Delay definition:

CrimeBB

date post at CrimeBB - date NVD published CVE

Russian Market

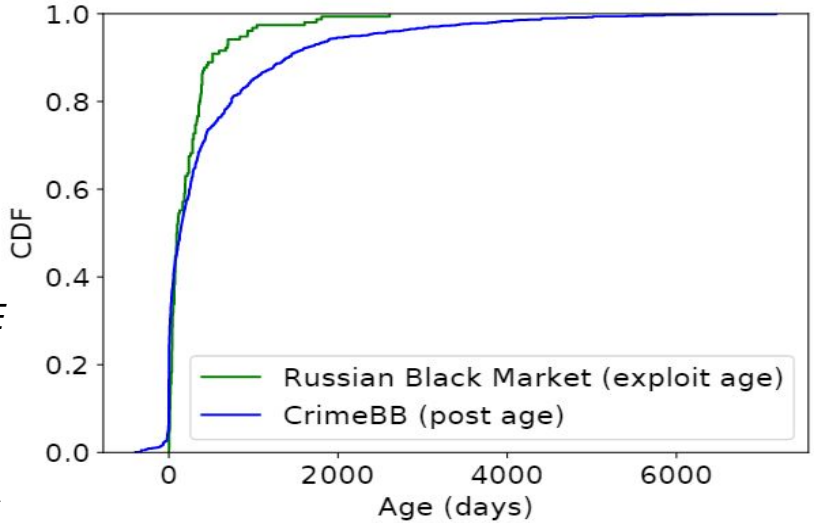
date exploit publication at market - date NVD published CVE

Delays at CrimeBB are larger: why?

Russian Market is closed market

Exploits are published at Russian market and activity ceases

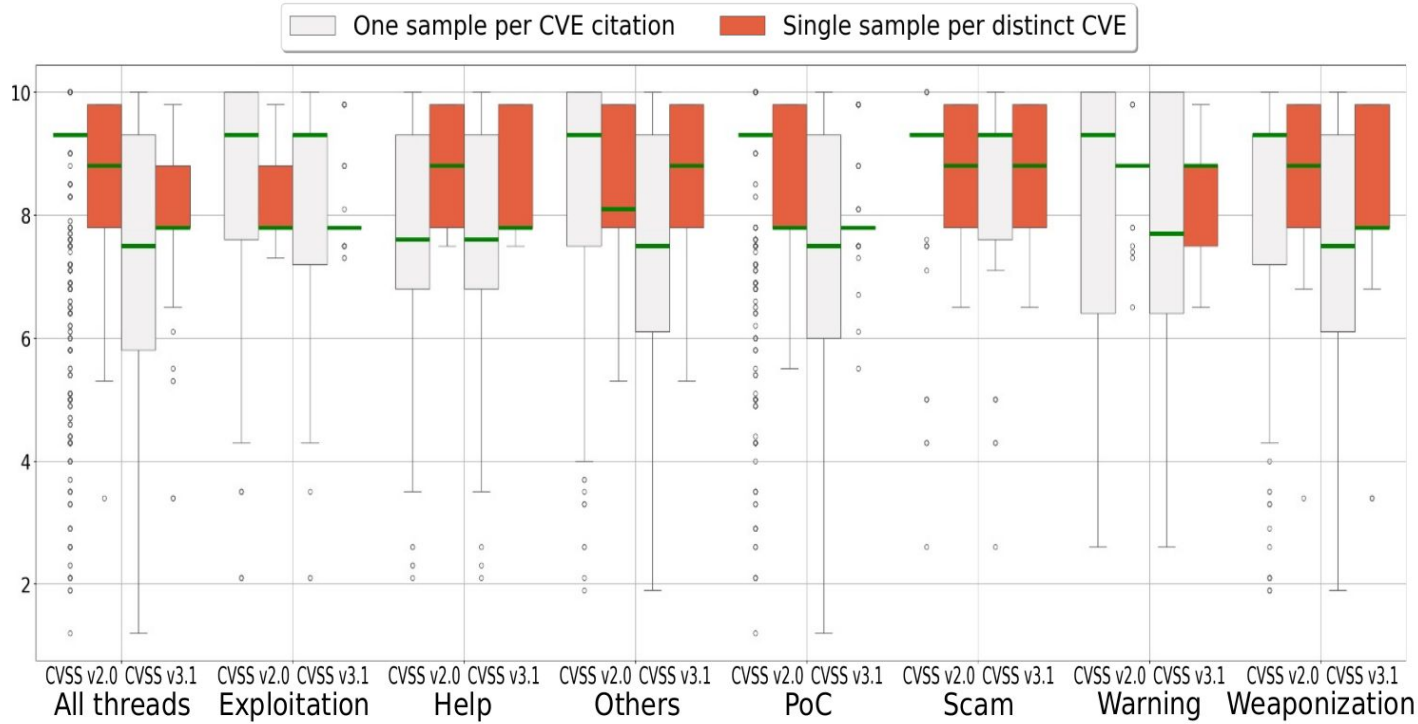
At CrimeBB, continuous discussion *of exploitation strategies*



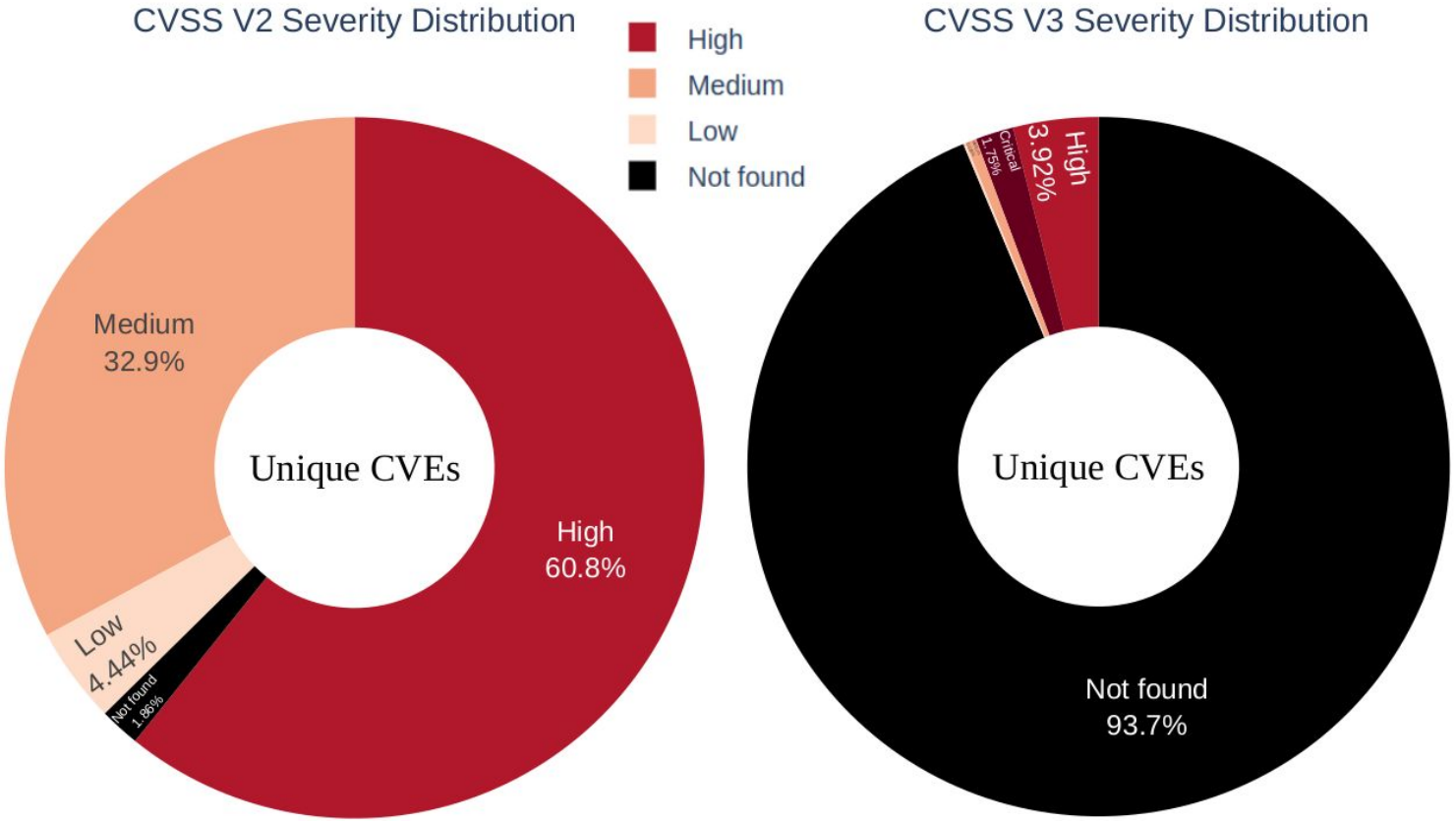
CDF of the difference in days between CrimeBB citation and NVD publish date

CVSS scores of CVE codes

CVSS ranges between 0 and 10, and values above 8 correspond to high risk.

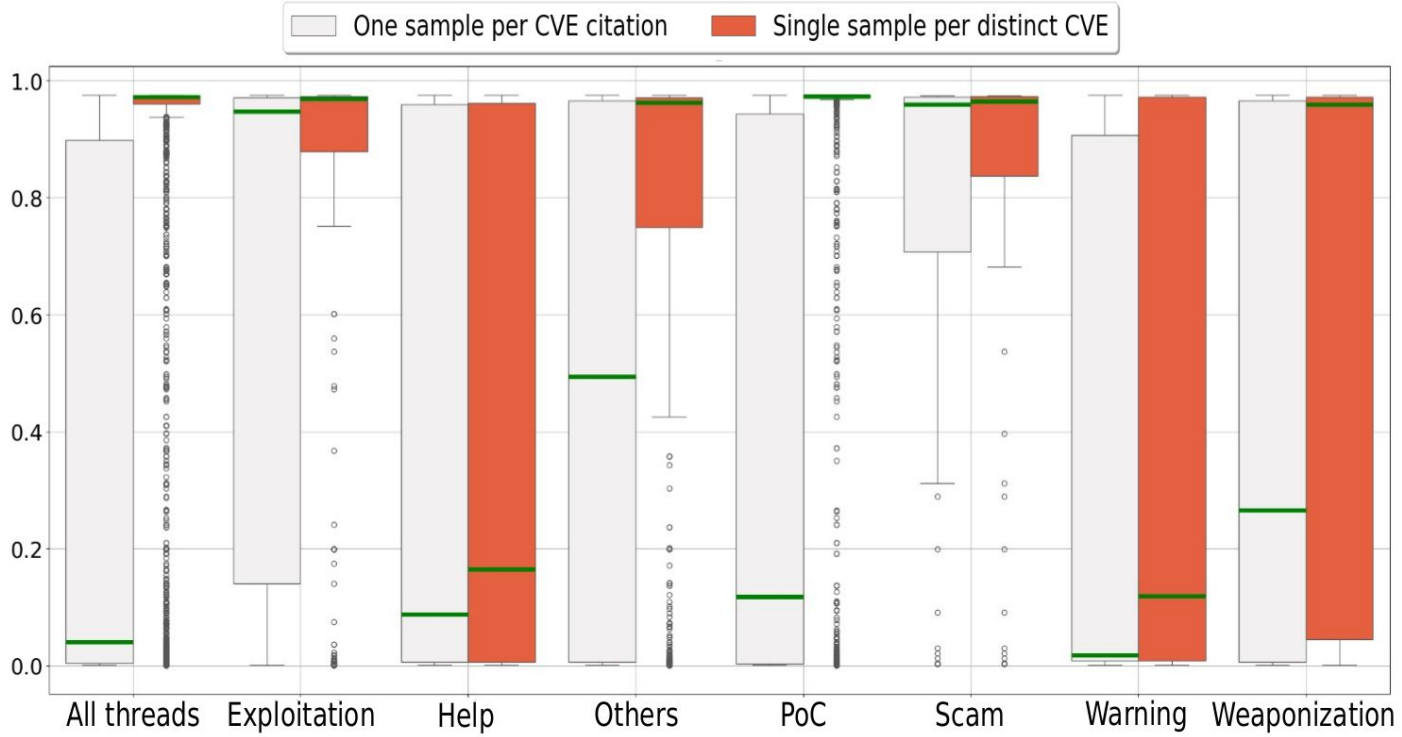


CVSS scores of CVE codes



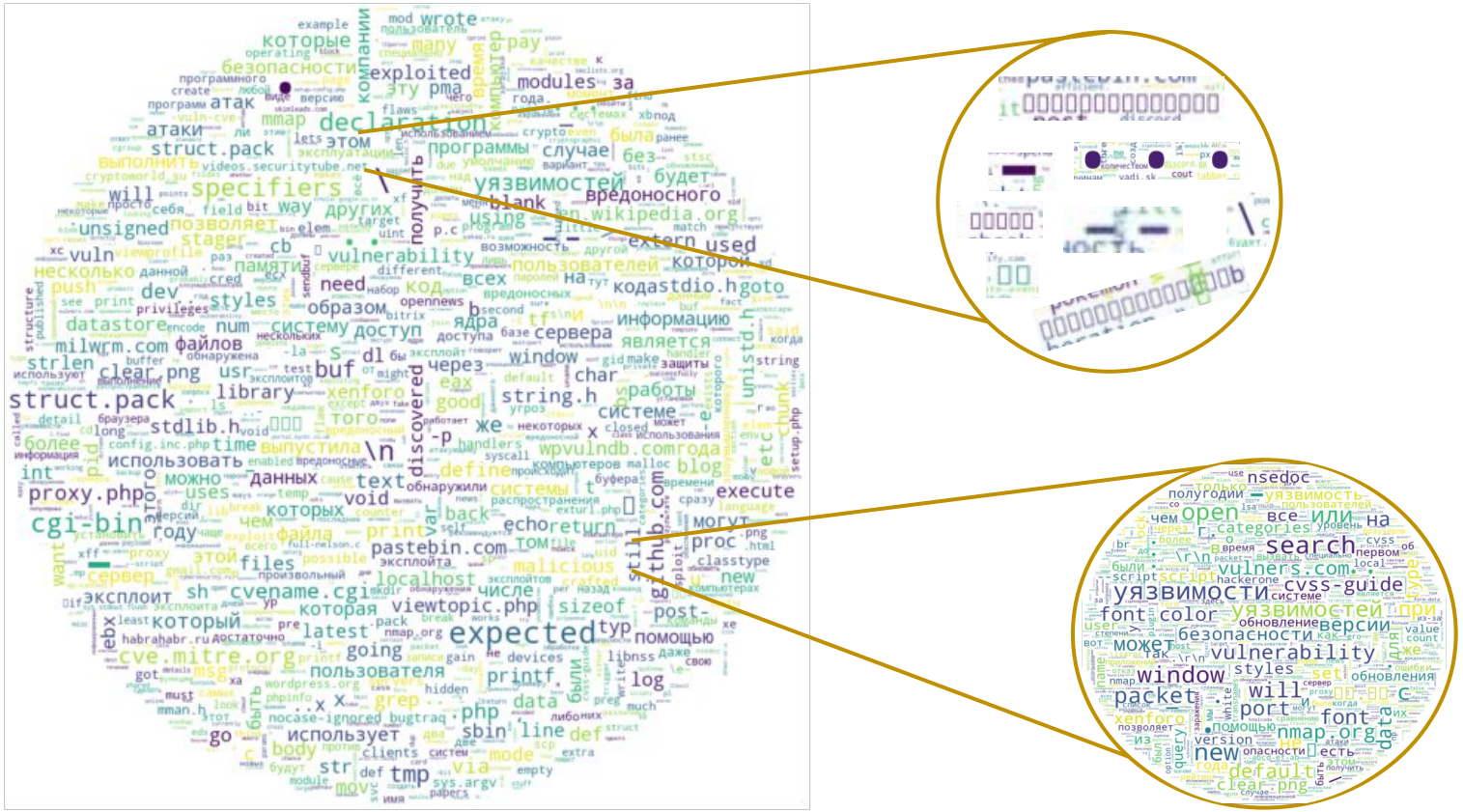
EPSS scores of CVE codes

EPSS represents the probability of exploitation in the wild in the next 30 days.

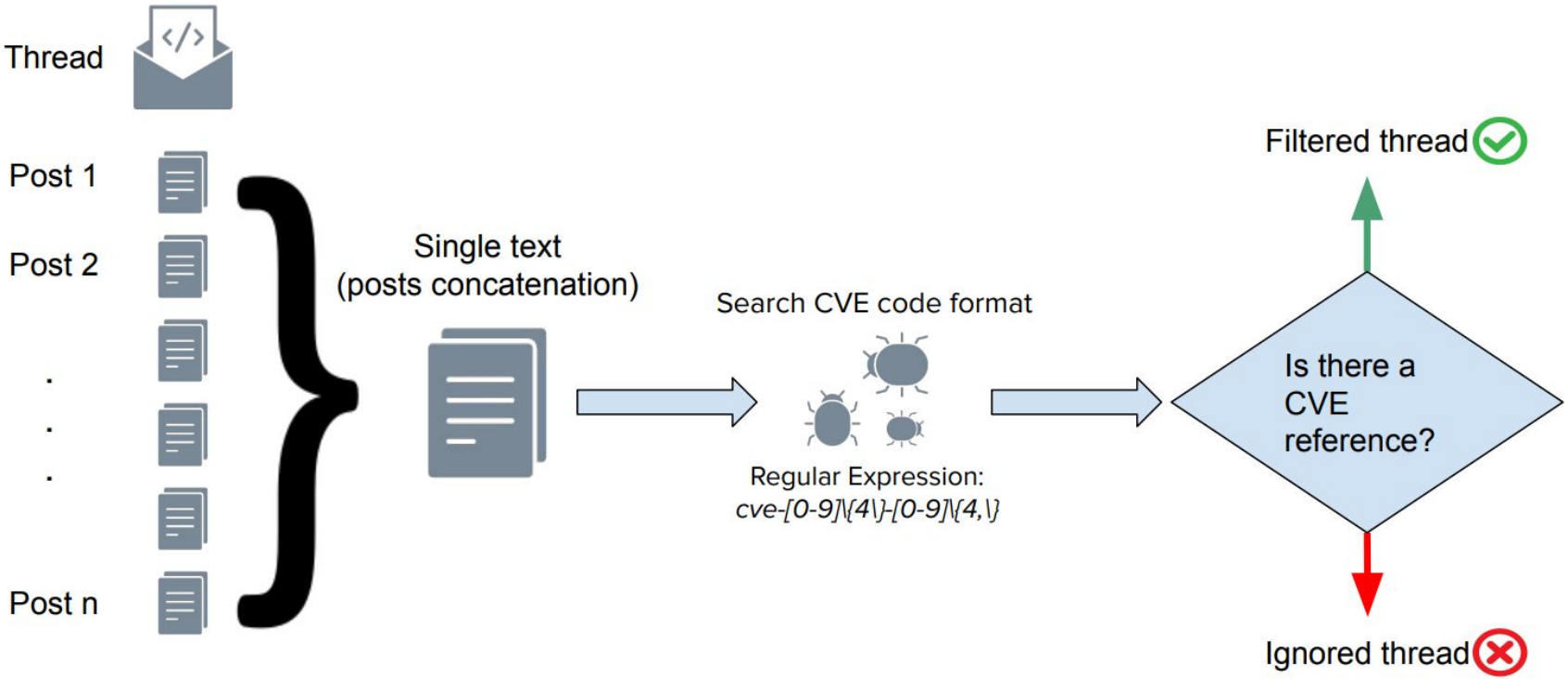


Data Pre-processing

Unreadable content



Threads and Post Concatenation



Language evaluation

- We define an Indicator Language function (ilf) as:

$$\mathbb{1}_{ilf}(word, language) = \begin{cases} 1, & \text{if word belongs to language} \\ 0, & \text{otherwise} \end{cases}$$

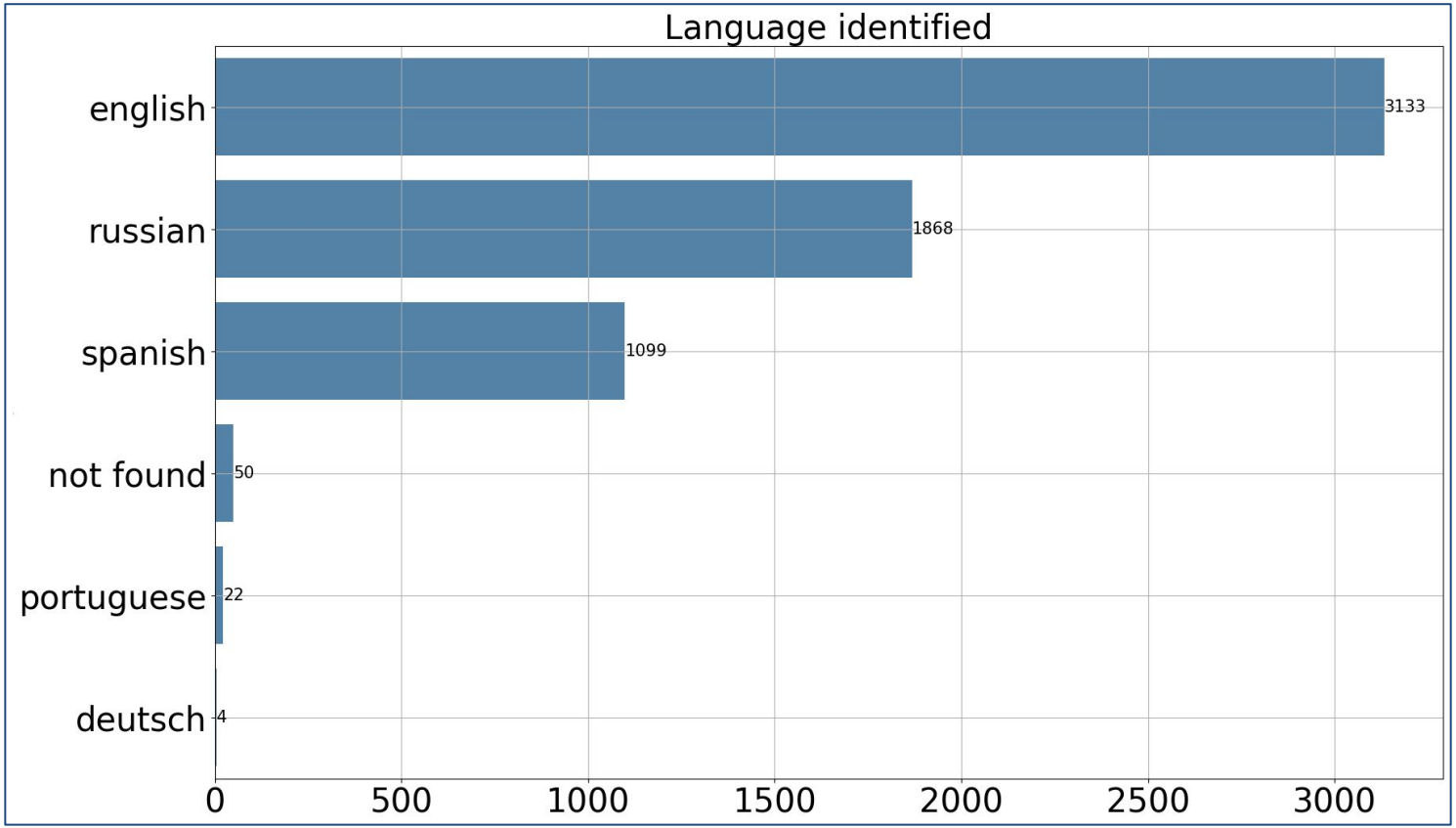
- We define a Language Ratio Function (lrf) as:

$$Ratio_{lrf}(text, language_j) = \frac{1}{\text{Total words in text}} \sum_{\substack{i=1 \\ \text{word}_i \in \text{ text}}}^{n} \mathbb{1}_{ilf}(word_i, language_j)$$

- We determine which language is the most probable to be after evaluate a text as:

$$language(text) = \max_{\forall lang \in \text{languages list}} Ratio_{lrf}(text, lang)$$

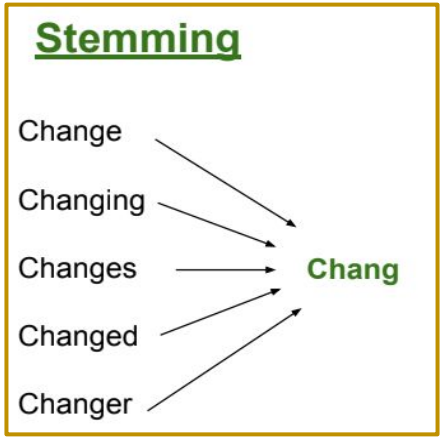
Languages identified



Text normalization

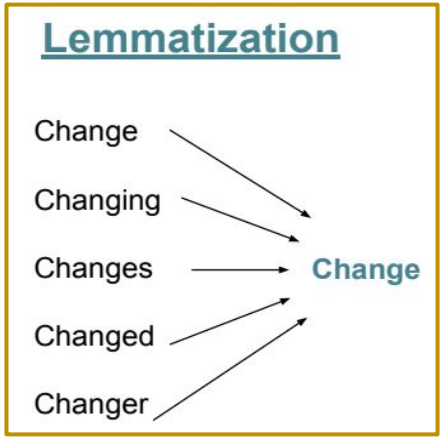
Stemming:

- Keeps the roots base form (stem) of the word.
- Content analysis without knowledge of the context.
- Simpler and faster.



Lemmatization:

- Keep the meaningful base form (lemma) of the word.
- Morphological analysis leveraging the context.
- Accurate and slower.



Text embedding

Corpus	
Document 1	I like cats
Document 2	cats are the best, they are awesome
Document 3	also dogs are nice

Document-Term Matrix

Words	I	like	cats	are	the	best	they	awesome	also	dogs	nice
Document 1	1	1	1	0	0	0	0	0	0	0	0
Document 2	0	0	1	2	1	1	1	1	0	0	0
Document 3	0	0	0	1	0	0	0	0	1	1	1

Doc2Vec

Vectors	x1	x2	x3	x4	x5	x6	x7	…
Document 1	0.35	0.86	1.82	3.48	1.05	10.15	8.63	…
Document 2	0.84	0.45	3.45	4.49	2.64	2.87	13.97	…
Document 3	0.39	1.0	0.98	7.92	5.14	6.19	20.98	…

Bag-Of-Words (1-2-gram)

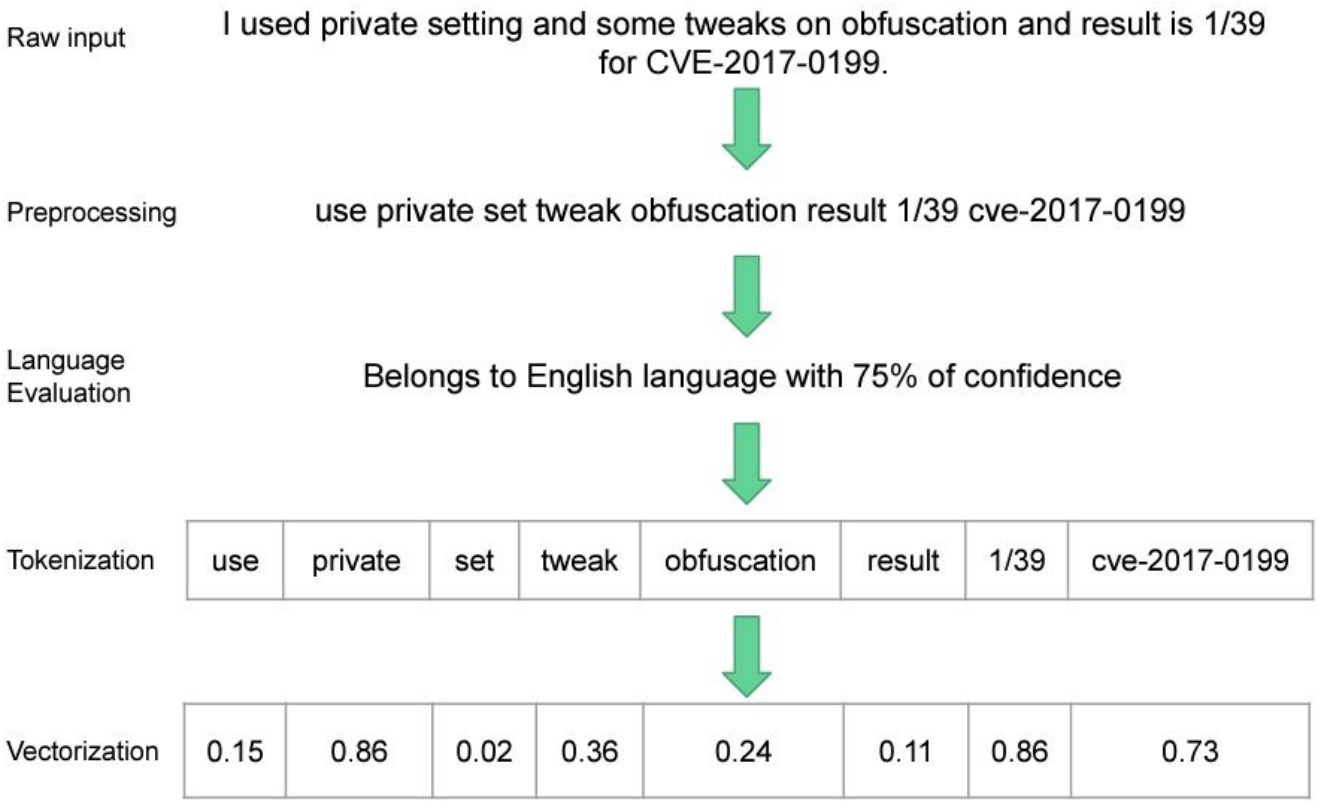
Words	I	like	cats	I like	like cats	are	the	best	they	awesome	cats are	the best	they are	…
Document 1	1	1	1	1	1	0	0	0	0	0	0	0	0	…
Document 2	0	0	1	0	0	2	1	1	1	1	1	1	1	…
Document 3	0	0	0	0	0	1	0	0	0	0	0	0	0	…

TF-IDF (1-2-gram)

Words	I	like	cats	I like	like cats	are	the	best	they	awesome	cats are	the best	they are	…
Document 1	0	0	0.47	0.62	0.62	0	0	0	0	0	0	0	0	…
Document 2	0	0	0.33	0	0	0.56	0.43	0	0	0	0.43	0.43	0.13	…
Document 3	0	0	0	0	0	0.35	0	0	0	0	0	0	0	…

We also apply standard NLP pre-processing techniques, e.g., filtering stopwords and punctuation

Text processing pipeline



Model Configurations

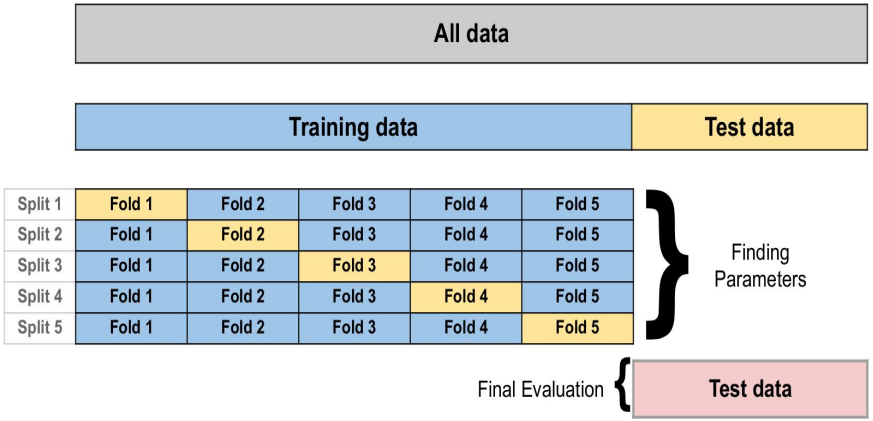
Configurations

- **For BoW and TF-IDF:**
  - top 30,000 most frequently occurring words
  - A word should appear at least 5 times
  - Appear in at least 90% of the posts in the corpus are considered for analysis
- **Doc2Vec:**
  - encode threads into 5000-dimensional vectors
- **RandomForest:**
  - Regularization parameter, learning rate.
  - Tree depth, number of features to consider at each tree split
  - Minimum samples required to split an internal node
  - Maximum node degree.

Data split

- Oversampling method to balance classes and split data into 75% and 25%, respectively

- Hyperparameters tuning: Grid search using Stratified 5 Cross-Validation



Metrics

- Accuracy — What percent of the data were predicted correct?
- Precision — What percent of your predictions were correct?
- Recall — What percent of the positive cases did you catch?
- F1 score — What percent of positive predictions were correct?

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}$$

$$Recall = \frac{T_P}{T_P + F_N} \qquad Precision = \frac{T_P}{T_P + F_P}$$

$$F1_{score} = 2\frac{Precision * Recall}{Precision + Recall}$$

Experiments and Results

Publications

- **Beneath the Cream: Unveiling Relevant Information Points from CrimeBB with Its Ground Truth Labels**
  Felipe Moreno-Vera, Menasché, D.S., and Cabral Lima.
  In International Symposium on Cyber Security, Cryptology and Machine Learning (CSCML), 2024.

- **Cream Skimming the Underground: Identifying Relevant Information Points from Online Forums**
  Felipe Moreno-Vera, Mateus Nogueira, Cainã Figueiredo, Daniel S. Menasché, Miguel Bicudo, Ashton Woiwood, Enrico Lovat, Anton Kocheturov, and Leandro Pfleger de Aguiar.
  In IEEE International Conference on Cyber Security and Resilience (IEEE CSR), 2023.

- **Inferring Discussion Topics about Exploitation of Vulnerabilities from Underground Hacking Forums**
  Felipe Moreno-Vera.
  In IEEE International Conference on Information and Communication Technology Convergence (IEEE ICTC), 2023.

Identifying Relevant Information Points from Online Forums

Dataset

Board	Number of posts (threads) citing CVEs							
	Posts tagged as							
	PoC		Weapon		Exploit		All posts	
Pentesting and Forensics	271	(55)	210	(57)	11	(3)	557	(166)
Premium Tools and Programs	198	(1)	28	(3)	142	(4)	433	(20)
Website and Forum Hacking	93	(34)	139	(43)	16	(12)	333	(132)
Hacking Tools and Programs	10	(7)	57	(28)	174	(7)	260	(59)
Premium Sellers Section	–	–	81	(28)	89	(26)	210	(66)
Beginner Hacking	86	(43)	58	(47)	6	(6)	219	(143)
Botnets, IRC, and Zombies	24	(4)	85	(34)	22	(5)	160	(62)
Hacking Tutorials	58	(21)	8	(4)	3	(3)	74	(33)
Secondary Sellers Market	8	(4)	33	(21)	–	–	91	(40)
News and Happenings	9	(9)	11	(5)	1	(1)	75	(54)
Total, all boards	757	(244)	710	(397)	464	(102)	3,037	(1,162)

Decision Tree

	Text encoding	Target classes	Accuracy	Precision	Recall	F1
DT	BoW	PoC, Weaponization, Exploitation	0.71	0.71	0.72	0.70
DT	TF-IDF	PoC, Weaponization, Exploitation	0.73	0.73	0.74	0.72
DT	doc2vec	PoC, Weaponization, Exploitation	0.74	0.74	0.74	0.73
DT	BoW	Exploitation vs Non-exploitation	0.85	0.86	0.85	0.85
DT	TF-IDF	Exploitation vs Non-exploitation	0.91	0.91	0.91	0.91
DT	doc2vec	Exploitation vs Non-exploitation	**0.92**	**0.93**	**0.92**	**0.92**
DT	BoW	PoC vs Non-PoC	0.75	0.75	0.75	0.75
DT	TF-IDF	PoC vs Non-PoC	0.77	0.78	0.77	0.77
DT	doc2vec	PoC vs Non-PoC	0.70	0.71	0.70	0.70
DT	BoW	Weaponization vs Non-weapon.	0.68	0.68	0.68	0.68
DT	TF-IDF	Weaponization vs Non-weapon.	0.63	0.64	0.63	0.62
DT	doc2vec	Weaponization vs Non-weapon.	0.59	0.59	0.59	0.59

Decision Tree

	Text encoding	Target classes	Accuracy	Precision	Recall	F1
DT	BoW	PoC, Weaponization, Exploitation	0.71	0.71	0.72	0.70
DT	TF-IDF	PoC, Weaponization, Exploitation	0.73	0.73	0.74	0.72
DT	doc2vec	PoC, Weaponization, Exploitation	0.74	0.74	0.74	0.73
DT	BoW	Exploitation vs Non-exploitation	0.85	0.86	0.85	0.85
DT	TF-IDF	Exploitation vs Non-exploitation	0.91	0.91	0.91	0.91
DT	doc2vec	Exploitation vs Non-exploitation	**0.92**	**0.93**	**0.92**	**0.92**
DT	BoW	PoC vs Non-PoC	0.75	0.75	0.75	0.75
DT	TF-IDF	PoC vs Non-PoC	0.77	0.78	0.77	0.77
DT	doc2vec	PoC vs Non-PoC	0.70	0.71	0.70	0.70
DT	BoW	Weaponization vs Non-weapon.	0.68	0.68	0.68	0.68
DT	TF-IDF	Weaponization vs Non-weapon.	0.63	0.64	0.63	0.62
DT	doc2vec	Weaponization vs Non-weapon.	0.59	0.59	0.59	0.59

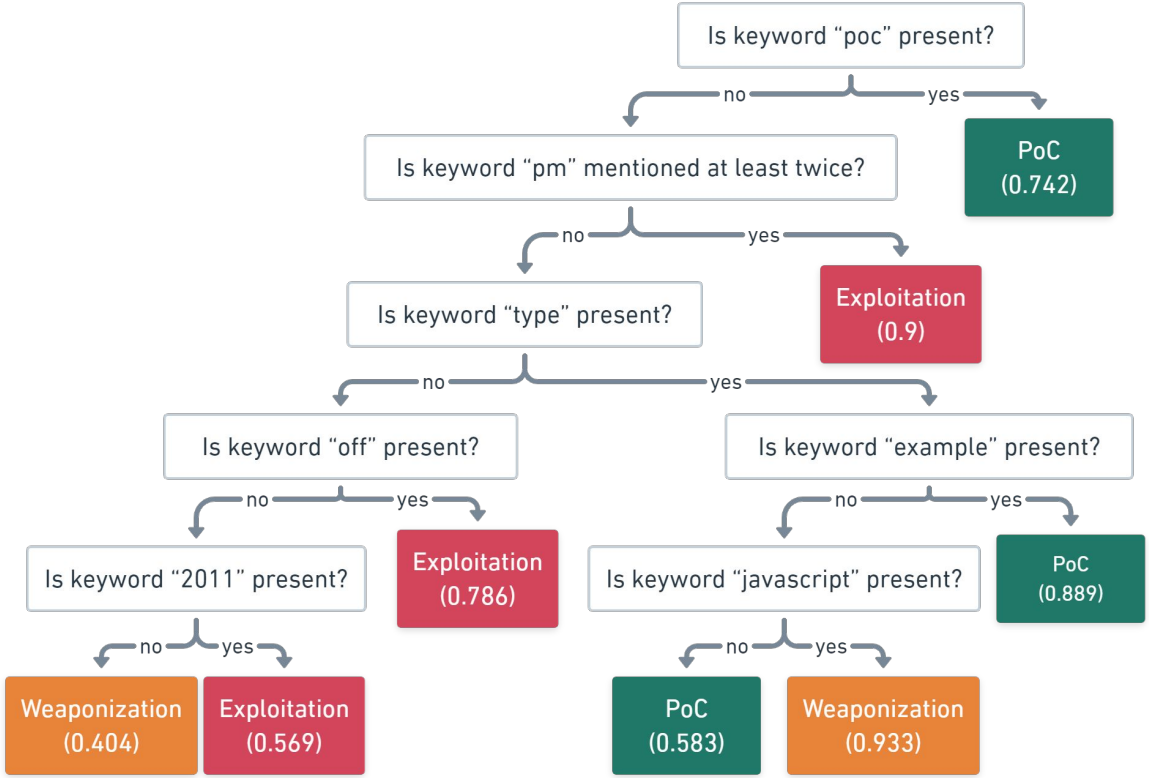
Decision tree performance: easier to distinguish exploitation from rest

Decision Tree

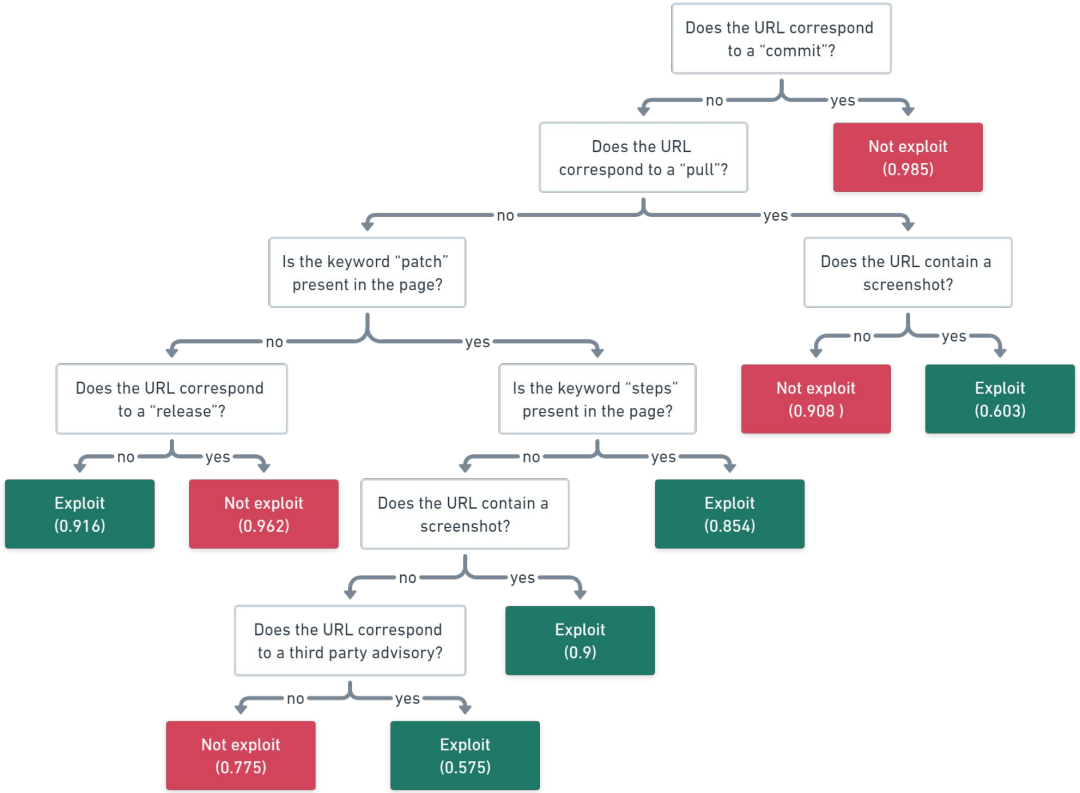
	Text encoding	Target classes	Accuracy	Precision	Recall	F1
DT	BoW	PoC, Weaponization, Exploitation	0.71	0.71	0.72	0.70
DT	TF-IDF	PoC, Weaponization, Exploitation	0.73	0.73	0.74	0.72
DT	doc2vec	PoC, Weaponization, Exploitation	0.74	0.74	0.74	0.73
DT	BoW	Exploitation vs Non-exploitation	0.85	0.86	0.85	0.85
DT	TF-IDF	Exploitation vs Non-exploitation	0.91	0.91	0.91	0.91
DT	doc2vec	Exploitation vs Non-exploitation	**0.92**	**0.93**	**0.92**	**0.92**
DT	BoW	PoC vs Non-PoC	0.75	0.75	0.75	0.75
DT	TF-IDF	PoC vs Non-PoC	0.77	0.78	0.77	0.77
DT	doc2vec	PoC vs Non-PoC	0.70	0.71	0.70	0.70
DT	BoW	Weaponization vs Non-weapon.	0.68	0.68	0.68	0.68
DT	TF-IDF	Weaponization vs Non-weapon.	0.63	0.64	0.63	0.62
DT	doc2vec	Weaponization vs Non-weapon.	0.59	0.59	0.59	0.59

Decision tree performance: can't determine weaponization from rest

Decision Tree: PoC, Weaponization, and Exploitation



Decision Tree: Exploitation vs non-exploitation



Takeaways

In this experiment, we leveraged CrimeBB and machine learning methods to learn textual content and distinguish between:

- *(1) potential threat (proof of concept),*
- *(2) eminent threat (weaponization),*
- *(3) criminals chatting about a threat (exploitation in the wild).*

We found that the most cited CVEs are typically related to higher risks and that it is feasible to filter exploitation threads, with an accuracy above 99% automatically

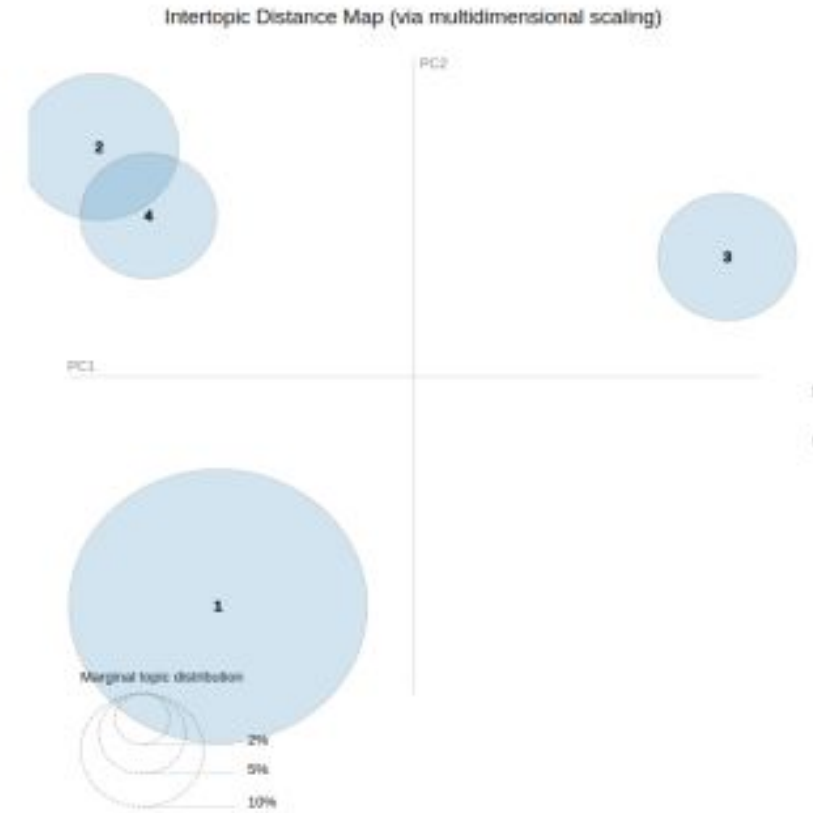
Inferring Discussion Topics about Exploitation from Online Forums

Dataset

- *We employ the labels as topics:*
  Proof of Concept (PoC),
  weaponization,
  “other” (including others, scam, warning, and help),
  exploitation,
  labeled with corresponding IDs 1, 2, 3, and 4, respectively.

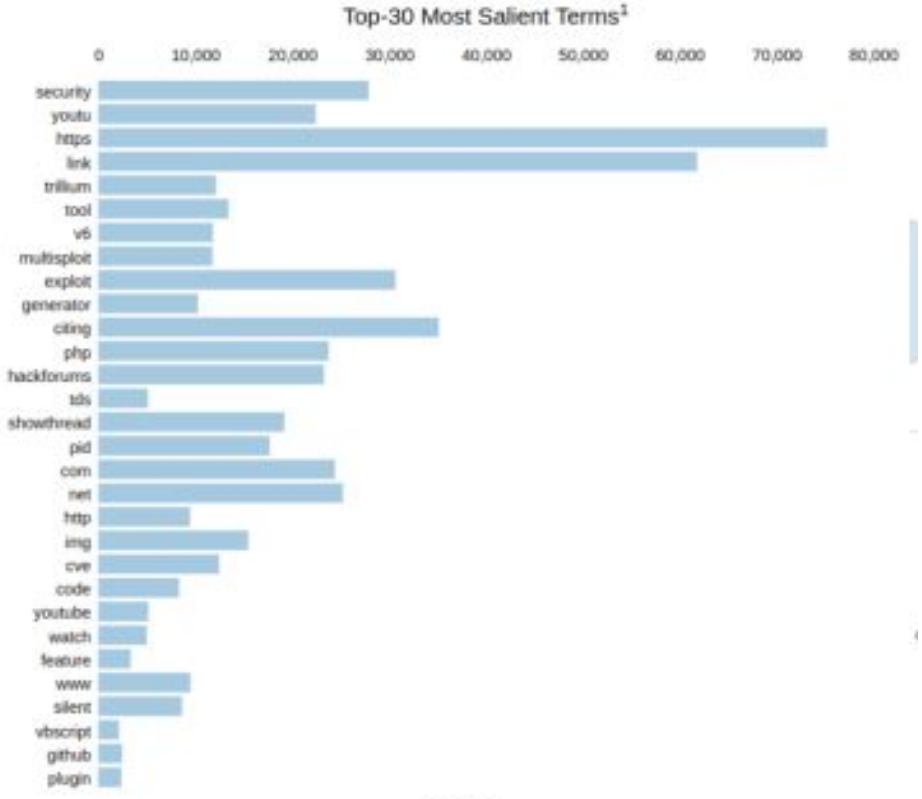
- *The goal is to identify key themes related to exploit techniques, vulnerabilities, and potential targets, providing valuable insights into the landscape of vulnerability exploitation.*

Topic Modeling

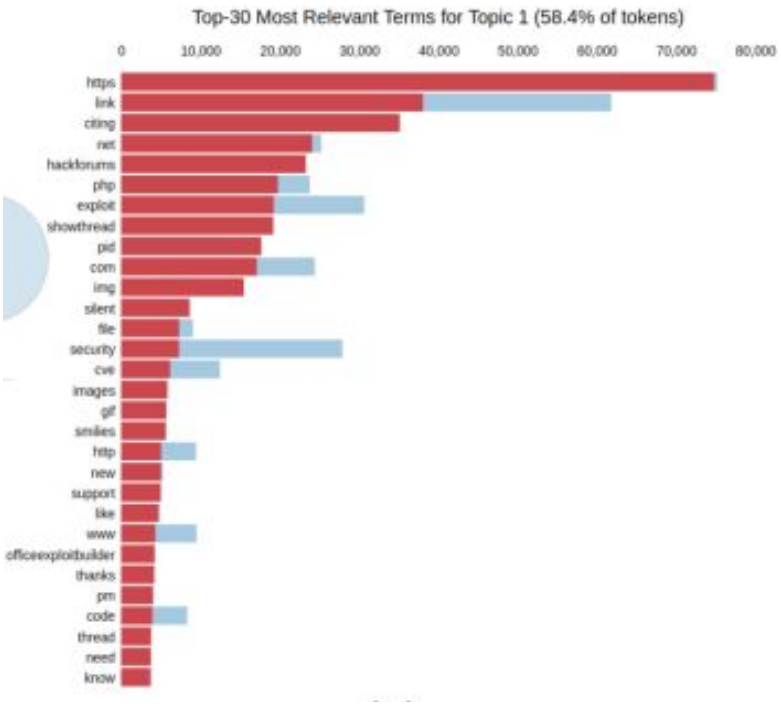


- The radius of each group determines the marginal topic distribution
- *1 - Proof of Concept (PoC),*
- *2 - weaponization,*
- *3 - “other” (including others, scam, warning, and help)*
- *4 - exploitation*

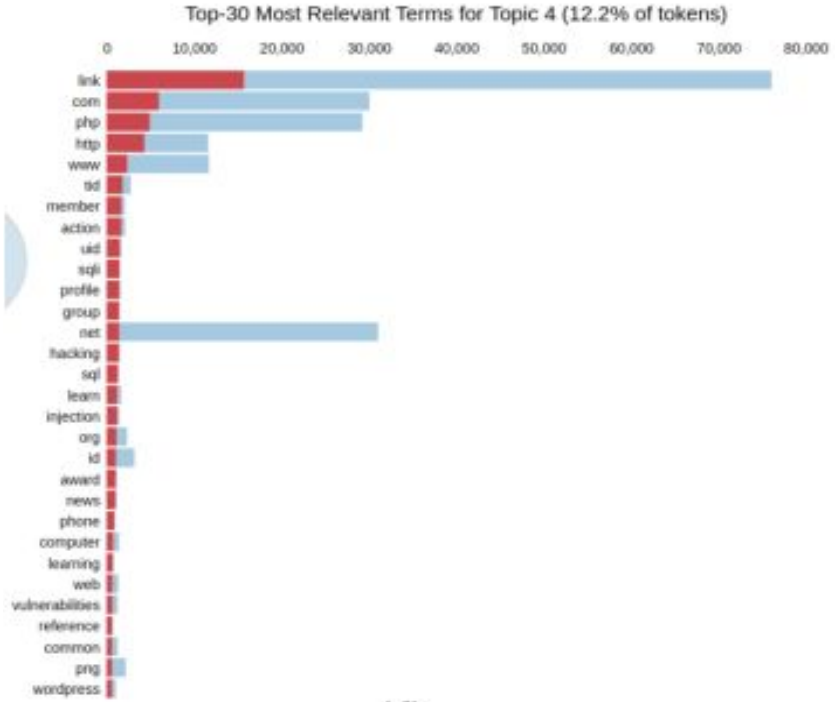
Topic Modeling: Top 30 terms



Topic Modeling

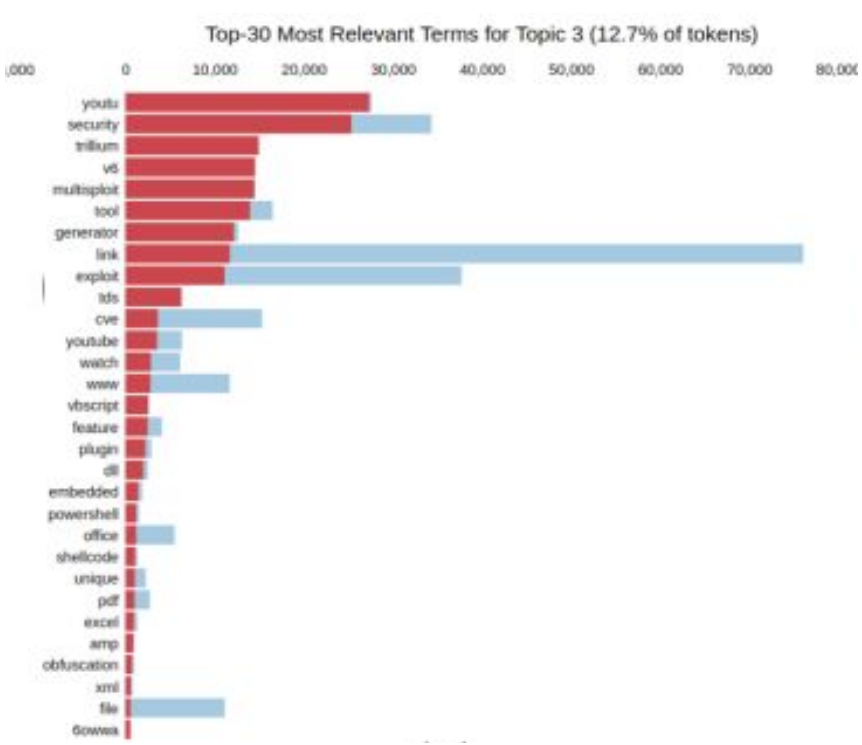


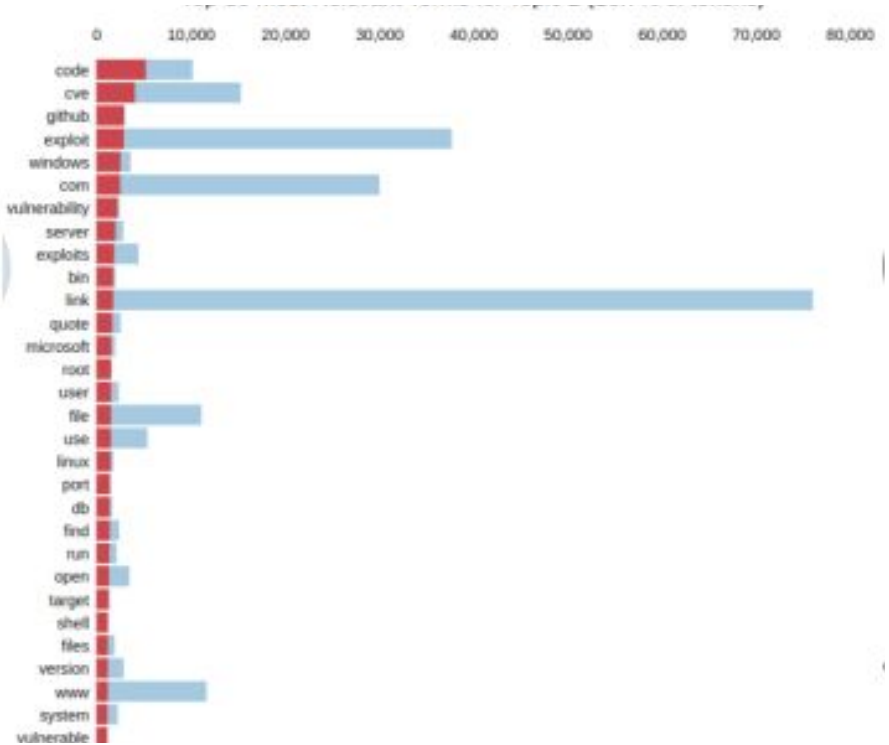
PoC



Others

Topic Modeling





Exploitation

Weaponization

Takeaways

- *We apply an unsupervised learning technique, topic modeling, to analyze the exploitation in the wild.*

- *We use experts annotations and re-categorize them into PoC, exploitation, weaponization, and others.*

- *We were able to identify separate clusters of words corresponding to each topic, we also identified relevant words such as “code”, “cve”, “exploit”, “link”, and “vulnerability” are commonly used in exploitation and weaponization discussions.*

- *We found that it is possible to identify emerging threats, helping security professionals, and researchers stay informed and prioritize their defense strategies accordingly.*

Unveiling Relevant Information Points from Online Forums

GPT Labeling

We use the GPT-4o model to labels our threads using prompts given the context and content of each thread concatenated:

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), . . . , (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$_label$", $newlabel$), ...]

GPT Labeling

- We ask to GPT to re-assign in group of categories for crime type, post type, and intent:

- **Intent:**
  - Sentiment (emotions or attitudes)
  - Expression of Interaction (ways of communicating or expressing oneself)
  - Intensity (levels of strength or forcefulness)

- **Post type:**
  - Requests (seeking information or services)
  - Offers/Exchanges (providing services or trading)
  - Communication/Interaction (forms of interaction or content type)

GPT Labeling

- We ask to GPT to re-assign in group of categories for crime type, post type, and intent:

- **Crime type:**
  - Cybercrime Activities (illegal activities related to cybercrime)
  - Cybercrime Support Services (services that facilitate cybercrime activities)
  - Non-Criminal (activities that are not considered criminal)

- **Annotations:**
  - Malicious Activity (weaponization, exploitation, and scam)
  - Support and Assistance (help)
  - Informational (poc and warning)
  - Others

GPT Labeling

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

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

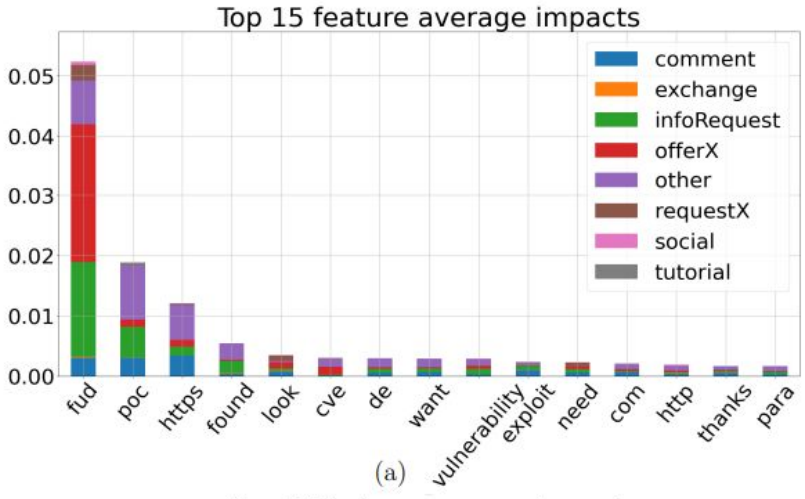
Post type labels	
Labels	Samples
Communication/Interaction	1050
Requests	1049
Other	494
Offer/exchanges	627

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

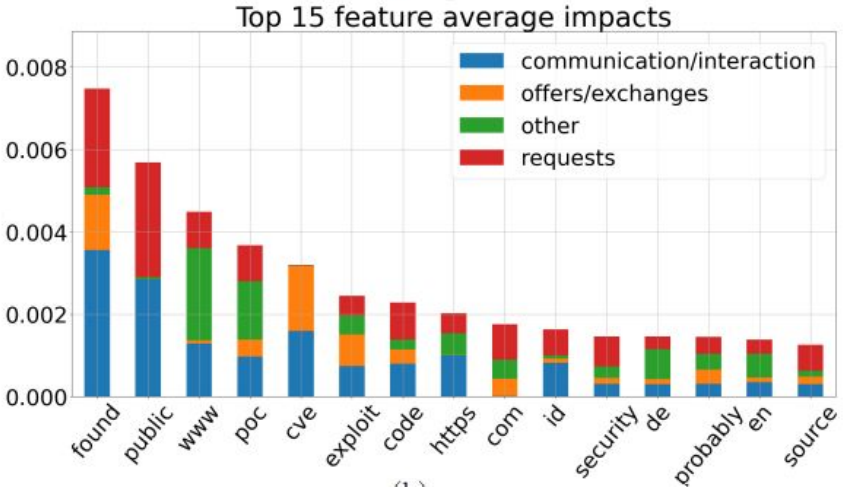
Model Results

	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 (SIU; COLLIER; HUTCHINGS, 2021)	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 (CAINES et al., 2018b)	–	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 (CAINES et al., 2018b)	–	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 (MORENO-VERA et al., 2023)	0.86	0.87	0.86	0.86

Model explanations

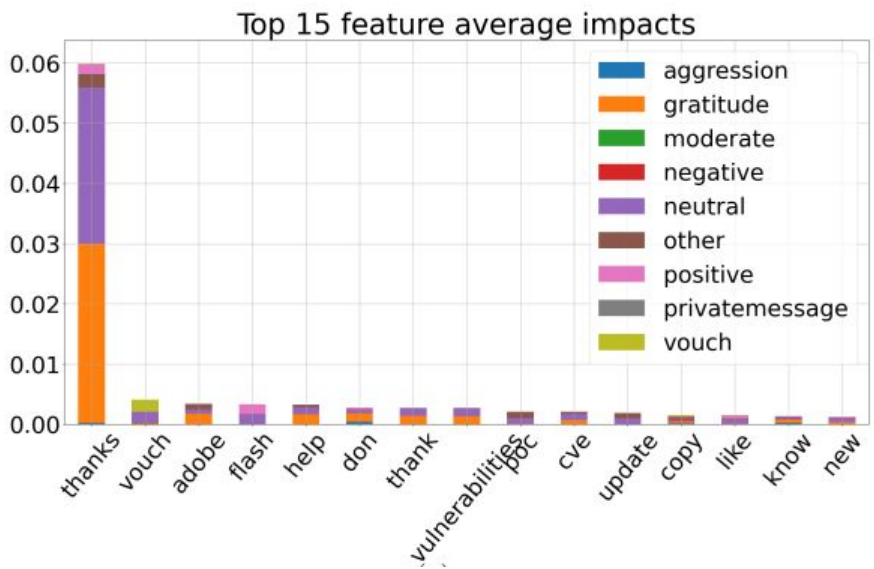


PostCog labeling - posttype

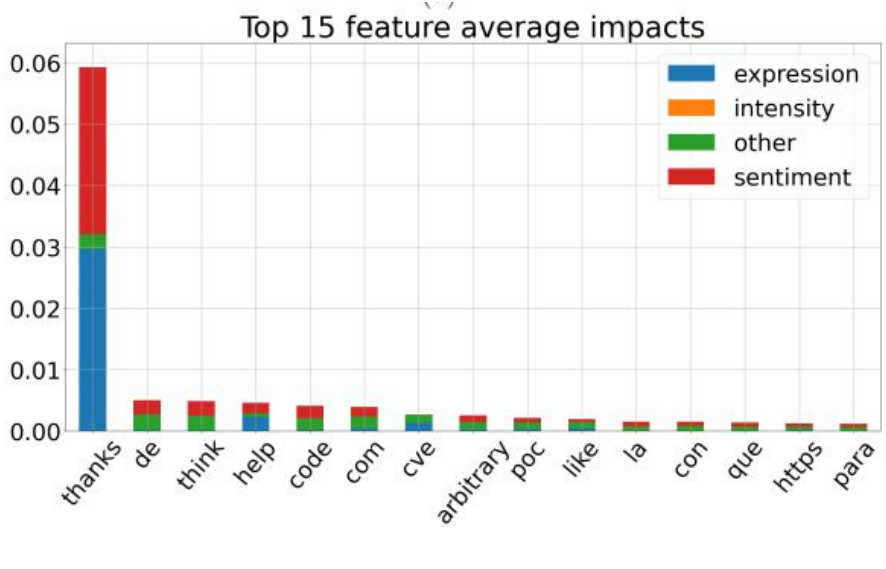


GPT labeling

Model explanations

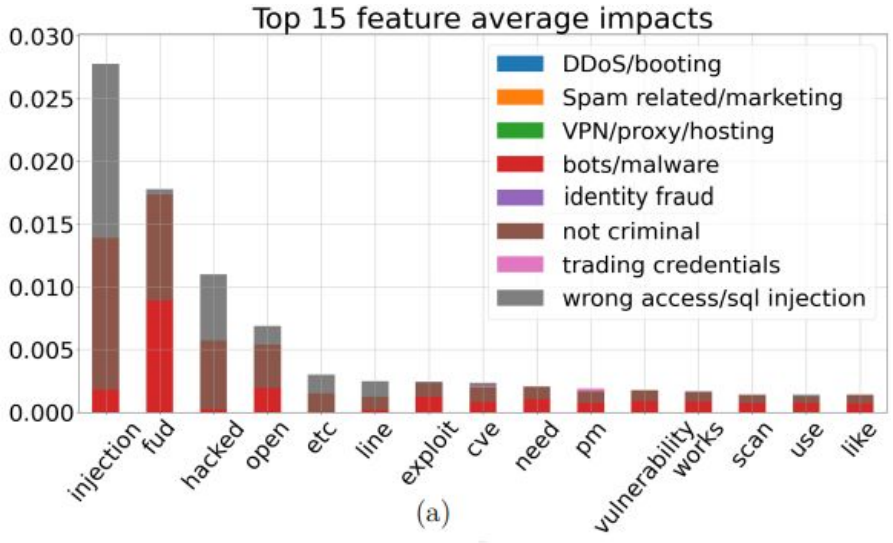


PostCog labeling - intention

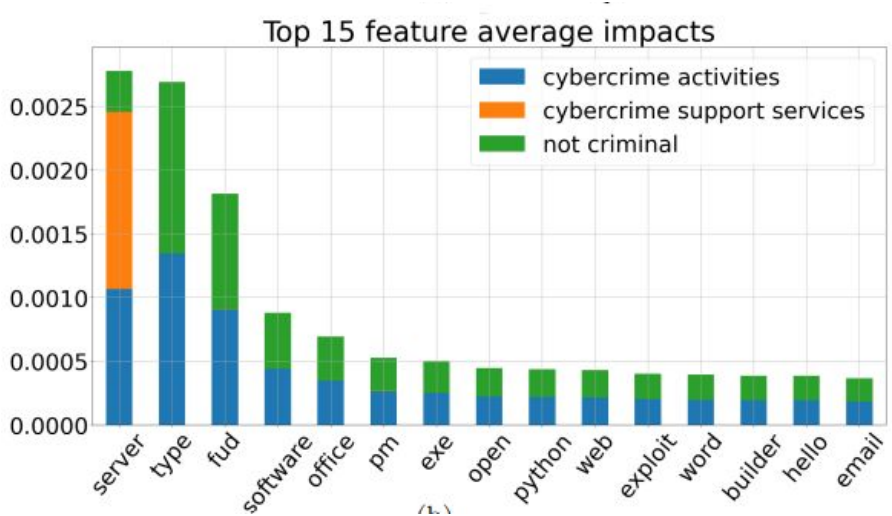


GPT labeling

Model explanations

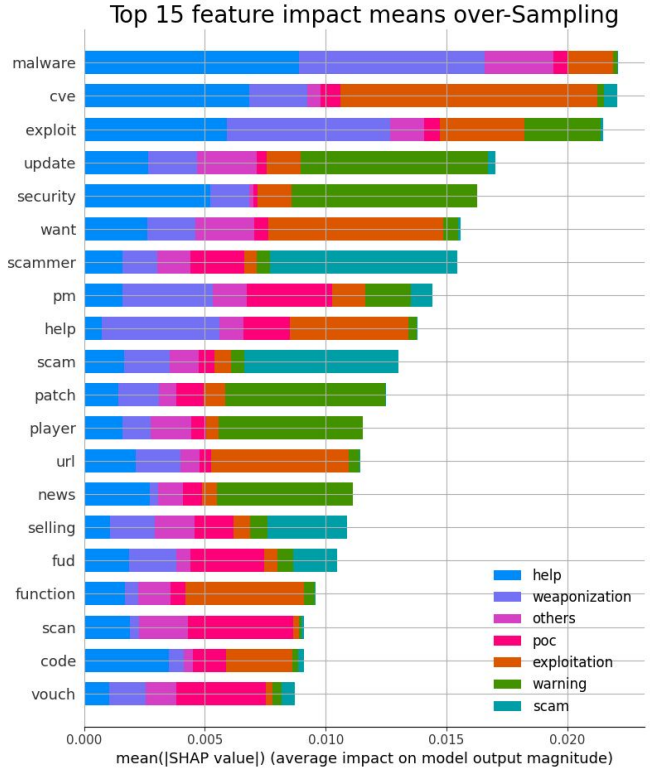


PostCog labeling - crime type

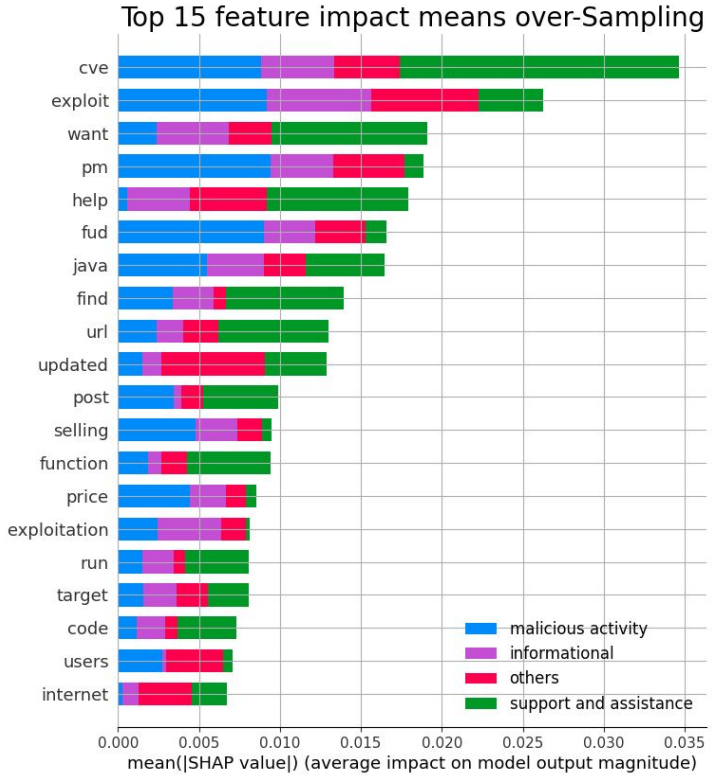


GPT labeling

Model explanations

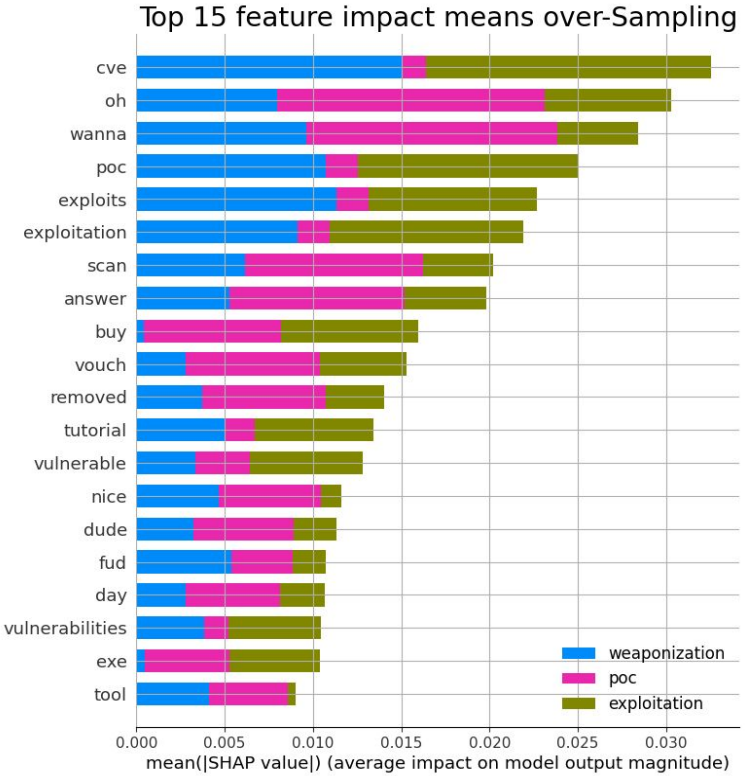


Manual labeling

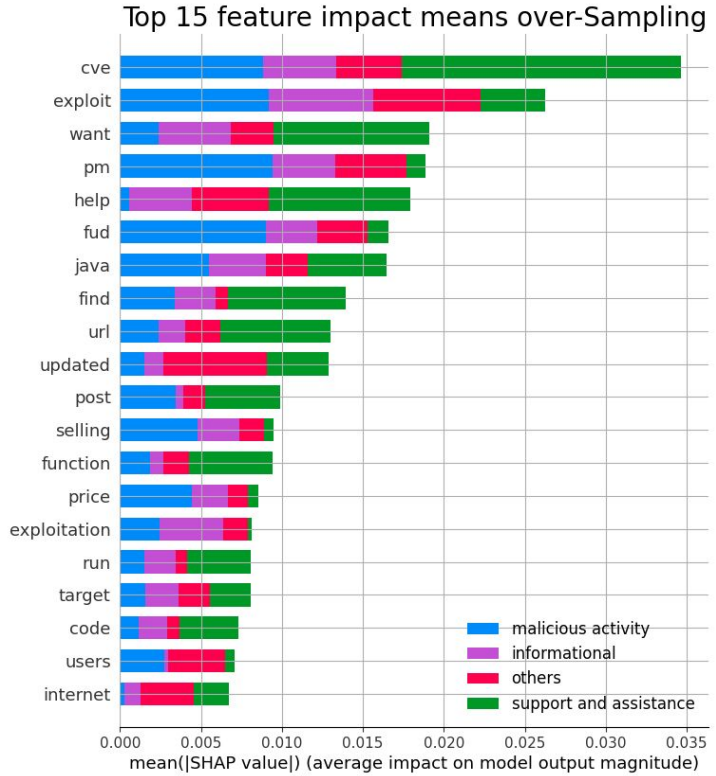


GPT labeling

Model explanations

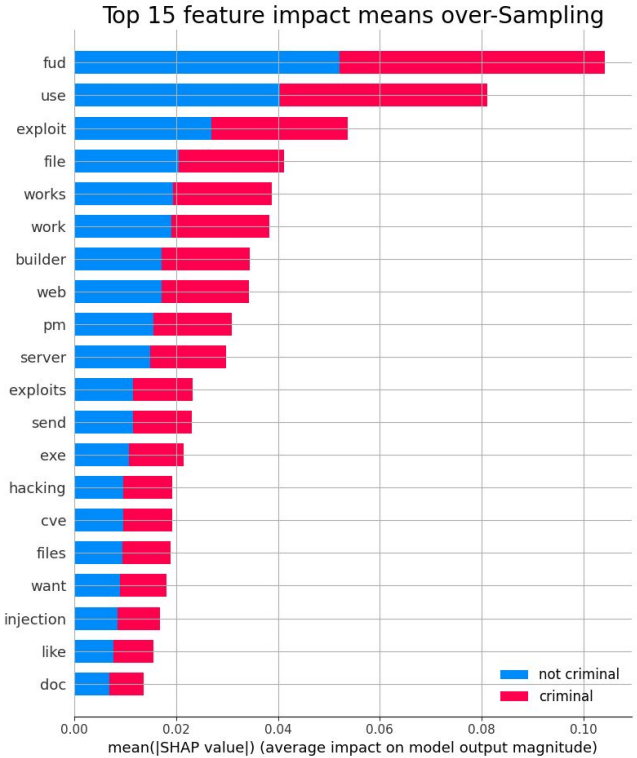


Manual labeling

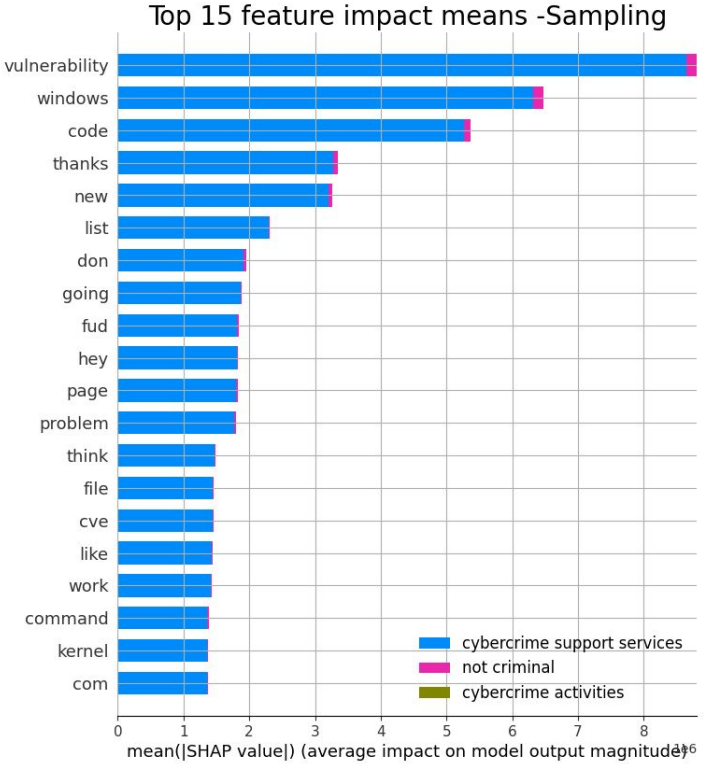


GPT labeling

Model explanations



Manual labeling



GPT labeling

Takeaways

- *We classified underground forum posts by cybercriminal activity, textual intentions, and content type using TF-IDF, and a RandomForest classifier.*

- *We used the SHAP method for model explanation analysis.*

- *We observed that ChatGPT introduced noisy labels in several instances. Its tendency to generate plausible but not always accurate or relevant information can lead to misleading or irrelevant features in the training data.*

- *We observe a little degrade classifier performance by skewing feature importance or introducing bias. Thus, while ChatGPT can be a valuable tool, it is crucial to use it carefully and to complement it with rigorous data validation and preprocessing steps to mitigate potential drawbacks.*

Conclusions

Conclusions

- We were able to analyze and join pertinent annotations related to **CVEs thread-posts**

- It is feasible to **train** a classifier to **infer** the **maturity level and type of threads.**

- **Black-box random forests** help in understanding word relevance.
  - It performs better than decision trees, SVM, ridge regression, booster models, etc.
  - We won't be able to use complex architectures, such as transformers, due to limited computational resources.

- It has **high performance** in distinguishing categories, it is possible to **understand** predictions using explanation methods such as SHAP.

Thanks! Any questions?

felipe.moreno@ppgi.ufrj.br

THANKS!