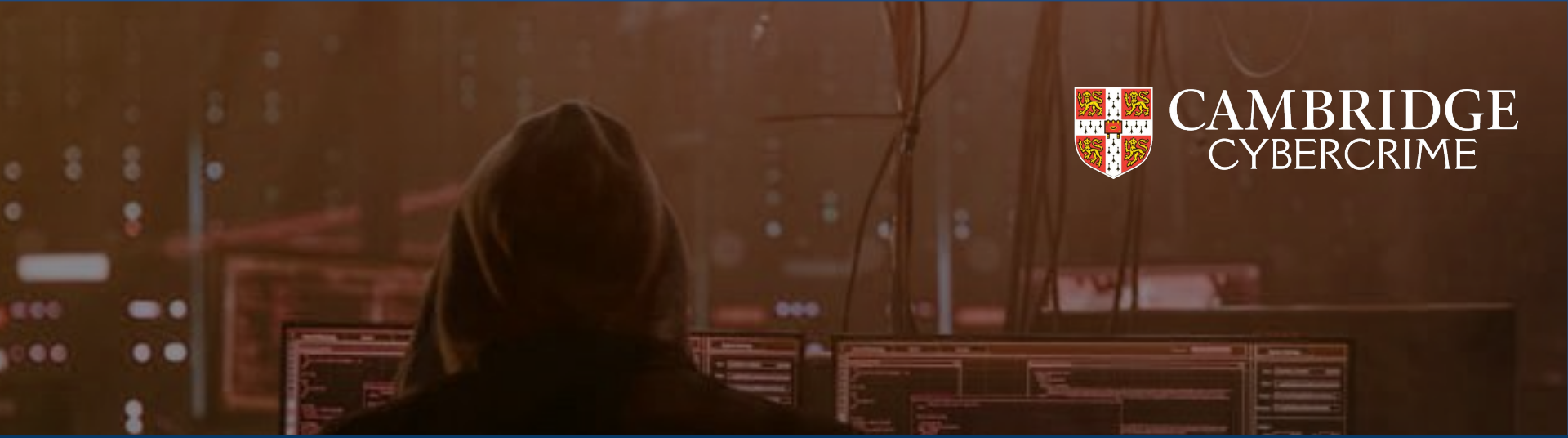
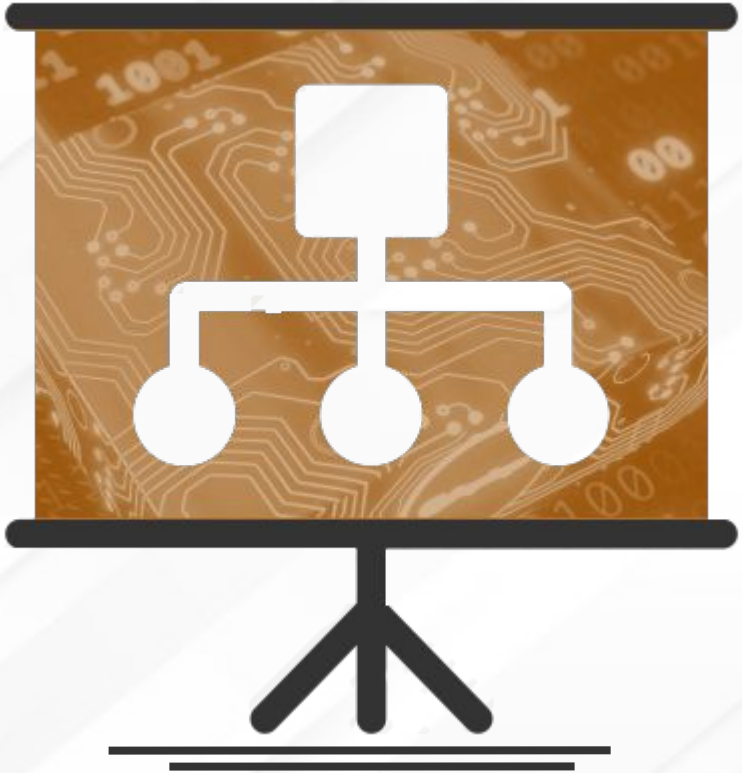


CAMBRIDGE CYBERCRIME

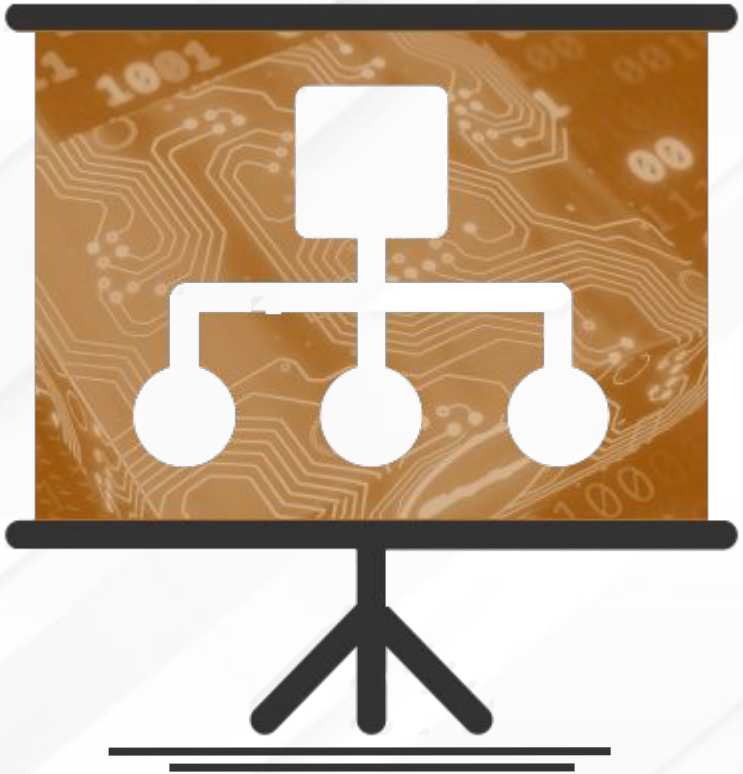
Beneath the Cream

Unveiling Relevant Information Points from CrimeBB with Its Ground Truth Labels

Felipe Moreno-Vera, Cabral Lima, and Daniel S. Menasché



1. Introduction
2. Dataset
3. Beneath the Cream Methodology
4. Experimental Results
5. Conclusion



1. Introduction
2. Dataset
3. Beneath the Cream Methodology
4. Experimental Results
5. Conclusion

Motivation

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.

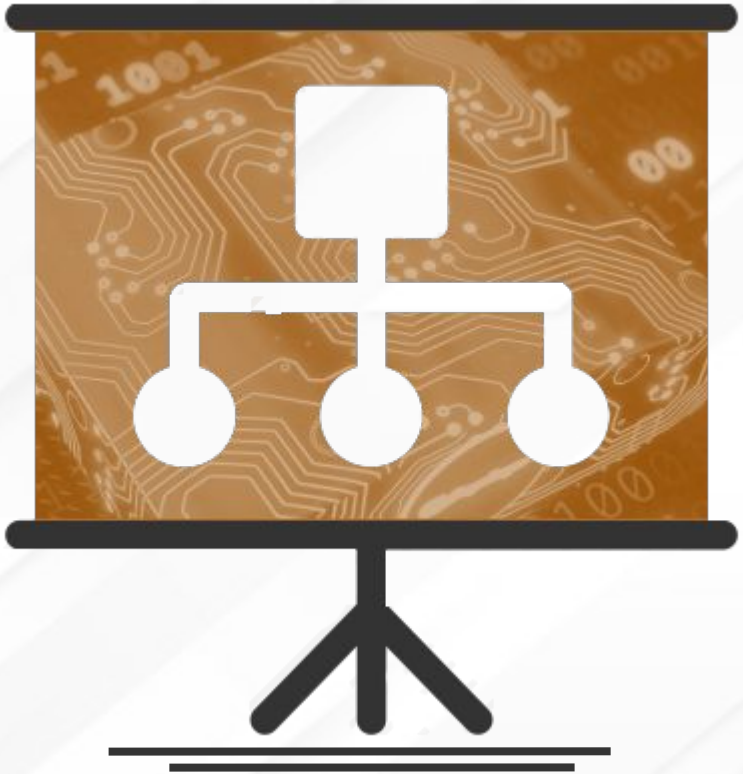
There is a lack of effective methods to process discussions about threats and identify potential exploitation in underground forums.

Context

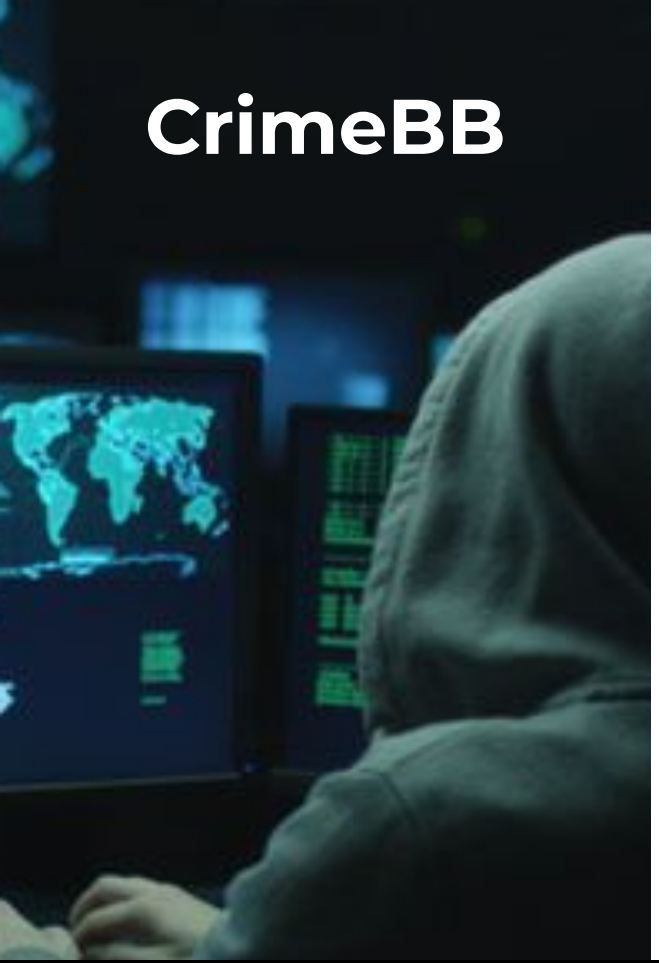
Developing methods to analyze these forums can help predict and prevent cyber attacks, safeguarding critical systems and data.

Analyzing these forums allows for tracking

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



1. Introduction
2. Dataset
3. Beneath the Cream Methodology
4. Experimental Results
5. Conclusion



CrimeBB

Dataset Description

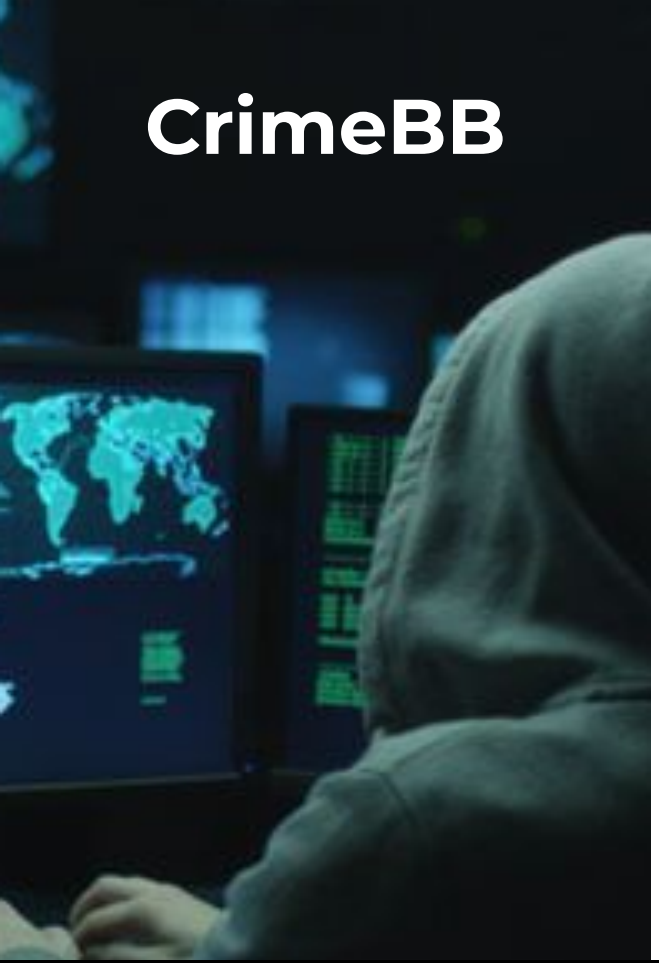
Made available by Cambridge Cybercrime Centre

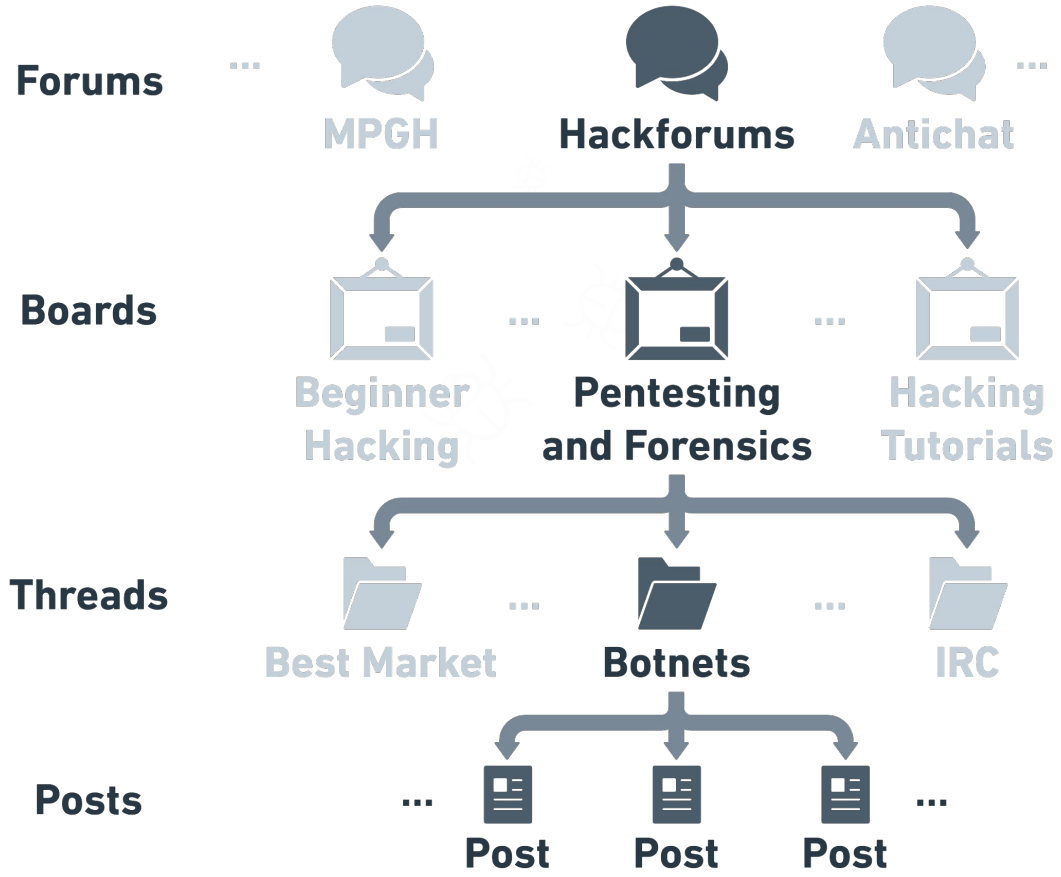
Contains data scraped from multiple underground forums (37 studied)

Organized in forums, boards, threads and posts

Provide about 45,2 GB of textual information.

CrimeBB




Forums
MPGH
Hackforums
Antichat
Boards
Beginner Hacking
Pentesting and Forensics
Hacking Tutorials
Threads
Best Market
Botnets
IRC
Posts
Post
Post
Post

CrimeBB

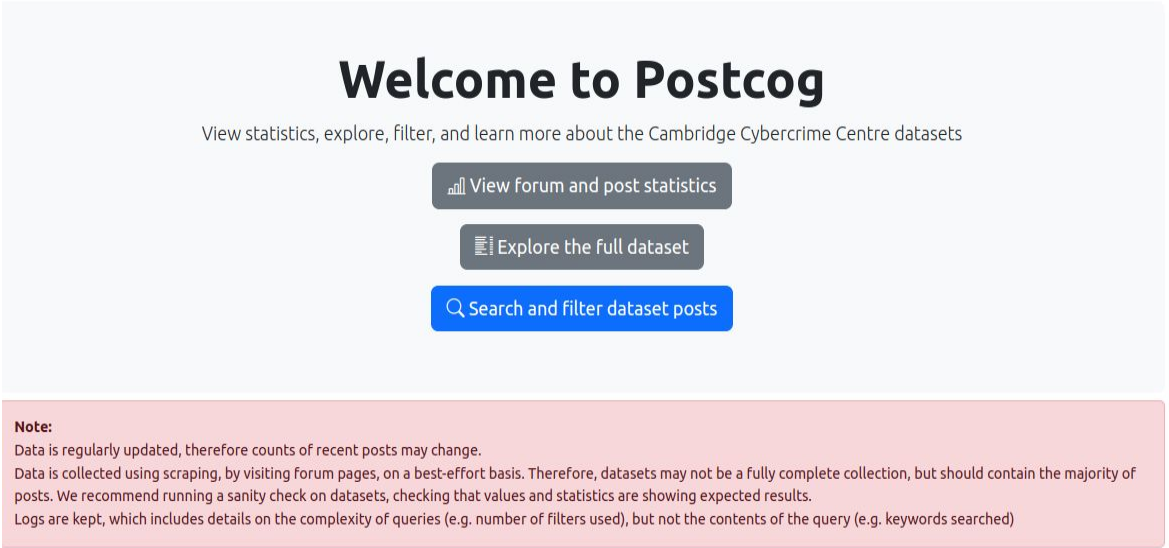
As of August 28, 2024, CrimeBB have:

- **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
Underc0de	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		

PostCog

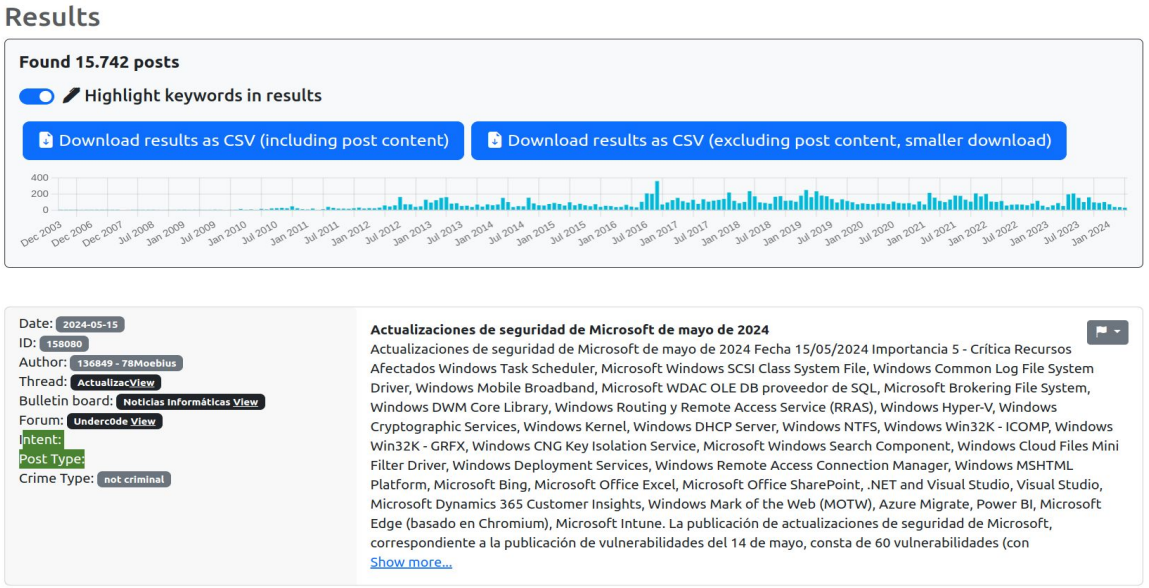
PostCog, a framework to navigate through CrimeBB data.

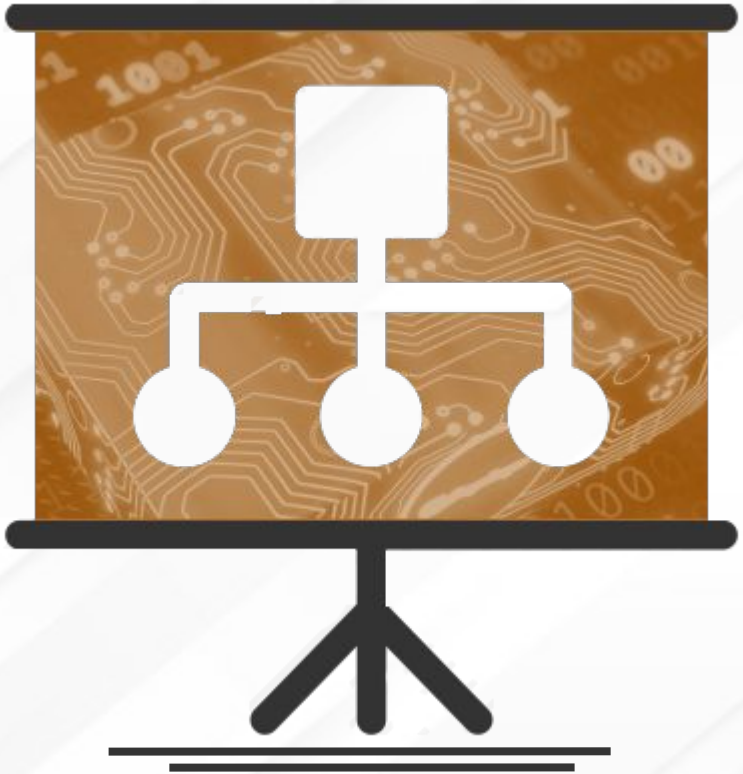


PostCog - Labels

- 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**.

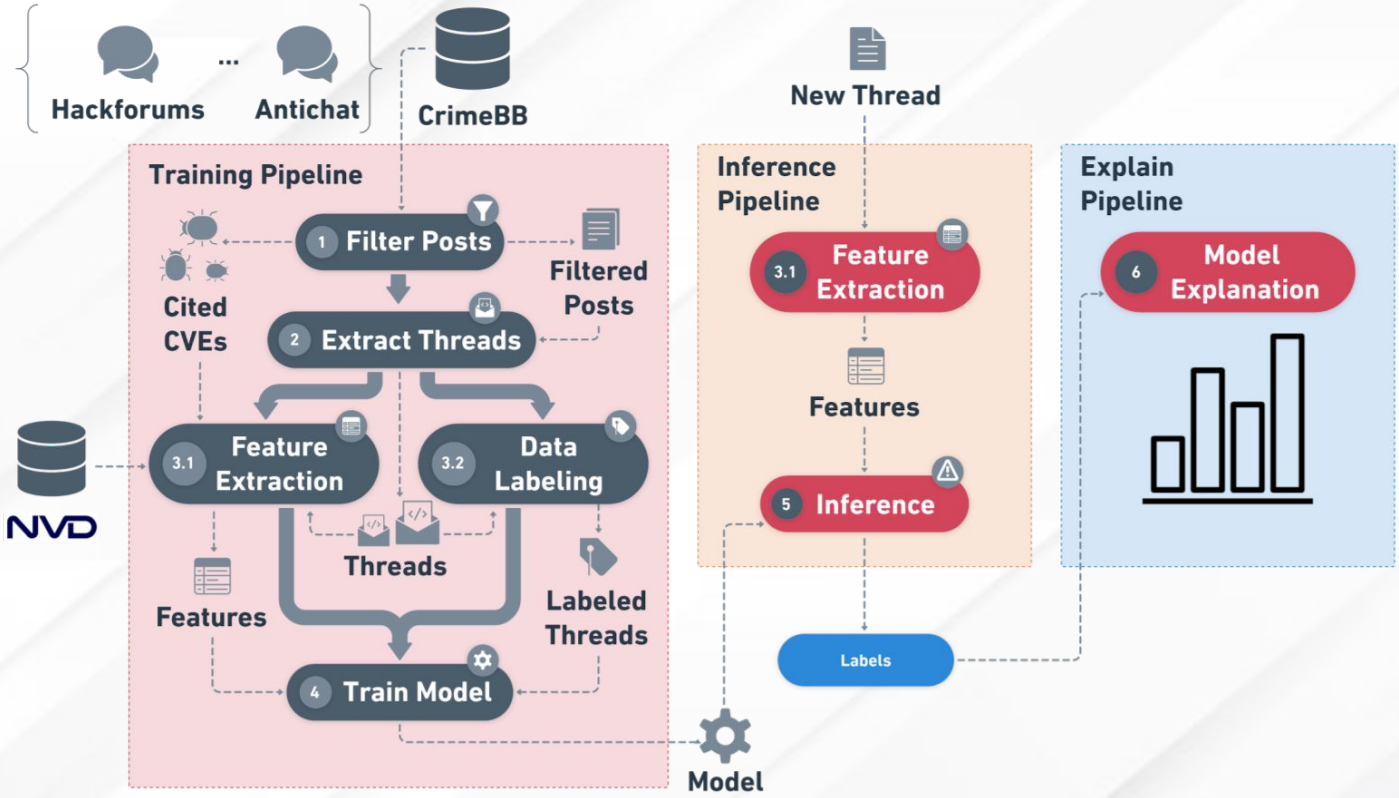
- We identify that only post scrapped from **HackForums** has **crime type**, **post type**, and **intent** tags included.



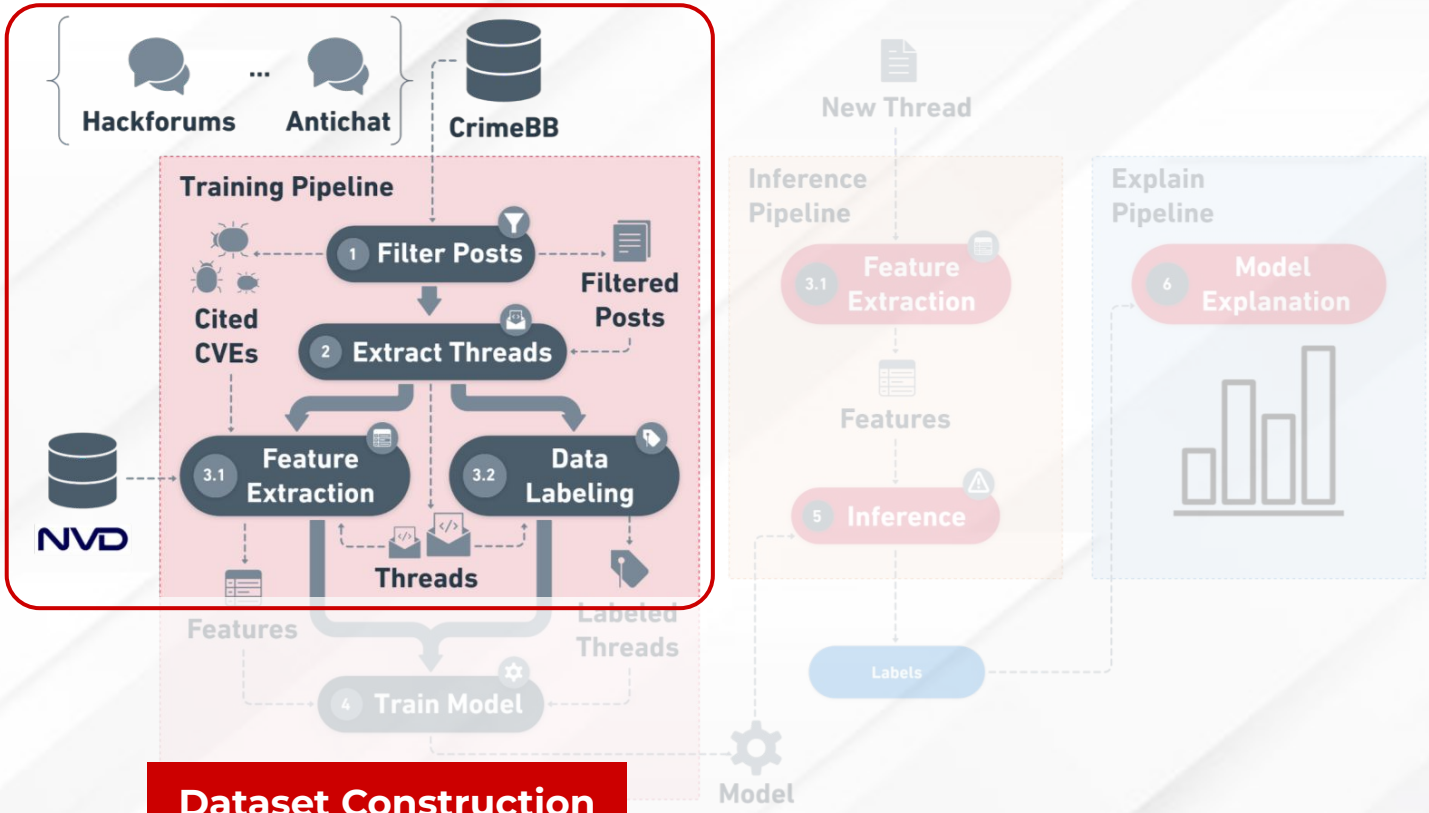


1. Introduction
2. Dataset
3. Beneath the Cream Methodology
4. Experimental Results
5. Conclusion

Pipeline

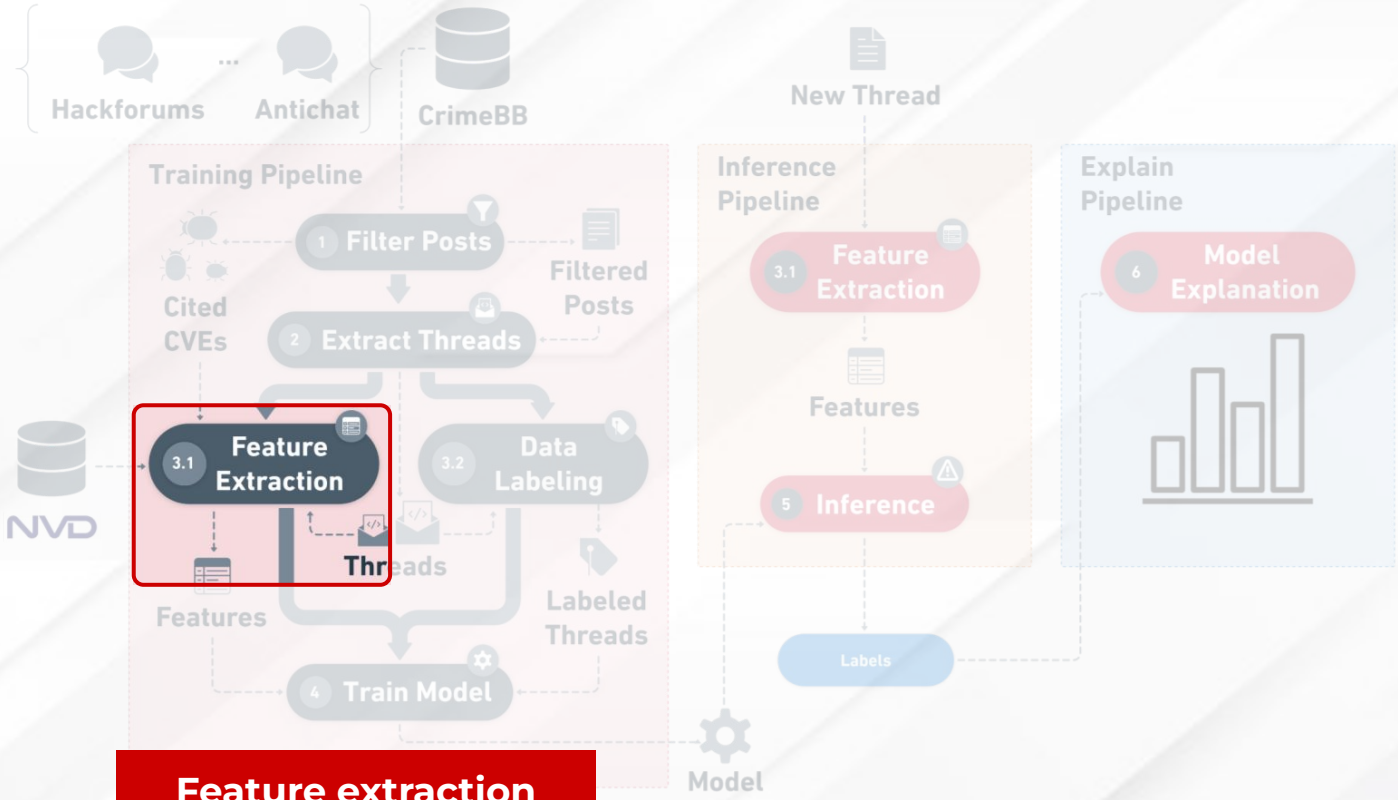

Hackforums
...
Antichat
CrimeBB
New Thread
Training Pipeline
Inference Pipeline
Explain Pipeline
1 Filter Posts
Filtered Posts
Cited CVEs
2 Extract Threads
3.1 Feature Extraction
3.2 Data Labeling
NVD
Threads
Features
Labeled Threads
4 Train Model
Model
3.1 Feature Extraction
Features
5 Inference
Labels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

Pipeline



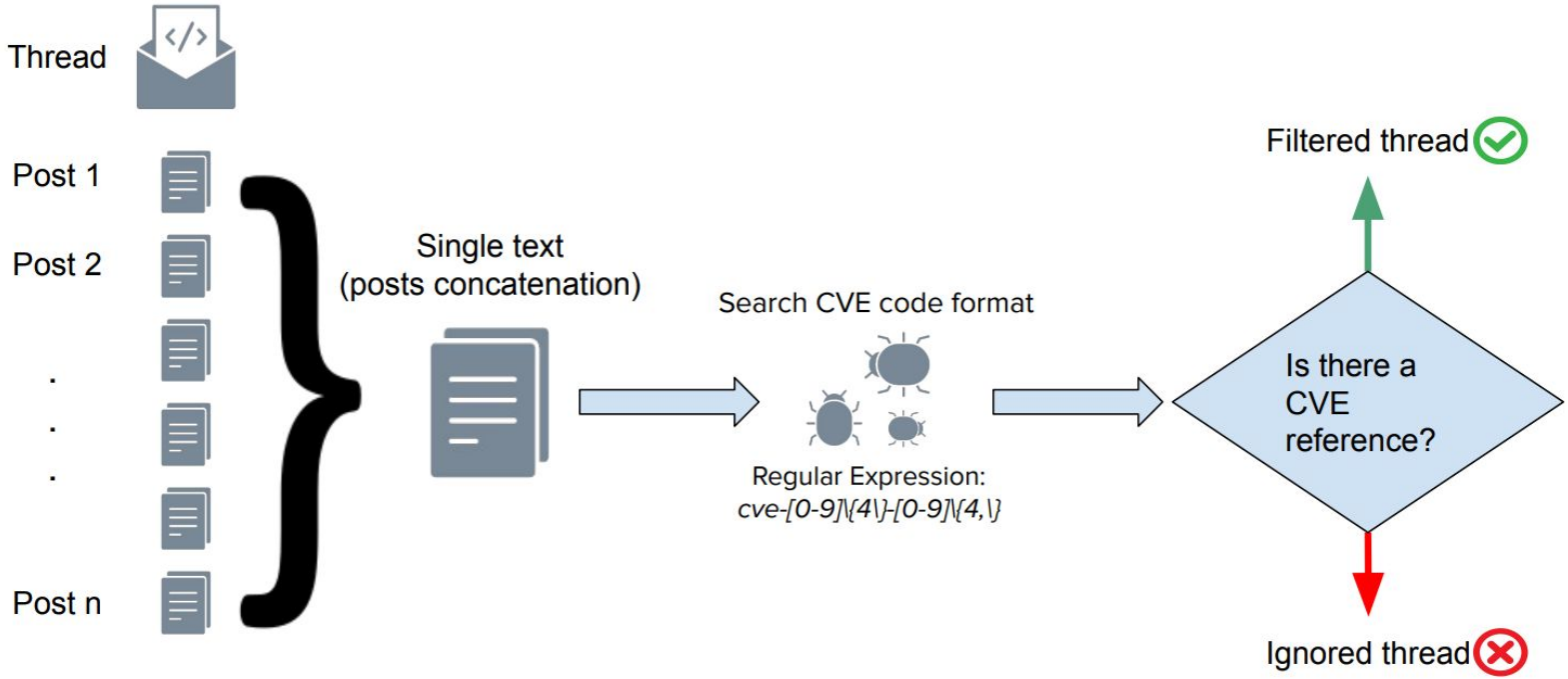
Dataset Construction

Pipeline



Feature extraction

Data Preparation - Filtering threads



Data Preparation - Feature Extraction

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	…

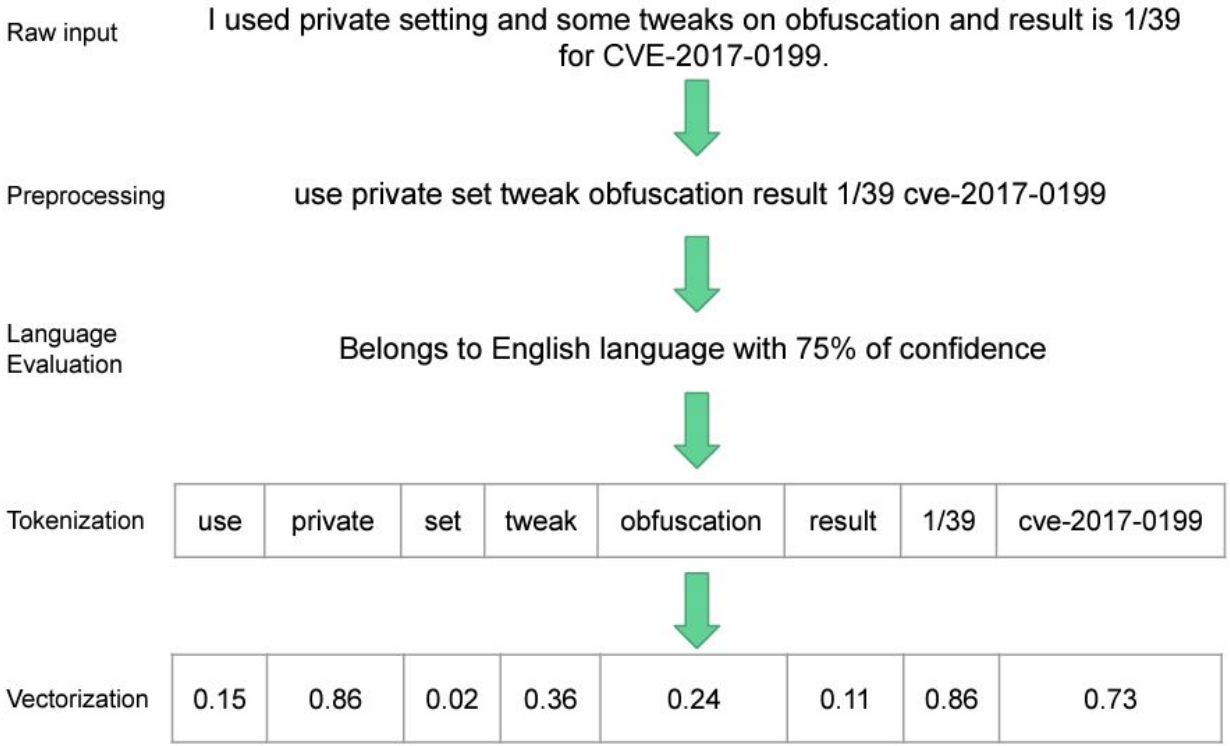
Bag-Of-Words (1-2-gram)

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	…

Data Preparation - Feature Extraction



Data Preparation - 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:

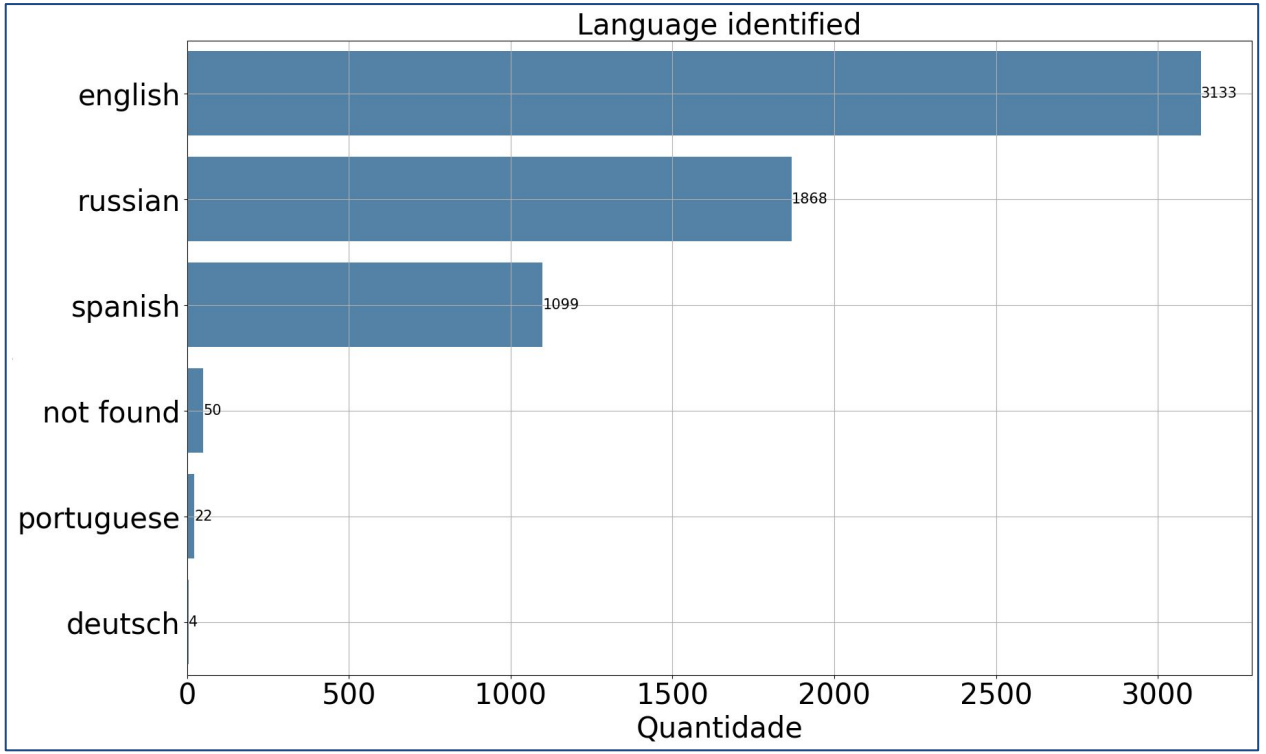
$$Ratio_{lrf}(text, language_j) = \frac{1}{\text{Total words in text}} \sum_{\substack{i=1 \\ \text{word}_i \in \text{ text}}}^{n} \mathbb{1}_{ilf}(word_i, language_j)$$

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

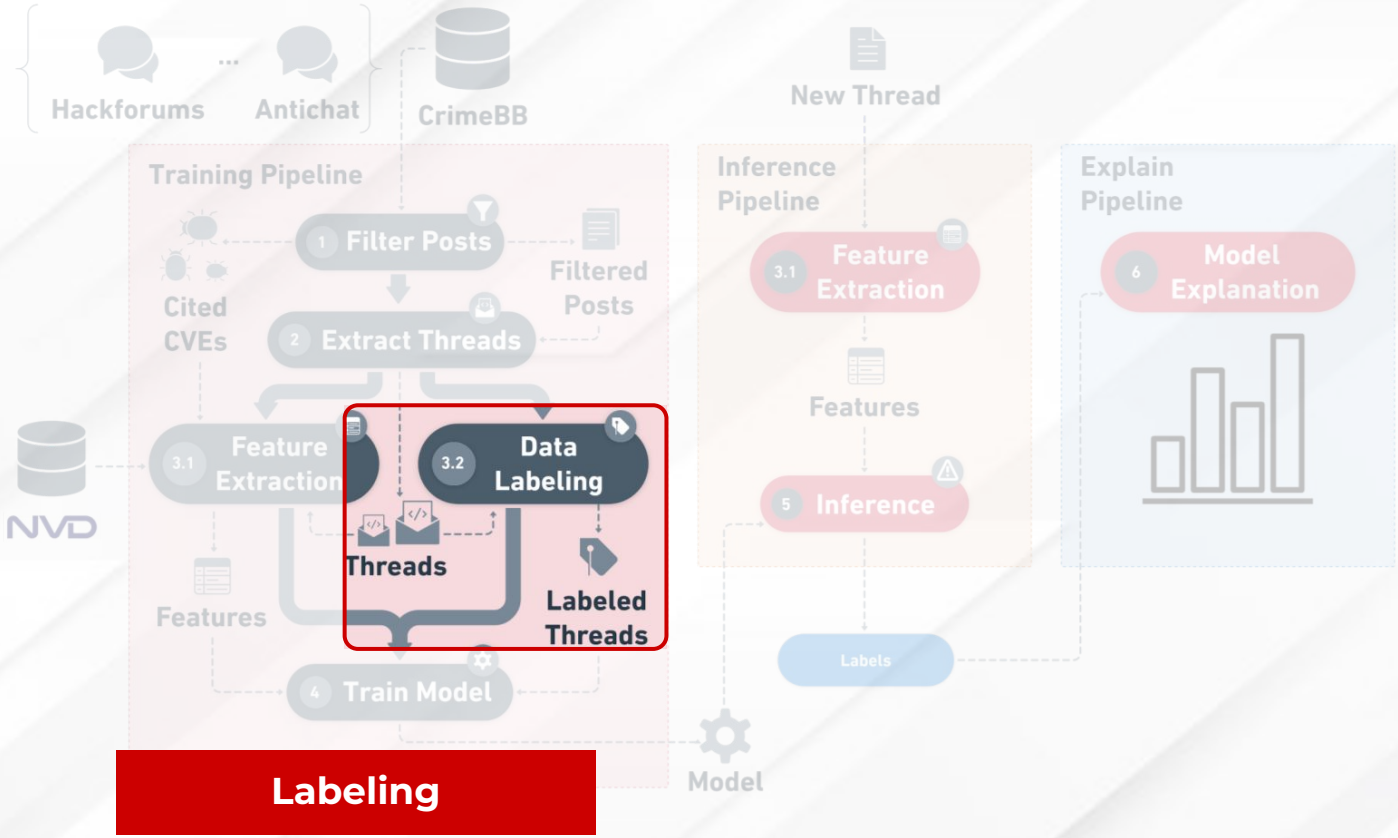
$$language(text) = \max_{\forall lang \in \text{languages list}} Ratio_{lrf}(text, lang)$$

Data Preparation - PostCog

- From PostCog, we search the word “**CVE**” and found about **14,857** posts since .
- We identify that only post scrapped from **HackForums** has **crime type**, **post type**, and **intent** tags.



Pipeline



Labeling

CrimeBB - 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

Data Preparation - PostCog-HF

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

Data Preparation - ChatGPT labeling

- In order to aggregate labels, we kindly 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)
- 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**

Prompt

- 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$_l$*abel*", *newlabel*), ...]

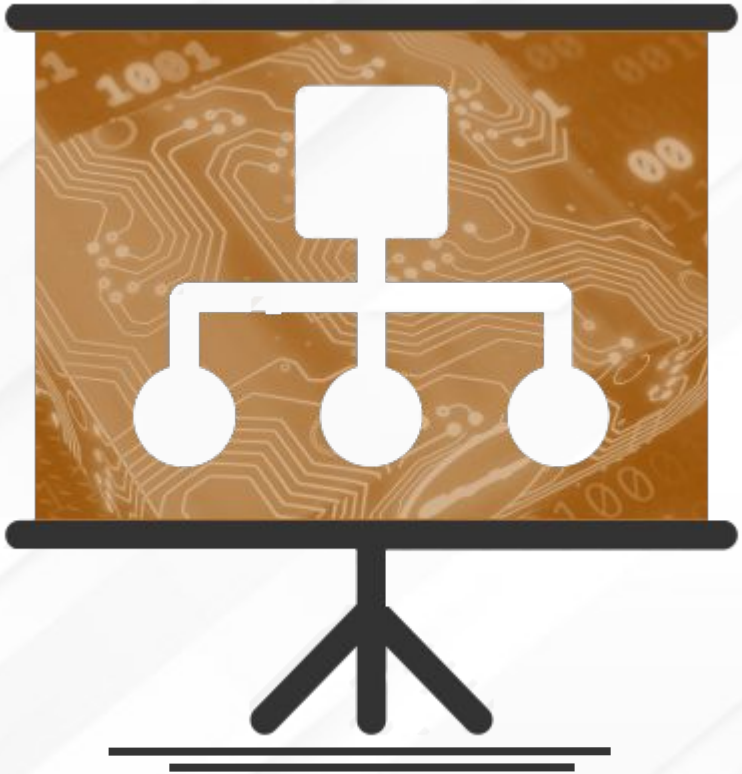
Data Preparation - PostCog-HF

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

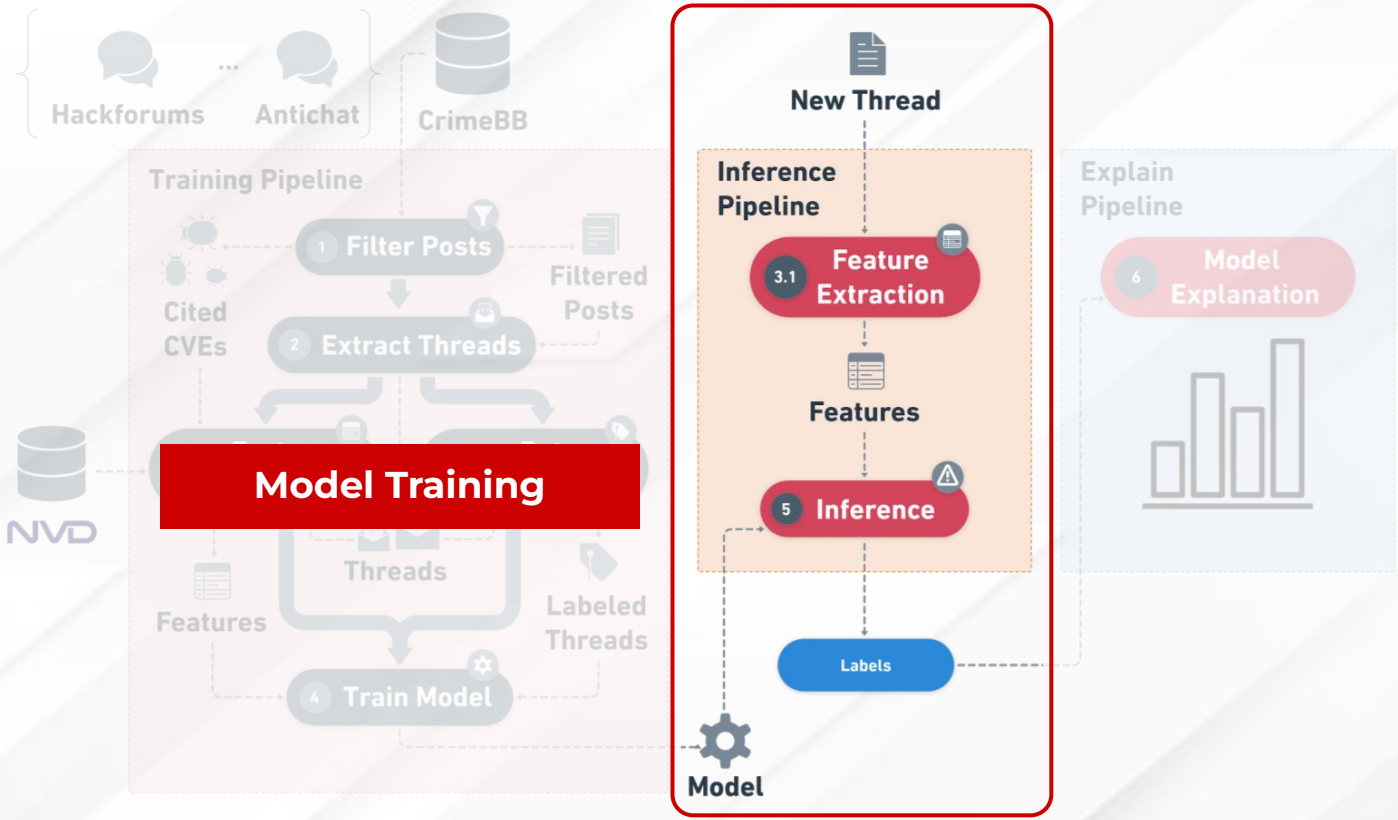
Post type labels	
Labels	Samples
Communication/Interaction	1050
Requests	1049
Other	494
Offer/exchanges	627

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



1. Introduction
2. Dataset
3. Beneath the Cream Methodology
4. Experimental Results
5. Conclusion

Pipeline


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

Experimental Setting

- Train and test split
  - 75% and 25%, respectively
  - Oversampling method to balance classes
  - Stratified split in order to preserve the original distribution
- Hyperparameters tuning
  - Grid Search
  - 5-fold Stratified Cross-Validation on the training set
- Evaluation metrics
  - Accuracy, Precision, Recall, and F1-score

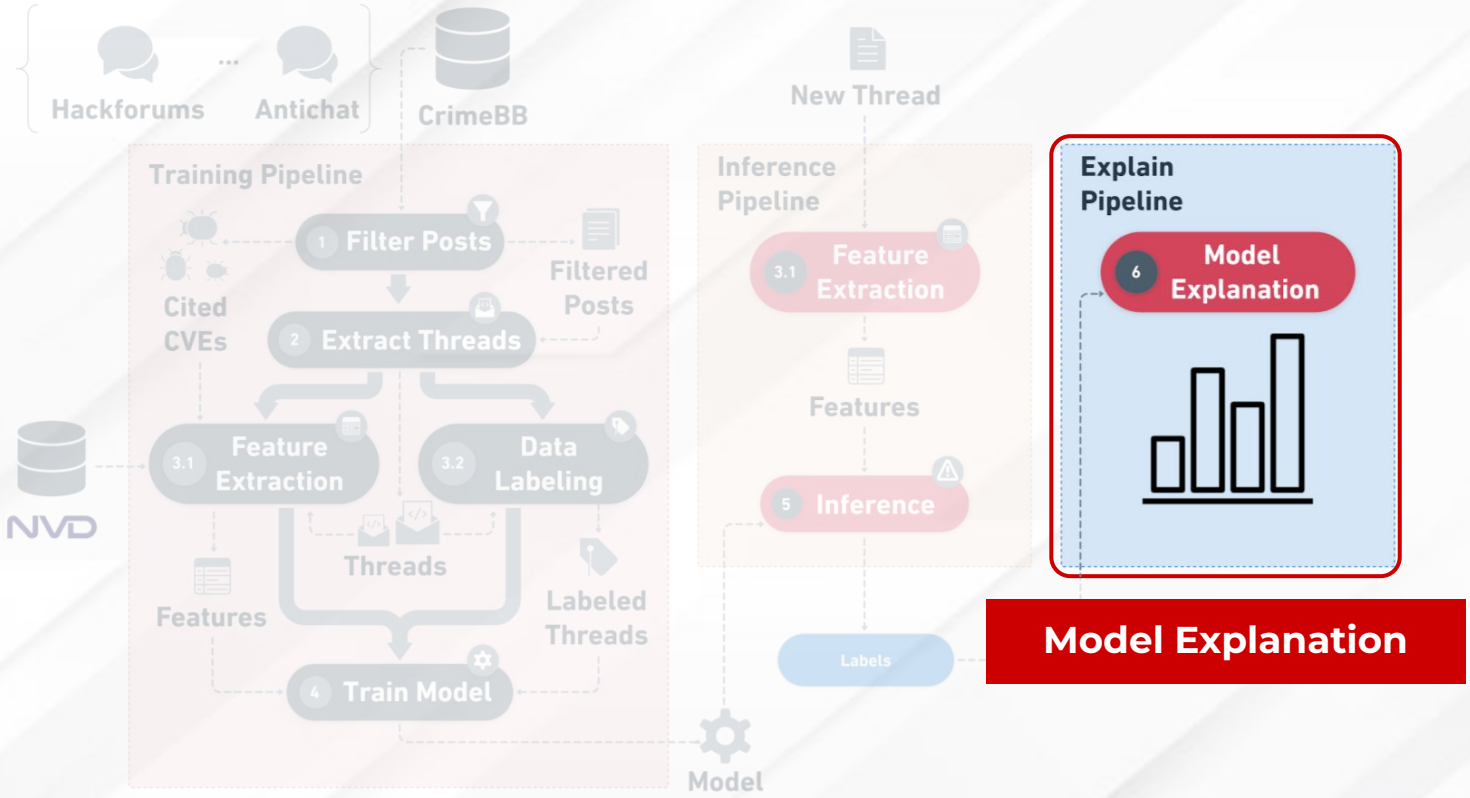
Experimental Setting

- **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.

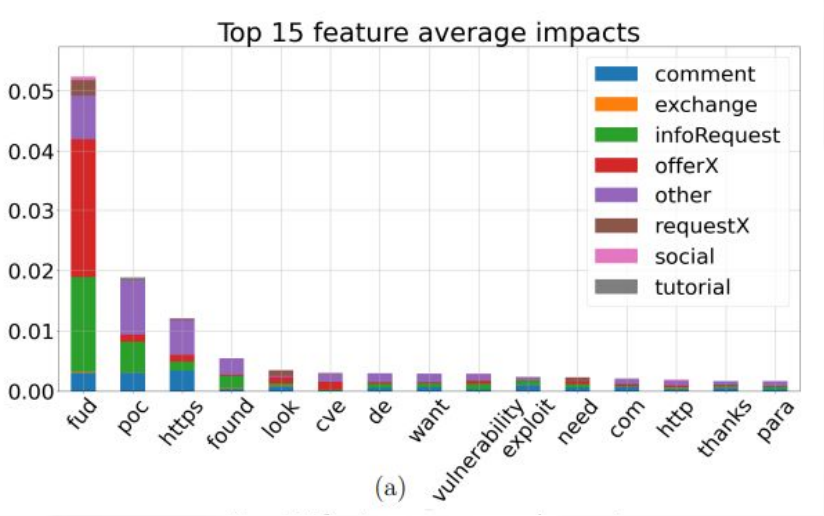
RandomForest

	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

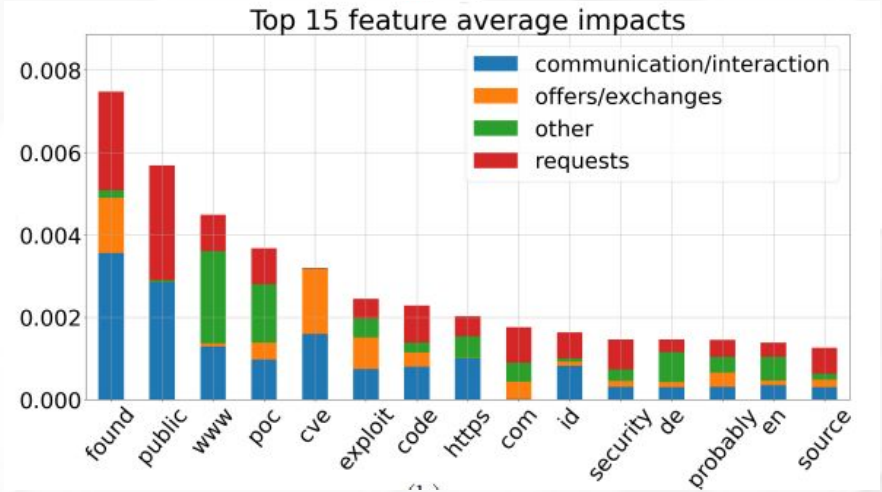
Pipeline


Hackforums
Antichat
CrimeBB
Training 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
New Thread
Inference Pipeline
3.1 Feature Extraction
Features
5 Inference
Labels
Explain Pipeline
6 Model Explanation
Model Explanation

Explanation

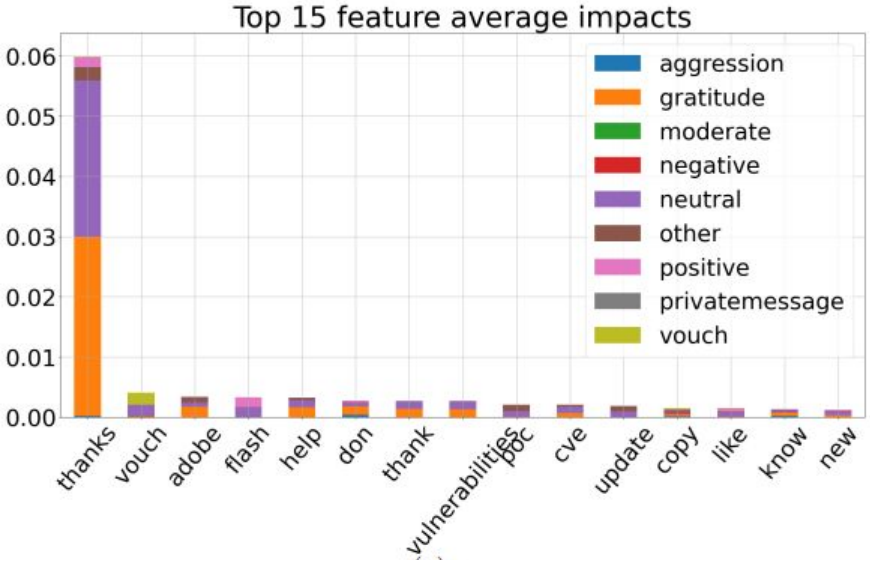


PostCog labeling - posttype

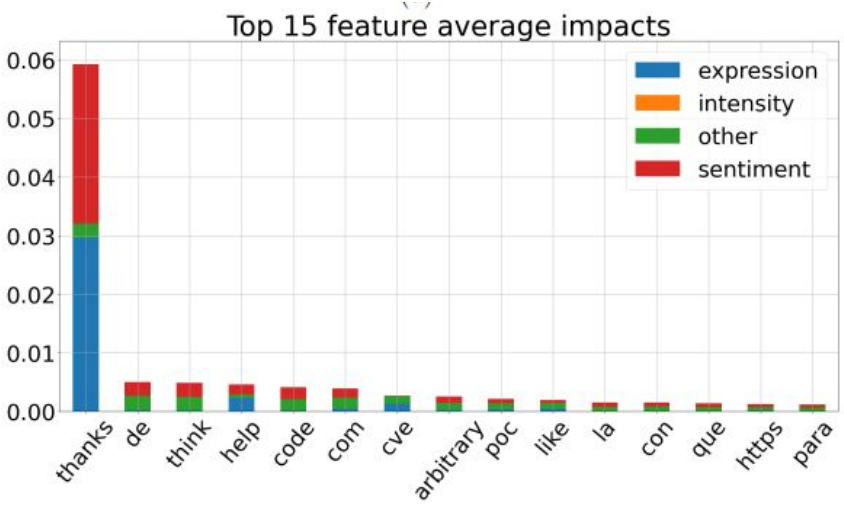


GPT labeling

Explanation

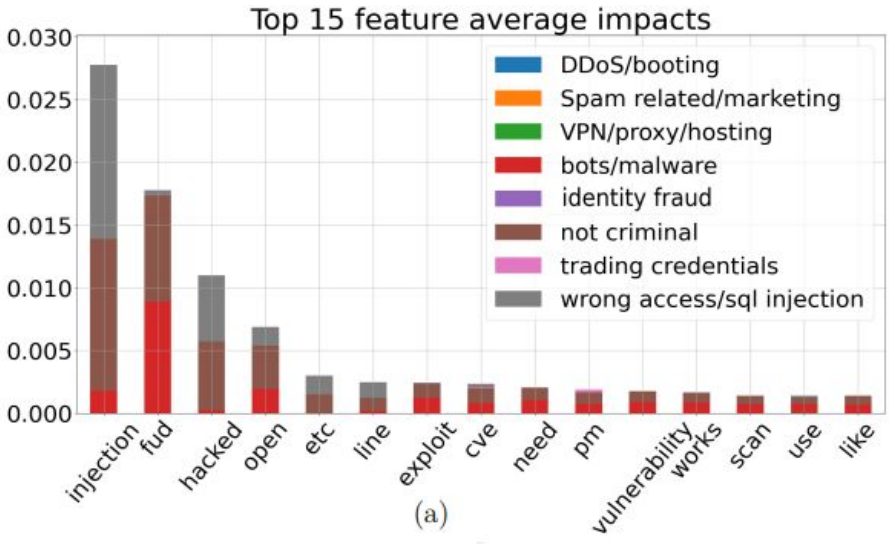


PostCog labeling - intention

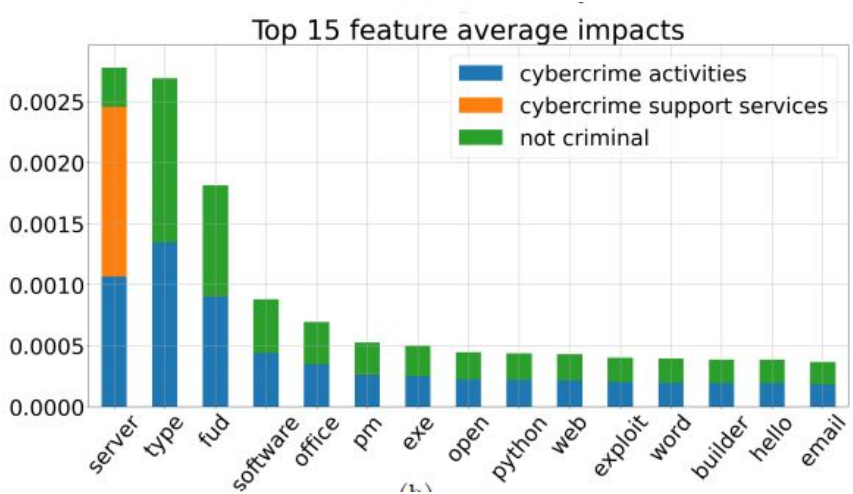


GPT labeling

Explanation

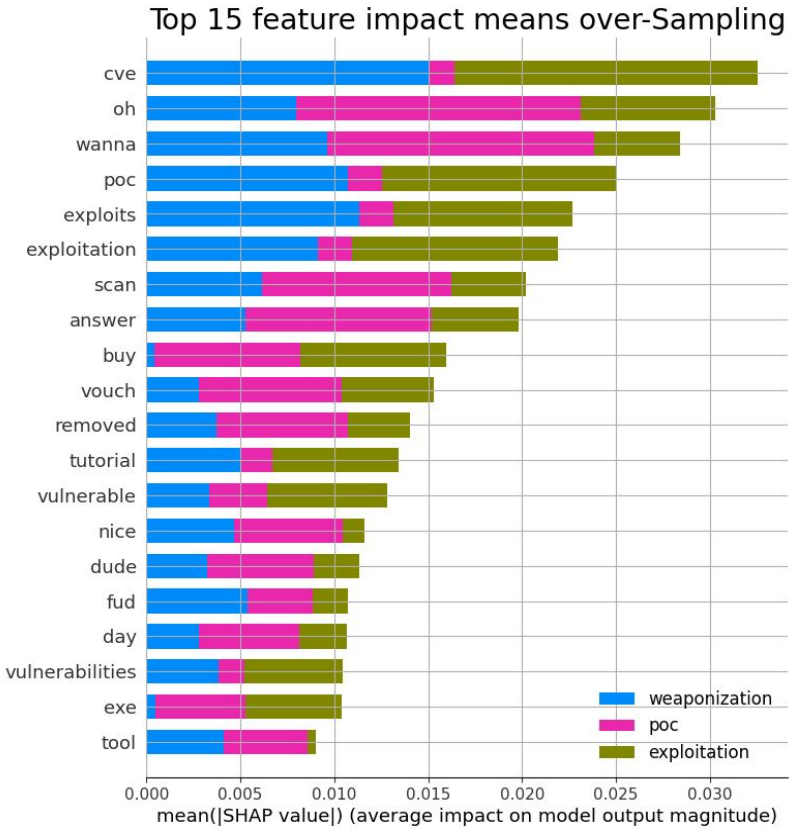


PostCog labeling - crime type

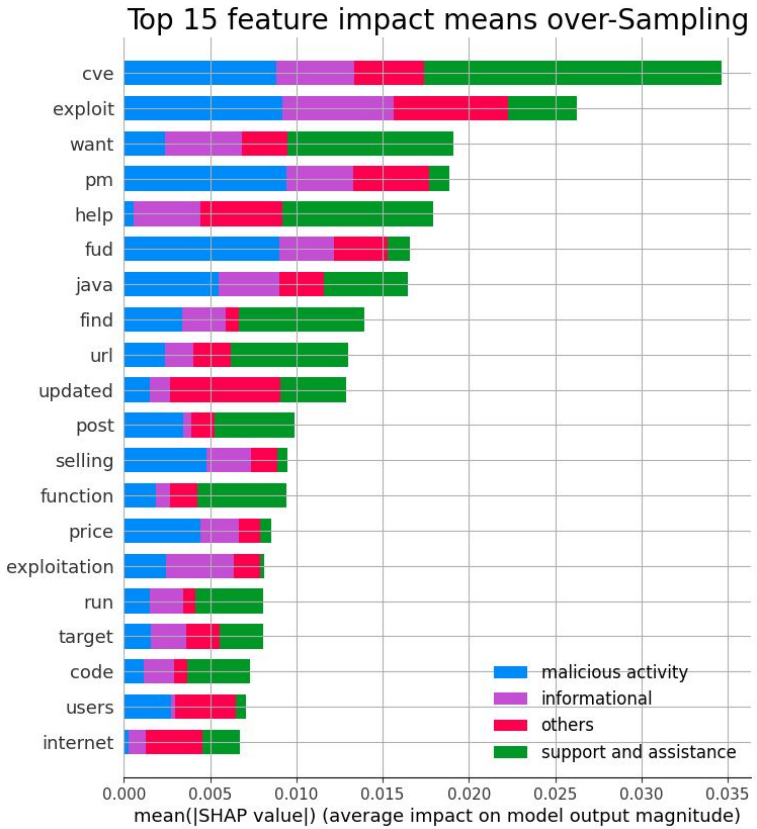


GPT labeling

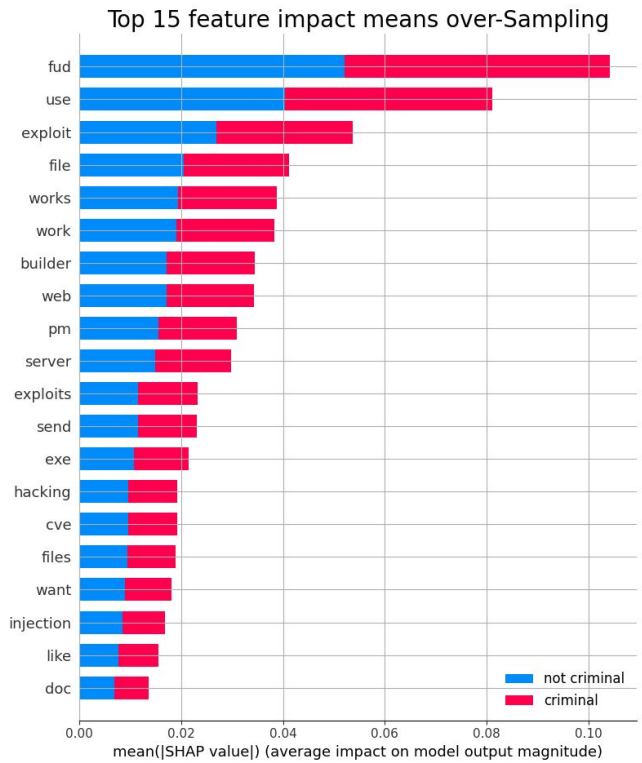
Manual labeling



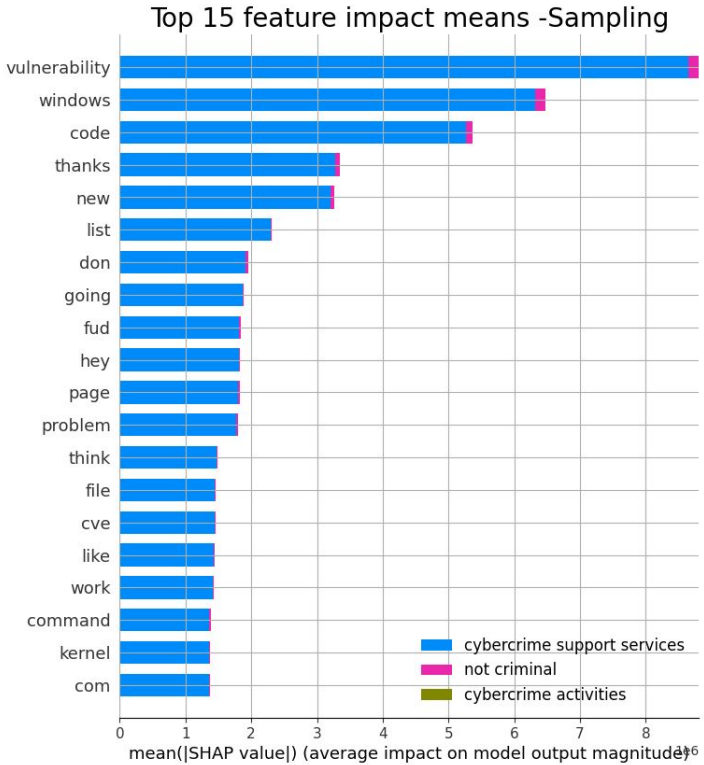
GPT labeling



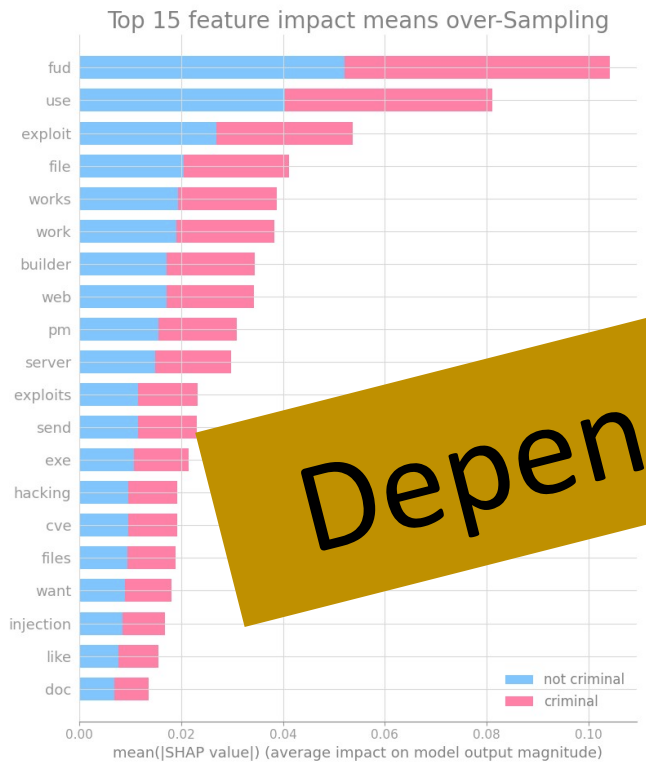
Binary labeling



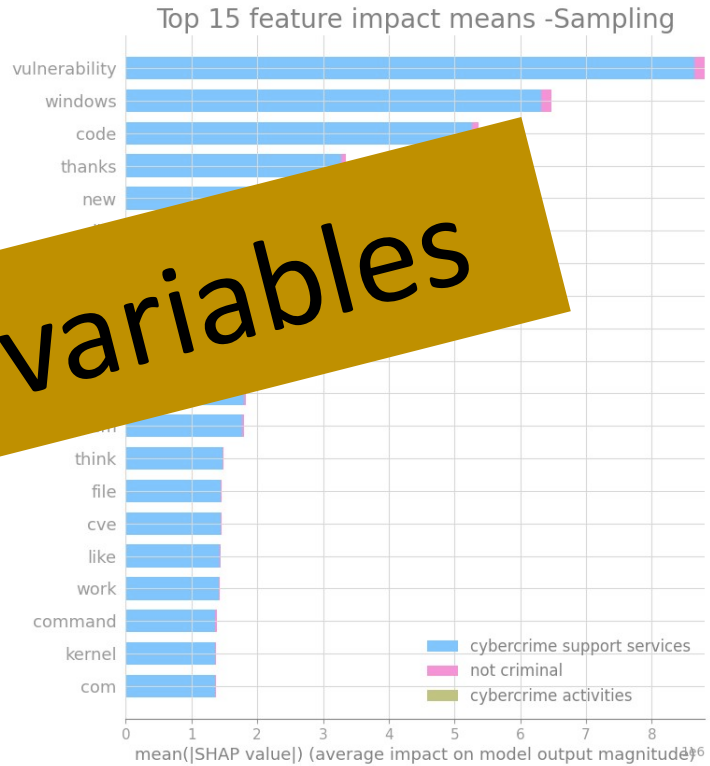
GPT labeling



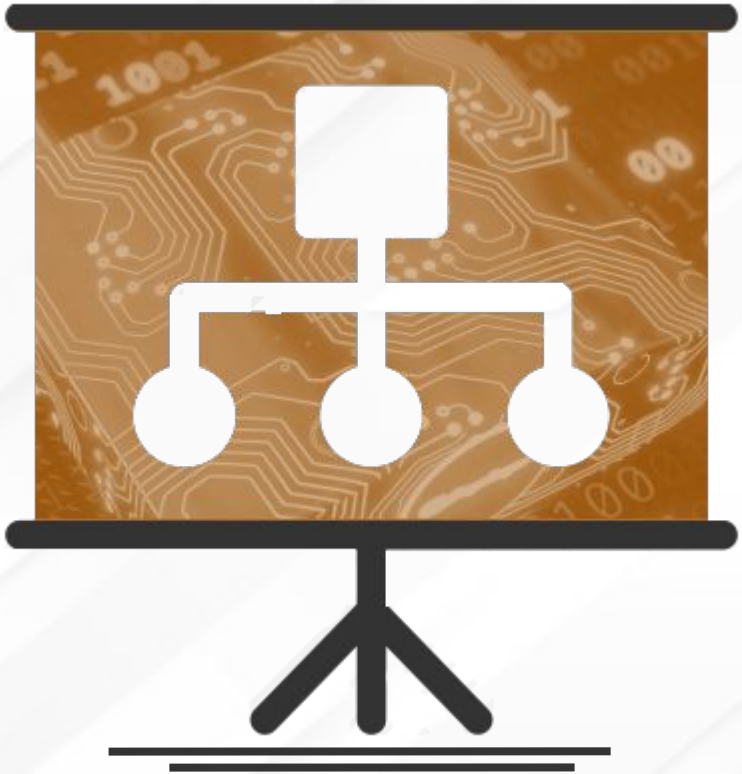
Manual labeling



GPT labeling



Dependent variables



1. Introduction
2. Dataset
3. Beneath the Cream Methodology
4. Experimental Results
5. Conclusion

Conclusion

- We were able to analyze and join pertinent annotations related to **CVEs thread-posts**
  - studying the HackForums underground forum.
- It is feasible to **train** a classifier to **infer** the **maturity level and type of threads.**
  - next step is to analyze all of them together.
- **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, but in some cases, the results are **not explainable**.
  - We will use other explainers such as LIME which works better than SHAP.

Thanks! Any questions?

felipe.moreno@ppgi.ufrj.br

THANKS!

Any Questions?