

Ca1SYS Lab Senior Project

By Sidharth Basam



INTRODUCTION

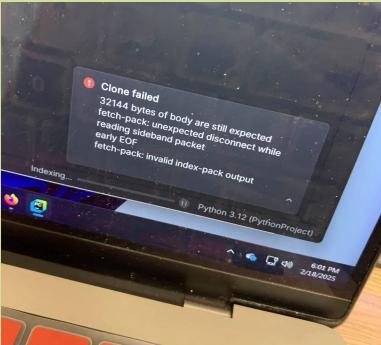
- Environment Setup
- Scraped Pitch and NZ Darknet (New Zealand Darknet)
- Binary Probabilistic Classifier



Environment Setup



- Had to use VMware Fusion instead of Virtualbox since I used a Apple Silicon Mac
- Ran into a few problems with the setup
 - Cloning git error was fixed by restarting VM
 - Certain install commands didn't work because of dependency issues
 - Usually fixed most issues by uninstalling and reinstalling modules



```
Traceback (most recent call last): @ Explain with AI
  File "C:\Users\jus\PycharmProjects\dw_pipeline_test\Forums\Initialization\forums_mining.py", line 8, in <module>
    from Forums.BestCardingWorld.crawler.selenium import crawler as crawlerBestCardingWorld
  File "C:\Users\jus\PycharmProjects\dw_pipeline_test\Forums\BestCardingWorld\crawler.selenium.py", line 7, in <module>
    from selenium import webdriver
ImportError: cannot import name 'webdriver' from 'selenium' (unknown location)
```

Pitch

- The entry link into the Pitch forum is the Popular forums tab
- The Pitch forum has many different topics
- I crawled relevant forums like OpSec, Markets, and Hacking

The screenshot shows the Pitch forum homepage at https://pitchprash4aqjlf7sbmuwe3pnkpylwqybj2q5o4szcfeea6d27yd.onion. The page title is "Pitch". The top navigation bar includes links for Rules, FAQ, BBCode, Donate, and Contact. A pinned message from "im wamus" (@wamus) says: "I've just spent a small fortune to speed up @DIG's response time by 1000%. Enjoy! 🌟 42 🌟 75". Below this is a sidebar titled "Popular People" listing various users with their follower counts. To the right is a "Popular" section featuring a post from "DIG AI" (@dig) announcing its private assistant. It also highlights three models: DIG-Uncensored, DIG-GPT, and DIG-Vision, each with a brief description and a "Amnesic" button.

The screenshot shows the Pitch forum homepage at https://pitchprash4aqjlf7sbmuwe3pnkpylwqybj2q5o4szcfeea6d27yd.onion. The page title is "Pitch". The top navigation bar includes links for Rules, FAQ, BBCode, Donate, and Contact. A pinned message from "im wamus" (@wamus) says: "Now out of beta with 3 new models: DIG-Uncensored, DIG-GPT, and DIG-Vision. Announcing DIG AI: Your private assistant - created by @Pitch". Below this is a sidebar titled "Popular People" listing various users with their follower counts. To the right is a "Topics" section featuring several discussion categories: Pitch (All discussion of Pitch and it's future), Markets (Discussion about darknet markets), Privacy (Tips, news, and resources to help you protect your privacy), Hacking (Everything related to hacking, opsec, and programming. Malware, phishing, DDoS, coding, research and news), HiddenService (Discussion of onion, i2p, and darknet sites), HarmReduction (A substance testing community geared towards harm reduction for end users globally), News (New relevant to the darknet), Guides (Guides & Tutorials), and Art (A place to showcase various forms of user created and inspired art).

NZ Darknet

- NZ Darknet is a forum for New Zealand's Dark Web Community
- The site have forums not only on illegal software, but also on drug reviews and other drug discussions
- I crawled the General Discussion Forum only since all the other forums on the site discussed only drugs.
- The General Discussion Forum talked more about software than drugs

NZ Darknet Market Forums

Darknet Market education and discussion

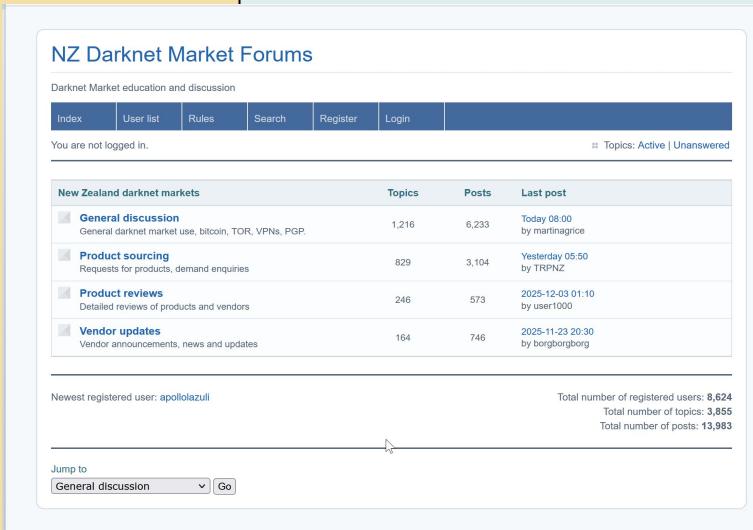
Index User list Rules Search Register Login Topics: Active | Unanswered

You are not logged in.

New Zealand darknet markets	Topics	Posts	Last post
General discussion	1,216	6,233	Today 08:00 by martnagrice
Product sourcing	829	3,104	Yesterday 05:50 by TRPNZ
Product reviews	246	573	2025-12-03 01:10 by user1000
Vendor updates	164	746	2025-11-23 20:30 by borgborgborg

Newest registered user: apollo lazuli Total number of registered users: 8,624
Total number of topics: 3,855
Total number of posts: 13,983

Jump to General discussion Go



Binary Classifier

- The first two cells are just for imports and loading the CSV file
- Cells 3 and 4 ensure that the columns are correct and the labels are binary
 - I also printed out the distribution of labels so we can see how imbalance the dataset is after modeling
 - The classifier needs to be at thread level.

This means that all posts that belong to the same thread need to be grouped.

```
!pip install -q scikit-learn pandas matplotlib

import numpy as np
import pandas as pd
import re
import matplotlib.pyplot as plt

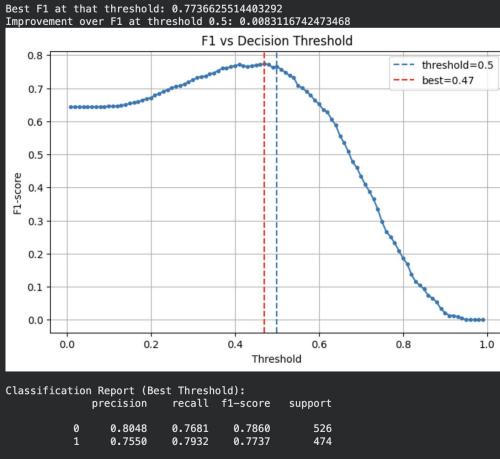
from sklearn.model_selection import StratifiedKFold, cross_val_predict
from sklearn.metrics import f1_score, classification_report, confusion_matrix
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.pipeline import Pipeline

data_path = "/content/forums_labeled_data.csv"
df = pd.read_csv(data_path)

print(df.head())
print(df.columns)
```

Binary Classifier

- Cell 5 builds the actual text input for the classifier
- For every thread, we clean the title, every post, and concatenate them into a single string
- The result is a list of X_test which is a list of 1000 thread texts and y which is the corresponding labels
- In cell 6, the model is defined. Raw text is turned into numbers by TF-IDF
- In cell 7, I use the 5 fold stratified cross-validation where Each fold uses 80% of the data to train and 20% to test and its repeated 5 times
 - If the probability is ≥ 0.5 , then we classify it as positive.
 - Cell 8 trains the final model and adds prediction helpers.



```
    """
    return float(final_pipeline.predict_proba([text])[0, 1])

def predict_thread_label(text):
    """
    Returns 1 (criminal hacking) or 0 (not)
    using the optimized threshold found in Cell 5.
    """
    prob = predict_thread_probability(text)
    return int(prob >= best_threshold)

# ---- Test with an example ----
example_title = "Any working exploit packs?"
example_posts = [
    "Looking for browser exploit kits.",
    "PM me if you have 0day or fresher packs."
]

example_text = example_title + " " + ".join(example_posts)

p = predict_thread_probability(example_text)
label = predict_thread_label(example_text)

print("Example probability:", p)
print("Predicted label:", label)

...
Final model trained on all data.
Using optimized threshold: 0.4700000000000003
Example probability: 0.5005176141600262
Predicted label: 1
```



THANK YOU

