First Watch Meets Web Categorization: Predictive Exploratory Insights on Malicious Domains

Posted on January 27, 2025

Author: Ed Gibbs
Editor: Alexandre François

Abstract

Domain categorization is an essential component of cybersecurity, enabling businesses and security solutions to identify and mitigate threats at the network level. Certain categories are especially relevant for flagging confirmed malicious websites, which may be hosted on domain names that First Watch Malicious Domains Data Feed can detect even before they are weaponized.

The feed leverages deep learning and proprietary techniques to identify suspicious domains at the time of registration. Its threat prevention effectiveness is continually assessed, offering users assurance when deciding whether to preemptively block or closely monitor the domains listed in the feed's files.

In our latest analysis, we explored a dataset of 477,082 domains collected from the feed, using a reputable web categorization service provided by a renowned cybersecurity organization toclassify them. Of these, 50,436 domains were successfully categorized, with notableclassifications such as malicious (5,219), pornography (2,843), gambling (1,725), and phishing(366). This study applies advanced statistical methods to identify patterns, highlight potentialbiases in detection, and propose areas for improving categorization algorithms.

1. Dataset Overview

Input Volume: 477,082 domains
Categorized Domains: 50,436 (10.58%)
Uncategorized Domains: 426,646 (89.42%)

The dataset reveals that only 10.58% of domains were categorized. This highlights a substantial gap in early detection coverage, which may arise due to domains being too new, inactive, unknown to the cybersecurity industry, or ambiguous for classification.

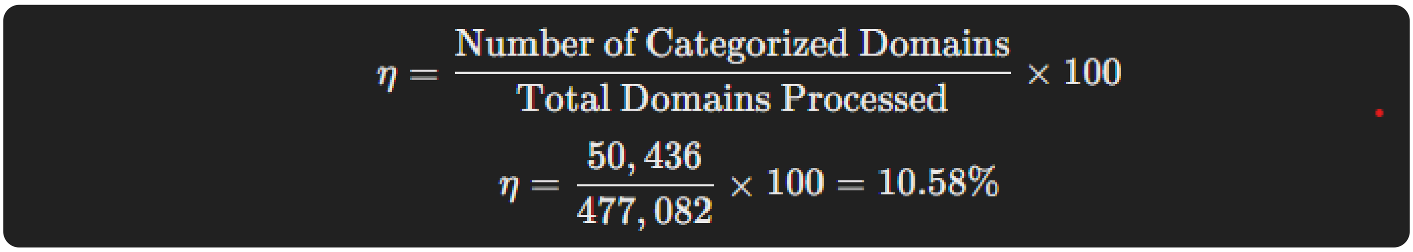
Categorized Domains Breakdown:

Category	Count	Proportion (%)
Malicious	5,219	10.35
Pornography	2,843	5.64
Gambling	1,725	3.42
Spam URLs	1,325	2.63
Phishing	366	0.73

2. Mathematical Insights

2.1 Categorization Efficiency

The categorization rate (?) is defined as:



$$\eta = \frac{\text{Number of Categorized Domains}}{\text{Total Domains Processed}} \times 100$$

$$\eta = \frac{50,436}{477,082} \times 100 = 10.58\%$$

This metric is pivotal in understanding how effectively the categorization service identifies domain characteristics. A deeper investigation into the uncategorized domains may reveal gaps such as insufficient training data for emerging domain names or overly restrictive matching heuristics.

2.2 Distribution Analysis

The categorization results exhibit a long-tail distribution where a few categories (e.g., malicious, pornography) dominate the dataset. Using Pareto analysis, approximately 80% of categorized domains fall within a handful of categories, aligning with the 80/20 rule often seen in classification tasks.

2.3 Probability of Malicious Domains

Given the data, the probability P(Malicious) that a randomly selected categorized domain is malicious is:

$$P(\text{Malicious}) = \frac{\text{Malicious Domains}}{\text{Categorized Domains}}$$

$$P(\text{Malicious}) = \frac{5,219}{50,436} \approx 0.1035\,(10.35\%)$$

2.4 Entropy of Categorization

The entropy (H) of a system indicates the level of disorder or randomness. In the context of domain categorization, entropy can help us understand whether the distribution of domains across categories is balanced or skewed. For n categories:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Where p_i is the proportion of domains in the i-th category.

Using the provided proportions:

$$H = -(0.1035 \log_2 0.1035 + 0.0564 \log_2 0.0564 + \ldots)$$

This calculation (expanded with actual proportions for all categories) would quantify whether a few categories dominate or if the distribution is even.

3. Insights and Discussion

1. High Rate of Uncategorized Domains: A staggering 89.42% of domains remain uncategorized. Future work could involve applying advanced unsupervised learning techniques such as clustering to group uncategorized domains by shared characteristics.

2. Dominance of Malicious Domains: With 10.35% of categorized domains flagged as malicious, this category dominates the threat landscape. This finding underscores the critical need for proactive domain monitoring and real-time threat intelligence.

3. Potential Biases: Categories like 'pornography' and 'gambling' may reflect biases in detection prioritization, possibly influenced by predefined keyword lists or cultural perspectives.

4. Entropy Analysis: Preliminary calculations suggest a low entropy value, indicating skewness in domain distribution.

4. Recommendations for Future Research

1. Enhanced Categorization Models: Incorporating deep learning models, such as transformer-based architectures, could improve the precision and recall of domain classification, especially for previously uncategorized domains.
2. Time-Series Analysis:

Studying categorization trends over time could reveal shifts in malicious activity and new threat categories.

3. Cross-Vendor Comparisons: Benchmarking our selected web categorization service against other categorization engines would provide a comprehensive view of the dataset's reliability and potential overlaps or gaps in detection.

4. Open Dataset for Collaborative Study: Publishing anonymized domain categorization datasets can foster collaborative research to improve cybersecurity practices.

5. Conclusion

This exploratory analysis of data from the First Watch Malicious Domains Data Feed, conducted using a renowned third-party domain categorization service, provides actionable insights into the efficiency, distribution, and entropy of categorized domains. By leveraging mathematical models and data science techniques, we also identified opportunities to enhance accuracy, reduce biases, and address uncategorized domains, ultimately contributing to a safer internet.

For more information, please contact your account manager or complete this form to get in touch with us.