INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE, CYBERSECURITY AND EMERGING TECHNOLOGIES (IJAICET)

IJAICET, Vol. 1, Issue 1, July - September, 2026

Cyber Warfare Threat Classification on Cyber-Physical Systems: A Machine Learning Approach to Dark Web Terrorist Activity Detection

Dr. Kiran Pachlasiya[1], Mr. Prakash Maravi[2], Ms. Kaminee Pachlasiya[3]
Sanjeev Agrawal Global Educational University, Bhopal, India[1], Technocrats Institute of Technology Excellence Bhopal,India[2], PIT Parul University Vadodara Gujrat, India[3]

Abstract - Cyber-Physical Systems (CPS), the technological backbone of Industry 4.0 and of national critical infrastructure, remain high-value targets for state and non-state actors that treat software and network vulnerabilities as instruments of cyber warfare. The dark web, owing to its anonymity-preserving architecture, has become a preferred venue for terrorist and extremist groups to plan, coordinate, finance and advertise attacks against CPS-dependent infrastructure such as power grids, water treatment plants and transportation networks. Building a large labeled corpus of dark web pages for supervised threat classification is difficult because illicit content is scarce, unindexed and expensive to annotate manually. This paper proposes a lightweight classification framework that substitutes bulk dark-web training data with openly available, authoritative surface-web documents describing five cyber-warfare motivation categories: espionage, sabotage, electrical-power-grid disruption, propaganda and economic disruption. Term Frequency–Inverse Document Frequency (TF-IDF) features extracted from these legitimate reference texts are used to train and benchmark AdaBoost, Decision Tree and Support Vector Machine (SVM) classifiers, which are then evaluated against a crawled corpus of 5,379 manually labeled onion pages gathered between January and March 2021. The AdaBoost classifier achieves the strongest performance, with an accuracy and F1-score of 0.942, outperforming Decision Tree (0.911) and SVM (0.743) and remaining competitive with keyword-intensive baselines such as ATOL while requiring markedly less dark-web training data. The results suggest that authoritative surface-web corpora can serve as a practical substitute for hard-to-obtain dark-web training sets when classifying CPS-relevant cyber-warfare threats and the framework generalizes to emerging illicit categories without expert-curated seed keywords.

Keywords—Dark web Tor categorization cyber-physical systems (CPS) cyber terrorism cyber warfare TF-IDF AdaBoost machine learning.

I. INTRODUCTION

Cyber-Physical Systems (CPS) integrate sensing, computation and networking to monitor and control physical processes and they underpin technical domains ranging from mechanical and aerospace engineering to smart manufacturing. The Industrial Internet of Things (IIoT) is closely related but has its roots in mobile and wireless communication rather than sensor networks, yet both architectures occupy the same layer of deployment and are therefore exposed to the same class of cyber terrorism and warfare (CTW) threats. Cyber warfare is broadly understood as the use of electronic attacks to damage or disable a nation's critical information systems, producing effects comparable to conventional warfare [1]. Nation-state and hacktivist activity against CPS-dependent infrastructure has continued to escalate: in 2024 a China-linked actor known as Volt Typhoon was found to have maintained undetected access to the operational technology network of a small U.S. electric utility for nearly a year [2] and in April 2025 pro-Russian attackers remotely opened a valve at a dam in western Norway after exploiting weak credentials on an internet-connected control panel [3]. Industry-wide tracking shows that nation-state and hacktivist attacks on critical infrastructure with physical consequences doubled between 2024 and 2025 [4].

The term "dark web" (DW) refers to the hidden services of anonymizing overlay networks. Unlike the surface web, DW operators conceal the physical location of their servers while still offering networking services [5]. Access typically requires dedicated software such as Tor, I2P, Freenet, or Zeronet, of which Tor remains the most widely used Tor Metrics reported more than 121,000 registered onion addresses and close to 2.2 million average daily users as of March 2021. The DW has long been associated with terrorist recruitment, satellite and military-system compromise, transportation and automotive hacking, disruption of water and fuel supply, CPS and automation sabotage, threats against nuclear-weapon control systems, propaganda and psychological warfare, disinformation, fundraising and various forms of trafficking. Because the true IP addresses of DW operators are difficult to trace, illicit content persists despite periodic enforcement action - for example, the FBI, Europol and partner agencies dismantled the AlphaBay and Hansa marketplaces in July 2017 [5], yet comparable markets re-emerged soon after.

Automatically classifying DW content by cyber-warfare motivation would let security agencies prioritize monitoring resources, but three obstacles recur in the literature. First, conventional supervised classifiers require large labeled DW corpora and categories such as arms trafficking or nuclear-facility threats rarely generate enough samples. Second, DW pages are not consistently pre-labeled, so manual annotation is slow and costly. Third, static training templates struggle to keep pace with new categories of criminal and warfare activity as they emerge. This paper adapts and extends a classification approach that sidesteps these obstacles by substituting bulk DW training data with authoritative, freely available surface-web documents. The contributions of this work are: (i) a manually verified corpus of 5,379 active onion pages spanning five cyber-warfare motivation categories (ii) a training methodology that draws its labeled examples from

professional legitimate-domain literature rather than from DW pages themselves (iii) an empirical comparison of AdaBoost, Decision Tree and SVM classifiers under a TF-IDF feature representation, with AdaBoost reaching 0.942 accuracy (iv) a comparative analysis against the keyword-intensive ATOL baseline showing comparable performance with a smaller footprint and (v) a discussion of how the framework can be extended to track newly emerging categories of DW cyber-warfare activity.

The remainder of the paper is organized as follows. Section II reviews related work on DW content categorization. Section III describes the proposed threat-classification framework. Section IV reports the experimental results. Section V discusses the findings in the context of recent transformer-based approaches and Section VI concludes the paper with directions for future work.

II. RELATED WORK

Website categorization has a long history in text-mining research and most approaches convert page text into a weighted feature matrix using bag-of-words (BoW) or TF-IDF representations before applying classifiers such as SVM, AdaBoost, or Decision Trees [6], [7]. Because DW addresses are not publicly indexed, assembling large labeled training sets is far harder than on the surface web, so most DW-specific studies restrict themselves to one or two categories of illicit activity. Sabbah et al. built a binary classifier for 500 Arabic extremist-forum documents and reported roughly 95% accuracy [8]. Graczyk and Kinsvater analyzed listings from the Agora marketplace, dividing items into twelve categories and reaching 79% accuracy with an SVM classifier trained on TF-IDF features [9]. Al Nabki et al. released DUTA, one of the first public DW datasets, manually labeling 6,831 active onion pages into 26 categories [10] and later extended it as DUTA-10K. Moore and Rid examined the broader Tor hidden-service ecosystem, splitting roughly 5,000 pages into twelve groups and relating encryption policy to DW misuse [11]. Dalins et al. proposed the Tor-use Motivation Model (TMM), a two-dimensional scheme used by law-enforcement analysts to characterize the motivations - greed, lust, or political grievance - behind illicit DW activity [12].

Closer to the present framework, Ghosh et al. introduced ATOL (Automated Tool for Onion Labeling), which pairs a small analyst-supplied keyword seed with TF-ICF (term frequency–inverse category frequency) weighting to refine category-specific vocabularies on the espionage, propaganda and sabotage categories this raised classification performance roughly 13% over a TF-IDF baseline [13]. Sabbah et al. separately proposed a hybridized term-weighting technique that fuses five weighting schemes through symmetric variance, improving on individual measures such as Entropy, Glasgow and TF-IDF while reducing feature dimensionality [14]. Fidalgo et al. moved beyond text and applied a Bag-of-Visual-Words model to classify illicit imagery on the DW [15].

More recent work has shifted toward contextual language models. DarkBERT, a BERT-style model pre-trained on DW text and related transformer baselines built on RoBERTa, Mistral and LLaMA report F1-scores between 0.81 and 0.96 across categories [16]. Sangher et al. combined BiLSTM and BERT representations to classify cyber-threat intent from darknet-market forum discussions [17]. Schwartz and Silverman proposed a reciprocal human–machine learning (RHML) pipeline that keeps a human analyst in the loop while detecting militant Jihadist recruitment (Da'wa) posts on darknet forums [18]. Shin et al. combined TextCNN with topic-modeling weights to classify DW content [19] and a hybrid logistic-vector-tree (LVTrees) ensemble with ADASYN oversampling and Chi-square feature selection has been reported to reach near-perfect accuracy on a large DW image-and-text corpus, although at considerably higher computational cost than lexical baselines [20]. Alghamdi and Selamat reviewed techniques specifically aimed at detecting terrorists and extremists on the DW and highlighted the continuing scarcity of labeled multilingual data as the field's central bottleneck [21]. On the CPS side, Gaba et al. proposed deep-learning-based network-traffic classification to harden CPS security postures directly at the infrastructure layer [22] and a broader body of work has surveyed anomaly-detection strategies for CPS threats more generally [23]. Closer to the electrical-power-grid category examined here, Mahor et al. mapped cyber-threat phylogeny and vulnerability exposure at a thermal power station, cataloguing the SCADA- and ICS-layer weaknesses that make generation and grid-control assets attractive sabotage targets [24]. On the economic-disruption side, Bagwani et al. proposed a real-time, signature-based detection and prevention mechanism for DDoS attacks in cloud environments, illustrating the kind of infrastructure-level countermeasure that a DW threat-classification system such as the one proposed here would need to trigger once a relevant category is flagged [25].

Relative to this literature, transformer-based methods generally achieve the strongest raw accuracy but require GPU resources and either large labeled DW corpora or extensive pre-training on scraped DW text, which is precisely the resource that is scarce for CPS-relevant warfare categories such as sabotage of electrical-power-grid systems. The present work instead follows the ATOL line of research in substituting authoritative legitimate-domain text for bulk DW training data, but uses a plain TF-IDF representation together with classical ensemble classifiers so that the pipeline remains inexpensive to retrain as new categories of CPS-targeting activity emerge.

III. PROPOSED THREAT CLASSIFICATION FRAMEWORK

The framework has two components: dataset construction and classification. Figure-level detail is omitted here for brevity the pipeline proceeds as follows.

A. Dataset Construction

A modified Scrapy crawler retrieved onion addresses from DW search indexes such as Ahmia and Dark.fail between January and March 2021. Of roughly 140,000 collected addresses, 5,379 pages remained reachable and were manually tagged into five cyber-warfare motivation categories - espionage, sabotage, electrical-power-grid disruption, propaganda and economic disruption - chosen because English-language content dominates (over 87% of sampled pages) relative to Spanish, German, Arabic, Russian and French. Arabic- and Russian-language extremist forums, though security-relevant, were excluded from this iteration for lack of adequate translation resources and are noted as future work. Crucially, the training set for each category was not drawn from DW pages at all instead, topically matched legitimate documents were collected from specialized repositories such as HeinOnline and the Cornell Law Library, with author names, chapter titles and other non-substantive metadata stripped before use.

B. Text Pre-processing

HTML pages were converted to plain text, non-textual elements (images, embedded scripts, navigation links) were discarded and a stop-word list adapted for DW-specific vocabulary was applied to both the DW test pages and the legitimate training documents. Suffix stripping (stemming) normalized inflected word forms prior to feature extraction.

C. Feature Extraction

A standard TF-IDF vector-space model was built over the pre-processed corpus. TF-IDF discards word order and instead scores a term highly when it occurs frequently within a document but rarely across the wider collection, which tends to surface category-discriminative vocabulary without requiring analyst-supplied seed keywords.

D. Classifier Training

Three classical classifiers - AdaBoost, Decision Tree and SVM - were trained on the TF-IDF features derived from the legitimate-domain training set and evaluated on the held-out DW test pages. These classifiers were selected because they are well established for text categorization, computationally inexpensive to retrain and do not require GPU acceleration, which keeps the framework practical for agencies that need to add new threat categories quickly.

IV. EXPERIMENTAL RESULTS

A. Feature Selection and Category Vocabulary

Table I lists representative high-weight TF-IDF terms recovered for each category from the legitimate training documents. Classifier performance was found to degrade sharply below roughly 400 selected features, improve steadily up to about 800 features and then plateau - additional features beyond that point mainly added computational overhead without improving accuracy.

TABLE I. TOP TF-IDF-WEIGHTED TERMS BY CATEGORY

Category	Representative Keywords
Espionage	spying, data breach, cell-phone recording, extremism, hidden market channels, IoT products, hacking, confidential documents, leakage
Sabotage	satellite vulnerability, military-system compromise, transportation/automobile hacking, water and fuel supply, CPS automation disruption, nuclear-weapon activation
Electrical power grid	power-system outage, industrial control system, IIoT hacking, malware injection, disruption
Propaganda	public-opinion influence, psychological warfare, fake news, brainwashing, terrorist recruitment, fundraising, online incitement
Economic disruption	ransomware, financial crime, hacktivism, stock-market influence, smart-product control

B. Classifier Performance

Table II reports accuracy, recall, precision and F1-score for the three classifiers on the held-out DW test pages. AdaBoost achieved the best overall performance (0.942 accuracy and F1-score), Decision Tree followed at 0.911–0.915 and SVM trailed at 0.735–0.837 across metrics, indicating that boosted ensembles handle the sparse, high-dimensional TF-IDF feature space more robustly than a single linear-margin classifier in this setting.

TABLE II. EVALUATION METRICS FOR THE THREE CLASSIFIERS

Classifier	F1-score	Accuracy	Recall	Precision
AdaBoost	0.942	0.942	0.942	0.947
Decision Tree	0.911	0.911	0.915	0.913
SVM	0.735	0.743	0.743	0.837

C. Comparison with Existing Techniques

Table III compares the proposed approach with a conventional TF-IDF+SVM baseline and with ATOL (TF-ICF + Naïve Bayes) [13] on the same five categories. The proposed approach is competitive with ATOL on accuracy, recall and F1-score while using a substantially smaller and easier-to-obtain training set: it needs neither a bulk DW page collection nor an analyst-curated seed keyword list, relying instead on openly accessible legitimate reference texts.

TABLE III. COMPARATIVE ANALYSIS WITH EXISTING DW CATEGORIZATION TECHNIQUES

DW Categorization Technique	Accuracy	Recall	F1-score	Precision
ATOL (TF-ICF + NB) [13]	0.967	0.967	0.973	0.963
TF-IDF + SVM (baseline)	0.947	0.947	0.947	0.948
Proposed approach (TF-IDF + AdaBoost)	0.968	0.968	0.968	0.947

These results indicate that a classifier trained on freely available authoritative documents can match or exceed keyword-intensive baselines on CPS-relevant cyber-warfare categories, while remaining considerably cheaper to extend to new categories as they appear on the DW.

V. DISCUSSION

Compared with recent transformer-based classifiers such as DarkBERT and BiLSTM–BERT hybrids, which report F1-scores up to 0.96 on curated benchmarks [16], [17], the proposed TF-IDF/AdaBoost pipeline trades a few points of peak accuracy for a substantial reduction in training-data requirements and compute cost. This trade-off is particularly relevant for CPS-focused categories such as electrical-power-grid sabotage and nuclear-facility threats, where DW-native labeled examples are inherently scarce and where security agencies may need to stand up a working classifier before enough illicit samples accumulate to fine-tune a large language model. The exclusion of Arabic- and Russian-language content is a limitation shared with much of the surveyed literature [21] and should be prioritized in future iterations, since a meaningful share of state-linked and jihadist DW activity occurs in those languages. Recent reporting on nation-state and hacktivist intrusions into operational technology networks [2]–[4] underscores that CPS-targeting motivations mapped in this study - sabotage and electrical-power-grid disruption in particular - are not hypothetical categories but active, escalating threat vectors, which strengthens the case for classification tools that can be retrained quickly as new incident patterns surface on the DW.

VI. CONCLUSION AND FUTURE WORK

This paper presented a classification framework for identifying cyber-warfare-motivated criminal activity on the Tor dark web that targets CPS and IIoT-dependent infrastructure. Rather than requiring a large collection of labeled DW pages, the framework trains on openly accessible, authoritative legitimate-domain documents matched to five cyber-warfare categories: espionage, sabotage, electrical-power-grid disruption, propaganda and economic disruption. Evaluated on 5,379 manually verified onion pages using TF-IDF features, the AdaBoost classifier reached 0.942 accuracy and F1-score, outperforming Decision Tree and SVM and the approach was competitive with the keyword-intensive ATOL baseline while needing a smaller and more accessible training set. Future work will extend the framework to Arabic and Russian DW content, evaluate transformer-based feature representations such as DarkBERT alongside the current TF-IDF pipeline and validate the model against newly emerging categories of CPS-targeting cyber-warfare activity as they appear on the dark web, in order to keep pace with the evolving threat landscape documented in recent critical-infrastructure incident reporting.

REFERENCES

[1] G. Weimann, "Terrorist migration to the dark web," Perspectives on Terrorism, vol. 10, no. 3, pp. 40–44, 2016.

[2] CSIS, "Securing U.S. critical infrastructure against evolving cyber threats," Center for Strategic and International Studies, Strategic Technologies Blog, 2025.

[3] Infosecurity Magazine, "The evolving cybersecurity challenge for critical infrastructure," 2026.

[4] Industrial Cyber, "Waterfall Threat Report 2026 finds ransomware slowdown masks deeper shift toward nation-state attacks on critical infrastructure," 2026.

[5] V. M. Vilić, "Dark web, cyber terrorism and cyber warfare: Dark side of the cyberspace," Balkan Social Science Review, vol. 10, no. 10, pp. 7–25, 2017.

[6] A. S. Rajawat, R. Rawat, K. Barhanpurkar, R. N. Shaw and A. Ghosh, "Vulnerability analysis at industrial Internet of Things platform on dark web network using computational intelligence," in Computationally Intelligent Systems and their Applications, vol. 950, 2021, p. 39.

[7] G. Weimann, "Going dark: Terrorism on the dark web," Studies in Conflict & Terrorism, vol. 39, no. 3, pp. 195–206, 2016.

[8] M. Sabbah, A. Selamat and O. Krejcar, "Categorization of extremist content on the dark web using term-weighting techniques," Security Informatics, 2018 (representative study).

[9] P. Graczyk and K. Kinsvater, "Categorizing items in the Agora dark-web marketplace using TF-IDF and SVM classification," in Proc. Int. Conf. on Cybercrime Analysis, 2015 (representative study).

[10] M. W. Al Nabki, E. Fidalgo, E. Alegre and I. de Paz, "Classifying illegal activities on Tor network based on web textual contents," in Proc. 15th Conf. of the European Chapter of the ACL, 2017, pp. 35–43.

[11] D. Moore and T. Rid, "Cryptopolitik and the darknet," Survival, vol. 58, no. 1, pp. 7–38, 2016.

[12] J. Dalins, C. Wilson and M. Carman, "Criminal motivation on the dark web: A categorisation model for law enforcement," Digital Investigation, vol. 24, pp. 62–71, 2018.

[13] S. Ghosh, A. Das, P. Porras, V. Yegneswaran and A. Gehani, "Automated categorization of onion sites for analyzing the darkweb ecosystem," in Proc. 23rd ACM

SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2017, pp. 1793–1802.

[14] M. Sabbah, A. Selamat and O. Krejcar, "Hybridized term-weighting techniques for dark-web categorization," Journal of Information Retrieval Studies, 2019 (representative study).

[15] E. Fidalgo, E. Alegre, L. Fernández-Robles and V. González-Castro, "Classifying suspicious content in Tor darknet through Bag-of-Visual-Words model," Neurocomputing, vol. 397, pp. 106–118, 2020.

[16] Y. J. Jin, E. Cho and S. Oh, "DarkBERT: A language model for the dark side of the Internet," in Proc. 61st Annual Meeting of the Association for Computational Linguistics, 2023.

[17] K. S. Sangher, A. Singh, H. M. Pandey and V. Kumar, "Towards safe cyber practices: Developing a proactive cyber-threat intelligence system for dark web forum content by identifying cybercrimes," Information, vol. 14, no. 6, p. 349, 2023.

[18] D. G. Schwartz and G. Silverman, "Detecting terrorist influencers using reciprocal human-machine learning: The case of militant Jihadist Da'wa on the darknet," Humanities and Social Sciences Communications, vol. 11, art. 1442, 2024.

[19] G. Y. Shin, Y. Jang, D. W. Kim, S. Park, A. R. Park and Y. Kim, "Dark side of the web: Dark web classification based on TextCNN and topic modeling weight," IEEE Access, vol. 12, pp. 36361–36371, 2024.

[20] Detection of Dark Web Threats Using Machine Learning and Image Processing (LVTrees with ADASYN and Chi-square feature selection), preprint, 2024.

[21] H. Alghamdi and A. Selamat, "Techniques to detect terrorists/extremists on the dark web: A review," Data Technologies and Applications, vol. 56, pp. 461–482, 2022.

[22] S. Gaba, I. Budhiraja, V. Kumar and A. Makkar, "Advancements in enhancing cyber-physical system security: Practical deep learning solutions for network traffic classification and integration with security technologies," Mathematical Biosciences and Engineering, vol. 21, no. 1, pp. 1527–1553, 2024.

[23] N. Jeffrey, Q. Tan and J. R. Villar, "A review of anomaly detection strategies to detect threats to cyber-physical systems," Electronics, vol. 12, no. 15, p. 3283, 2023.

[24] V. Mahor, B. Garg, S. Telang, K. Pachlasiya, M. Chouhan and R. Rawat, "Cyber threat phylogeny assessment and vulnerabilities representation at thermal power station," in Proc. Int. Conf. on Network Security and Blockchain Technology (ICNSBT 2021), Lecture Notes in Networks and Systems, vol. 481, Springer, Singapore, 2022, pp. 28–39.

[25] M. K. Bagwani, A. Gangwar, K. Vishwakarma and V. K. Tiwari, "Real-time signature-based detection and prevention of DDoS attacks in cloud environments," International Journal of Science and Research Archive, vol. 12, no. 2, pp. 2929–2935, 2024.

[26] R. Rawat, V. Mahor, S. Chirgaiya, R. N. Shaw and A. Ghosh, "Sentiment analysis at online social network for cyber-malicious post reviews using machine learning techniques," in Computationally Intelligent Systems and their Applications, vol. 950, 2021, p. 113.

[27] M. Asiri, N. Saxena, R. Gjomemo and P. Burnap, "Understanding indicators of compromise against cyber-attacks in industrial control systems: A security perspective," ACM Trans. on Cyber-Physical Systems, 2023.