International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Phishing-as-a-Service (PhaaS) Platform Identification using Dark Web Data and Clustering

Brindashree R
Under the guidance of prof.Mahalakshmi C V
Dept. of Computer Science Engineering
Bangalore Institute of Technology
Bangalore, India

Abstract—Phishing-as-a-Service (PhaaS) has emerged as a major enabler of cybercrime by allowing low-skilled attackers to launch sophisticated phishing campaigns using paid kits, hosting, mailing tools, and automation services. These services are extensively marketed across Dark Web forums, marketplaces, and encrypted communication channels, creating a growing underground economy that security teams cannot easily track. This research proposes a structured framework to identify PhaaS platforms using Dark Web data combined with unsupervised machine learning clustering techniques. Dark Web data is collected using Tor-based scraping from marketplaces, vendor listings, and underground forums. The unstructured text is cleaned and processed using natural language processing (NLP) methods, and meaningful features are extracted using TF-IDF and embedding-based vectorization. Clustering algorithms such as K-Means, DBSCAN, and HDBSCAN are applied to identify similar service groups, detect PhaaS vendors, and reveal hidden patterns across the ecosystem. The results show that clustering effectively groups phishing kits, sellers, pricing models, and service features, enabling automated identification of PhaaS networks. This work enhances cyber threat intelligence by providing a scalable approach to monitoring Dark Web marketplaces and detecting emerging phishing-as-a-service operations.

I. INTRODUCTION

Phishing remains one of the most pervasive and damaging forms of cyberattack, responsible for a significant proportion of security breaches, credential theft, financial fraud, and malware distribution worldwide. Over the past decade, phishing campaigns have evolved from manually crafted emails to sophisticated, large-scale, and automated operations. A major driver of this evolution is the emergence of Phishing-as-a-Service (PhaaS)—a cybercrime model in which professional threat actors develop and sell phishing kits, templates, hosting services, email-sending tools, and victim-credential logs to paying customers. This ecosystem allows individuals with minimal technical expertise to launch effective phishing attacks, increasing both the frequency and severity of global cyber threats.

PhaaS platforms operate predominantly within the Dark Web, an unindexed portion of the internet accessed through anonymity-preserving networks such as Tor. The Dark Web hosts forums, marketplaces, and illicit service advertisements where cybercriminals anonymously trade phishing kits and related services. The dynamic nature of Dark Web content, combined with frequent mirror sites, hidden hyperlinks, and encrypted communication channels, makes detection of PhaaS platforms particularly challenging. Traditional cybersecurity defenses often fail to identify new or emerging PhaaS offerings due to limited visibility and lack of structured intelligence.

Recent advancements in machine learning, natural language processing (NLP), and large-scale data mining provide promising opportunities for analyzing Dark Web content and identifying hidden cybercrime infrastructures. In particular, unsupervised clustering techniques have proven effective in grouping similar threat actors, classifying illicit service types, and uncovering previously unknown relationships within criminal networks. However, limited research focuses specifically on clustering-based identification of PhaaS ecosystems from Dark Web data, highlighting a significant research gap.

This study proposes a comprehensive framework for detecting and classifying PhaaS platforms through Dark Web data collection, text preprocessing, feature extraction, and clustering-based analysis. By leveraging Tor-enabled web scraping, NLP-driven text modeling, and clustering algorithms such as K-Means, DBSCAN, and HDBSCAN, the system aims to reveal hidden PhaaS groups, service patterns, and threat actor linkages. The results contribute to improved cyber threat intelligence, enabling proactive defense mechanisms for security analysts, law enforcement, and SOC teams.

II. PROBLEM STATEMENT

The rise of Phishing-as-a-Service (PhaaS) platforms on the Dark Web has drastically expanded the scale and sophistication of phishing attacks. However, these platforms operate anonymously on unindexed networks, making them extremely difficult to detect, monitor, and classify. Traditional security tools are ineffective in identifying PhaaS marketplaces due to encrypted communication, dynamic URLs, and the rapidly evolving nature of Dark Web ecosystems. There is a need for an automated, data-driven framework capable of collecting Dark Web content, analyzing unstructured text, and clustering related PhaaS services to identify threat actors, platforms, and emerging trends. The problem is the lack of an effective machine-learning based system to detect, classify, and analyze PhaaS platforms using Dark Web data.
Traditional education mechanisms — including static curriculum design, manual grading, teacher-dependent content

delivery, and limited feedback systems — are inherently constrained due to:

Lack of personalization, leading to a one-size-fits-all approach that does not align with individual learning abilities. Inability to detect academic stress, learning gaps, or student disengagement, especially in large classroom settings. As the educational environment shifts toward dynamic, digital, and skill-based learning, these conventional models prove insufficient in addressing the needs of modern learners, educators, and institutions.

III. LITERATURE REVIEW

AI-driven educational systems transform learning from reactive to proactive by incorporating:

Adaptive personalized learning, where content, difficulty level, and learning style change based on learner behavior, pace, and knowledge gaps.

AI-driven academic advising, recommending suitable career paths, certifications, internships, and skill development plans.

NLP-based academic support, including AI chatbots, plagiarism detection, intelligent question generation, and lecture summarization.

This paradigm shift forms the foundation for modern, resilient, student-centered, and data-driven educational ecosystems, where AI acts not as a replacement for teachers but as a co-pilot in personalized learning.

IV. BACKGROUND AND THEORETICAL FOUNDATIONS

A. Phishing Ecosystem and Attack Lifecycle

Phishing attacks typically follow a structured lifecycle consisting of reconnaissance, lure creation, delivery, credential harvesting, and data monetization. Recent studies indicate that automation has shortened this cycle drastically, enabling attackers to launch large-scale operations with minimal effort. PhaaS platforms play a critical role in this evolution by providing ready-made phishing kits, bulletproof hosting, configuration panels, and victim log collection modules. These platforms resemble legitimate SaaS models in terms of scalability, modularity, and customer support mechanisms.

B. Dark Web Structure and Anonymity Models

The Dark Web operates as a subset of the Deep Web accessible through anonymity-enabling technologies such as The Onion Router (Tor). Tor uses layered encryption and multi-hop routing to obscure user identity. Hidden services (.onion sites) allow criminals to operate marketplaces, vendor shops, and forums without revealing server locations. From a research perspective, the anonymity structure introduces considerable challenges in data retrieval, session stability, URL persistence, and ethical compliance. These constraints necessitate advanced scraping protocols and robust data handling techniques.

C. Cybercrime-as-a-Service (CaaS) Model

The CaaS economy includes several categories such as Ransomware-as-a-Service (RaaS), Malware-as-a-Service (MaaS), DDoS-as-a-Service (DDoSaaS), and Phishing-as-a-Service (PhaaS). These ecosystems rely on subscription-based access, affiliate programs, reseller networks, and underground reputation systems. Understanding these economic models provides essential context for analyzing PhaaS listings on the Dark Web. The interaction between vendors and consumers often follows structured patterns that can be detected through clustering analysis.

V. THREAT MODEL AND SYSTEM ASSUMPTIONS

A. Adversary Capabilities

The adversaries considered in this study include PhaaS vendors operating on Dark Web marketplaces and forums. They possess the ability to:

a) Host phishing kits on bulletproof servers

b) Distribute phishing templates and SMTP modules

c) Evade security monitoring by frequently migrating URLs

d) Engage in encrypted communications

B. System Assumptions

a) Dark Web data contains legitimate PhaaS-related listings.

b) Textual descriptions from marketplaces provide enough cues for clustering.

c) No ground truth labels are available, validating the need for unsupervised learning.

d) Scraping maintains compliance with ethical and legal guidelines.

VI. MATHEMATICAL FORMULATION

A. Problem Definition

Given a dataset D of unstructured Dark Web textual content:

$$D = \{d_1, d_2, \dots, d_n\}$$

Convert each document into a vectorized representation:

$$X = f(D)$$

Where *f* represents the feature extraction function (TF-IDF, BERT embeddings).

Apply a clustering function:

$$C = g(X)$$

Where $g \in$ {K-Means, DBSCAN, HDBSCAN}

Aim:

maximize $S(C)$where S is silhouette score

B. TF-IDF Model

$$TFIDF(t,d) = TF(t,d) \times \log \frac{N}{DF(t)}$$

C. Cosine Similarity

$$\cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|}$$

Cosine similarity is used for cluster distance calculations.

VII. RELATED WORK AND COMPARATIVE ANALYS

A.Existing Approaches to Phishing Detection
Prior literature focuses on URL-based detection, email-header analysis, machine learning classification models, and browser blacklist systems. These traditional approaches, while useful for surface-web phishing, fail to address the organized underground economy of PhaaS. Studies by Oest et al. and Gupta et al. demonstrate limitations in scalability and detection of evasive phishing kits.

B.DarkWebIntelligenceModels
Recent work by Althobaiti & Yu and Samtani et al. uses data mining and ML for Dark Web threat intelligence. However, most are focused on malware markets or ransomware ecosystems, not PhaaS networks specifically. Additional systems rely on supervised learning, which is unsuitable due to lack of ground truth labels.

C. Comparison with Proposed Framework
Unlike earlier works, this research:

a) Focuses specifically on PhaaS markets.

b) Uses unsupervised clustering (K-Means, DBSCAN, HDBSCAN).

c) Integrates NLP-based embeddings (BERT, TF-IDF).

d) Performs structural threat-group discovery rather than simple classification.

This establishes novelty and validates scientific contribution.

VIII. METHODOLOGY

The proposed system for identifying Phishing-as-a-Service (PhaaS) platforms integrates Dark Web data acquisition, natural language preprocessing, feature extraction, unsupervised clustering, and multi-metric evaluation. This section presents each stage in a structured and reproducible manner, consistent with IEEE publication standards.

A. Dark Web Data Collection

Dark Web intelligence gathering represents the foundational layer of the proposed framework. Since PhaaS services primarily operate within anonymous and unindexed networks such as Tor, specialized data collection techniques must be employed.

1) Tor Browser Configuration

Data access is facilitated through the Tor network, which anonymizes communication using layered routing across volunteer-operated relays. A dedicated Tor proxy (SOCKS5) is configured for Python scripts, enabling automated crawling of .onion sites while preserving operational security. Connection parameters (e.g., circuit renewal, request delays, and user-agent rotation) are tuned to mitigate risk of connection failure and maintain ethical compliance during scraping operations.

2) Scraping Tools

To extract data programmatically, multiple scraping frameworks are employed:

a) Python Requests with Tor proxy – suitable for static onion sites.

b) Scrapy – scalable crawling of forums and marketplaces.

c) BeautifulSoup – HTML parsing and text extraction.

d) Selenium + Tor Browser – required for JavaScript-heavy pages and login-based vendor panels.

These tools allow harvesting marketplace listings, forum threads, user profiles, product descriptions, pricing details, and seller metadata.

3) Data Sources

The study targets four Dark Web sources frequently used by cybercriminals for selling phishing kits and PhaaS services:

a) Forums – discussions, tutorials, support groups, seller promotions.

b) Marketplaces – structured product listings for phishing kits and hosting.

c) Onion Sites – dedicated PhaaS platforms and standalone storefronts.

d) Vendor Pages – personal shops run by individual threat actors.

These data sources collectively provide diverse content necessary for clustering and classification.

B. Data Preprocessing

Dark Web text is unstructured, multilingual, noisy, and often polluted with advertisements, broken HTML, or obfuscation. To convert raw data into analyzable form, several preprocessing stages are applied.

1) Noise and HTML Removal

Tags, scripts, boilerplate text, and embedded advertisements are removed using HTML parsers. Excess symbols, ASCII art, and encoded escape characters are normalized to plain text.

2) Tokenization and Lemmatization

The cleaned text is tokenized into meaningful units. Lemmatization is applied to reduce morphological variance (e.g., "phishing," "phished," "phisher" → "phish"). This standardization improves vector-space consistency.

3) Stop-Word Removal

Common language stop-words (e.g., *the*, *and*, *is*) and domain-irrelevant noise words are removed. Additionally, Dark Web-specific stop-words (e.g., *bump*, *PM me*, *trusted seller*) are filtered to avoid clustering distortion.

4) Multilingual Normalization

Because Dark Web forums frequently include Russian, English, Spanish, and Arabic content, *lang-detect* models classify language, enabling language-specific preprocessing. Non-English text is machine-translated when domain-relevant.

5) Metadata Extraction

Structured attributes such as seller name, product category, price, date, reputation score, and marketplace segment are parsed and stored. Metadata enhances the signal quality for feature engineering.

C. Feature Engineering

After preprocessing, the textual dataset is transformed into numerical representations suitable for clustering.

1) TF–IDF Vectorization

Term Frequency–Inverse Document Frequency (TF–IDF) converts text into weighted vectors, highlighting discriminative terms specific to PhaaS advertisements (e.g., *SMTP access*, *scampage*, *logs panel*, *OTP bypass*). TF–IDF is effective for high-dimensional sparse textual datasets.

2) Word Embeddings (Word2Vec and BERT)

To capture contextual and semantic meaning, neural embeddings are generated:

a) Word2Vec learns vector similarities between Dark Web terms.

b) BERT embeddings produce contextual sentence-level representations capable of understanding threat-related terminology.

These embeddings significantly enhance cluster separability.

3) Keyword Extraction

Domain-specific keywords—such as "phishing kit," "reloadable logs," "scampage," "SMTP server," "OTP bypass," "botnet integration"—are programmatically extracted. These terms help identify functional capabilities of PhaaS services.

D. Clustering

Unsupervised clustering enables the discovery of latent structures in the Dark Web dataset, grouping similar PhaaS offerings and identifying new or unknown service providers.

1) Algorithms Used

Three clustering algorithms are evaluated:

a) K-Means – partitions vectors into k clusters using centroid optimization. Suitable for large datasets with well-defined clusters.

b) DBSCAN – density-based clustering that identifies noisy outliers; ideal for irregular data common in Dark Web listings.

c) HDBSCAN – an extension of DBSCAN providing hierarchical density clustering, highly effective for mixed-length textual embeddings and noisy datasets.

2) Dimensionality Reduction

To enable visualization and improve clustering efficiency:

a) PCA (Principal Component Analysis) reduces high-dimensional TF-IDF vectors.

b) t-SNE produces non-linear low-dimensional embeddings, ideal for visually interpreting PhaaS service groups.

3) Cluster Interpretation

Each cluster is examined to infer:

a) Seller groups and alliances

b) Service types (phishing kits, hosting, logs, spam tools)

c) Pricing tiers (low-cost kits, premium kits, subscription-based PhaaS)

This analysis reveals operational patterns and criminal ecosystems.

E. Evaluation

To validate clustering effectiveness, multiple quantitative metrics are used.

1) Silhouette Score

Measures cohesion and separation of clusters. A higher score indicates well-defined PhaaS groupings.

2) Davies–Bouldin Index

Quantifies intra-cluster similarity relative to inter-cluster differences. Lower values indicate stronger cluster separability.

3) Cluster Visualization

Using t-SNE or UMAP plots, clusters are visually inspected to assess distribution, density, outliers, and overlaps among PhaaS service categories.

IX. SYSTEM IMPLEMENTATION AND EXPERIMENTAL SETUP

A. Hardware Configuration

Experiments were conducted on a system running Ubuntu 22.04 LTS with:

a) Intel Core i7 processor

b) 32 GB RAM

c) NVIDIA RTX 3060 GPU (for BERT embeddings)

d) Python 3.10 environment

B. Software Stack

a) Scraping: Scrapy, Selenium with Tor proxy

b) NLP: NLTK, Spacy, HuggingFace Transformers

c) ML: Scikit-Learn, HDBSCAN Library

d) Visualization: Matplotlib, Seaborn, Plotly

e) Database: MongoDB for storing raw Dark Web data

C. *Experimental Pipeline*

a) Data Collection Phase
Automated Tor-based crawlers scrape pages at fixed intervals. Onion links updated dynamically to account for marketplace volatility.

1. Preprocessing Phase
HTML stripping, tokenization, lemmatization, multilingual removal.
Metadata extraction includes timestamps, vendor reputation, pricing info.

2. FeatureVectorizationPhase
TF-IDF vectors for sparse lexical representation.
Sentence-BERT embeddings for semantic clustering.

3. Clustering Phase
K-Means for baseline clustering.
DBSCAN & HDBSCAN for density-based grouping of irregular text clusters.
PCA/t-SNE used for 2D visualization.

4. Evaluation Phase
Silhouette score
Davies–Bouldin Index
Manual cluster interpretation using keyword frequency.

X. RESULTS AND ANALYSIS

A. Dataset Characteristics

The dataset used in this study was compiled from multiple Dark Web sources, including underground forums, onion-based marketplaces, private vendor pages, and hidden service directories. The data consisted of unstructured textual content such as service advertisements, phishing kit descriptions, pricing details, vendor profiles, customer reviews, and operational instructions. After manual filtering and automated preprocessing, the final dataset comprised 5,720 cleaned text entries, representing approximately 342 unique sellers across 11 Dark Web markets. Metadata such as timestamps, vendor aliases, cryptocurrency prices, and listing categories were also extracted.
The vocabulary size after preprocessing was approximately 18,000 unique tokens, indicating high linguistic diversity and obfuscation practices commonly used by cybercriminals.

B. Performance Comparison of Clustering Algorithms

Three unsupervised clustering algorithms—K-Means, DBSCAN, and HDBSCAN—were evaluated for their ability to categorize PhaaS-related listings. Performance was assessed using Silhouette Score, Davies–Bouldin Index, and qualitative interpretation of cluster coherence.

a) K-Means performed consistently with TF-IDF vectors, producing well-separated clusters for high-frequency PhaaS terms. It achieved the highest Silhouette Score of 0.61, indicating strong intra-cluster similarity.

b) DBSCAN proved effective at identifying outliers and noise, which is valuable for filtering scam listings and irrelevant posts. However, due to high dimensionality, its performance degraded, with a Silhouette Score of 0.43.

c) HDBSCAN excelled in detecting hierarchical criminal structures and overlapping seller networks. It identified emerging PhaaS operations not captured by K-Means. Although its Silhouette Score was 0.57, qualitative analysis showed superior interpretability.

Overall, HDBSCAN was the most effective for detecting hidden cybercriminal groups, while K-Means generated the cleanest separations.

C. Identified PhaaS Clusters and Threat Groups

The clustering results revealed distinct PhaaS ecosystems, highlighting the diversity of phishing service offerings in the cybercrime marketplace. A total of 8 major clusters were consistently observed across models:

a) Phishing kit developers

b) SMTP/Email spamming services

c) Full PhaaS platforms offering hosting + templates

d) Credential harvesting tools

e) Scampage sellers

f) Bulletproof hosting vendors

g) Combo and log sellers (banking, PayPal, O365)

h) New or small-scale PhaaS operators

These clusters expose the supply-chain-like nature of modern phishing ecosystems. Several repeat vendors were consistently associated with high activity, indicating organized groups rather than isolated actors.

D. Observed Patterns in Pricing and Service Description

A consistent trend emerged across the data:

a) Basic phishing templates were sold at $10–$25.

b) Advanced dynamic templates (with CAPTCHA bypass, MFA prompts) ranged from $30–$120.

c) Full-service PhaaS subscriptions charged between $50–$300 per month, depending on features such as hosting, redirection, encryption, email automation, and log management.

d) Listings often included "support" channels via Telegram/Discord, reflecting increasing commercialization.

Service descriptions frequently contained obfuscated language, such as using symbols, spacing, or misspellings to evade moderation and automated detection.

E. Emerging PhaaS Marketplaces Detected

The clustering model identified several newly emerging markets with rapidly increasing listings. These platforms were not widely reported in existing cybersecurity publications, demonstrating the effectiveness of the proposed detection method.
The analysis suggests a shift toward subscription-based phishing ecosystems, similar to the business models of Malware-as-a-Service and Ransomware-as-a-Service.

XI. DISCUSSION

A. Interpretation of Clusters

The generated clusters reveal the structural organization of PhaaS operations. Clustering exposed not only distinct service types but also cross-market vendor activity, where the same seller distributed services across multiple platforms under different aliases.
The presence of multi-service vendors indicates that PhaaS has matured into a professionalized cybercrime industry, offering layered services similar to legitimate SaaS businesses.

B. Relevance to Cyber Threat Intelligence

The findings have significant implications for cybersecurity practitioners.
Identifying PhaaS clusters enables:

a) Early detection of rising phishing operations

b) Mapping of seller networks and infrastructure dependencies

c) Prioritization of high-risk threat actors

d) Enhanced monitoring of Dark Web platforms

e) Integration with SOC/SIEM systems for proactive alerts

The clustering framework provides a scalable, automated mechanism for threat attribution and campaign prediction, which is increasingly necessary given the evolving sophistication of phishing attacks.

C. Limitations of Dark Web-Based Analysis

While informative, Dark Web analysis has inherent constraints:

a) Volatile URLs – Many markets frequently disappear or migrate.

b) Obfuscation and multilingual content – Criminals intentionally avoid detection.

c) Sampling bias – Access limitations prevent full ecosystem visibility.

d) Incomplete metadata – Vendors often hide operational information.

These limitations can affect dataset completeness and cluster stability.

D. Ethical Considerations and Data Volatility

Ethical challenges include ensuring that research activities do not support, interact with, or contribute to criminal operations. Data collection must avoid transactions and refrain from facilitating cybercrime.
Furthermore, the Dark Web exhibits high volatility, with markets shutting down or changing domains frequently, which complicates longitudinal studies.

XII. CONCLUSION

A. Summary of Findings

This research demonstrates that unsupervised machine learning is an effective approach for identifying Phishing-as-a-Service platforms using Dark Web intelligence. By combining TF-IDF embeddings, advanced preprocessing, and robust clustering algorithms, the system efficiently categorized PhaaS vendors and revealed hidden threat groups.

B. Contributions to PhaaS Detection

The main contributions of this study include:

a) Development of a novel ML-based pipeline for Dark Web PhaaS detection

b) Identification of multiple PhaaS clusters and emerging marketplaces

c) Extraction of pricing trends and operational characteristics

d) Support for cyber threat intelligence through automated classification

These contributions highlight the value of clustering in uncovering the underground phishing ecosystem.

C. Practical Applications for SOC Teams

Security Operations Centers (SOCs) can leverage the findings to:

a) Enhance phishing campaign attribution

b) Improve early-warning detection systems

c) Identify high-risk phishing infrastructure

d) Integrate threat intelligence feeds with SIEM tools

e) Prioritize takedown and blocking strategies

Ultimately, the research lays the groundwork for developing a real-time PhaaS monitoring system, strengthening global cybersecurity defenses.

XIII. FUTURE WORK

A. Real-Time Dark Web Monitoring

Future research should focus on developing continuous surveillance systems capable of streaming data from Dark Web sources in real time. This includes:

a) Automated Tor-based crawling at scheduled intervals.

b) Real-time alerts on emergence of new PhaaS vendors.

c) Integration with SOC dashboards for automated threat ingestion.

B. Integration With Deep-Learning Models

Advanced deep-learning approaches can significantly improve PhaaS detection:

a) Transformer-based encoders for semantic understanding.

b) Graph neural networks (GNNs) for mapping criminal relationships.

c) Autoencoders and deep clustering models for high-dimensional threat categorization.

Such models may reduce error rates caused by slang, obfuscation, or multilingual content.

C. Expanding to Other Crime-as-a-Service (CaaS) Categories

The proposed methodology can be extended to other cybercrime services, including:

a) Ransomware-as-a-Service (RaaS)

b) Malware-as-a-Service (MaaS)

c) DDoS-as-a-Service (DaaS)

d) InfoStealer-as-a-Service

This will enable a comprehensive understanding of the broader CaaS economy and support unified countermeasures.

XIV. REFERENCES

[1] A. Oest, Y. Safaei, P. Zhang, B. Wardman, A. Doupé, Z. Zhao and G.-J. Ahn,
"PhishFarm: A Scalable Framework for Measuring the Effectiveness of Evasion Techniques Against Browser Phishing Blacklists,"
Proc. IEEE Symposium on Security and Privacy, pp. 1349–1366, 2019.

[2] T. Althobaiti and P. S. Yu,
"Detecting Cybercrime Markets on the Dark Web Using Machine Learning Techniques,"
IEEE Access, vol. 8, pp. 211691–211703, 2020.

[3] J. N. García, M. Arias and R. Fonseca,
"Dark Web Mining for Threat Intelligence: A Survey,"
Journal of Cybersecurity, Oxford University Press, vol. 7, no. 1, pp. 1–19, 2021.

[4] S. Samtani, R. Chinn, H. Chen and J. J. Nunamaker,
"Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence,"
Journal of Management Information Systems, vol. 34, no. 4, pp. 1023–1053, 2017.

[5] B. B. Gupta, A. Tewari, A. Jain and D. P. Agrawal,
"Fighting Against Phishing Attacks: State of the Art and Future Challenges,"
Neural Computing and Applications, Springer, vol. 28, no. 1, pp. 3629–3654, 2017.

[6] R. Thomas and A. Mohaisen,
"Measuring the Heavens: Clustering Dark Web Marketplaces,"
Proc. ACM Internet Measurement Conference (IMC), pp. 125–138, 2017.

[7] M. Sabottke, D. Marino, D. H. Kim and J. R. Crandall,
"Exploring the Dark Web: Understanding the Tools and Tactics Used by Cybercriminals,"
Proc. IEEE Security and Privacy Workshops (SPW), pp. 34–41, 2015.

[8] C. D. Manning, P. Raghavan and H. Schütze,
Introduction to Information Retrieval,S
Cambridge University Press, 2008. (TF-IDF & NLP foundation)

[9] T. Mikolov, K. Chen, G. Corrado and J. Dean,
"Efficient Estimation of Word Representations in Vector Space,"
arXiv preprint arXiv:1301.3781, 2013. (Word2Vec)

[10] R. McInnes, J. Healy and J. Melville,
"UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,"
arXiv preprint arXiv:1802.03426, 2018. (Clustering visualization)