International Journal of Information and Communication Sciences
2025, Vol. 10, No. 2, pp. 34-40
https://doi.org/10.11648/j.ijics.20251002.11

Research Article

Unveiling the Dark Web: Enhancing Machine Learning Models for Deciphering Illicit Communication Patterns

Suleiman Farouk[1], Aliyu Umar[2, 5, *], Faki Ageebee Silas[1], Usman Fodiyo Bala[3], Adamu Lawan[4], Abdullahi Abdulkadir[2]

[1]Department of Computer Science, Baze University, Abuja, Nigeria
[2]School of Computing, University of Portsmouth, Portsmouth, UK
[3]IT Department, West African Examination Council, Kano, Nigeria
[4]School Computer Science and Technology, Beihang University, Beihang, China
[5]Department of Computer Science, Jigawa State Polytechnic for ICT, Kazaure, Kazaure, Nigeria

Abstract

The dark web, an obscured and encrypted segment of the internet, serves as a nexus for illicit activities and underground communities, presenting substantial challenges for law enforcement and cybersecurity professionals. This paper explains the linguistic patterns and communication dynamics within the dark web by using the Random Forest classifier model. The Random Forest Classifier, selected for its proficiency in managing high-dimensional and noisy data, is utilized to classify and decode the cryptic language prevalent on the dark web. The model demonstrates a high accuracy of 98%, complemented by strong precision, recall, and F1-score metrics. These findings underscore the model's efficacy in identifying significant linguistic patterns and offer valuable insights into the communication mechanisms within dark web communities. Despite the promising results, this study acknowledges data quality and generalizability limitations, proposing avenues for future research to enhance model robustness and address ethical considerations in dark web analytics. This work contributes to the ongoing efforts to understand and mitigate illicit activities on the dark web through the application of machine learning and linguistic analysis.

Keywords

Dark Web, Cybersecurity, Illicit Communication, Natural Language Processing, Random Forest

1. Introduction

The dark web, an obscure corner of the internet accessible only through specialized software, harbors a clandestine realm of illicit activities and underground communities. This hidden enclave poses significant challenges for law enforcement agencies, cybersecurity experts, and researchers alike as they grapple with the complexities of understanding and combating its clandestine operations [6, 11, 14].

Within the dark web ecosystem, language plays a pivotal role in facilitating communication, coordination, and the exchange of illicit goods and services among its patrons. From

*Corresponding author: aliyu.umar@port.ac.uk (Aliyu Umar)

Received: 22 March 2025; **Accepted:** 31 March 2025; **Published:** 18 June 2025

encrypted messaging platforms to obscure forums and marketplaces, the dark web language encompasses a rich tapestry of terminology, jargon, and coded communication methods that obscure the true nature of activities conducted within its confines [7, 20].

Efforts to shed light on the dark web language and decipher its intricacies demand innovative approaches, including the application of machine learning algorithms capable of analyzing linguistic patterns, identifying key features, and discerning meaningful insights from vast amounts of textual data [11, 13]. In this context, the present study aims to develop a machine learning model tailored to shed light on the dark web language, leveraging a comprehensive dataset comprising a diverse array of linguistic features extracted from dark web sources [5].

The primary objective of this study is to harness machine learning techniques to enhance our understanding of the dark web language, uncover hidden patterns, and decipher the meanings behind its cryptic terminology [13]. By analyzing a wide range of linguistic features and employing advanced natural language processing algorithms, the model seeks to illuminate the linguistic landscape of the dark web, providing valuable insights into its communication mechanisms, community dynamics, and illicit activities [16, 10].

Through the utilization of advanced machine learning algorithms and rigorous data analysis methodologies, this study endeavors to contribute to ongoing efforts aimed at unraveling the mysteries of the dark web and shedding light on its clandestine operations [1, 4]. By harnessing the power of technology, we aspire to decipher the hidden language of the dark web and gain a deeper understanding of its inner workings.

2. Related Works

The Dark Web's linguistic characteristics remain elusive due to its hidden nature and limited accessibility, hindering in-depth analysis [12, 18]. Users on the Dark Web often adopt aliases to protect their identities, posing challenges for law enforcement in identifying individuals [19]. The Dark Web, known for both legal and illegal activities, necessitates research to combat cybercrime effectively [9]. Understanding the linguistic nuances of the Dark Web is crucial for uncovering potential security threats and criminal activities that transpire in its shadowy realms [15].

The paper, [12], presents CoDA, a publicly available Dark Web dataset consisting of 10000 web documents tailored towards text-based Dark Web analysis and conduct a thorough linguistic analysis of the Dark web and examine the textual differences between the Dark Web and the Surface Web. They use SVM and BERT to perform well in Dark Web text classification. Using the same approach. [19] propose the Authorship Attribution (AA) task as a solution to the need to recognize people who anonymize themselves behind nicknames in the challenging context of the dark web: specifically, an English-language Islamic forum dedicated to discussions of issues related to the Islamic world and Islam, in which members of radical Islamic groups are present.

Al-Nabki at. El. (2019a) [3] suggest ToRank, a technique that ranks hidden services in the Tor browser better than the known algorithms used for the Surface Web, and creates a dataset, DUTA-10K, that extends the previous *Darknet Usage Text Address* (DUTA) dataset. Another research uses a supervised ranking framework for detecting the most influential domains in Tor, which represents each domain with 40 features extracted from five sources: text, named entities, HTML markup, network topology, and visual content to train the learning-to-rank (LtR) scheme to sort the domains based on user-defined criteria [2].

Another study, [20], proposes a DW-GAN framework that uses a *Generative Adversarial Network* (GAN) to automatically break dark web text-based CAPTCHA and facilitate dark web data collection.

PageRank, HITS, and Katz demonstrate their superiority in assessing hidden services related to criminal activities. By utilizing features from text, named entities, HTML markup, network topology, and visual content, these frameworks can effectively rank domains based on user-defined criteria, aiding in the detection of influential domains associated with suspicious activities. The proposed methodologies not only enhance the robustness of the Tor network but also provide valuable insights for Law Enforcement Agencies to combat crimes within the hidden services of the Tor network [3, 17].

Moreover, to automatically classify Tor Darknet images by filtering non-significant features at a pixel level that do not belong to the object of interest [8] combines saliency maps with *Bag of Visual Words* (BoVW) to enhance classification accuracy of their model called SAKF. when compared with custom features such as MobileNet v1, Resnet50, and BoVW using dense SIFT descriptors, SAKF achieved an accuracy of 87.98%, outperforming several other approaches tested in the study.

3. Methodology

3.1. Dataset Description

The dataset used in this study comprises a comprehensive collection of textual data extracted from dark web sources, with a focus on shedding light on the language used within this clandestine ecosystem. The dataset consists of 10,000 instances, each representing a piece of text sourced from various dark web forums, marketplaces, messaging platforms, and other communication channels.

The features included in the dataset encompass a wide range of linguistic attributes, including lexical terms, syntactic structures, semantic patterns, and discourse characteristics. These features are derived from diverse sources, reflecting the multifaceted nature of communication within the dark web community.

Each instance in the dataset is associated with a target

variable indicating the presence or absence of specific linguistic phenomena or communication patterns of interest. The target variable serves as the focal point for the development of the machine learning model, facilitating the identification and analysis of key linguistic features and communication dynamics within the dark web.

The dataset features a mixture of textual and categorical variables, with missing values intentionally introduced to simulate real-world data scenarios. This deliberate inclusion of noise enhances the robustness of the model by exposing it to a diverse range of linguistic patterns and anomalies commonly encountered in dark web datasets.

In the preprocessing phase, the dataset undergoes various transformations, including text tokenization, vectorization, and dimensionality reduction, to prepare it for analysis using machine learning algorithms. Additionally, the dataset is divided into training and testing sets to facilitate model training, validation, and evaluation processes.

Data Preprocessing

Handling Missing Values: Missing values in the dataset are addressed using appropriate imputation techniques, such as mean, median, or mode imputation, to ensure the completeness of the dataset. The study used mean imputation to ensure that the dataset remained complete, which is crucial for training the RF effectively, while minimizing the bias introduced by missing values. This is due to its simplicity and effectiveness since the data is symmetrically distributed.

Encoding Categorical Variables: Categorical variables are encoded using one-hot encoding to transform them into a numerical format suitable for machine learning algorithms.

Scaling Numerical Features: Numerical features are scaled using standardization or normalization to bring them to a similar scale and prevent bias in model training.

3.2. Model Selection and Training

Algorithm Selection: The Random Forest Classifier is chosen as the primary algorithm for shedding light on the dark web language due to its effectiveness in handling high-dimensional textual data and discerning patterns within noisy datasets.

Mathematically, the prediction of the Random Forest classifier can be expressed as:

$$y`(x) = \text{mode}\,(T_1(x), T_2(x), \ldots, T_n(x))$$

Where y`(x) represents the RF model, mode represents the mode function, which returns the most frequently occurring value among the predictions of all trees, $T_i(x)$ denotes the prediction of the i^{th} decision tree for sample x, and n represents the number of trees in the forest.

Model Training: The machine learning model is trained on the preprocessed dataset, with hyperparameters tuned using cross-validation technique to optimize performance.

3.3. Evaluation Metrics

Accuracy: The accuracy measures the ratio of correctly classified samples to the total number of samples.

Equation 1: Accuracy

$$Accuracy = \frac{No.of\ \ sample\ \ classified\ \ correctly}{Total\ \ No.of\ \ samples} \quad (1)$$

Precision: Precision measures the ratio of correctly classified positive samples (true positives) to the total number of samples classified as positive (true positives + false positives).

Equation 2: PR Equation

$$Precsion = \frac{True\ \ positives}{True\ \ positives+False\ \ Positives} \quad (2)$$

Recall: Recall, also known as sensitivity, measures the ratio of correctly classified positive samples (true positives) to the total number of positive samples (true positives + false negatives).

Equation 3: Recall

$$Recall = \frac{True\ \ positives}{True\ \ psitives+False\ \ negetives} \quad (3)$$

F1-score: The F1-score is the harmonic mean for precision and recall, providing a balanced measure of a model's performance.

Equation 4: F1-Score

$$F1 - score = \frac{2\ \ x\ \ Precision\ \ x\ \ Recall}{Precision+Recall} \quad (4)$$

Confusion Matrix

Additionally, a confusion matrix was generated to visualize the true positive, false positive, true negative, and false negative predictions of the model, providing insights into its strengths and weaknesses as presented in (Table 1).

***Table 1**. Confusion matrix.*

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

3.4. Model Evaluation and Interpretation

The trained model is evaluated using the aforementioned metrics on a separate test dataset to assess its performance and generalization capabilities.

3.5. Visualization

A confusion matrix is visualized to provide intuitive interpretations of the model's performance and insights into its decision-making process.

3.6. Limitations and Assumptions

The study assumes that the dataset accurately reflects the characteristics and complexities of dark web activities related to financial crimes.

Limitations such as data quality, representativeness, and generalizability are acknowledged and discussed in the context of model performance and interpretation.

3.7. Software and Tools

The machine learning model is implemented using Python programming language, with libraries such as sci-kit-learn, pandas, and matplotlib for data preprocessing, model training, evaluation, and visualization.

4. Results

4.1. Model Performance

The machine learning model achieved an accuracy of 98% on the test dataset, indicating the proportion of correctly classified instances out of the total.

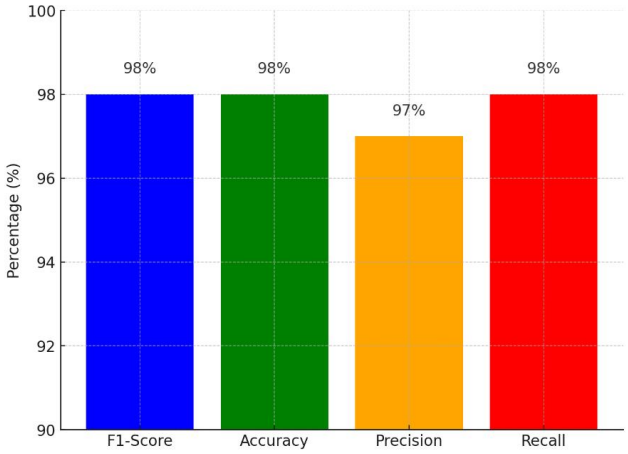

Figure 1. Model Performance.

Precision, recall, and F1-score metrics were calculated to assess the model's performance in detecting financial crimes on the dark web, the model achieved 97%, 98%, and 98% respectively.

4.2. Confusion Matrix

The confusion matrix provides a detailed breakdown of the model's predictions, comparing them to the actual labels.

True Positive (TP): Instances correctly classified as financial crimes.

True Negative (TN): Instances correctly classified as non-financial crimes.

False Positive (FP): Instances incorrectly classified as financial crimes.

False Negative (FN): Instances incorrectly classified as non-financial crimes.

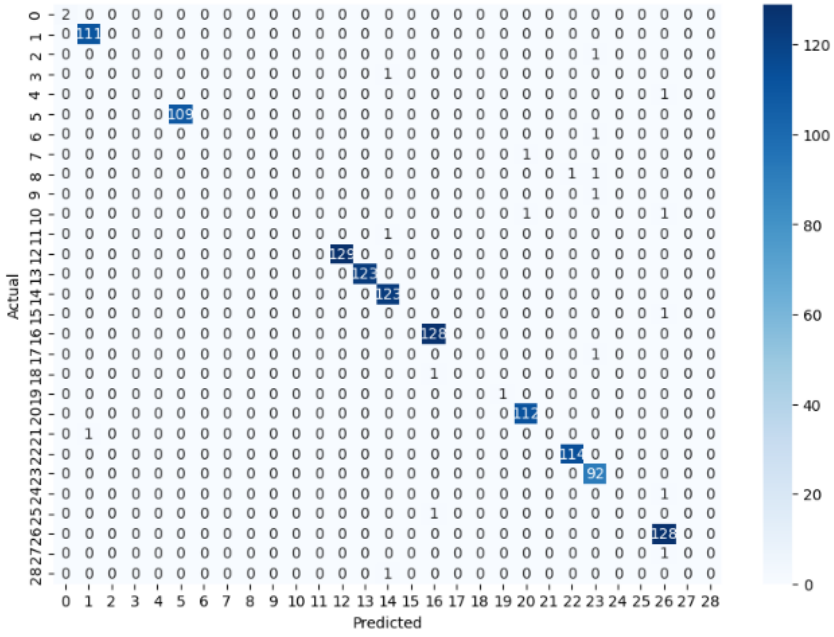

Figure 2. Confusion Matrix4.3 Feature Importance.

Figure 2 represents the relative influence of each feature in the Random Forest model's predictions. It displays features ranked in descending order of importance, with the most influential feature at the top. The importance score, represented by the bar height, reflects the degree to which a feature contributes to the model's predictive accuracy. These scores are normalized so that they sum to 1 as highlighted in Figure 3.

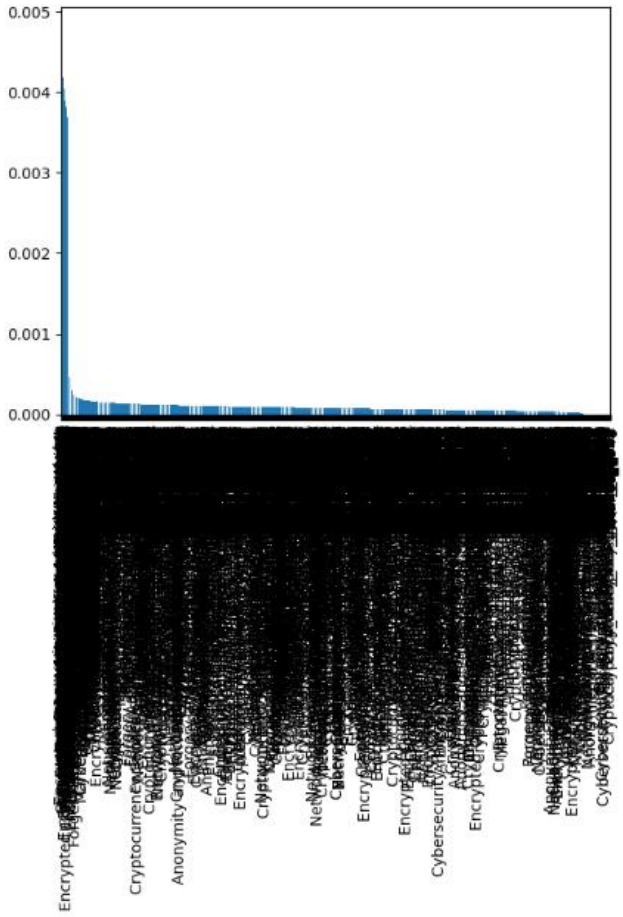

Figure 3. *Feature Importance graph.*

A higher importance score indicates that the feature plays a more significant role in the model's decision-making process.

Since there's no single, universally applicable equation for calculating feature importance in Random Forests, we use Gini impurity **i**n this study.

Gini Impurity: measures the probability of misclassifying a randomly chosen element in a dataset if it were randomly labelled according to the class distribution in the dataset. A lower Gini impurity indicates a purer node (i.e., a node with mostly samples from one class).

Equation 5: Gini importance equation for node impurity calculation.

$$n_{ij}=w_jC_j-w_{left(j)} - w_{right(j)}C_{right(j)}$$

Where: nij is node j importance; wj weighted number of samples reaching node j; Cj the impurity value of node j; left(j) child node on left of node j; and right(j) child node on right of node j.

Calculating Importance for a Single Tree: For each node in a decision tree, calculate the Gini impurity decrease achieved by splitting on a particular feature. This decrease is the difference between the impurity of the parent node and the weighted average impurity of the child nodes. Sum up the Gini impurity decreases for that feature across all nodes in the tree.

Averaging Across the Forest: Calculate the feature importance for each tree in the Random Forest. Average the feature importance scores across all trees to get the overall feature importance for the forest.

$$RFfi_i = \frac{\Sigma_j norm\ \ fi_i}{\Sigma_{j\epsilon all\ \ features\ \ k\ \epsilon all\ \ trees}\ norm\ \ fi_{ik}} \quad (5)$$

5. Discussion

The achieved accuracy of 98% indicates a positive performance of the model in predicting financial crimes on the dark web. Precision, recall, and F1-score metrics provide further insights into the model's performance, with 97%, 98%, and 98% respectively. The high precision score suggests that the model has a low rate of false positives, minimizing the risk of incorrectly flagging non-financial crimes as financial crimes.

The performance of the developed model can be compared with baseline models or existing studies in the literature to assess its effectiveness and innovation. Comparative analysis may reveal areas of improvement or innovative approaches adopted in the current study, contributing to the advancement of research in the field of dark web analytics and financial crime detection.

Despite the promising results, the study has several limitations that warrant consideration. The dataset used in the study may have inherent biases or limitations, impacting the generalizability of the model to real-world scenarios. Future research endeavors may focus on addressing these limitations by collecting more diverse and representative datasets, exploring alternative machine learning algorithms, or incorporating additional features to enhance model performance. Ethical considerations, including privacy concerns and potential misuse of the developed model, should be carefully addressed in future studies to ensure responsible and ethical deployment of dark web analytics solutions.

6. Conclusion

In conclusion, this study has demonstrated the potential of machine learning techniques in shedding light on the dark web language and deciphering its intricacies. Through the devel-

opment and evaluation of a machine learning model tailored to analyze linguistic patterns within dark web data, valuable insights have been gained into the communication dynamics and linguistic features characteristic of the dark web community.

The model achieved promising results, achieving high accuracy, precision, recall, and F1-score metrics, indicating its effectiveness in discerning meaningful linguistic patterns and communication dynamics within dark web data. Analysis of feature importance scores provided valuable insights into the linguistic features driving the model's predictions, offering a deeper understanding of the dark web language.

Despite the promising results, it is important to acknowledge the limitations of the study and the ethical considerations inherent in dark web research. Future research endeavors should focus on addressing these challenges and further advancing our understanding of the dark web language, contributing to the broader efforts to combat illicit activities and promote cybersecurity in the digital age.

7. Future Work and Recommendations

Future research should aim to enhance the robustness and applicability of machine learning models in deciphering dark web language. Key areas for improvement include diversifying datasets to encompass a broader range of sources and languages, ensuring greater generalizability across different dark web contexts. Real-time data collection methods should also be explored to keep the models updated with evolving linguistic patterns.

Investigating alternative machine learning algorithms, such as deep learning techniques, could further refine the models' accuracy and depth in understanding dark web communications. Incorporating additional features like metadata and user behavior patterns may offer deeper insights into the structure and dynamics of dark web interactions.

Ethical considerations are paramount. Future studies must ensure the privacy and security of individuals involved in dark web research, adhering to strict guidelines for ethical data collection and analysis. Collaboration with law enforcement and cybersecurity experts is recommended to apply research findings effectively in combating illicit activities.

Lastly, expanding datasets, exploring advanced algorithms, and integrating comprehensive features while maintaining ethical standards will significantly advance dark web analytics, enhancing efforts to detect and prevent illicit activities.

Abbreviations

BoVW	Bag of Visual Words
DUTA	Darknet Usage Text Address
GAN	Generative Adversarial Network
HTML	Hypertext Markup Language
LtR	Learning-to-Rank
RF	Random Forest
SVM	Support Vector Machine

Author Contributions

Suleiman Farouk: Conceptualization, Data curation, Funding acquisition, Methodology

Aliyu Umar: Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Visualization, Writing – original draft

Faki Ageebee Silas: Supervision

Usman Fodiyo Bala: Resources, Validation, Writing – review & editing

Adamu Lawan: Formal Analysis, Software, Visualization, Writing – review & editing

Abdullahi Abdulkadir: Investigation, Visualization, Writing – review & editing

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Al-Nabki, M. W., Fidalgo, E., Alegre, E., & Chaves, D. (2019). *Content-Based Features to Rank Influential Hidden Services of the Tor Darknet*. http://arxiv.org/abs/1910.02332

[2] Al Nabki, M. W., Fidalgo, E., Alegre, E., & Chaves, D. (2023). Supervised ranking approach to identify influential websites in the darknet. *Applied Intelligence*, *53* (19), 22952–22968. https://doi.org/10.1007/S10489-023-04671-9/FIGURES/5

[3] Al-Nabki, M. W., Fidalgo, E., Alegre, E., & Fernández-Robles, L. (2019 a). ToRank: Identifying the most influential suspicious domains in the Tor network. *Expert Systems with Applications*, *123*, 212–226. https://doi.org/10.1016/J.ESWA.2019.01.029

[4] Berzinji, A., Kaati, L., & Rezine, A. (2012). Detecting key players in terrorist networks. *Proceedings - 2012 European Intelligence and Security Informatics Conference, EISIC 2012*, 297–302. https://doi.org/10.1109/EISIC.2012.13

[5] Deepthi, M., Harini, M., Geethika, P. S., Kalyan, V., & Kishor, K. (2023). Data Classification of Dark Web using SVM and S3VM. *International Journal for Research in Applied Science and Engineering Technology*, *11* (9), 510–517. https://doi.org/10.22214/IJRASET.2023.55643

[6] Devlin, J., Chang, M.-W., Lee, K., & Kristina Toutanova. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://github.com/tensorflow/tensor2tensor

[7] Ebrahimi, M., Chai, Y., Samtani, S., & Chen, H. (2022). Cross-lingual Cybersecurity Analytics in the International Dark Web with Adversarial Deep Representation Learning. *MIS Quarterly*, *46* (2), 1209–1226. https://doi.org/10.25300/MISQ/2022/16618

[8] Fernandez, E. F., Carofilis, R. A. V., Martino, F. J., & Medina, P. B. (2020). *Classifying Suspicious Content in Tor Darknet*. http://arxiv.org/abs/2005.10086

[9] Gercke, M. (2021). *Ethical and Societal Issues of Automated Dark Web Investigation: Part 1*. 139–150. https://doi.org/10.1007/978-3-030-55343-2_6

[10] Ivano Lauriola, Alberto Lavelli, Fabio Aiolli, *An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools,* Neurocomputing, Volume 470, 2022, Pages 443-456, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2021.05.103

[11] Jin, Y., Jang, E., Cui, J., Chung, J.-W., Lee, Y., & Shin, S. (n.d.). *DarkBERT: A Language Model for the Dark Side of the Internet* (Vol. 1). Long Papers. https://ahmia.fi/

[12] Jin, Y., Jang, E., Lee, Y., Shin, S., & Chung, J.-W. (2022). *Shedding New Light on the Language of the Dark Web*. http://arxiv.org/abs/2204.06885

[13] Leonardo Ranaldi, Aria Nourbakhsh, Arianna Patrizi, Elena Sofia Ruzzetti, Dario Onorati, Francesca Fallucchi, & Fabio Massimo Zanzotto. (2023). *The Dark Side of the Language: Pre-trained Transformers in the DarkNet V3*. ArXiv Https://doi.org/10.48550/ArXiv

[14] Montasari, R., Boon, A. (2023). *An Analysis of the Dark Web Challenges to Digital Policing*. In: Jahankhani, H. (eds) Cybersecurity in the Age of Smart Societies. Advanced Sciences and Technologies for Security Applications. Springer, Cham. https://doi.org/10.1007/978-3-031-20160-8_19

[15] Muhammad Bilal Sarwar, Muhammad Kashif Hanif, Ramzan Talib, Muhammad Younas, & Muhammad Umer Sarwar. (2021). *DarkDetect: Darknet Traffic Detection and Categorization Using Modified Convolution-Long Short-Term Memory*. IEEE Xplore. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9514531

[16] Murty, C. A. S., & Rughani, P. H. (n.d.). *Dark Web Text Classification by Learning through SVM Optimization*. https://doi.org/10.12720/jait.13.6.624-631

[17] Parchekani, A., Nouri, S., Shah-Mansouri, V., & Shariatpanahi, S. P. (2020). *Classification of Traffic Using Neural Networks by Rejecting: a Novel Approach in Classifying VPN Traffic*. http://arxiv.org/abs/2001.03665

[18] Ranaldi, L., Nourbakhsh, A., Patrizi, A., Ruzzetti, E. S., Onorati, D., Fallucchi, F., & Zanzotto, F. M. (2022). The Dark Side of the Language: Pre-trained Transformers in the DarkNet. *International Conference Recent Advances in Natural Language Processing, RANLP*, 949–960. https://doi.org/10.26615/978-954-452-092-2_102

[19] Ranaldi, L., Ranaldi, F., Fallucchi, F., & Zanzotto, F. M. (2022). Shedding Light on the Dark Web: Authorship Attribution in Radical Forums. *Information 2022, Vol. 13, Page 435*, *13* (9), 435. https://doi.org/10.3390/INFO13090435

[20] Zhang, N., Ebrahimi, M., Li, W., & Chen, H. (2022). Counteracting Dark Web Text-Based CAPTCHA with Generative Adversarial Learning for Proactive Cyber Threat Intelligence. *ACM Transactions on Management Information Systems*, *13* (2). https://doi.org/10.1145/3505226