International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

DARK SIDE OF THE WEB: DARK WEB CLASSIFICATION BASED ON TEXTCNN AND TOPIC MODELING WEIGHT

MITTA SAI CHARAN , Ms. S. FARHEEN TAJ
MCA Student, Assistant Professor
DEPT OF MCA
PVKK INSTITUTE OF TECHNOLOGY(AUTONOMOUS), Anantapuramu – 515001 (A.P)
mscharan249@gmail.com, farheentaj.2001@gmail.com

ABSTRACT

The DarkWeb is an internet domain that ensures user anonymity and has increasingly become a focal point for illegal activities and a repository for information on cyberattacks owing to the challenges in tracking its users. This study examined the classification of the Dark Web in relation to these cyber threats. We processed Dark Web texts to extract vector types suitable for machine learning classification. Traditional methods utilizing the entirety of Dark Web texts to generate features result in vectors including all words found on the DarkWeb. However, this approach incorporates extraneous information in the vectors, diminishing learning effectiveness and extending processing duration. The research aimed to optimize the classification process by selectively focusing on keywords within each class, thereby curtailing word vector dimensions. This optimization was facilitated by leveraging the anonymity characteristic of the Dark Web and employing topic-modeling-based weight generation. These methods enabled the creation of word vectors with a constrained feature set, enhancing the distinction of Dark Web classes. To further improve classification performance, we integrated TextCNN with topic modeling weights. For validation, we employed two datasets and compared the performance of the model with other text classification algorithms, where the proposed model demonstrated superior effectiveness in Dark Web classification.

1. INTRODUCTION

1.1 Purpose

The purpose of the system is to enhance the detection and analysis of cyber threats within the Dark Web. By leveraging advanced machine learning methods such as TextCNN integrated with topic modeling-based weights, the system offers a robust mechanism to identify and categorize malicious activities and information. This enables researchers, cybersecurity professionals, and law enforcement to track and mitigate cyber threats more effectively.

1.2 Scope

The scope of this study lies in classifying the Dark Web texts into distinct categories relevant to cyber threats, enabling a deeper understanding of illicit activities and their patterns. By focusing on optimizing classification through selective keyword-based vectorization, the study aims to overcome the inefficiencies of traditional methods. This innovation ensures that the generated word vectors are concise and relevant, significantly improving processing speed and accuracy.

1.3 Need for System

The need for such a system arises from the growing prevalence of cyber threats originating from the Dark Web and the limitations of existing methods in accurately classifying its content. Traditional approaches often include unnecessary information, leading to inefficient processing and poor model performance. The proposed system addresses these challenges by employing a focused and optimized classification methodology, ensuring better resource utilization, faster processing

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

times, and higher accuracy in identifying potential threats. This system is critical in enhancing cyber defense capabilities and safeguarding digital ecosystems.

2. SOFTWARE REQUIREMENT ANALYSIS AND SPECIFICATION

2.1. RELATED WORK

1. S. Dolev and S. Lodha, In Proceedings of the First International Conference, CSCML 2017, Beer-Sheva, Israel, June 29-30, (2017).

This book constitutes the proceedings of the first International Symposium on Cyber Security Cryptography and Machine Learning, held in Beer-Sheva, Israel, in June 2017. The 17 full and 4 short papers presented include cyber security; secure software development methodologies, formal methods semantics and verification of secure systems; fault tolerance, reliability, availability of distributed secure systems; game-theoretic approaches to secure computing; automatic recovery of self-stabilizing and self-organizing systems; communication, authentication and identification security; cyber security for mobile and Internet of things; cyber security of corporations; security and privacy for cloud, edge and fog computing; cryptography; cryptographic implementation analysis and construction; secure multi-party computation; privacy-enhancing technologies and anonymity; post-quantum cryptography and security; machine learning and big data; anomaly detection and malware identification; business intelligence and security; digital forensics; digital rights management; trust management and reputation systems; information retrieval, risk analysis, DoS

2. G. A. Wang, M. Chau, and H. Chen., Proceedings. Cham, Switzerland: Springer, May 23, (2017).

The recent rapid growth in big data, networking, and machine learning is due to exponential advances in processing, storage, and network technologies. As the world becomes more digitalized, there is a greater need for comprehensive and sophisticated security technologies and strategies to address the increasingly complex nature of cyber-attacks. This project examines how machine learning, big data is being used in cyber security, both in defence and offense, with a focus on cyber-attacks against machine learning models. Machine learning can be used to carry out cyber-attacks, such as smart botnets, sophisticated spear fishing, and evasive malware. In the field of defence, big data analytics refers to the ability to collect large volumes of digital data in order to analyse, visualize, and derive knowledge that can help predict and prevent cyber-attacks. It gives us a stronger cyber defence stance when combined with security technologies. They allow businesses to identify patterns of behaviour that indicate network threats. Machine learning is used in cyber security for threat identification and prevention, malware detection and classification, and network risk rating, among other things.

2.2. PRODUCT ARCHITECTURE

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

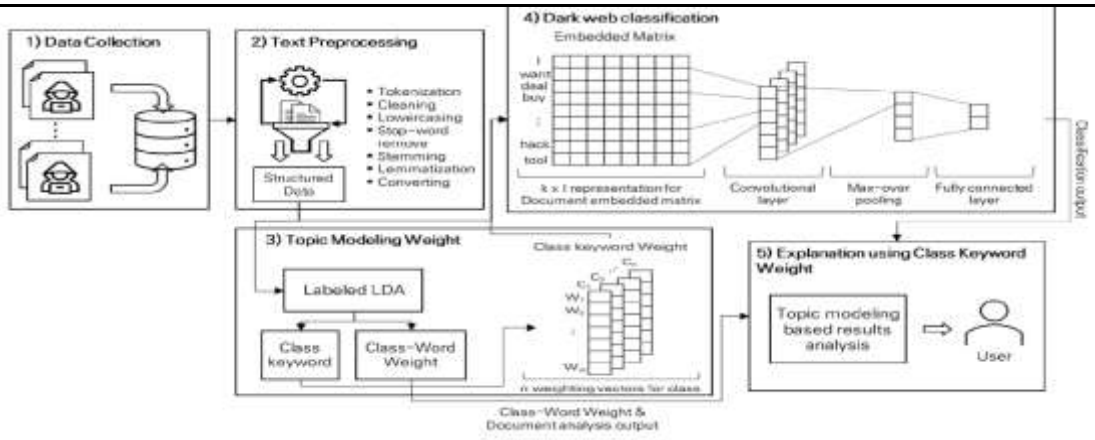


Fig.1.1: System Architecture

Existing System

- Utilizes traditional machine learning methods with all words in Dark Web texts as features.
- Generates high-dimensional word vectors that include irrelevant information.
- Processing duration is extended due to redundant data in feature vectors.
- Classification accuracy is limited by noise and extraneous data.

Disadvantages

- Inefficient processing due to high-dimensional vectors.
- Difficulty in distinguishing relevant content from irrelevant data.
- Poor classification performance for complex Dark Web datasets.
- Increased computational overhead.

Proposed System

- Employs keyword-based feature selection to reduce word vector dimensions.
- Leverages topic modeling for weight generation to enhance feature relevance.
- Integrates TextCNN with topic modeling weights for improved classification.
- Focuses on enhancing classification accuracy and processing efficiency.

Advantages

- Reduces computational complexity by constraining word vector dimensions.
- Improves classification performance with targeted keyword focus.
- Accelerates processing times by excluding extraneous data.
- Facilitates better distinction of Dark Web classes and categories.

PRODUCT FUNCTIONS

- **Data Collection Module**: Gathers Dark Web text data from relevant sources.
- **Preprocessing Module**: Cleans and preprocesses text data for analysis.
- **Feature Extraction Module**: Extracts keyword-based vectors using topic modeling.
- **Model Integration Module**: Implements TextCNN with topic-based weight vectors.
- **Classification Module**: Categorizes Dark Web data into relevant classes.
- **Validation and Performance Analysis Module**: Compares proposed model with existing algorithms for accuracy and efficiency.

3.PROPOSED ALGORITHMS:

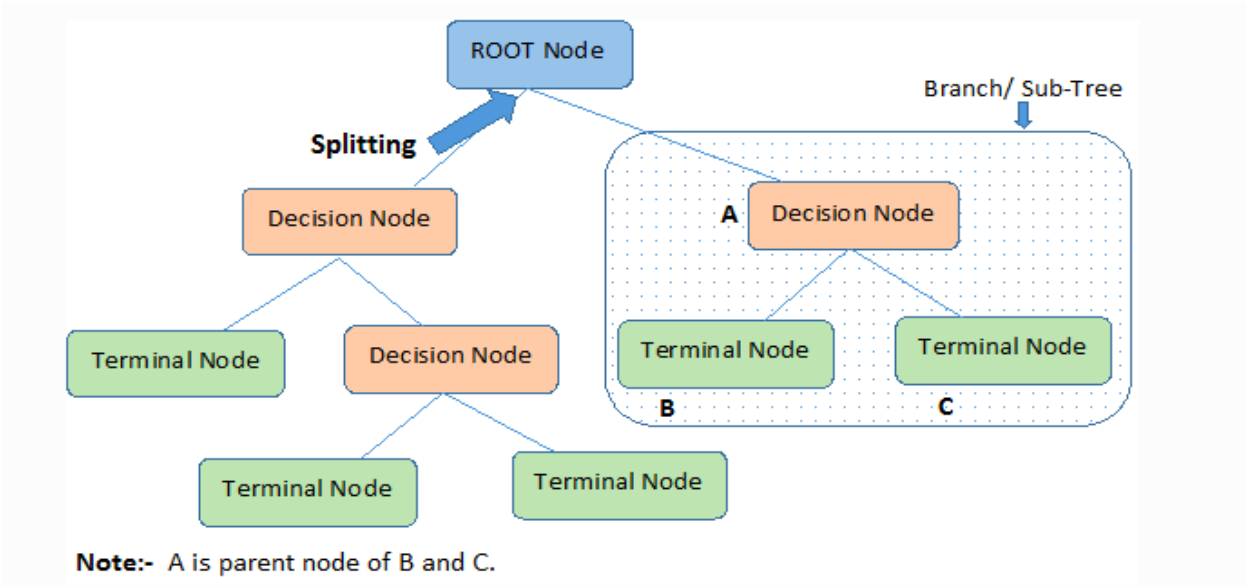
DECISION TREE:

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Decision trees are non-parametric supervised learning Method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

A decision tree is drawn upside down with its root at the top. In the image on the left, the bold text in black represents a condition/internal node, based on which the tree splits into branches/ edges. The end of the branch that doesn't split anymore is the decision/leaf, in this case, whether the passenger died or survived, represented as red and green text respectively.

Block Diagram for Decision Tree Algorithm:



Root Node: It represents the entire population or sample and this further gets divided into two or more homogeneous sets.

Splitting: It is a process of dividing a node into two or more sub-nodes.

Decision Node: When a sub-node splits into further sub-nodes, then it is called the decision node.

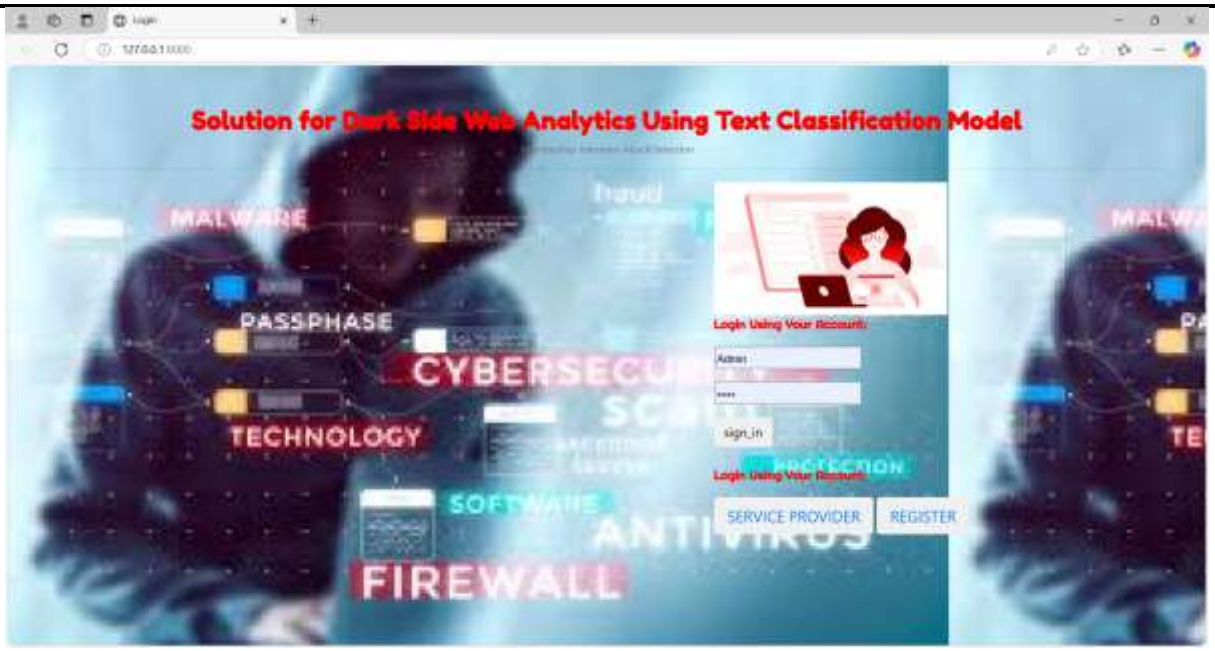
Leaf / Terminal Node: Nodes do not split is called Leaf or Terminal node.

Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.

Branch / Sub-Tree: A subsection of the entire tree is called branch or sub-tree.

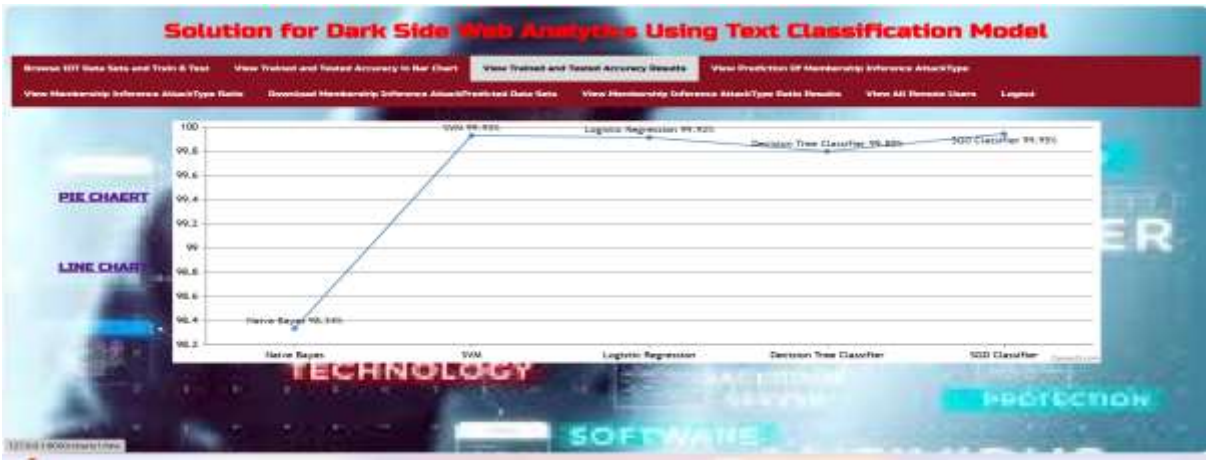
4. SAMPLE SCREENS

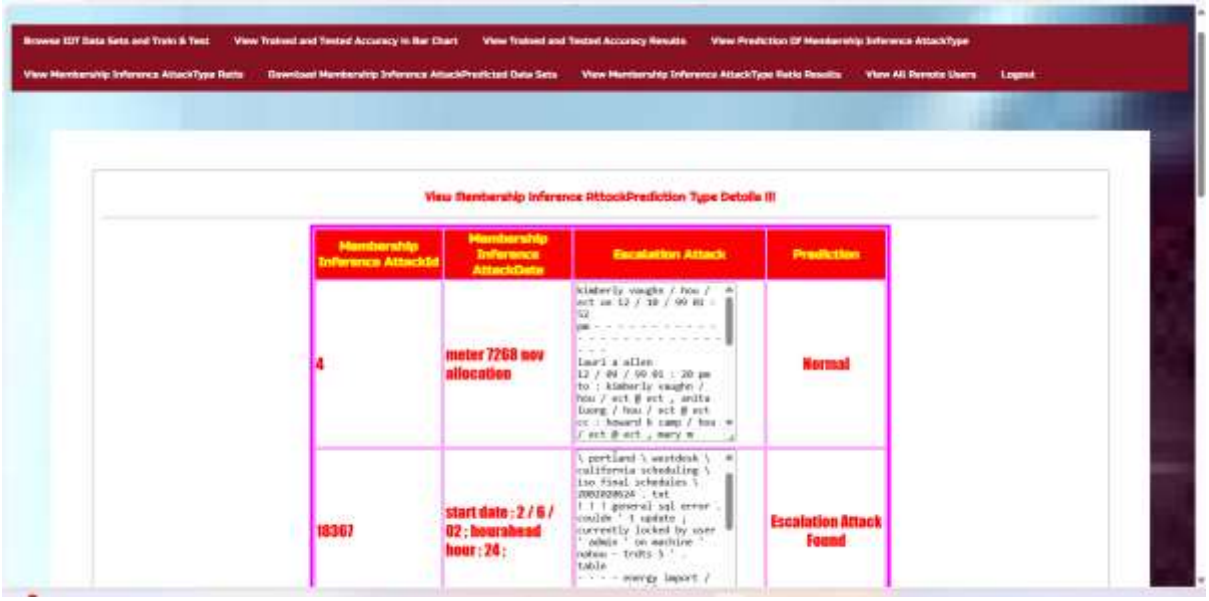
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper



International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper





5.CONCLUSION

This research introduces a DarkWeb classification methodology integrating topic modeling weights with TextCNN. The Dark Web data collected included an extensive vocabulary, not all of which was conducive to effective classification. Thus, a preprocessing step to eliminate extraneous words was essential. We tackled this challenge using topic modeling weights. Initially, structured data were created via text preprocessing, followed by the identification of principal keywords for each class using LLDA. These keywords were then applied to the DarkWeb data to exclude irrelevant words, resulting in a significant reduction of the word vector size by approximately 300 times. The word vectors were instrumental in forming an embedded matrix, which served as the input for the DarkWeb classification model. We utilized topic modeling weights in our classification process to

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

discern the influential keywords and their respective impact. The efficacy of our proposed approach was assessed using two labeled Dark Web datasets, validated through comparative analysis with various text classification algorithms and prior studies. The experimental outcomes verified the efficiency of the proposed method in reducing the vocabulary size required for learning and its superior performance.

BIBLIOGRAPHY

[1] 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. Eur. Chapter Assoc. Comput. Linguistics, vol. 4, 2017, pp. 35–43, doi: 10.18653/v1/E17-1004.

[2] G. Cascavilla, D. A. Tamburri, and W.-J. Van Den Heuvel, ‘‘Cybercrime threat intelligence: A systematic multi-vocal literature review,’’ Comput. Secur., vol. 105, Jun. 2021, Art. no. 102258, doi: 10.1016/j.cose.2021.102258.

[3] J. Saleem, R. Islam, and M. A. Kabir, ‘‘The anonymity of the dark Web: A survey,’’ IEEE Access, vol. 10, pp. 33628–33660, 2022, doi: 10.1109/ACCESS.2022.3161547.

[4] M. Balduzzi andV. Ciancaglini, ‘‘Cybercrime in the deep web,’’ Black Hat, Amsterdam, The Netherlands, Tech. Rep., 2015. [5] S. Rahamat Basha and J. K. Rani, ‘‘A comparative approach of dimensionality reduction techniques in text classification,’’ Eng., Technol. Appl. Sci. Res., vol. 9, no. 6, pp. 4974–4979, Dec. 2019, doi: 10.48084/etasr.3146.

[6] K. N. Singh, S. D. Devi, H. M. Devi, and A. K. Mahanta, ‘‘A novel approach for dimension reduction using word embedding: An enhanced text classification approach,’’ Int. J. Inf. Manage. Data Insights, vol. 2, no. 1, Apr. 2022, Art. no. 100061, doi: 10.1016/j.jjimei.2022.100061.

[7] W. Alkhatib, C. Rensing, and J. Silberbauer, ‘‘Multi-label text classification using semantic features and dimensionality reduction with autoencoders,’’ in Proc. 1st Int. Conf. Lang., Data, Knowl., 2017, pp. 380–394, doi: 10.1007/978-3-319-59888-8_32.

[8] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B.-W. On, ‘‘Fake news stance detection using deep learning architecture (CNN-LSTM),’’ IEEE Access, vol. 8, pp. 156695–156706, 2020, doi: 10.1109/ACCESS.2020.3019735.

[9] R. Basheer and B. Alkhatib, ‘‘Threats from the dark: A review over dark Web investigation research for cyber threat intelligence,’’ J. Comput. Netw. Commun., vol. 2021, pp. 1–21, Dec. 2021, doi: 10.1155/2021/1302999.

[10] A. Alharbi, M. Faizan, W. Alosaimi, H. Alyami, A. Agrawal, R. Kumar, and R. A. Khan, ‘‘Exploring the topological properties of the tor dark Web,’’ IEEE Access, vol. 9, pp. 21746–21758, 2021, doi: 10.1109/ACCESS.2021.3055532.

[11] L. Ouyang and Y. Zhang, ‘‘Phishing Web page detection with HTMLlevel graph neural network,’’ in Proc. IEEE 20th Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom), Oct. 2021, pp. 952–958, doi: 10.1109/TrustCom53373.2021.00133.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

[12] H. Alnabulsi and R. Islam, ‘‘Identification of illegal forum activities inside the dark net,’’ in Proc. Int. Conf. Mach. Learn. Data Eng. (iCMLDE), Dec. 2018, pp. 22–29.