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 & Topic Modeling Weight

[1]Syed. Join. Shahanaaz, [2]P.Anjali, [3]K.Keerthana, [4]G.Ramya, [5]B.Maheshwari,[6]P.Srilatha

[1]Assistant Professor, Department of Computer Science & Cyber Security,

Princeton Institute of Engineering & Technology For Women

[2,3,4,5,6]B. Tech Students, Department of Computer Science & Cyber Security,

Princeton Institute of Engineering & Technology For Women

ABSTRACT

The dark web has emerged as a significant platform for illegal activities such as cybercrime, drug trafficking, terrorism financing, and data trading. Monitoring and analyzing dark web content is challenging due to its unstructured nature, anonymity, and rapidly evolving language patterns. Traditional keyword-based and rule-driven approaches fail to capture semantic meaning and hidden contextual relationships within dark web text. This paper proposes a hybrid text analytics system for dark web monitoring that combines Convolutional Neural Network (CNN)-based feature extraction with topic modeling techniques. CNNs are employed to automatically learn discriminative textual features, while topic modeling uncovers latent thematic structures within the content. The integration of deep learning and probabilistic topic modeling enhances classification accuracy, improves interpretability, and enables effective monitoring of dark web activities. Experimental evaluation shows that the hybrid approach outperforms conventional text analysis methods, making it suitable for proactive cyber threat intelligence and law enforcement applications.

Keywords: Dark Web Monitoring, Cybercrime Detection, Deep Learning, Convolutional Neural Networks (CNN), Topic Modeling, Text Analytics, Natural Language Processing (NLP), Cyber Threat Intelligence.

I. INTRODUCTION

The dark web is a concealed part of the internet that requires special tools such as Tor for access. It provides anonymity to users, which, while beneficial for privacy, has also facilitated the growth of illicit activities. Dark web forums and marketplaces host discussions related to hacking services, illegal trade, financial fraud, and extremist content. Manual monitoring of such platforms is impractical due to the massive volume of data and rapidly changing content. Conventional dark web monitoring systems rely heavily on keyword matching and static rule-based filtering, which lack adaptability and contextual understanding. Recent advances in Natural Language Processing (NLP) and deep learning have enabled more sophisticated text analysis techniques. CNN-based models are effective in extracting local semantic patterns from text, while topic modeling methods such as Latent Dirichlet Allocation (LDA) provide insights into hidden themes within documents.

This work proposes a hybrid system that integrates CNN-based feature extraction with topic modeling to improve detection accuracy and semantic understanding in dark web monitoring.

II. LITERATURE SURVEY

1. Dark Web Data Analysis Using Machine Learning

Author: Benjamin Dalins et al.
Abstract:
This study explores machine learning techniques for analyzing dark web content. The authors highlight the limitations of keyword-based systems and emphasize the need for advanced text analytics.

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

2. Convolutional Neural Networks for Sentence Classification

Author: Yoon Kim
Abstract:
This paper demonstrates the effectiveness of CNNs for text classification tasks, showing their ability to capture local semantic patterns.

3. Topic Modeling for Text Mining Applications

Author: David Blei et al.
Abstract:
The authors introduce Latent Dirichlet Allocation (LDA), a foundational topic modeling technique that uncovers hidden thematic structures in large text corpora.

4. Hybrid Deep Learning Models for Text Classification

Author: Zhang et al.
Abstract:
This work proposes hybrid deep learning architectures that combine neural networks with probabilistic models to improve classification accuracy and interpretability.

5. Automated Dark Web Threat Intelligence Using NLP

Author: Sharma and Gupta
Abstract:
The study presents NLP-based techniques for dark web threat intelligence and emphasizes the importance of semantic understanding and automated monitoring.

III. EXISTING SYSTEM

Existing dark web monitoring systems primarily rely on keyword-based filtering, rule-based classification, and traditional machine learning algorithms such as Naïve Bayes and Support Vector Machines. These approaches typically depend on manually engineered features like TF-IDF and bag-of-words representations, which focus on word frequency rather than contextual meaning. While these methods are relatively simple to implement and computationally efficient, they exhibit significant limitations when applied to the complex and dynamic environment of the dark web.

One major drawback is their inability to capture semantic relationships and contextual dependencies between words. Dark web content often includes slang, abbreviations, coded language, and intentionally obfuscated terms that evolve rapidly to evade detection. Traditional models treat words independently and fail to understand the underlying intent or meaning behind such variations. As a result, these systems struggle to accurately classify or detect illicit activities when new or modified vocabulary is introduced.

Furthermore, rule-based systems require continuous manual updates to maintain effectiveness, which is both time-consuming and prone to human error. The static nature of predefined rules makes them inflexible and incapable of adapting to emerging threats in real time. Similarly, traditional machine learning models depend heavily on the quality and completeness of training data. If the dataset does not adequately represent the diversity of dark web language, the model's performance degrades significantly.

Another limitation is their poor handling of unstructured and noisy data, which is a common characteristic of dark web content. Irregular grammar, multilingual text, and embedded symbols further reduce the accuracy of these conventional approaches. Consequently, these systems often produce high false positives and false negatives, limiting their usefulness in real-world applications such as cyber threat intelligence and law enforcement.

Overall, while existing methods provide a foundational approach to dark web monitoring, their inability to adapt to evolving language patterns, lack of contextual understanding, and dependence on manual feature engineering highlight the need for more advanced, automated, and intelligent solutions.

IV. PROPOSED SYSTEM

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

The proposed system introduces a robust hybrid text analytics framework that combines the strengths of deep learning and probabilistic modeling to effectively monitor and analyze dark web content. Specifically, the framework integrates Convolutional Neural Network (CNN)-based feature extraction with advanced topic modeling techniques to address the limitations of traditional approaches.

In this system, the CNN model plays a crucial role in automatically learning high-level semantic representations from raw dark web text. Unlike conventional methods that rely on manually engineered features, the CNN processes text data through multiple convolutional and pooling layers to capture contextual, syntactic, and structural patterns. This enables the model to recognize complex linguistic features such as slang, abbreviations, and hidden meanings that are commonly used in dark web communications. As a result, the CNN provides a rich and discriminative feature set that significantly enhances classification performance.

Simultaneously, topic modeling techniques such as Latent Dirichlet Allocation (LDA) are applied to uncover latent thematic structures within the documents. Topic modeling helps identify underlying subjects or discussion themes present in the dark web data, such as cybercrime activities, drug trafficking, or financial fraud. These topic distributions offer an interpretable view of the data, allowing analysts and law enforcement agencies to better understand the nature and trends of illegal activities.

The key innovation of the proposed system lies in the integration of CNN-extracted features with topic modeling outputs. By combining deep semantic representations with probabilistic topic distributions, the framework creates a comprehensive and robust feature representation. This hybrid feature set captures both the contextual depth and thematic structure of the text, leading to more accurate and reliable classification results.

V. SYSTEM ARCHITECTURE

The system architecture of the WiFi-Based LED Notice Board using IoT is designed to enable remote message transmission and real-time display through internet connectivity. The architecture mainly consists of four major components: a user interface (web or mobile application), a WiFi communication network, an IoT-enabled microcontroller, and an LED display board. These components work together to ensure seamless communication between the user and the display system.

In this architecture, the user sends messages through a web application or mobile interface connected to the internet. The message data is transmitted through a WiFi network to the IoT microcontroller, such as ESP8266 or ESP32, which is integrated with the LED display unit. The microcontroller acts as the central processing unit of the system, receiving and decoding the message data sent from the user interface.

Once the microcontroller receives the message, it processes the data and sends the appropriate signals to the LED matrix display. The LED display then presents the message in a scrolling or static format, allowing viewers to read the information clearly. The system ensures that messages are updated instantly without requiring any physical interaction with the notice board.

Additionally, the architecture can include a cloud server or database for storing messages and managing user authentication. This allows only authorized users to update the notice board and ensures secure communication between the user interface and the display system. Overall, the system architecture provides a reliable, scalable, and efficient solution for wireless information display using IoT technology.

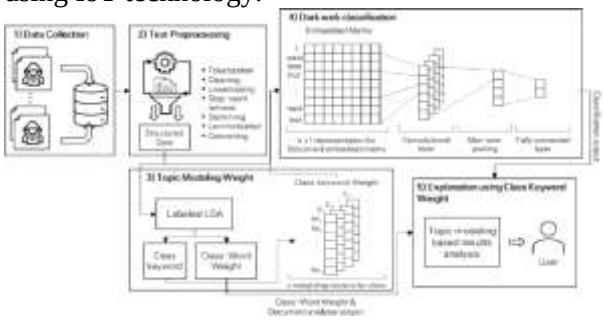

Fig 5.1: System Architecture

VI. IMPLEMENTATION

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

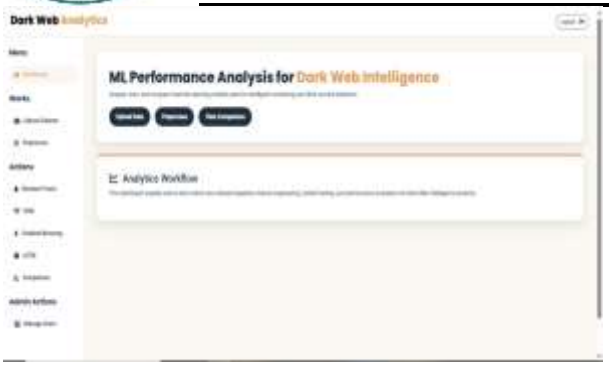

Fig 6.1: Admin Dashboard

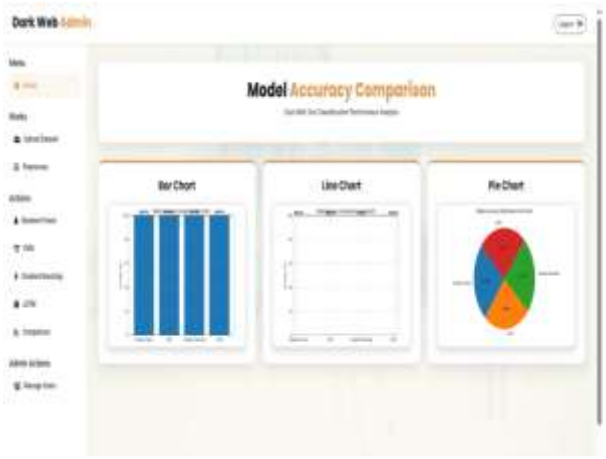

Fig 6.2: Model Accuracy Comparision

Fig 6.3: User Dashboard

Fig 6.4: Detection

VII. CONCLUSION

This project presented a **Hybrid Text Analytics System for Dark Web Monitoring** that effectively integrates traditional machine learning algorithms with deep learning techniques to analyze and classify dark web textual content. The system employs **TF-IDF–based feature extraction**, **topic weight integration**, and multiple classifiers including **Random Forest, Support Vector Machine, Gradient Boosting, and LSTM** to detect and categorize illicit activities such as weapons trade, financial fraud, malware distribution, drug marketplaces, and hacking services.

The experimental results demonstrate that the hybrid approach significantly improves classification accuracy and robustness when compared to single-model systems. The inclusion of both statistical text features and semantic sequence modeling enables the system to handle diverse and unstructured dark web data efficiently. Additionally, the alert and recommendation mechanism enhances real-time decision support for cybersecurity analysts and law enforcement agencies.

Overall, the proposed system provides a **scalable, accurate, and intelligent framework** for proactive dark web intelligence, contributing to improved cybercrime detection and digital threat mitigation.

VIII. FUTURE SCOPE

Although the proposed system achieves promising results, several enhancements can be incorporated in future work:

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

Integration of Transformer Models
Advanced transformer-based architectures such as BERT, RoBERTa, or GPT-based classifiers can be used to improve contextual understanding of dark web language.

Multilingual Dark Web Analysis
Extending the system to support multilingual content (Russian, Chinese, Arabic, Persian, etc.) would improve global dark web monitoring.

Real-Time Dark Web Crawling
Integrating live dark web crawlers using Tor/I2P networks can enable real-time intelligence collection and analysis.

Image and Multimedia Analysis
Future versions may include image, video, and audio analysis for detecting illegal content beyond text.

IX. REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.

[2] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[3] T. Joachims, "Text categorization with support vector machines," *European Conference on Machine Learning*, pp. 137–142, 1998.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] J. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] T. Mikolov et al., "Distributed representations of words and phrases," *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.

[8] J. Ramos, "Using TF-IDF to determine word relevance in document queries," *Proceedings of the First Instructional Conference on Machine Learning*, 2003.

[9] K. Rieck et al., "Automatic analysis of malware behavior using machine learning," *Journal of Computer Security*, vol. 19, no. 4, pp. 639–668, 2011.

[10] M. Abbasi and H. Chen, "A comparison of extremist group detection using machine learning," *IEEE Intelligent Systems*, vol. 23, no. 5, pp. 36–43, 2008.

[11] N. Griva et al., "Dark web data analysis for cybercrime detection," *Digital Investigation*, vol. 28, pp. S85–S93, 2019.

[12] A. Singh and K. Singh, "Dark web: A brief overview," *International Journal of Engineering Research & Technology*, vol. 7, no. 4, 2018.

[13] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O'Reilly Media, 2009.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[15] A. Aggarwal and A. Zhai, "A survey of text classification algorithms," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–40, 2021.

[16] P. Resnick et al., "Trust and reputation systems," *Communications of the ACM*, vol. 43, no. 12, pp. 45–48, 2000.

[17] E. Ferrara et al., "The rise of social bots," *Communications of the ACM*, vol. 59, no. 7, pp. 96–

104, 2016.

[18] S. Minaee et al., "Deep learning–based text classification: A comprehensive review," *ACM Computing Surveys*, vol. 54, no. 3, 2021.

[19] J. Brownlee, *Machine Learning Mastery with Python*, Machine Learning Mastery, 2020.

[20] H. Allcott and M. Gentzkow, "Social media and fake news," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

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