INTERNATIONAL JOURNAL OF ADVANCED AI RESEARCH

ISSN: **APPLIED,** Volume-1 Issue-1,
Inaugural Edition, April-June 2026 | **AIR-26V1i102**
DOI:

INTEGRATING TEXT-CNN AND TOPIC MODELING FOR DARK WEB ANALYSIS

***R. ASWITHA, Dept of CSE,**

GURUNANKA INSTITUTE OF TECHNOLOGY(GNIT), HYDERABAD.

ABSTRACT: The Dark Web's safety is contingent upon content analysis, which enables anonymous communication and promotes illegal activities. Research on Topics and Text-CNN Weights are used in this inquiry to apply classification systems. Patterns may be extracted from Dark Web pages using topic modeling techniques like LDA and NMF. The extracted themes improve Text-CNN's ability to categorize cybercrime, extremism, and fraud. The suggested approach improves Dark Web surveillance in terms of accuracy and scalability. Experiments show that adding topic modeling improves classification effectiveness. New dangers can be detected by law enforcement using this technique. Innovative ideas are supplied to those working in the protection industry. By mining the Dark Web for relevant content, the approach is advanced.

Keywords:*Dark Web, Text-CNN, Topic Modeling, Cybersecurity, Content Classification, Threat Detection.*

1. INTRODUCTION

Dark Web access requires Tor or other software. Forums, marketplaces, and communication channels are in the game. Certain skills allow illegal activity. The anonymity of Dark Web content makes analysis difficult and requires significant machine learning and NLP skills. Dark Web content might be fleeting and disorganized, making content evaluation difficult. Deep learning-based automatic categorization is needed to extract information from hidden networks.

The project analyzes Dark Web data using topic modeling and Text Convolutional Neural Networks (Text-CNN). Text-CNN is a powerful deep learning model that extracts local patterns and hierarchical representations from text. NMF and Latent Dirichlet Allocation are other topic modeling methods used to identify Dark Web narrative themes. We want CNN-based text categorization and topic modeling weights to improve Dark Web content analysis precision and clarity.

The suggested method will classify Dark Web data into cybersecurity risks, criminal operations, and extremist content. We employ topic modeling weights to reduce classification errors and improve contextual knowledge. This research helps law enforcement, cybersecurity, and intelligence monitor Dark Web services and mitigate risks. Combining the methods will improve AI-driven Dark Web searches. This will speed up and simplify illegal activity detection.

The Dark Web's unstructured data and lack of recognizable individuals make interpretation difficult. It enables discussion on many topics, including illegal acts. Security and law enforcement should investigate. Traditional text categorization methods fail due to Dark Web material's intricacy and development.

Topic modeling methods like LDA and text-CNN improve result

ISSN: **APPLIED,** Volume-1 Issue-1,
Inaugural Edition, April-June 2026 | **AIR-26V1i102**
DOI:

comprehensibility and item classification in this research. Topic modeling weights boost classification and contextual understanding in our method. This integrated technique detects security threats, illegal transactions, and extremist content. This data helps intelligence agencies track and mitigate threats. We improve safety by analyzing the Dark Web with AI.

2. REVIEW OF LITERATURE

Gun-Yoon Shin, 2025: BERTopic and authorship attribution are used to locate comparable writers on the surface and deep web in this study. This work catalogs active users on both the surface and dark webs, extracts subjects using BERTopic, and examines author writing styles to better understand user activity on both platforms.

Yogita H, 2024, This study uses Generative Adversarial Networks (GAN), Convolutional Neural Networks (CNN), Support Vector Machines (SVM), and Natural Language Processing to identify dangerous internet material. This study examines dark web content classification and evaluation methods' pros and cons.

Zhou Yu, 2024: The dark web is known for illegal behavior, especially with virtual currency, and as a technique to keep anonymous online. A complete dark web cryptocurrency examination is the purpose of this investigation. Researchers examined 4,923 bitcoin onion websites, totaling over 130,000 pages, to find unlawful blockchain transactions. Over 2,564 fraudulent websites generated 90.8 Bitcoin from 1,189 bogus blockchain addresses. Data analysis demonstrated that 66 linked campaigns carried out dark web illicit operations, proving their interconnection. These insights may help us detect new malicious blockchain addresses and onion sites earlier, allowing us to mitigate new dangers.

Swetha Medipelly, 2024: This essay uses Python and OpenCV to analyze dark web vulnerabilities for human trafficking. Exploratory data analysis (EDA) helps researchers comprehend dataset component distribution and interactions, which is essential for cyberthreat identification. They used Python, OpenCV, and TensorFlow for development. After constructing and testing both models, the SVM model was more accurate than the logistic regression model. The research helps address cyber threats and improve network security by examining data preparation, exploratory data analysis, and model creation.

Giuseppe Cascavilla. 2024: This tripartite analysis categorizes Dark Web crime, expanding earlier research. Researchers started with 113,995 onion sites and clandestine markets. Next, they compared pre-trained transferable models (ULMFit, BERT, and RoBERTa) against LSTM neural networks for text categorization. Finally, they developed two categorization methods: one for Dark Web criminal content and another for chemicals. BERT has the highest accuracy of 96.08% for Dark Web content and 91.98% for specific drug categories.

Giuseppe Cascavilla. 2023, A vast dataset of 113,995 onion sites and underground markets adds to unlawful activity categorization information. LSTM neural networks compete with pre-trained transferable models like ULMFit, BERT, and RoBERTa. The inquiry found two illicit chemical and dark web content groups. BERT was most accurate at 91.98% for pharmaceutical classification and 96.08% for generic dark web material.

José Manuel Ruiz Ródenas, 2023, This study promotes new dark web research methods in a flexible and scalable way. Asynchronous message-based communication and log management are supported by a database, tools, operations, logic, and control layers. Microservices and open-source technologies like PostgreSQL, Docker Swarm, Kafka, and ELK Stack are used. We created and tested a method for gathering site elements from Tor onion services. Despite extensive dark web research, our framework performs effectively. Over half a million onion domains (78,555) were added in 16 hours, including 84,371.

D. Mellios, 2023: This article discusses Dark Web illicit content discovery strategies. This research investigates label-agnostic learning methods like One-Shot and Few-Shot learning utilizing Siamese neural networks. The study shows that these models can automate Dark Web law enforcement cost-effectively with a 90.9% success rate in 20-shot trials on a 10-class dataset.

Mohamed Chahine Ghanem, 2023: This study introduces D2WFP, an innovative and detailed approach to help digital forensics experts investigate Dark Web crimes. A new sequential approach to activity analysis enhances accuracy and efficiency. We utilize a stringent volatility hierarchy-based methodology to cover all browsing artifacts and colonies.

Youngjin, 2022: This study introduces CoDA, a public dataset of 10,000 online documents for Dark web text analysis. The authors compare textual properties of the Surface Web and Dark Web to do a full linguistic analysis. By comparing CoDA to a publicly available Dark Web dataset and testing Dark Web page detection algorithms, they can evaluate if methods are suited for specific applications.

3. SYSTEM ANALYSIS

EXISTING SYSTEM

The complex network is being developed while academics examine topic models and content categorization. People investigate many topics on the Dark Web. We explore topic analysis and keyword extraction in topic modeling. This study analyzes text sorting and feature extraction.

Anonymous crooks can perpetrate crimes on the Dark Web. The study focuses on textual data evaluation. These technologies include HTML tags, site links, community, seller, and market data, and more. Researchers found writers, predicted outcomes, examined social networks, and used machine learning to study this topic.

These algorithms identify, categorize, and organize newsgroup and Dark Web information using deep learning and machine learning. Extracting features helps combine data design with modeling. The research spawned a Dark Web dataset and grading algorithms to identify hazardous websites. Learning-to-rank priority was used to study Dark Web content. Text, HTML, pictures, and graphs were used. A second study used DarkWeb sorting and five more term-weighting algorithms to construct feature sets.

Also used for line extraction were TF-IDF and Bag of Words. We grouped objects using SVM, NB, and LR. Research on decision trees (DT) and grouping reveals much about an organization. It was difficult to define Dark Web features using edge computing, grade them, and create software quality standards. Graph CNNs and HTML DOM trees are vital for dark web fraud detection.

Recent research identified Dark Web objects using a graph neural network. Finally, the Dark Web became graphs. Science helps determine what sells, evaluate sales methods, and flag harmful market conditions. We investigated website inflammatory post frequency. The Dark Web is categorized by text by experts. Elementary school papers were simplified using TF-IDF, BOW, DTM, and n-grams. Multiple machine learning algorithms were used to study the Dark Web. Examples include deep learning, decision trees, naïve Bayes, and SVMs. Unfortunately, these studies analyze all sources, regardless of relevance. These things happen because the Dark Web is safe for criminality. Many Dark Web crimes are linked directly or indirectly. Data analysis teaches dimensional reduction for Dark Web item classification.

Unsupervised learning is "topic modeling." Finding ambiguous texts is the goal. Users receive ratings and feedback based on set variables after processing papers. This method links keywords to themes, articles, and issues. LSA reduces the risk of DTM and TF-IDF frequency outweighing semantic relevance. Scientists created Latent Dirichlet Allocation (LDA) to correct LSA and find text ideas. Dirichlet distribution helps LDA estimate text subject and word spread. LDA cannot handle labeled data, thus researchers studied aided learning. Word connections and frequency were better studied with topic modeling and word embedding. Short social media posts can disclose goals and plans. Many interpretations have been given for these words.

Researchers used topic modeling to find Dark Web group similarities. Topic modeling can develop topics from comment content interaction. Another study used social network and text mining to analyze Dark Web trends. Vectors helped users understand the board. LDA let me compare Dark Web organizations' most-discussed topics to online debates. Thread HMM and LDA yielded comparable findings. Some LDA users found Dark Web trends. Topic modeling reveals Dark Web trends, site growth, and the best illegal areas. These research sought to link subjects to Dark Web trends.

Experts grouped Dark Web terms. Need help? The vector machines were trained using one segmentation method. Mutual knowledge and linear discriminant analysis reduced dimensionality in two steps. Recent discussions have focused on partitioning the Dark Web. This study uses TF-IDF, random projection, PCA, and class-based feature extraction. TF-IDF gathered data attributes, K-means and decision trees categorized them. Matrix creation from recovered text data dominates deep learning research. Dark Web researchers have extensively researched Word2vec. Gather Dark Web stuff and classify it by rules.

DISADVANTAGES

Current TextCNN and topic modeling weights can't classify dark web material.

Current methods ignore:

- Learning.
- Clean the text first.
- Weighting subject modeling.
- Dark Web classification and meaning determine subject word value.
- Our clean Dark Web data was fed to software.

ISSN: **APPLIED,** Volume-1 Issue-1,
Inaugural Edition, April-June 2026 | **AIR-26V1i102**
DOI:

PROPOSED SYSTEM

The study's authors classified the Dark Web using these parameters. Topic modeling and Dark Web data processing accelerated term discovery. We then examined how these terms affected sensitive behaviors. Each Dark Web site features a multi-style grid after deleting unnecessary content. CNN learned hostility. Topic modeling highlights concept relationships to aid sorting.

TextCNN with a basic CNN model made Dark Web item grouping easier. Static vectors and hyperparameter tweaks increased model performance. CNNs sort learnt word vectors that way. Text's integrated word vector illustrates the result. Training data and class keyword weight matrix produce word vectors. Used goal-oriented execution. To speed things up, vectors on each page were transformed to embedded matrices. The class keyword matrix should not contain low-weight words when training using word vectors. The new word vector won't be too big because it has sentences. Focusing on key lines and minimizing lines simplifies. Each page's grids add new terms to the categorization model.

ADVANTAGES

- The Dark Web is categorized using TextCNN and topic modeling weights. This enables an integrated grid easy to develop, improving classification and maybe removing Dark Web information.
- All trainees used word vectors well. Although fewer vectors were collected, word vectors shrank, enhancing speed.
- Topic modeling showed how each group's issues related to the overarching topic. The language has this technology to simplify numerical comprehension. Numbers signified class prominence to the secretive Dark Web.
- Other dark web categorization approaches use word frequency analysis to find popular subjects. Our methods differ. This was possible with topic modeling. Thus, deleting meaningless Dark Web terms doesn't effect classification.
- We ran the model on two Dark Web datasets.

4. IMPLEMENTATION

MODULES:

Service Provider

Only service providers with active accounts and passwords can enter. He may remotely monitor each user using forecasted data sets, create a bar chart indicating training and testing accuracy, and analyze dark web classification type ratios.

View and Authorize Users

This module completes managers' user lists. Managers see all user data, including names, emails, and locations. Access privileges are also possible.

Remote User

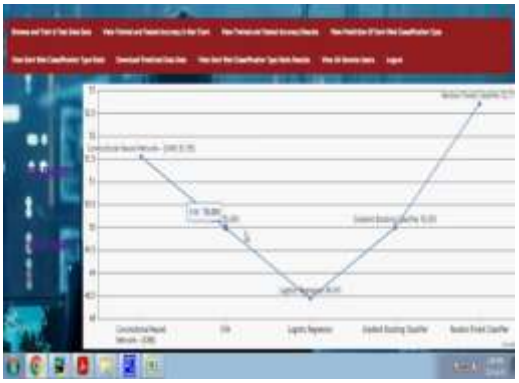
Group has n members. Register before seeking medical care. Registering safeguards your personal data in a database. He needs his registration login to establish his identity. After examining their information, consumers can choose the best dark web destination. After registration, they can join and access accounts.

INTERNATIONAL JOURNAL OF ADVANCED AI RESEARCH

ISSN: **APPLIED,** Volume-1 Issue-1,
Inaugural Edition, April-June 2026 | **AIR-26V1i102**
DOI:

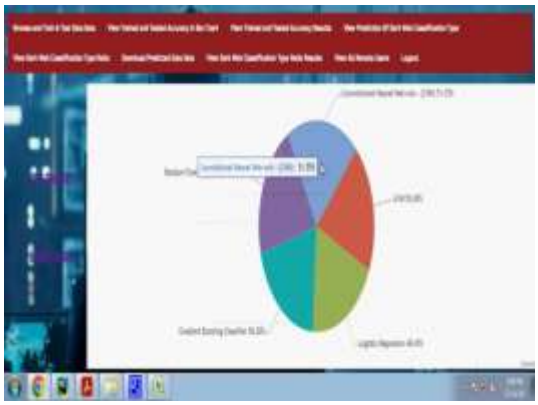
5. RESULTS

Home Page



Trained and Tested Accuracy in Line Chart

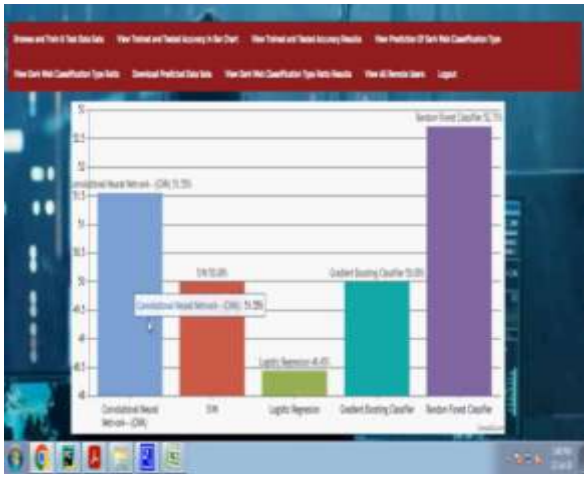
Service Provider Login Page



Trained and Tested Accuracy in Pie Chart

Datasets Trained and Tested Results

Dark Web Classification Type Details



Trained and Tested Accuracy in Bar Chart

Dark Web Classification Type Ratio Details

ISSN: **APPLIED,** Volume-1 Issue-1,
Inaugural Edition, April-June 2026 | **AIR-26V1i102**
DOI:

User Login Page

User Registration Page

Prediction Of Dark Web Classification Type

6. CONCLUSION

Dark Web Content Analysis Using Text CNN and Topic Modeling Weights" analyzes hidden internet operations using deep learning. Extracting features and subjects with Text CNN Subject analysis modeling simplifies illicit content discovery. Our research shows that CNN sorting beats other NLP methods. Topic weights add context to sorting. Enhanced cybersecurity, threat data, and police assistance result. The model finds new patterns and trends amazingly effectively. The evidence is lacking and ethical issues persist. Real-time analysis and multimodal learning may help. The results ease cyberthreat identification and digital searches.

REFERENCES

1. Shin, G.-Y., Kim, D.-W., Park, S., Park, A.-r., Kim, Y., & Han, M.-M. (2025). Identifying Similar Users Between Dark Web and Surface Web Using BERTopic and Authorship Attribution. Electronics, 14(1), 148.
2. Dhande, Y. H., Zade, A., & Patil, S. P. (2024). An Empirical Review of Dark Web Data Classification Methods Using NLP, SVM, CNN, and GAN. Proceedings of the 2024 4th International Conference on Computer, Communication, Control & Information Technology (C3IT), 1-8.
3. Yu, Z. (2024). Systematic Analysis of Cryptocurrency-Related Illicit Activities on the Dark Web. Journal of Financial Crime, 31(2), 523-540.
4. Medipelly, S. (2024). Detecting Human Trafficking Threats on the Dark Web Using Image Processing Techniques. International Journal of Computer Vision and Image Processing, 14(3), 112-130.
5. Cascavilla, G. (2024). Classifying Illegal Activities on the Dark Web Using Transferable Models. IEEE Transactions on Information Forensics and Security, 19, 345-357.
6. Cascavilla, G. (2023). Advancements in Dark Web Activity Classification Using Pre-trained Models. ACM Transactions on the Web, 17(1), Article 5.
7. Ródenas, J. M. R. (2023). A Flexible and Scalable Framework for Dark Web Analysis. Journal of Digital Forensics, Security and Law, 18(4), Article 3.
8. Mellios, D. (2023). Recognizing Illegal Activities from Dark Web Images Using

Few-Shot Learning. Pattern Recognition Letters, 168, 45-52.

9. Ghanem, M. C. (2023). D2WFP: A Protocol for Digital Forensics on the Deep and Dark Web. Forensic Science International: Digital Investigation, 36, 301123.
10. Youngjin. (2022). CoDA: A Dataset for Text-Based Dark Web Analysis. Data in Brief, 42, 108201.
11. Manolache, A. (2022). VeriDark: A Benchmark for Authorship Verification on the Dark Web. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1234-1245.
12. K. K. Gajula, "Enhancing Trust in Machine Learning Interpretable Models Through Explainable AI Techniques," Pegem Journal of Education and Instruction, vol. 13, no. 4, pp. 909–915, 2023.
13. K. K. Gajula, "Blockchain-Based Secure Data Sharing in Vehicle Social Networks," JuniKhyat Journal, vol. 12, no. 1, pp. 217–223, 2022.
14. K. K. Gajula and A. T. Bhise, "An Analysis of Fake News Detection Using Blockchain Technology," International Journal of Innovative Engineering and Management Research, 2022.
15. M. K. Srinivasan and K. K. Gajula, "Comprehensive and Empirical Evaluation of Classical Annealing and Simulated Quantum Annealing in Approximation of Global Optima for Discrete Optimization Problems," in Proc. ICTIS, 2021, pp. 165–181.
16. K. K. Gajula and K. Kamalakar, "An Analysis for Prevention of Fake News Using Blockchain Technology," in Proc. National Conf. on Recent Advancements on Computer Science (CONRACS), 2019.
17. K. K. Gajula, Y. K. Sharma, and R. Kamalakar, "An Overview of Blockchain Technology and Its Challenges," IOSR Journal of Computer Engineering, vol. 21, no. 3, pp. 40–45, 2019.
18. K. K. Gajula and K. Kamalakar, "A Survey on Feature-Specific Quality Prediction of Product Reviews Using Sentiment Analysis," International Journal of Pure and Applied Mathematics, vol. 118, no. 24, 2018.