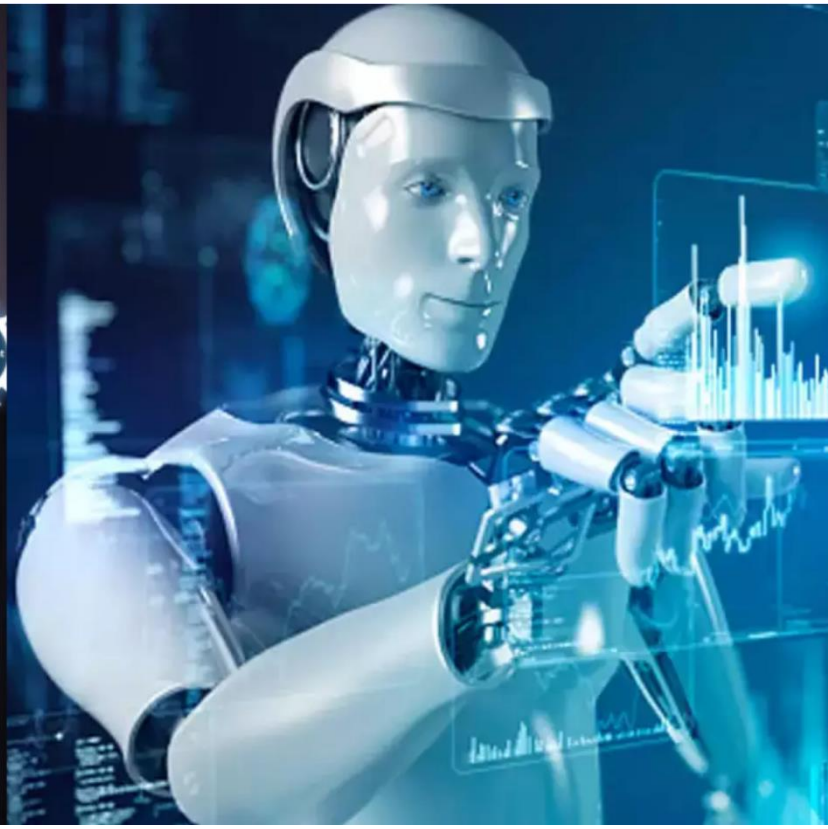
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404542|

Classification of Dark Web Using Text Based CNN and Topic Weight Model

Burle Sudharani, Karuturi Eeshika, Sangireddy Ramu, Gadepalli Sai Gayathri, Gollu Thilok Naidu

UG Scholar, Dept. of Computer Science Engineering (Artificial Intelligence and Data Science), Satya Institute of Technology and Management, Vizianagaram, India

Kuppili Yasoda Krishna

Assistant Professor, Dept. of Computer Science Engineering (Artificial Intelligence and Data Science), Satya Institute of Technology and Management, Vizianagaram, India

ABSTRACT: It is difficult to monitor its users, the DarkWeb, an online domain that guarantees user anonymity, has grown to be a hub for illicit activity and a source of information about cyberattacks. This study looked at how the Dark Web is categorised in connection with various online dangers. To identify vector types appropriate for machine learning categorisation, we analysed words from the Dark Web. Conventional techniques that build features by using all Dark Web texts produce vectors that contain every word on the DarkWeb. Nevertheless, this method adds unnecessary information to the vectors, which reduces learning efficiency and lengthens processing time. By concentrating on certain keywords within each class, the study sought to reduce the size of the word vectors and improve the categorisation process. Utilising the Dark Web's anonymity feature and topic-modeling-based weight creation made this optimisation possible. These techniques improved the differentiation of Dark Web classes by enabling the construction of word vectors with a limited feature set. We combined TextCNN with topic modelling weights in order to enhance classification performance even further. We used two datasets for validation and evaluated the model's performance against alternative text classification methods; the suggested model outperformed the others in Dark Web categorisation.

KEYWORDS: Dark Web, Dark Web analysis, text classification, topic modeling, model explanation

I. INTRODUCTION

Data interchange has increased globally as a result of the widespread usage of internet services and the availability of reasonably priced gadgets. Even while the surface web, or visible part of the internet, seems like a vast informational resource, it is only a small part of the whole web. This surface web makes up a negligible portion of the whole web; more than 90% of it is hidden within the Deep Web, also known as the Hidden Web, which contains a wide variety of comprehensive information. One subsection of the Deep Web, the Dark Web, is well known for being used for illegal purposes [1]. Specialised browsers that guarantee user anonymity, such as Tor, I2P, and Freenet, are necessary to access the Dark Web [2]. Even though anonymity was previously seen favourably, it has since changed into a cover for illegal activity, providing a space where criminal activity may go unhindered. The Dark Web is currently the source of 57% of malicious internet activity, including drug trafficking, pornography, personal data theft, and hacking [3].

There are real-world negative effects from the illegal activities on the Dark Web. According to investigations, malware, such as Trojan horses and ransomware, is distributed through phishing emails or utilised in cryptocurrency transactions on the Dark Web, where anonymity is guaranteed [4]. Users are encouraged to actively participate in and support malicious acts because of the secrecy and anonymity, knowing that their actions in this secret realm would stay undetected. For example, the Dark Web plainly shows information like medicine names, transaction methods, and pricing, or even promotes malware sales, transaction processes, and source codes for public access, in contrast to the surface web, which uses coded language or acronyms for information transmission. Text mining has been used in earlier research to categorise and detect these kinds of behaviours, and more recently, different machine learning methods have been used for categorisation.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404542|

Dark Web text must be preprocessed and converted into a vector format in order for machine learning to be used to Dark Web categorisation. This procedure, called word embedding, converts natural language into a machine-understandable format. The document-term matrix (DTM), GloVe, word2vec, latent semantic analysis (LSA), and term frequency-inverse document frequency (TF-IDF) are examples of pertinent technology. These techniques seek to vectorise every word in the text as efficiently as possible, but they neither identify or eliminate words that have a major learning impact nor omit words that are not crucial to learning. Furthermore, this embedding technique raises the dimensional count concurrently with the growing text word count. In order to extract information from the complete text, text learning often entails analysing every sentence. On the other hand, the Dark Web's anonymity makes it easier to expose cyberthreat information directly, without the need of coded words or acronyms. By removing unnecessary Dark Web elements and turning only the most important parts of each class into vectors, this feature makes learning possible.

In this context, [5] investigated and evaluated different dimensional reduction methods to improve text classification performance, while [6] applied dimensional reduction to word embedding in an attempt to increase classification accuracy. Additionally, [7] sought to enhance performance through the use of auto encoders (AE) for dimension reduction, while [8] sought to improve the accuracy of models for detecting false news by utilising principal component analysis (PCA) and chi-square prior to convolutional neural network (CNN)-based learning. Therefore, dimension reduction becomes a feasible approach to enhance performance regardless of the text classification method chosen.

Tor software can be used to achieve this anonymity. To prevent users involved in criminal activity from being discovered by the government or a vendor, the Tor network encrypts and conceals data from the entrance node to the exit node [9]. The second is the evolution of the online landscape. In addition to the ease with which equipment may now be purchased for nefarious purposes, internet users are growing more interested in accessing the dark web [10]. Moreover, transactions have increased since the emergence of cryptocurrencies [11]. Studies are being carried out to address these problems by gathering information by crawling dark web marketplaces, forums, and websites; classifying the types of malicious activities [12] by analysing the images, files, HTML, and text that are contained within them; identifying the topics associated with malicious activities in the forums; and gathering the IDs of the primary forum users [13]. Current techniques categorise the many kinds of harmful activity occurring on the dark web and use the quantity of articles or posts to determine the primary writers in the dark web forums [14]. Nevertheless, these techniques only classify the data from the dark web based on the type of hazardous action; as a result, it is hard to get precise information about the authors or individuals who engaged in these activities. Additionally, because the techniques that identify the primary writers or sellers through dark web forums or marketplace analysis depend on the quantity of articles produced, they are unable to determine the grammatical traits or writing styles of certain authors. Furthermore, it is hard to ascertain the relationship between articles published on the surface web and those on the dark web as the majority of dark web analysis methods exclusively employ data from the dark web. Furthermore, authors who make postings while engaging in harmful activity and switch between the surface and dark webs are hard to detect. Thus, by examining and eliminating important phrases and unnecessary words for each category, this study sought to improve word dimensions. As a result, extraneous vector dimensions were removed, improving the quality of the data. This prevented the learning model from absorbing pointless word vectors and improved performance by adding highly significant word vectors for classification. We also offered relevant classifications and unique site data to help users understand the categorised Dark site results. With this method, topic modelling was used to identify the major words for each class, and weights indicating the word's class significance were calculated. This makes it possible to calculate, using word distribution and importance, the degree to which particular texts correspond with a class.

There were many phases in the process used to categorise the Dark Web domain in this investigation. In order to determine important phrases and their influence on each harmful behaviour, we first preprocessed the obtained Dark Web data and applied topic modelling. After that, we converted each Dark Web website into a topic-embedded matrix by removing any unnecessary material. After that, we used a CNN model to train the matrix in order to categorise harmful actions. The categorisation was then highlighted by the frequency and influence of key terms that were identified using topic modelling. The following summarises this study's main contributions:

- We presented a Dark Web categorisation method that combines Text CNN with topic modelling weights. This approach guarantees accurate categorisation by eliminating pointless Dark Web information using topic modelling weights and creating an integrated matrix.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404542|

- Only the keyword vectors relevant to each class were chosen as part of the learning process. Even with a lower amount of vectors learnt, performance was enhanced by this reduction in word vector size.
- We determined the primary terms for each class and their corresponding weights in connection to the subject using topic modelling. In order to simplify the analysis findings for user comprehension, this formula was performed to the text and gave users a numerical indicator of the class's relevance to the categorised Dark Web.
- Our strategy used topic modelling to decipher the meanings of words found in the Dark Web, in contrast to conventional Dark Web classification techniques that concentrate on word frequency quantification. This allowed us to maintain classification performance without adding terms from the Dark Web that aren't necessary.

We assessed the model's performance using two different datasets that were obtained from the Dark Web and used as the ground truth.

II. METHODOLOGY

The methodology for classifying the website of the Dark Web in this study involved several steps. Initially, we preprocessed the acquired Dark Web data and utilized topic modeling to identify significant words and their impact on each malicious activity. Thereafter, we filtered out non-essential information from each dark website of the Dark Web and transformed it into a topic-embedded matrix. Following this, we trained the matrix using a CNN model to classify malicious activities, highlighting the classification through the influence and frequency of key words determined *via* topic modeling.

In the field of Dark Web classification, we developed a classifier utilizing both a basic CNN model and the TextCNN model. This model demonstrated exceptional performance through hyperparameter tuning and the use of static vectors. In this approach, pre-trained word vectors are processed using a CNN for classification; the performance is dependent on the word vector employed for document embedding. Consequently, we generated pre-trained word vectors based on the class keyword weight matrix devised in the proposed system. Thereafter, each document was converted into an embedded matrix using these vectors. The pre-training of word vectors is exclusively conducted with words specified in the class keyword weight (CKW) matrix, which diminishes the size of the word vector created from thousands of existing words. This reduction accelerates processing speed as well as enhances performance by focusing on key terms. Newly trained word vectors are then utilized to transform each document into an embedded matrix, which serves as input for the classification model.

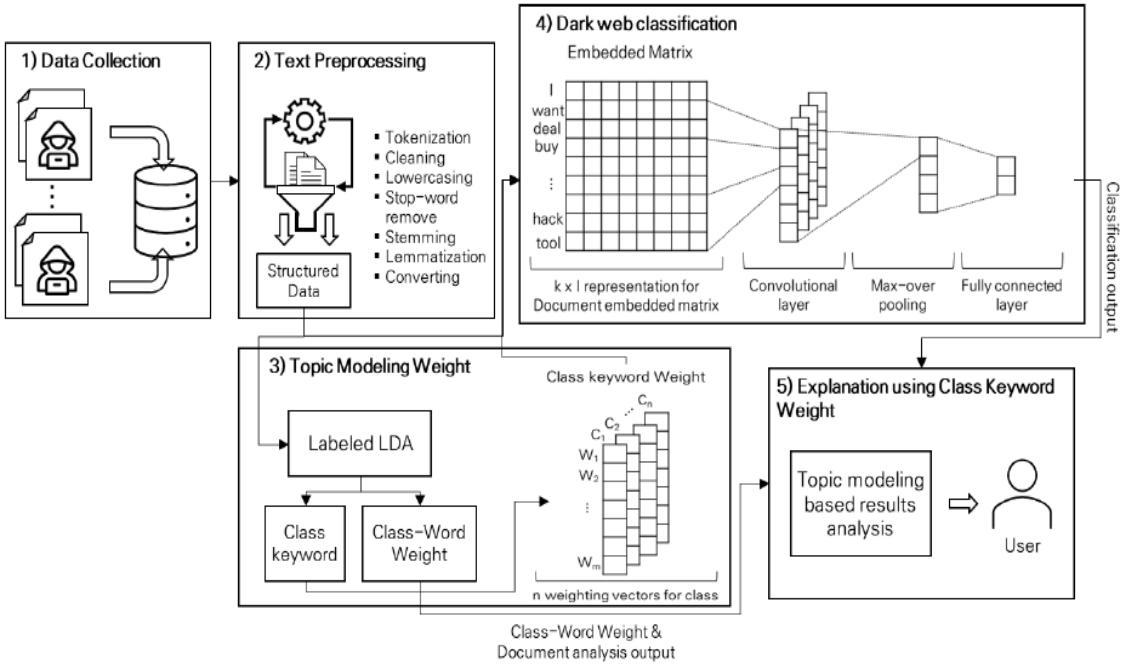


Figure 1. Architecture of proposed work

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404542|

A. Data Collection

Data with established classifications are necessary for efficient Dark Web categorisation. Two Dark Web datasets that have already been verified by analysis were used in this investigation. While CoDA has 10 classes and 8,855 samples, DUTA-10k has 25 primary classes and 10,367 samples. We limited the data we collected from the Dark Web to English-language content. Some of the classes in the DUTA-10k were noticeably smaller. Consequently, we removed classes that accounted for the lowest 1% of the whole Dark Web dataset as well as the "empty" class, which primarily consisted of sparse or non-textual information.

B. Text Preprocessing

It was essential to transform the gathered data into a standard format because it was written by different users in different ways. The unstructured data was optimised through the use of text preprocessing. We were able to obtain data appropriate for model input through this preprocessing step. First, we found and removed any material that was devoid of any useful information. We next deleted stop words and special symbols that were unrelated to or only loosely associated with criminal activity, erased white spaces, and used tokenisation to break and construct tokens. We converted all terms to lowercase in order to account with the programming language's distinction between capital and lowercase letters. We used lemmatisation and stemming to group terms with related meanings in order to decrease the quantity of the vocabulary.

C. Topic Modeling Weight

The keywords for each class and their corresponding effects on each class were identified in this study using LLDA. The drawbacks of conventional LDA, which excludes learning from data with preset classes, were addressed by the development of LLDA. LLDA differs from LDA in that it learns the label set beforehand, even though it uses Dirichlet distributions for probability computations. Fig. 2 shows the whole LLDA procedure for Nwords in a document, documents D, and themes K.

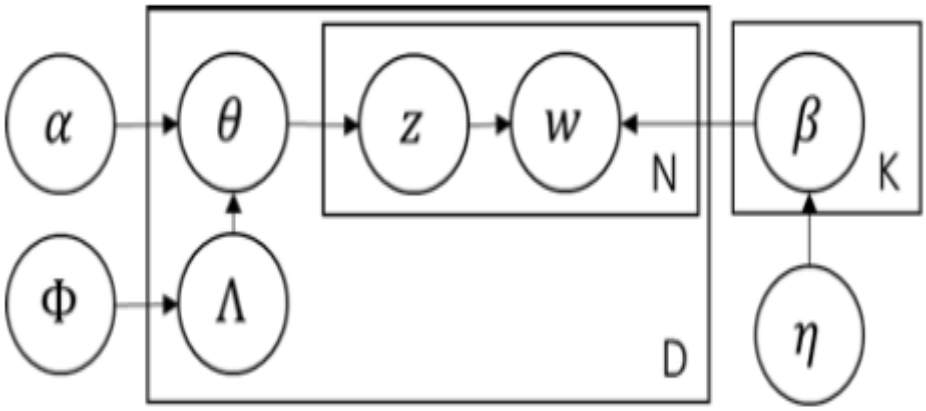


Figure 2. Graphical model of LLDA process

D. Dark Web Classification

We created a classifier for Dark Web categorisation that uses the TextCNN model in addition to a simple CNN model. Through the use of static vectors and hyperparameter adjustment, this model showed remarkable performance. This method uses a CNN to process pre-trained word vectors for classification; the word vector used for document embedding determines the performance. As a result, we used the class keyword weight matrix developed to create pre-trained word vectors. These vectors were then used to transform each document into an embedded matrix. In order to reduce the size of the word vector generated from thousands of existing words, pre-training of word vectors is only done using terms included in the class keyword weight (CKW) matrix. By concentrating on important phrases, this reduction improves performance and speeds up processing. Each page is subsequently converted into an embedded matrix, which is used as input for the classification model, using freshly trained word vectors.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404542|

E. Explanation using Topic Keyword Weight

We employed class-word weights and word frequencies for providing user-centric explanations. We ascertained the predicted class of the classified documents and assessed the keywords from the predicted class within the document, along with their proportion relative to the total word count in the document.

III. EXPERIMENTAL FINDINGS

A. System Environment

The proposed model is developed using said hardware system configuration and software required to implement the design is shown below.

Hardware System Configuration:

- Processor: Intel i5
- RAM : 8 GB
- Hard Disk : 256 GB
- Key Board : Standard Windows Keyboard
- Mouse : Two or Three Button Mouse
- Monitor : SVGA

Software Requirements:

- Operating system : Windows 7 Ultimate.
- Coding Language: Python.
- Front-End : Python.
- Back-End : Django-ORM
- Designing : Html, css, javascript.
- Data Base : MySQL (WAMP Server).

B. Experimental Results

The processed results and user interface results are shown in below figures.

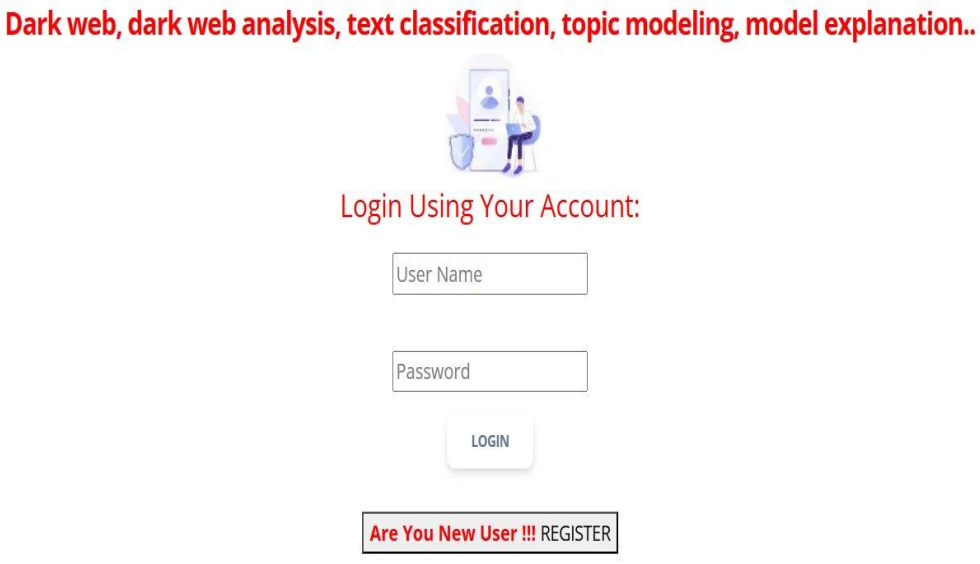


Figure3. Input login interface

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404542|

Figure 4. Interface status after login

Figure 5. Dark Web page

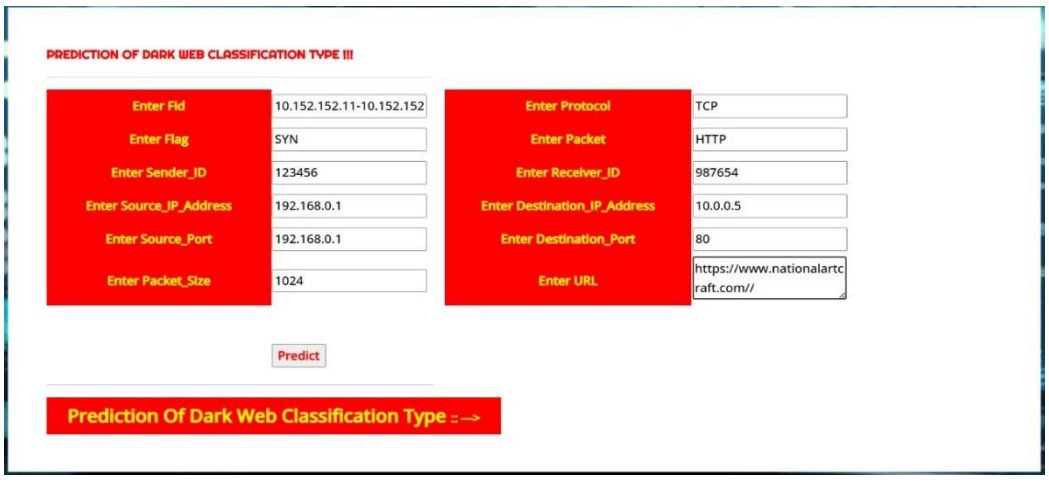


Figure 6. Entering of web details for prediction

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404542|

Figure 7. Prediction of dark web

After performing the classification and processing of details as shown in figure 6, the output predicted is threat found and is shown in figure7.

IV. CONCLUSION

This study presents a Dark Web classification approach that combines Text CNN with topic modelling weights. There was a large vocabulary in the Dark Web data that was not all suitable for efficient categorisation. Preprocessing was therefore necessary to remove unnecessary words. We used topic modelling weights to address this problem. Text preprocessing was used to first produce structured data, and then LLDA was used to identify the major keywords for each class. The word vector size was significantly reduced by almost 300 times when these keywords were applied to the Dark Web data to eliminate unnecessary terms. The word vectors played a key role in creating an embedded matrix that was used as the Dark Web categorisation model's input. In order to identify the influential keywords and their corresponding impact, we applied topic modelling weights in our classification procedure. Two labelled Dark Web datasets were used to evaluate the effectiveness of our suggested method, which was confirmed by comparison with other text categorisation algorithms and earlier research. The results of the trial confirmed the suggested method's superior performance and effectiveness in lowering the vocabulary size needed for learning. Given the dynamic nature of the Dark Web, we plan to create real-time categorisation methods for it in further studies. In real-time situations, processing speed and accuracy are crucial. Consequently, it is expected that topic modelling weights would produce complementing advantages when superfluous words are eliminated. Furthermore, we want to develop a method that combines deep learning and topic modelling weights into a single model in order to clarify which parts of the neural network affect the final outcomes.

REFERENCES

1. M. W. Al Nabki *et al.*, “Classifying illegal activities on tor network based on web textual contents” in *Proc. 15th Conference of the European Chapter of the Association for Computational Linguistics, Long [Papers]*, vol. 1, no. 2017, Apr., pp. 35-43.
2. G. Cascavilla *et al.*, “Cybercrime threat intelligence: A systematic multi-vocal literature review,” *Comput. Sec.*, vol. 105, p. 102258, 2021.
3. J. Saleem *et al.*, “The anonymity of the dark web: A survey,” *IEEE Access*, vol. 10, pp. 33628-33660, 2022.
4. M. Balduzzi and V. Ciancaglini, 2015, “Cybercrime in the deep web,” Black Hat. Amsterdam: EU.
5. S. R. 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, 2019.
6. K. N. Singh et al., “A novel approach for dimension reduction using word embedding: An enhanced text classification approach,” *International Journal of Information Management Data Insights*, vol. 2, no. 1, p. 100061, 2022.
7. W. Alkhatib et al., “Multi-label text classification using semantic features and dimensionality reduction with autoencoders” in *Language, Data, and Knowledge: First International Conference, LDK 2017*, 2017, pp. 380-394.
8. M. Umer et al., “Fake news stance detection using deep learning architecture (CNN-LSTM),” *IEEE Access*, vol. 8, pp. 156695-156706, 2020.
9. Parkar, A.; Sharma, S.; Yadav, S. Introduction to deep web. Int. Res. J. Eng. Technol. **2017**, 4, 229–234.
10. Basheer, R.; Alkhatib, B. Threats from the dark: A review over dark web investigation research for cyber threat intelligence. J. Comput. Netw. Commun. **2021**, 2021, 1302999.
11. Kommineni, M., & Chundru, S. (2025). Sustainable Data Governance Implementing Energy-Efficient Data Lifecycle Management in Enterprise Systems. In Driving Business Success Through Eco-Friendly Strategies (pp. 397-418). IGI Global Scientific Publishing.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404542|

12. Lee, S.; Yoon, C.; Kang, H.; Kim, Y.; Kim, Y.; Han, D.; Son, S.; Shin, S. Cybercriminal minds: An investigative study of cryptocurrency abuses in the dark web. In Proceedings of the 26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, CA, USA, 24–27 February 2019.
13. Ruiz Ródenas, J.M.; Pastor-Galindo, J.; Gómez Mármol, F. A general and modular framework for dark web analysis. Clust. Comput. **2024**, 27, 4687–4703.
14. Pete, I.; Hughes, J.; Chua, Y.T.; Bada, M. A social network analysis and comparison of six dark web forums. In Proceedings of the 2020 IEEE European symposium on security and privacy workshops (EuroS&PW), Genoa, Italy, 7–11 September 2020; pp. 484–493.
15. Sabbah, T.; Selamat, A.; Selamat, M.H.; Ibrahim, R.; Fujita, H. Hybridized term-weighting method for dark web classification. Neurocomputing **2016**, 173, 1908–1926.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details