iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67998

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

Performance Analysis of Enhanced Reinforcement Algorithms While Web Content Retrieval in Topic Categorization

M. Karthica[1], Dr. K. Meenakshi Sundaram[2], Dr. J. Vandarkuzhali[3]
[1]*Ph.D, Research Scholar, PG & Research Dept. of Comp.Sci, Erode Arts and Science, College, Erode, Tamilnadu, India.*
[2]*Former Associate Professor & Head, PG & Research Dept. of Comp.Sci, Erode Arts and Science College, Erode, Tamilnadu, India*
[3]*Assistant Professor, PG & Research Dept. of Comp.Sci, Erode Arts and Science College, Erode, Tamilnadu, India*

Abstract: ***Web content mining is the process of extracting or mining knowledge or useful information from web pages. The purpose of this work is to investigate web content extraction technology enhanced Reinforcement algorithm which anticipates user interest by analyzing the page according to user view related topic. Information seekers rely heavily on search engines to extract relevant information because of the Internet's exponential development in users and traffic. The availability of a vast amount of textual, audio, video, and other content has expanded search engines' duty. Users of the Internet can obtain pertinent information about their query from the search engine by using factors like content and link structure. The three improved algorithms like ANB-MLM, CBARM-MLM was then applied to a web text mining system and used to mine and collect information from various users. To test the performance of the proposed algorithms, the study compared the precision, recall, accuracy and F1-score values of the three algorithms under TREC dataset. The proposed algorithms was found and compared to find the better performance with maximum recall value, accuracy value and F-value in the dataset. Finally, it was found that the latter could achieve a maximum prediction accuracy which is much more accurate than the traditional algorithm for customer information collection. This ERA-MLM procedure involves locating web sites linked to user queries and using hyperlinks to locate a collection of related web pages and find the topic categorization using machine learning method compare existing and proposed methods.***

Keywords: Web Content Based Similarity, Topic Categorization, Machine Learning, Enhanced Reinforcement, Advanced Naïve Bayes Machine learning system, Content Based Automated Parsing.

I. INTRODUCTION

Web content mining is a technique used to extract valuable information from the web using various techniques to address heterogeneity and lack of unique representation. It identifies hidden web data and is becoming the primary source of information for individuals. Web mining involves adjusting algorithms to better meet web data requirements and using new approaches that align with web data characteristics. It automates extraction rules using suitable algorithms and extracts information from web data sources like user registration forms and transactions. Clustering algorithms group similar data points based on common characteristics. This method helps extract valuable information from unstructured materials like press releases, emails, and contracts. The World Wide Web has revolutionized human life, increasing information and performance measurement, necessitating consistent updates and strategies. Furthermore, the use of data mining and big data approaches is rapidly growing in all facets of real world situations. Thereis a greatdeal of Web exploration involved in these tactics. Generally speaking, the term "World Wide Web" (WWW)refers to the process of gathering relevant data in structured, unstructured, and semi- structured formats, including documents, text, files, and images.

These forms include a higher level of diversity, accuracy, and a dynamic modular framework. The goal of this process is to increase the rate of scalability while simultaneously minimizing the quantity of redundant data. Users add additional pages that are not relevant to the search. As a result list, different ranking algorithms are applied to the search results to make it easier for users to navigate the result list. Finding pertinent information resources from a collection of information resources is the process of information retrieval. Web content mining combines text mining, data mining, information retrieval, and machine learning to improve search results and facilitate sophisticated online searches.

It aims to filter information and integrate data on the web, combining valuable data from web pages. Data mining techniques are related to both text mining and data mining, as they are used on virtually free text found online.

II. RELATED WORK

The Enhanced Reinforcement Learning Method (ERA-MLM) is a machine learning technique that enables agents to learn through trial and error and feedback from their actions. It maximizes the effectiveness of reward signals by allowing agents to make decisions through interaction with their environment. The objective is to maximize cumulative rewards over time. Reinforcement learning (RL) can be applied to sentiment analysis by training models to predict a text's sentiment based on a reward signal. The model's parameters are adjusted to optimize the reward. Metrics like precision, recall, accuracy, and F1-score serve as the basis for the reward signal.

Over time, the RL algorithm learns to fine-tune the model's parameters to enhance the sentiment analysis's quality. The authors[9] discorved RL- based sentiment analysis can be challenging due to class imbalance and creating an appropriate reward function that matches the model's intended performance. After that, you need to identify the main types of accessibility or user interests, and you need to conduct a specific event based on the user patterns discovered in order to determine the accessibility configuration and behavior of the user.

The authors [11] discovered This study introduces Digital Chaotic Mapping, a new approach based on digital image encryption, which offers the best results in encrypted image accuracy, enhancing the transmission of multimedia content and protecting image data.

With the use of web content mining techniques and analytical approaches, Lang Chunmin and colleagues offered an analysis for the investigation of the consumer's exposures to fashion- oriented content. The benefits and drawbacks of online mass modifications based on focus groups or individual customer experiences were noted by the authors. Additionally, the authors used several web content mining techniques, from data validation to data cleansing, to anticipate cost analysis and estimation [10].

Online Content Management (WCM) duties are complicated by the diverse structure of online resources. Depending on the type of data contained in the web page content, several extraction techniques are used. Certain criteria may be satisfied by extracting relevant information, such as product descriptions and forum conversations. Using wrapped webpages as training data could improve extractor performance and make it a viable option for further research [18].

Numerous approaches to different strategies for obtaining and analyzing web content are proposed by scientists. Excellent recommendations include using crawlers to locate news articles in Spanish, enhancing document representations using conceptual graphs or logic predicates, and developing generalization algorithms to uncover patterns such as deviations, associations, and clusters [14]. Adapting these procedures to particular fields, such as politics or the economy, may undoubtedly result in more insightful and useful study. Furthermore, the creative approach of visualizing online content to extract its structure looks interesting and may improve web data understanding and utilization [5].

In recent years, anomaly detection has become a major difficulty in social network research because of its significance in a number of applications, such as fraud and spammer identification. Aberrant nodes in static attributed networks can be more precisely recognized by incorporating both graph and node attributes [13].

We developed an ensemble model that integrates models using auto encoders (AEs), variational auto encoders (VAEs), and generative adversarial networks (GANs) to more effectively identify abnormalities in social networks. In the inference (detection) stage, we employ a new method to allocate weights to the several models that make up our ensemble model [7].

III. METHODOLOGY

A. Pre processing of Proposed Methods N-Gram Stemmer

Stemming is the process of reducing words to their basic form, which is crucial for natural language processing pipelines. It is made easier by stemmers and stemming algorithms. Words are divided into n-length chunks using "n-grams," which are then examined for trends using statistical analysis. An n-gram is a collection of n consecutive characters extracted from a word, and related words share many n-grams. Computational convolution enhances recommender performance. Text preparation is essential, with tokenization using punctuation and white space as the first stage. A stop-word filters out useless symbols like interjections, auxiliary verbs, prepositions, and conjunctions. Stemming is also necessary for decaying grammatical versions of expressions that differ from their root form. The most advanced stemming method uses the Porter.

1) NLTK Stop words

Stop words are frequently used words in a language but are rarely used in natural language processing (NLP) tasks. They are often removed during text preparation and may include domain- specific terms like "patient" or "treatment" in the medical industry. These words are not crucial to understanding a document's meaning.Numbers and numeric characters may occasionally be considered stop words, especially if the analysis focuses more on the meaning of the text than on specific numerical values. Stop words can be single characters such as "a," "I," "s," or "x," particularly if they don't mean anything on their own. Words that make sense in one context but are stop words in another are known as contextual stop words. The word "will" might be a stop word in the context of general language processing, but it might be essential for future prediction.

2) Feature Weighing

Explicit Semantic Analysis (ESA)

ESA leverages explicit characteristics present in an existing knowledge base instead than looking for latent patterns. ESA is mostly utilized as a feature extraction method for assessing the semantic similarity of text documents and for explicit topic modeling. The primary use of ESA as a classification algorithm is in text document classification. Both categorical and numerical input data can be used with the feature extraction and classification versions of ESA.It uses attribute vectors as input for feature extraction or classification, with each vector linked to a concept. For classification, the training set may contain multiple vectors for a specific target class, allowing for multiple classifications.

B. Advanced Naïve Bayes Machine Learning for Document Similarity (ANB-MLM)

The technology of data encryption is extensively used to safeguard the confidentiality of text data on the network; but, when users need to access the data, the layer of encryption becomes a barrier that prevents them from doing so. The plain text of a cipher cannot be decoded without the associated key, which is why cryptography is employed to prevent this from happening. The algorithm or the key can be cracked using brute force, although it is extremely difficult to do so if you use strong encryption. For the purpose of directly processing encrypted text, the Advanced Light GBM algorithm is suggested as a task. This algorithm describes both the symmetric and dissymmetric aspects of the SCDA dataset text document. For the purpose of securing the confidential data that is being communicated over the computer network key size value, the capability to directly extract in the decrypted state is of great assistance. The strategy to message communication that is recommended in this work is applied to handle a range of message types. This makes it feasible to interchange special characters and ASCII characters in a more safe and expedient manner.

C. Content Based Automated Parsing Machine Learning for Document Similarity (CBAP-MLM)

The transmission of multimedia content is crucial, and protecting image data is essential. This can be achieved through encryption and decryption, and various tactics should be implemented. Previous research has used deep learning algorithms for encryption, but this work presents a new approach called Digital Chaotic Mapping, based on digital image encryption. The proposed approach produces the best results in terms of encrypted image accuracy, but accuracy values are based on comparisons.

D. Enhanced Reinforcement Algorithm for Document Similarity (ERA-MLM)

The Enhanced Reinforcement Learning Method (ERA-MLM) is a machine learning technique that enables agents to learn through trial and error and feedback from their actions. This method maximizes the effectiveness of reward signals by allowing agents to make decisions through interaction with their environment. By obtaining feedback in the form of incentives or penalties, agents can learn the knowledge necessary to carry out the most effective course of action in various situations. The agent's objective is to maximize cumulative rewards over time.By assigning states to actions, the agent aims to learn a policy that maximizes its cumulative reward over time. RL can also be used to train models that prioritize and condense content by identifying key phrases or sentences. The model produces a summary that is compared to a reward signal, and the model's parameters are adjusted to optimize the reward. By training a model to predict a text's sentiment based on a reward signal, reinforcement learning (RL) can be applied to sentiment analysis. Based on how well it predicts, the model engages with a task environment, such as a dataset of labeled instances, and is rewarded with an output. Metrics reflecting the quality of the model's predictions, like as precision, recall, accuracy, F1- score, might serve as the basis for the reward signal. Over time, the RL algorithm learns to fine-tune the model's parameters in order to enhance the sentiment analysis's quality.

As the model learns to recognize pertinent elements and patterns that are helpful for predicting the sentiment of a text, this method may lead to more reliable and accurate sentiment analysis models. Nevertheless, RL-based sentiment analysis can also be difficult since it has to address the problem of class imbalance, which occurs when one sentiment class is more common than the others, and create an appropriate reward function that matches the model's intended performance.

E. Web Page Repository

A Web repository stores and manages Web pages, similar to file systems, database management systems, and information retrieval systems. However, it doesn't require features like transactions or directory naming schemes. Web repositories can deliver only the most necessary services scalably and effectively. Document scalability is crucial for handling large objects due to the web's size and growth. Distributing the repository over a cluster of disks and machines, like network disks, is ideal due to their low-cost and easy-to-build big data storage arrays.

1) *Streams:* The repository must allow access to individual web pages, but bulk access to large groups of pages is most demanding. Stream access, which scans the entire collection and sends it to a client for analysis, may also be necessary.
2) *Expunging Pages:* When an object is no longer required, it is usually explicitly destroyed in file or data systems. On the other hand, the repository is not informed when a webpage is taken down from a website. As a result, the repository needs a system in place for identifying and eliminating outdated pages. This is similar to "garbage collection," except it doesn't rely on reference counts.
3) *Softmax Regression:* The Softmax Regression module extends logistic regression to multi-class classification problems, utilizing tf-idf to represent training data sets and characterizing feature vectors and groupings of samples using softmax training x and y.
4) *Word sense Disambiguation:* Word sense disambiguation enhances document clustering and visualization, addressing synonymy and polysemy issues in text classification tasks, primarily sourced from the texts to be classified.
5) Vector Space Parsing Method: The Parsing Vector Space Model is used in content-based filtering approaches to recommend documents based on positive ratings. It represents text as a vector of identifiers, allowing it to identify similarities in meaning between texts, regardless of shared words.
6) *Parameters*: Testing the trained model, the method to quotation the feature vector x of a document is precisely the same as training. The multi-label classification results are utilized to recommend for that document, which means that the predicted class from model is the recommended category. In specific, the user prefer the Top 3 classes rendering to the probability $p(y|x)$ in its place of solitary one since our model's output as the ending classification result.

$V(s) = max\ [R(s,a) + \gamma V(s`)]$ ---------- *(1)*

V(s)= value calculated at a particular point.

R(s,a) = Reward at a particular states by performing an action.

γ = Discount factor

V(s`) = The value at the previous state.

An agent's behaviour at a specific moment in time is referred to as its policy. It links the actions performed in relation to the perceived situations of the environment. The fundamental component of RL is a policy since it is the only thing that can specify an agent's behavior. A straightforward function or lookup table may be used in certain situations, whereas general calculation acting as a search procedure may be required in others. The policy may be stochastic or deterministic:

For deterministic policy: $a = \pi(s)$

For stochastic policy: $\pi(a \mid s) = P[At = a \mid St = s]$ *(2)*

The reward signal determines the objective of reinforcement learning. A reward signal is an instantaneous signal that the environment gives to the learning agent at each state.

These incentives are granted based on the agent's good and negative deeds. Maximizing the overall amount of rewards for good deeds is the agent's primary goal. If an agent's action selection yields a poor reward, the reward signal may alter the policy to select a different course of action in the future.

Relevant things are documents that help the utilizer answer a question. Irrelevant stuff that does not provide real useful information. Each item has two options obtainable or not obtainable depends on the utilizer's query.

$$Precision = \frac{Relevant\ Document\ Retrieved}{Retrived\ Documents} \text{---------} (3)$$

The second measure is recall. This is the proportion of documents that are related to the query as well as have been found.

$$Recall = \frac{Relevant\ \ Document\ (Retrieved\ Emojipedia)}{Retrived\ DEmojipedia} \text{ -----(4)}$$

Accuracy is utilized as statistical measure of binary classification test correct identifies or excludes a condition is

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \text{ ----------(5)}$$

The accuracy of system is calculated using the statistical and cross validation in this method and calculate the values using the formula. Searching academic publications for evolving, undefined research area across disciplines presents a significant challenge for data collection.

IV. RESULTS AND DISCUSSIONS

The region delineated by the vertices of the hypercube becomes denser and more populated as documents are added to the collection. In contrast to Boolean, the inverse document frequencies of the terms in the novel document increase while those of the outstanding terms fall when a document is added utilizing term frequency-inverse document frequency weights. Typically, as more papers are added, the area in which they are located becomes more flexible, hence increasing the overall compactness of the assortment picture. Certain class documents or themes, such as research articles and text mining, yield less useful recall findings. In order to recall, precision values, F1-score and Accuracy parameters. The ERA-MLM gives very high accuracy while comparing the existing method as support vector machine as well as comparing the proposed ANB- MLM, CBAP-MLM methods. The average values obtained by applying the text document supports to the accuracy outcomes of the complete ERA-MLM method.

The last part of reinforcement learning is the model, which mimics behavior found in the environment. One can draw conclusions about the behavior of the environment with the aid of the model. Such as, if a state and an action are given, then a model can forecast the next state and reward given a state and an action.

Table 1.1. The Comparison Table of Proposed Methods using TREC Dataset

Methods	Precision %	Recall %	F1-Score %	Accuracy %
SVM	**93.72**	**94.46**	**93.27**	**93.29**
ANB-MLM	**95.16**	**95.22**	**95.41**	**94.23**
CBAP-MLM	**96.33**	**96.37**	**95.25**	**95.53**
ERA-MLM	**97.78**	**97.28**	**96.21**	**96.36**

From the table 7.2 compared the proposed methods for TREC dataset with four popular parameters are named as precision, Recall, F1-score and Accuracy. ERA-MLM performs the best across all metrics, while SVM performs the worst. The ANB-MLM and CBAP-MLM models are intermediate performers but still better than SVM. The increasing trend in values suggests that the newer methods (especially ERA-MLM) have improved precision, recall, and overall classification performance.

When the precision value is increased the value is decreased in the proposed methods. The precision value for the SVM method is 93.72, ANB- MLM is 95.16, CBAP-MLM is 96.33 and ERA-MLM is 97.78. The values for Recall are 94.46, 95.22, 96.37 and 97.28.

The values for F1-Score are 93.27, 95.41,95.25 and 96.21. Finally the accuracy values are 93.29, 94.23, 95.53 and 96.36. Based on the analyses of the Accuracy values the ERA-MLM values are increased compared to other two proposed methods.

suggested ERA-MLM efficiently determines the similarity of Topic categorization related articles for all metrics taken into consideration.

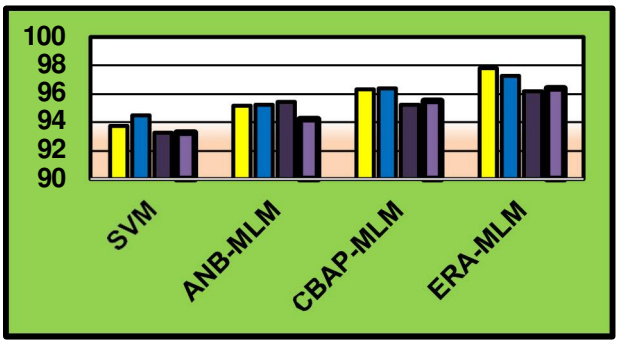


Fig 1.2 The Comparison chart of Proposed Machine learning methods using TREC dataset

The categorization of content-based document similarity mining using the Enhanced Reinforcement Machine learning Topic Document Similarity Method is explained in Fig. 1.2. The Enhanced Reinforcement Machine learning Topic Document Similarity Method is a superior approach to existing methods like Support Vector Machine, Advanced Naïve Bayes, CBAP-MLM, and ERA- MLM in terms of accuracy, precision, recall, and F1- Score measure. ERA-MLM outperforms other methods, with higher precision values, higher accuracy rates, and higher F1-Score measures. It achieves a maximum prediction accuracy of 96.36%, significantly higher than traditional algorithms for customer information collection.

V. CONCLUSION

The document collection and the region specified by the vertices of the hypercube become increasingly populated when the maximum Euclidean distance and the potential document representations are added. The term frequency- inverse document frequency method is used to add a document; as a result, the weights of some terms in the document increase while those of the other terms decrease. When new documents are added, the region in which they are located expands, controlling the density of the collection representation as a whole to find the topic categorization related articles from the TREC database. This is done using the ERA-MLM method. This study compares the precision, recall, F1-score, and accuracy numbers obtained from the Reuter's dataset with those obtained from other approaches. When compared to current methods, the suggested ERA-MLM efficiently determines the similarity of Topic categorization related articles for all metrics taken into consideration.

REFERENCES

[1] Alexandrino. M.V, G. Comarela, A. S. da Silva, and J. Lisboa-Filho, "A Focused Crawler for Web Feature Service and Web Map Service Discovering," in International Symposium on Web and Wireless Geographical Information Systems, 2020, pp. 111-124.
[2] Dhelim et al. (2020) Dhelim S, Ning H, Aung N, Huang R, Ma J. Personality-aware product recommendation system based on user interests mining and metapath discovery. IEEE Transactions on Computational Social Systems. 2020;8(1):86–98. [Google Scholar]
[3] Eirinaki et al. (2018) Eirinaki M, Gao J, Varlamis I, Tserpes K. Future generation computer systems. vol. 78. Elsevier; 2018. Recommender systems for large-scale social networks: a review of challenges and solutions; pp. 413–418. [Google Scholar]
[4] Ge et al. (2020) Ge J, Shi L-L, Wu Y, Liu J. Human-driven dynamic community influence maximization in social media data streams. IEEE Access.2020;8:162238–162251.doi: 10.1109/ACCESS.2020.3022096. [CrossRef] [Google Scholar]
[5] Haroon.M, Tripathi.M.M, and Ahmad.F, "Application of machine learning in forensic science," in Critical Concepts, Standards, and Techniques in Cyber Forensics, IGI Global, 2020,pp. 228-239.
[6] HudedM.S, Balutagi.S, and Ranjan.A, "Mapping of literature on data mining by j-gate database," 2019.
[7] Javed et al. (2021) Javed U, Shaukat K, Hameed IA, Iqbal F, Alam TM, Luo S. A review of content- based and context-based recommendation systems. International Journal of Emerging Technologies in Learning (iJET) 2021;16(3):274–306.doi: 10.3991/ijet.v16i03.18851. [CrossRef] [Google Scholar]
[8] Jin, J.; Lin, X. , "Web Log Analysis and Security Assessment Method Based on Data Mining," Computational Intelligence and Neuroscience, vol. 2022, pp. 1–9, Aug. 2022.
[9] Karthica. M, Meenakshi Sundaram. M, Vandarkuzhali. J, "EnhancedReinforcement Algorithm for Topic Categorization Using Machine Learning Method" Vol.10No.10s(2025). https://doi.org/10.52783/jisem.v10i10s.1411
[10] Khan.W and Haroon.M, "An unsupervised deep learning ensemble model for anomaly detection in static attributed social networks," International Journal of Cognitive Computing in Engineering, vol. 3, pp. 153-160, 2022. Available from: https://doi.org/10.1016/j.ijcce.2022.08.002
[11] Karthica.M,MeenakshiSundaram.K, J.Vandarkuzhali.,"A study on content based Automated parsing Machine learning Algorithm for Document Similarity"20 No.s14(2024).

[12] Karthica. M, Meenakshi Sundaram.K," Advanced Naïve Bayes Machine Learning System for Document Similarity checking"Volume3, issue1,2023.

[13] Li.C, "Research on an Enhanced Web Information Processing Technology based on AIS Text M-ining," Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering), vol. 14, pp. 29- 36, 2021.

[14] Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1889–1904, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1 /2021.acl-long.148

[15] Nikzad.E, Chenaghlu.N, M & Gao, J. (2020). Deep learning based text classification: A comprehensive review. arXiv preprint arXiv:200403705

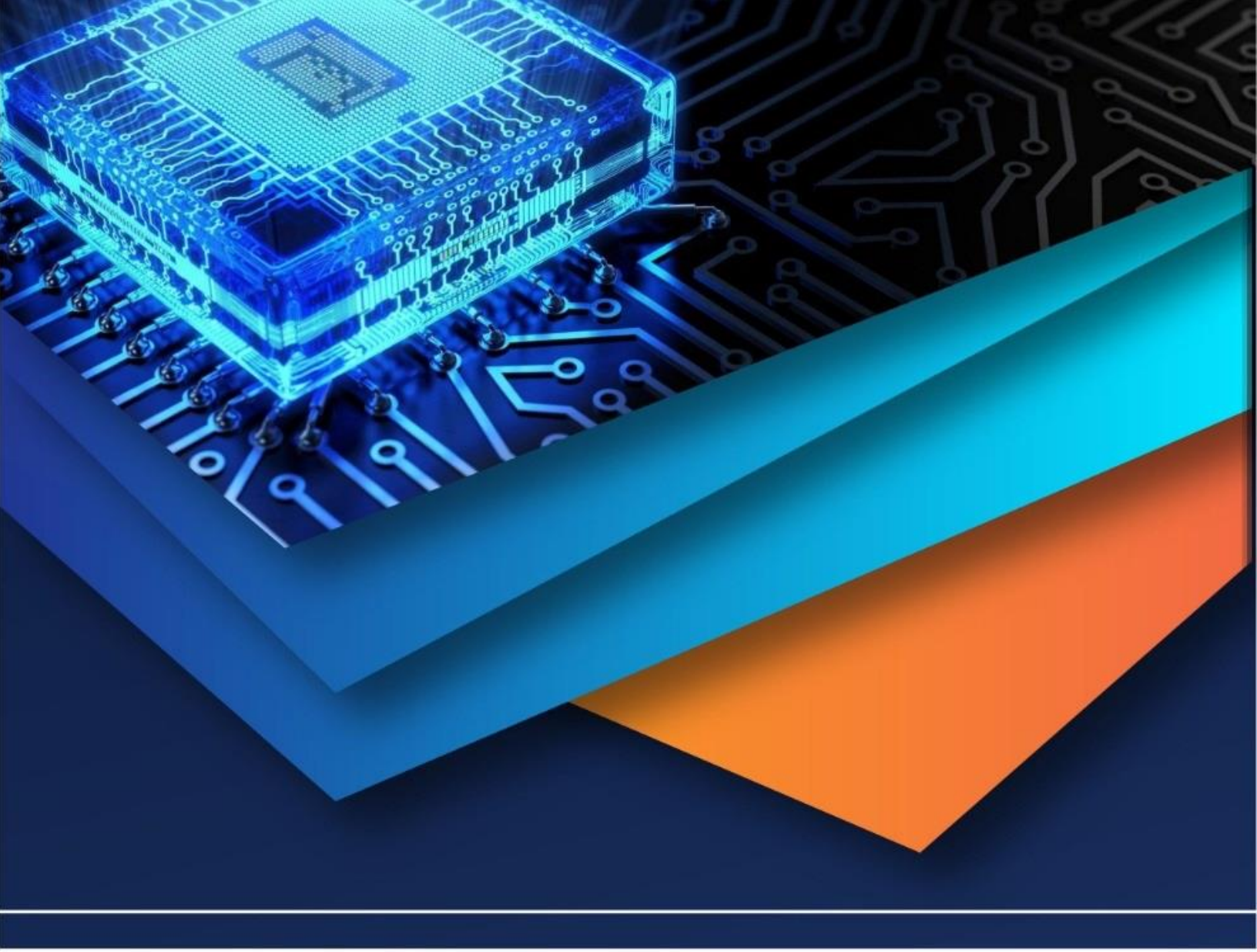
[16] Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. 2022. Docogen: Domain counterfactual generation for low resource domain adaptation. In Proceedings of the 60th Annual Meeting of the Association of Computational Linguistics(ACL). https://doi.org/10.18653/v1/2022.acl-long.533

[17] Phukon.K, "Incorporation of contextual information through Graph Modeling in Web content mining," Indian Journal of Science and Technology, vol. 13, pp. 4573-4578, 2020.

[18] Pradhan.N and Dhaka.V, "Comparison-based study of pagerank algorithm using web structure mining and web content mining," in Smart Systems and IoT: Innovations in Computing, ed: Springer, 2020, pp. 719-729.

[19] Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. 2020. Fairness and robustness in invariant learning: A case study in toxicity classification. arXiv preprint arXiv:2011.06485.

[20] Roudsari, A. H., Afshar, J., Lee, S., & Lee, W. (2021). Comparison and analysis of embedding methods for patent documents. In 2021 IEEE International Conference on Big Data and Sm.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY