



A Comparative Study of Various Text Summarization Methods

¹Chinni Mohith, ²Chinni Sai Raj Dheeraj, ³K Shiva Reddy

^{1,2,3}Reserch Scholar, Department of CSE

^{1,2,3}Apex Institute of Technology,

^{1,2,3}Chandigarh University, Mohali, Punjab

Abstract : Text summarization refers to the process of condensing long texts into short notes while keeping the most significant information, it is an application of natural language processing. This research provides an overview of text summarization methods that make use of different technologies such as natural language processing and machine learning. Considering many approaches, including extractive and abstractive summarising techniques, and discussing the benefits and drawbacks of each. Furthermore, the role that machine learning techniques such as convolutional neural networks (CNNs), transformer models like BERT and recurrent neural networks (RNNs) play in automating the summarization process is discussed. This study also highlights certain important factors, such as maintaining coherence and evaluating summary quality. Study concludes by discussing potential directions for creating text summarization techniques in future those making use of machine learning and natural language processing techniques. Factors such as Context Relevance, Keyword count, Accuracy, Framing and Decrease in the word count are manually evaluated for each technique. TF-IDF Method, Method based on Clusters, Neural Networks and UML based Text summarisation methods are some which are considered for evaluation. This paper presents the key findings and research gaps after studying the most cited research works.

IndexTerms – Text Summarization, Recurrent Neural Networks, BERT, Text Context, Natural Language Processing, Feature Extraction.

I. INTRODUCTION

The objective of the cutting-edge field of natural language processing (NLP) research is to automate the process of compressing enormous quantities of text into succinct summaries through the use of machine learning for text summarization. This technology has potential for a number of uses, including supporting document analysis and decision-making as well as information retrieval. Fundamentally, machine learning helps computers identify patterns in data and forecast future events, whereas natural language processing (NLP) helps computers comprehend and interpret human language. Researchers are creating algorithms that can efficiently extract important information from texts, articles, or other textual material by merging various fields. In real life, text summary algorithms use strategies like key phrase recognition, sentence structure analysis, and context awareness to produce summaries that successfully communicate the primary concepts of the original content. Large text and summary datasets may be used to train these algorithms to identify significant and pertinent patterns. Because of this, they are able to create summaries that are shorter overall while still retaining important information. This study topic is still developing, with deep learning and neural network advancements driving important advances in the accuracy and efficiency of text summarization systems.

Text summarization is a potential technology that combines natural language processing and machine learning to automatically reduce massive text volumes into brief summaries. NLP makes language interpretation easier, whereas machine learning makes it possible to learn patterns from data. To create summaries, algorithms employ methods such as context comprehension, phrase structure analysis, and keyword identification. They can identify significant trends and generate concise summaries that retain crucial information since they have been trained on large datasets. Deep learning and neural network advancements are constantly improving efficiency and accuracy. This area of research is dynamic and has the capacity to fully transform document analysis, data retrieval, and also the decision-making procedures. There are exactly 2 main types of text summary methods: extractive summarization and the other abstractive summarization. Without altering the original text, extractive summarization groups combined the key sentences or phrases from the source materials to create a summary. On the other hand, the abstractive

summarization entails comprehending the previous text through text analysis and interpretation utilizing the language technique. The goal of the abstractive summarization is to provide the generalized summary for communicating information succinctly. Figure 1 shows the steps of text summarization of the model.

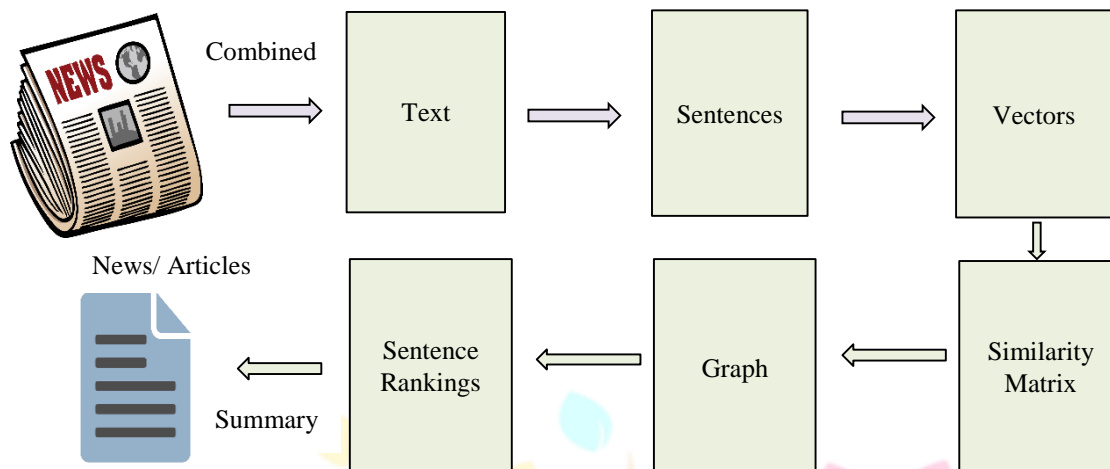


Figure 1. Model of text summarization

II. LITERATURE REVIEW

Text summaries cut down on word count without compromising meaning. For humans, summarizing long articles can be challenging. Methods include extractive (selecting important sections) and abstractive (rewriting). This survey focuses on extractive techniques for text summarization, crucial for understanding advancements in natural language processing. This literature review delves into the complexities of text summarization, particularly focusing on extractive techniques for effective natural language processing advancements.

Textual content analytics, traditionally done manually, is inefficient. To address this, software employing text mining and NLP algorithms has emerged as stated by D. Suleiman et al [1], enabling efficient analysis of large text volumes. These tools offer early issue detection by identifying customer complaints. Text analytics transforms unstructured data into valuable insights, making it indispensable for businesses. Within NLP, text summarization stands out, with its potential to significantly impact various aspects of our lives.

Text summarization involves condensing key information from a text document into a concise summary. Human comprehension of text is challenging. An extensive overview of abstractive text summarizing techniques is given in this study and categorizing them into structured and semantic approaches. It examines methodologies, challenges, and benchmarks, concluding that most methods yield cohesive coherent, and informative summaries as stated by S. H. B. Sri and S. R. Dutta [2].

The exponential growth of data across domains necessitates automatic text summarization. This study focuses on leveraging Natural Language Processing (NLP) for summarization, crucial in various fields. Extractive summarization methods are explored, proposing a Support Vector Machine (SVM)-based approach to enhance summary quality. Preprocessing involves tokenization, stop word removal, case folding, and stemming. Sentences are grouped into clusters using similarity measures and TF-IDF values, improving summarization performance as described by Gambhir et al. [3].

Content summarization condenses source documents while retaining key information. It employs abstractive or extractive methods. Abstractive summarization utilizes natural language processing tools, while extractive methods rely on statistical, linguistic, and heuristic techniques. Various strategies have been developed for summarizing text in different languages. In an effort to improve knowledge and comprehension of summary approaches, this research investigates extractive and abstractive text summarization strategies by M. R. Prathima et al. [4]

This paper by A. Joshi et al. [5] examines that in these recent years have seen a surge in text data, necessitating effective summarization. This paper presents an analysis of extractive text summarization techniques, addressing challenges in single and multiple document summarization. It discusses generic challenges and state-of-the-art techniques, vital for advancing text summarization.

Text Summarization extracts key information from text documents, crucial for human comprehension. It encompasses abstractive and extractive techniques. This paper reviews extractive methods, examining techniques, benchmark datasets, and challenges. It emphasizes producing concise, coherent, and informative summaries. This paper is written by A. Rajasekaran et al. [6] and they

dives into extractive text summarization, analyzing methods, benchmark datasets, and challenges, aiming for concise, coherent summaries.

In the era of abundant information, Text Summarization is crucial. This paper offers a detailed review of Automatic Text Summarization, covering various approaches and methods for extracting and abstracting key information from diverse sources. This paper meticulously reviews Automatic Text Summarization, detailing diverse extraction and abstraction methods.

III. RESEARCH METHODOLOGY

To evaluate the process there are two types of methods extractive and the abstractive text summarization method in this research paper we are mainly focused on the extractive methods and make the best of it In order to create a summary, the extractive-based summarization technique chooses informational sentences from the text exactly as they exist in the source according to predetermined criteria. Selecting the important and likely to be included sentences from the input material for the summary is the primary problem prior to extractive summarization. Sentence scoring is used for this purpose, taking into account the properties of sentences. It ranks sentences depending on their score after first giving each sentence a score based on a characteristic. Sentences that receive the final summary will most likely have the highest score. The following tactics are part of the extractive text summarizing technique.

3.1 Document Frequency-Term Frequency Inverse Method

A word's importance in a particular document is shown numerically by inverse document frequency (IDF) and term frequency (TF). The term's TF is determined by how many times it occurs in the text, and the IDF is a metric that makes rarer terms in the collection more important and less important terms in the collection. Subsequently, sentences are graded based on their product, and those with the highest are covered in the synopsis. A problem with this strategy is that it takes longer phrases occasionally receive higher scores since they have more words in them.

Neri-Mendoza et al [7] suggested a method of choosing terms and allocating weights using assistance from TF-IDF. An unsupervised learning technique was employed to provide a summary that was not redundant. Sarkar used sentence feature in addition to TF-IDF to boost the news summary result. In her study, the TF-IDF issue is covered and weighted item set based approach is used. This model links many important terms, extracts related item sets using TF-IDF, and then generates a summary. The premise of Kamal and Sultana's proposed method is the co-occurrence of biological terms in sentences. As an overview, the frequency of occurrence is calculated using three feature words. Probabilistic feature selection is called GSS, when it is multiplied by TF-IDF indicates how important a term is to have in the summary. The summary that is based on extraction approach creates a summary by choosing informational sentences from the text exactly as they exist in the source according to predetermined standards. Selecting the important and likely to be included sentences from the input material for the summary is the primary problem prior to extractive summarization. Sentence scoring is used for this purpose, taking into account the properties of sentences. It ranks sentences depending on their score after first giving each sentence a score based on a characteristic. Sentences that receive the final summary will most likely have the highest score.

3.2 Method Based on Clusters

Documents are written so that distinct concepts are covered in different parts. It seems sense to assume that summaries should include various topics that are divided into distinct document parts. If the material for which the summary is being provided covers completely separate topics, the summarizer incorporates this information by clustering. The TF-IDF of scores of words is used to represent the document. A cluster's topic is represented by a high frequency word. The selection of the summary phrase is determined by how well it connects to the cluster's subject. High-relevant summaries are produced by the cluster-based approach for the specified query or document subject. Yadav, Arun Kumar, et al [8] grouped texts together using the K-means clustering-algorithm. Based on sentence specifications, the cluster's core sentence is considered as the conclusion. Local and global search strategies are employed in the generation of summary sentences. Spectral clustering and the LexRank technique, which results in maximum coverage and minimal redundancy. The sparse matrix of similar texts is created by the k-nearest-neighbor method. In brief, the LexRank score is obtained via based on shared features. After the document is graphed, Text Rank is used to identify significant sentences, and clusters are created based on how similar phrases are to one another. In order to produce a summary, Semantically-grouped linked semi-structured items are assigned labels.

3.3 Using Neural Networks for Text Summarization

A processing system that emulates the way the human brain learns is called a neural network. Neural networks are an artificial neuron construct that is networked and processes data using a numerical model of computing. When it comes to text summarization, the plan entails training neural networks that identify the categories of sentences that belong in the synopsis. Sentences from test paragraphs are used to train neural networks, and each sentence's inclusion in or exclusion from the summary is verified. Training is provided based on the user's needs. Although a neural network can classify summary phrases effectively, it has trouble with long training times. Based on the properties each phrase has in the feature vector, and it is assigned a score to each sentence. Following

neural network training, a little weight is removed to get rid of unusual features. High scoring sentences are used to construct summaries. In order to decrease processing, G. Vijay Kumar et al [9] suggested neural network-based summary of Vietnamese text and semi-supervised learning. To create a summary, sentences are rated based on a word set. A recurrent neural network language model using word co-event auxiliary data was suggested by Zhang, Mengli, et al [10]. The probabilistic generative paradigm is used in language models to provide a summary by ranking sentences according to the frequency of each distinct term. The results of neural networks and other feature-selected summarization methods were examined. A basic word sum known as the phrase embedding is derived using an autoencoder on a binary parse tree that a recursive neural network. Gambhir et al [11] proposed employing recurrent neural networks for part of speech disambiguation. A bit vector of a voice segment is fed to a neural network that classifies phrases and outputs a summary.

3.4 Fuzzy Logic /UML Based Text Summarization

Du, Yan, and Hua Huo [12] designed a news text summarization system based on multi-feature and fuzzy logic and they discussed in their paper that fuzzy logic which is true-or-false (1 or 0), uses degrees of truth to simulate human reasoning. Performance in the fuzzy system is mostly determined by the fuzzy rules and membership function. For every sentence for the output, a value between zeros to one is determined based on the characteristics in each phrase and the rules defined in a knowledge base. IF-THEN rules are used to extract important sentences depending on feature criteria. Sentences are ordered based on their score. In conclusion, high-scoring sentences taken out. Fuzzy logic systems are simple, flexible, and capable of handling noisy, erroneous, and distorted input data.

Al Qassem et al [13] established fuzzy criteria and a triangle membership function to assess sentences according to their attributes. Utilising latent semantic analysis enhances the summary outcome. Learning automata for determining phrase similarity and particle swarm optimisation for feature weighting were proposed by Goularte et al [14]. The fuzzy approach uses rules created by humans to categorise summary phrases. A fuzzy inference system was created by them to score sentences according to characteristics. To create a summary, rules are triggered in accordance with each sentence's overall score. Key phrases are those with high membership values according to the maximum entropy model based on the degree of membership in the set. The final membership value is extracted using sentence features. Sharma Shikha et al [15] analyzed the Need for video summarization for online classes and created a system for video summarization.

Premakumara et al [16] provided sentence extraction based on fuzzy logic system using certain properties. Where each phrase is assigned a number between 0 and 1 according to its properties, and a sentence is selected using the criteria from the knowledge base. This study provides an overview of summarising methodology that includes many methods for measuring various elements that are crucial to the summary process. Word relevance and context, keyword count, framing, and word count reduction are some of them. Each of these factors are required to ensure the efficiency and uniformity of the summaries that are generated. Context relevance is determined by evaluating how effectively the extracted keywords capture the context and underlying meaning/pattern of the source material [17].

This evaluation includes examining the important keywords and determining the degree to which they malign the topic of the text. Keyword counts are also used to evaluate the importance and relevancy of specific terms within the text. By the use of keyword frequency analysis, we can describe the important topics and concepts to be included in the text [18][19].

To evaluate the correctness, cross-referencing of the retrieved keywords with a manually prescribed or ground truth collection of phrases is done. Recall, accuracy, and F1-score [20][21][22] are evaluation metrics that are used to determine how well the keyword extraction is performed. Framing analysis is the process of analysing how the keywords are presented and interpreted in context to the summary. By thoroughly monitoring the keywords usage and arrangement, we can ensure that they accurately pass the required message. Finally, techniques like text summarising are required to bring down word count without losing important information [23][24][25]. Table 1 shows the key finding of various method of text summarization.

Table 1 Summary of Key Findings of Various Methods

Parameter	TF-IDF Method	Method Based on Clusters	Using Neural Networks	Fuzzy Logic/ UML Based Text Summarization
Context Relevance	Measures how well extracted keywords capture the meaning and context of documents by examining top-ranked keywords.	Utilizes clusters to group similar documents or sentences, which may or may not reflect context relevance.	Uses neural networks to learn representations of input text and generate summaries based on learned context.	Utilizes UML diagrams to represent the structure and relationships of text components, potentially capturing context relevance through modelling.
Keyword Counts	Counts the occurrence of each keyword within documents to identify significant terms.	May or may not directly count keyword occurrences, depending on the clustering algorithm used.	Typically does not focus on counting individual keyword occurrences; instead, neural networks learn implicit representations.	May count occurrences of terms within UML diagrams or use keyword frequency as a factor in summarization.
Accuracy	Can be evaluated by comparing extracted keywords with ground truth or manually annotated keywords using metrics like precision, recall, and F1-score.	Accuracy depends on the clustering algorithm's ability to group similar documents or sentences effectively.	Accuracy is measured by comparing generated summaries with human-written summaries or ground truth using evaluation metrics.	Accuracy can be assessed by comparing generated summaries with manually created summaries based on UML diagrams.
Framing	Framing is determined by how well the extracted keywords are interpreted in the summary.	Framing may be influenced by the clustering algorithm's ability to group related content together.	The way a neural network uses its representations to produce summaries has an impact on framing.	Framing may be influenced by the structure and relationships depicted in the UML diagrams and how they are interpreted in the summary.
Decrease in Word Count (Minimization)	Minimizes word count by selecting the most informative keywords or sentences from the input documents.	Clustering may indirectly reduce word count by grouping similar content together, but it doesn't directly target word count minimization.	Minimizes word count by generating concise summaries using neural network-generated representations.	May minimize word count by focusing on key components represented in the UML diagrams and omitting redundant or less important information.

IV. RESULTS AND DISCUSSION

There are two primary categories of text summarization techniques: extractive and abstractive methods. For these kinds of tasks, a lot of research has been done on both machine learning (ML) and deep learning (DL) techniques. For extractive summarization, machine learning techniques such as Naive Bayes, Support Vector Machines (SVM), and decision trees are commonly employed. These methods pick out the most pertinent sentences or phrases from the text that are important. Although they are adept at recognizing factual information, they frequently have trouble deciphering subtler contextual cues.

In lengthy texts, ML models for extractive summarization typically yield extremely coherent summaries that lack fluency and relevance. Additionally, these models may be skewed toward choosing sentences that only include common terms, without taking the context into account. Abstractive summarization is using DL techniques like Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and, more recently, Transformer-based models like BERT and GPT-3 more and more. By creating new sentences instead of just picking them from the source text, these techniques give the summary a more flexible and human-like quality.

When it comes to producing more coherent, condensed, and context-aware summaries, DL models perform better than conventional ML techniques. Transformer-based models are particularly good at managing lengthy textual dependencies and comprehending the document's overall context, which results in summaries that are more pertinent.

The main drawback is that machine learning techniques are not able to understand the semantic relationships between words and instead mainly rely on handcrafted features. As a result, their effectiveness for abstractive summarization is typically lower. Abstractive summarization has advanced significantly with the use of DL techniques. They still struggle, though, to produce summaries that are factually correct and grammatically flawless. Furthermore, the computational cost of deep learning models and their need for extensive training datasets may restrict their practical applications.

Method	Context Relevance	Keyword Counts	Accuracy	Framing	Decrease in Word Count
TF-IDF Method	High	Medium	High	Medium	High
Method Based on Clusters	Medium	Medium	Medium	Medium	Medium
Using Neural Networks	High	High	High	High	High
Fuzzy Logic/UML Based Text Summarization	Medium	High	Medium	Medium	High

V. CONCLUSION

The above results determine the efficiency of Neural Networks in the context of text summarization, while considering many important aspects. Neural Networks are known for their ability to identify complex patterns and hence they outperformed other conventional techniques like TF-IDF, clustering at maintaining context relevance. Additionally, they are better in identification of important terms, which resulted in summaries with better keyword count. Neural Networks have also performed better than any other models in accuracy evaluation, which determine their ability to learn from large datasets.

Furthermore, Neural Networks are good at framing the summaries because of their ability to make connections between pieces of text in such a way that preserves the original context. And, lastly, they yield in a decrease in word count by giving priority to relevant and important information, guaranteeing small yet useful summaries. Neural networks is the best and most suitable option for text summarization as they outperformed other methods in terms of context relevance, keyword identification, precision, framing, and writing down of words

REFERENCES

- [1]. D. Suleiman and A. Awajan, "Deep Learning Based Abstractive Text Summarization: Approaches, Datasets, Evaluation Measures, and Challenges," *Math. Probl. Eng.*, vol. 2020, 2020, doi: 10.1155/2020/9365340.
- [2]. S. H. B. Sri and S. R. Dutta, "A survey on automatic text summarization techniques," *J. Phys. Conf. Ser.*, vol. 2040, no. 1, pp. 268–271, 2021, doi: 10.1088/1742-6596/2040/1/012044.
- [3]. Gambhir, Mahak, and Vishal Gupta. "Improved hybrid text summarization system using deep contextualized embeddings and statistical features." *Multimedia Tools and Applications* (2024): 1-30.
- [4]. M. R. Prathima and H. R. Divakar, "Automatic Extractive Text Summarization Using K-Means Clustering," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 6, pp. 782–787, 2018, doi: 10.26438/ijcse/v6i6.782787.
- [5]. A. Joshi, E. Fidalgo, and E. Alegre, "Deep Learning based Text Summarization: Approaches, Databases and Evaluation Measures," *Int. Conf. Appl. Intell. Syst.*, no. DI, pp. 1–4, 2018.
- [6]. A. Rajasekaran and D. R. Varalakshmi, "Review on automatic text summarization," *Int. J. Eng. Technol.*, vol. 7, no. 3.3, p. 456, 2018, doi: 10.14419/ijet.v7i2.33.14210.
- [7]. Neri-Mendoza, Verónica, et al. "Generic and Update Multi-Document Text Summarization based on Genetic Algorithm." *Computación y Sistemas* 27.1 (2023): 269-279.
- [8]. Yadav, Arun Kumar, et al. "Extractive text summarization using deep learning approach." *International Journal of Information Technology* 14.5 (2022): 2407-2415.
- [9]. G. Vijay Kumar, A. Yadav, B. Vishnupriya, M. Naga Lahari, J. Smriti, and D. Samved Reddy, "Text Summarizing Using NLP," *Adv. Parallel Comput.*, vol. 39, pp. 60–67, 2021, doi: 10.3233/APC210179.
- [10]. Zhang, Mengli, et al. "A comprehensive survey of abstractive text summarization based on deep learning." *Computational intelligence and neuroscience* 2022.1 (2022): 7132226.
- [11]. Gambhir, Mahak, and Vishal Gupta. "Deep learning-based extractive text summarization with word-level attention mechanism." *Multimedia Tools and Applications* 81.15 (2022): 20829-20852.
- [12]. Du, Yan, and Hua Huo. "News text summarization based on multi-feature and fuzzy logic." *IEEE Access* 8 (2020): 140261-140272.
- [13]. Al Qassem, Lamees, et al. "Automatic Arabic text summarization based on fuzzy logic." *Proceedings of the 3rd international conference on natural language and speech processing*. 2019.
- [14]. Goularte, Fábio Bif, et al. "A text summarization method based on fuzzy rules and applicable to automated assessment." *Expert Systems with Applications* 115 (2019): 264-275.
- [15]. Sharma, Shikha, and Madan Lal Saini. "Analyzing the Need for Video Summarization for Online Classes Conducted During

Covid-19 Lockdown." *Data, Engineering and Applications: Select Proceedings of IDEA 2021*. Singapore: Springer Nature Singapore, 2022. 333-342.

- [16]. Premakumara, Nilantha, et al. "Optimized Text Summarization method based on fuzzy logic." *2022 2nd International Conference on Image Processing and Robotics (ICIPRob)*. IEEE, 2022.
- [17]. S. Kulshrestha and M. L. Saini, "Study for the Prediction of E-Commerce Business Market Growth using Machine Learning Algorithm," *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, Jaipur, India, 2020, pp. 1-6, doi: 10.1109/ICRAIE51050.2020.9358275.
- [18]. Kavita Lal, Madan Lal Saini; A study on deep fake identification techniques using deep learning. *AIP Conf. Proc.* 15 June 2023; 2782 (1): 020155. <https://doi.org/10.1063/5.0154828>
- [19]. Y. Singh, M. Saini and Savita, "Impact and Performance Analysis of Various Activation Functions for Classification Problems," *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, Bangalore, India, 2023, pp. 1-7, doi: 10.1109/InC457730.2023.10263129.
- [20]. M. Sohail, M. Lal Saini, V. P. Singh, S. Dhir and V. Patel, "A Comparative Study of Machine Learning and Deep Learning Algorithm for Handwritten Digit Recognition," *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, Gautam Buddha Nagar, India, 2023, pp. 1283-1288, doi: 10.1109/IC3I59117.2023.10397956
- [21]. M. Lal Saini, B. Tripathi and M. S. Mirza, "Evaluating the Performance of Deep Learning Models in Handwritten Digit Recognition," *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, Tashkent, Uzbekistan, 2023, pp. 116-121, doi: 10.1109/ICTACS59847.2023.10390027.
- [22]. S. Chalechema, M. L. Saini, I. Perla and A. V. Shivanand, "Customer Segmentation Using K Means Algorithm and RFM Model," *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, Greater Noida, India, 2023, pp. 393-398, doi: 10.1109/ICCCIS60361.2023.10425556.
- [23]. S. Mittal, R. Agarwal, M. L. Saini and A. Kumar, "A Logistic Regression Approach for Detecting Phishing Websites," *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, Faridabad, India, 2023, pp. 76-81, doi: 10.1109/ICAICCIT60255.2023.10466221.
- [24]. V. Prabhas, M. Lal Saini, C. Mohith, R. Kumar and B. Tripathi, "Segmentation of E-Commerce Data Using K-Means Clustering Algorithm," *2023 Global Conference on Information Technologies and Communications (GCITC)*, Bangalore, India, 2023, pp. 1-6, doi: 10.1109/GCITC60406.2023.10426132.
- [25]. M. L. Saini, A. Patnaik, Mahadev, D. C. Sati and R. Kumar, "Deepfake Detection System Using Deep Neural Networks," *2024 2nd International Conference on Computer, Communication and Control (IC4)*, Indore, India, 2024, pp. 1-5, doi: 10.1109/IC457434.2024.10486659.

