Artificial Intelligence-Based Multi-Document Extractive Summarization Using Sentence-Level Features

Karina Patel[1]*, Shilpa Serasiya[2]
M Tech Student, Department of Computer Engineering, Sankalchand Patel College of Engineering, Visnagar, India[1]*
Assistant Professor, IT Department, Sankalchand Patel College of Engineering, Visnagar, India[2]

karinapatel5687@gmail.com[1*],shilpapatel84@gmail.com[2]

Abstract: The goal of multi-document text summarizing is to automatically generate a succinct yet thorough summary from multiple connected sources. The growing volume of digital text data is making the task of document summarization more challenging. More effective summarization is now possible because to significant advancements in artificial intelligence techniques and pre-trained language models; nonetheless, the majority of deep learning approaches require a significant amount of processing power and large training datasets. This work suggests a multi-document text summarizing method powered by artificial intelligence that makes use of supervised learning and sentence feature modeling.

The proposed system begins with heavy preprocessing that involves lemmatization, normalization, stop-word removal, and sentence segmentation. Statistical and structural features, including the TF–IDF representation, length of the sentence, normalized position of the sentence, and numeric token count, are extracted to denote sentence significance. Like other supervised extraction algorithms discussed in previous literature, a Logistic Regression classifier is created using a balanced dataset to classify sentences into significant or insignificant categories. Probabilities are used for ranking of selected sentences, while cosine similarity is applied under a certain word limit constraint to avoidredundancies.

The suggested approach demonstrates effective accuracy while needing very little processing effort, according to experimental examination of the DUC 2003 corpus. Based on the ROUGE measure, the model produces ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.1722, 0.0306, and 0.1048, respectively. It demonstrates how basic AI models combined with useful sentence-based features could be a reliable benchmark for information extraction from several publications. To increase the effectiveness of summaries, the authors intend to look into contextual representation algorithms like Sentence-BERT and implement human evaluation standards in subsequent studies.

Keywords: Multi-Document Summarization, Extractive Summarization, Artificial Intelligence, Logistic Regression, TF-IDF,Natural Language Processing (NLP),

I. INTRODUCTION

The enormous growth of digital text data in the current era of information explosion has made it necessary to develop automatic text summarizing systems that can extract the most crucial information from massive document libraries. Creating a condensed text summary while maintaining the primary sense of the original content is the aim of text summarization. While there are several methods used in text summarization, extractive text summarization is one of the most widely utilized methods in practice [5], [19].

Due to problems with redundancy, overlapping information, and semantic variety among the relevant papers, summary generation for multiple documents (MDS) is far more challenging than summary creation for individual documents. In a manner not necessary when working with a single document, MDS demands that sentences be chosen from several sources without duplication or redundancy. Because of this, MDS systems continue to be a topic of interest for research.

Conventional extractive summarization algorithms frequently rely on unsupervised ranking methods or surface-level statistical indicators, which may not adequately capture the real significance of sentences across document collections. Even though contemporary deep learning models have demonstrated encouraging outcomes, they are often less appropriate for lightweight or resource-constrained situations since they need a lot of processing power and training data [2], [3]. Consequently, machine learning-based methods that employ thoughtfully crafted sentence-level characteristics continue to be a workable and efficient alternative [13].

To address these challenges, this study proposes an artificial intelligence-based multi-document extractive summarization framework using supervised machine learning techniques. In particular, the proposed model makes use of the DUC 2003 corpus and employs preprocessing, involving CSV parsing, sentence extraction, stop word elimination, and lemmatization. For feature representation, both statistical and structural sentence-level features are employed, including TF-IDF encoding, sentence length,

http://doi.org/10.63766/spujstmr.26.000082

sentence position, and numeric token count.A logistic regression model is trained to detect extractable sentences based on ROUGE labeling [9]. Also, a topic-wise summary generator has been implemented by employing cosine similarity to avoid redundancy.

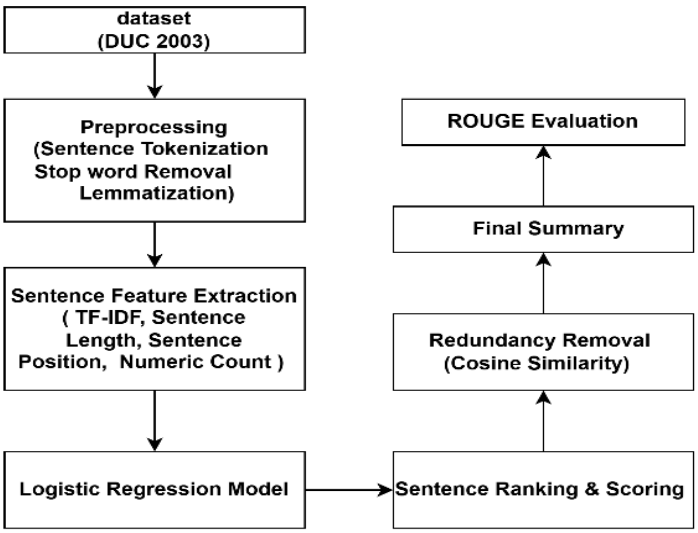

Fig.1: Architecture of the multi-document extractive summarization system.

II. LITERATURE REVIEW

In recent times, automatic text summarization has received significant attention owing to the exponential rise in digital content. Initially, extractive summarization techniques utilized purely statistical and heuristic sentence selection criteria like term frequency, sentence order, and lexical matching. Such methods concentrated solely on identifying key sentences using surface-level features [6].

With developments in machine learning algorithms, the supervised learning technique was brought into play to enhance sentence selection. Classical machine learning classifiers such as Logistic Regression, Naïve Bayes, and SVMs were used to find summary-worthy sentences by constructing sentence-specific features. The interpretable nature and efficient computation of feature-based models made them more applicable to real-world scenarios [8], [13].

While these deep learning techniques yield impressive results, they demand considerable computing power and extensive datasets. On the other hand, simple machine learning algorithms that use sentence-level statistical features continue to deliver optimal results and have practical significance [1], [7].For evaluation, ROUGE remains the most widely used metric in summarization research. Proposed by Chin-Yew Lin, ROUGE measures the overlap between system-generated summaries and reference summaries, providing a standard benchmark for performance comparison [18].

Based on the existing literature, there is a need for computationally efficient and interpretable multi-document extractive summarization systems that leverage sentence-level features while maintaining competitive performance. Therefore, this study proposes a supervised machine learning approach using Logistic Regression combined with TFIDF and structural sentence features to generate concise and informative summaries [14].

TABLE I
CRITICAL COMPARISON OF EXISTING MULTI-DOCUMENT EXTRACTIVE SUMMARIZATION METHODS

Ref.	Method	Type	Strengths	Limitations
[1]Shinde et al., 2022 [15] S. Serasiya et al., 2025	Hybrid Extractive–Abstractive	Deep Learning	Handles scientific articles well; improved semantic quality	Computationally expensive; complex pipeline
[4] Su et al., 2023	Heterogeneous Graph Neural Network	Graph-based	Captures multi-granularity relationships	High model complexity
[6] Gao et al., 2021 (SimCSE)	Contrastive Sentence Embedding	Representation Learning	Produces high-quality sentence embeddings	Not a complete summarization system

[7]Beltagy et al., 2021 (Longformer)	Long-document Transformer	Deep Learning	Handles long documents efficiently	Heavy computational cost
[8]Reimers&Gurevych, 2021 (SBERT)	Sentence Embedding	Representation Learning	Good semantic similarity performance	Needs downstream model for summarization
[13] Varma & Balamurugan, 2021	ML Comparative Study	Machine Learning	Shows effectiveness of classical ML	Limited multi-document focus
[14]Bandaru&Radhika, 2022	Hybrid Semantic Similarity	Extractive ML	Improved sentence ranking	Feature engineering required
[18] Chen et al., 2021 (SgSum)	Sub-graph Selection	Graph-based	Models document structure well	Resource intensive

III. PROBLEM STATEMENT

The rapid growth of digital information and online textual content has created significant challenges in extracting relevant information efficiently from multiple documents. Traditional manual summarization methods are time-consuming, inefficient, and impractical for handling large-scale document collections such as news articles, research papers, and online reports. Multi-document summarization becomes more difficult because related documents often contain redundant, overlapping, and semantically diverse information.

Existing extractive summarization approaches mainly depend on statistical techniques or deep learning models. Conventional statistical methods frequently fail to capture the true importance of sentences, while modern deep learning approaches require large training datasets, high computational power, and expensive hardware resources. These limitations make advanced summarization systems difficult to implement in lightweight and resource-constrained environments.

IV. METHODOLOGY

This chapter provides a description of the proposed machine learning algorithm for multi-document extractive summarization. The proposed system analyzes document clusters from the DUC 2003 dataset, performs feature extraction at the sentence level and uses Logistic Regression classifier to select summary sentences. The whole summarization process includes several steps such as data preparation, pre-processing, features extraction, supervised sentence classification and redundancy checking.

A. Dataset (DUC 2003)

In this paper, the DUC 2003 multi-document summarization dataset will be used for the experimentation. This benchmark dataset was developed specifically for summarization task [5].

Every topic in the dataset includes several news articles related to a particular topic together with manually generated reference summaries. Reference summaries are considered as ground truth for automatic performance estimation and sentence selection [18].In the proposed work, all documents within each topic cluster are processed to extract sentences, which are then used to build the supervised learning dataset.

TABLE II

DATASET STATISTICS (DUC 2003)

Parameter	Value
Dataset	DUC-2003
Total Sentences	3255
Feature Dimension	4003
Reference Summary Sets	90

B. Preprocessing

In order to guarantee data quality and uniformity, a number of pre-processing activities are undertaken on the document files. Firstly, the CSV files are parsed using the Beautiful Soup parser to get only the plain text. This text is divided into sentences through sentence tokenization [11].

Secondly, the sentences are cleaned by converting all letters to lower case as well as deleting any non-letter characters using regular expression operations. All sentences with less than five words are deleted to minimize the level of noise.

Thirdly, the sentences undergo word tokenization as well as stopword deletion using the NLTK list of English stop words. Lemmatization using the WordNetlemmatizer completes the process.

C. Sentence-Level Features

(a) TF-IDF Vector

The textual importance of each sentence is captured using the Term Frequency–Inverse Document Frequency (TF-IDF) representation [10]. For a term t in sentence s, the TF-IDF weight is computed as:

$$TF - IDF(t,s) = TF(t,s) \times \log\left(\frac{N}{\mathrm{DF(t)}}\right)$$

TF(t, s) TF(t, s)TF(t, s) is the term frequency

DF(t)DF(t)DF(t) is the document frequency

NNN is the total number of sentences

A TF-IDF vectorizer with a maximum vocabulary size of 4000 is used.

(b) Sentence Length

Sentence length measures the number of words in a processed sentence [13]:

$$L_s = |W_s|$$

where

$\mathbf{W}_s$represents the set of words in sentence s. Longer sentences often contain more informative content.

(c) Sentence Position

Sentence position captures the normalized position of a sentence within its document [3]:

$$P_s = \frac{index_s}{max\_index_d}$$

where:

• index_s is the sentence index.

• max_index_d is the total number of sentences in the document.

Earlier sentences in news articles are often more informative.

(d) Numeric Count Feature

The numeric feature counts the number of numeric tokens present in the original sentence [1]:

N_s =count of numeric tokens in sentence

Sentences containing numerical data may carry important factual information.

D. Sentence Classification Using Logistic Regression

A supervised classification approach is employed to identify summary-worthy sentences. Logistic Regression is selected due to its simplicity, computational efficiency, and strong performance on high-dimensional sparse data such as TF-IDF features [10].

Given a feature vector xxx, the Logistic Regression model estimates the probability that a sentence belongs to the summary class as:

$$P(y = 1/x) = \frac{1}{1 + e^{-(w^t x + b)}}$$

where www and b are model parameters.

The dataset is divided into training and testing sets using an 80:20 split. To address class imbalance between summary and non-summary sentences, class weighting is enabled during model training. The trained classifier outputs probability scores for each sentence, which are later used for ranking during summary generation.

E. Summary Generation

After classification, summaries are generated in a topic-wise manner using the predicted sentence probabilities [2].

(a) Probability Scoring

For each sentence, the trained Logistic Regression model produces a probability score indicating its likelihood of being summary-worthy. Sentences are ranked in descending order based on this probability [5].

(b) Threshold Selection

A probability threshold of 0.35 is used to select candidate summary sentences. Sentences with probability greater than the threshold are considered for inclusion. If no sentence satisfies the threshold, the top-ranked sentences are selected as a fallback strategy [13].

(c) Cosine Similarity-Based Redundancy Removal

To avoid repetition, cosine similarity is computed between candidate sentences using TF-IDF vectors. If the similarity between a candidate sentence and any previously selected sentence exceeds 0.8, the sentence is discarded as redundant [10].

(d) Word Limit Constraint

Finally, selected sentences are concatenated in ranked order while enforcing a predefined word limit of 100 words per topic. This ensures that the generated summary remains concise and comparable to the reference summaries [9].

V. EXPERIMENTAL RESULTS

In this section, the experiment configuration, evaluation criteria, and performance analysis of the proposed ML-based multi-document summarization model are discussed. Performance assessment is performed at two levels: sentence classification and the final summary.

A. Experiment Configuration

All experiments were carried out using the DUC 2003 dataset for multi-document summarization. This dataset consists of news articles clustered by topics, together with human-generated reference summaries. Preprocessing of all articles involved SGML parsing, sentence splitting, removal of stopwords, and lemmatization.

To represent features, both text and structure-based features were utilized. In terms of text, the TF-IDF vectorizer with a vocabulary size of 4000 was employed. Also, structural information in the form of sentence length, normalized sentence position, and number of numeric tokens was considered.

The labeled data set was split into training and test sets with proportions of 80% and 20%, respectively. The classifier implemented within the logistic regression algorithm with class balancing and a maximum of 1000 iterations was trained using the Scikit-learn library in Python.

B. ROUGE Evaluation

To evaluate the quality of the generated summaries, ROUGE metrics were computed against the human-written reference summaries provided in the DUC 2003 dataset.

TABLE III

ROUGE PERFORMANCE OF THE PROPOSED SUMMARIZATION SYSTEM

Metric	Score
ROUGE-1	0.1722

ROUGE-2	0.0306
ROUGE-L	0.1048

The ROUGE scores acquired reveal that the developed system can extract critical details from multi-document clusters. The ROUGE-1 score proves that there is adequate similarity between the extracted information and the reference summary regarding unigrams, meaning that the algorithm is successful in recognizing key content. On the other hand, the low ROUGE-2 score shows poor similarity when it comes to bigrams, implying that the generated summaries might not be coherent at the phrase level.

C. Comparison with Baseline Approach

In order to measure the success of the introduced method, it is necessary to compare it with the conventional approaches like Lead-3 and unsupervised ranking approaches.

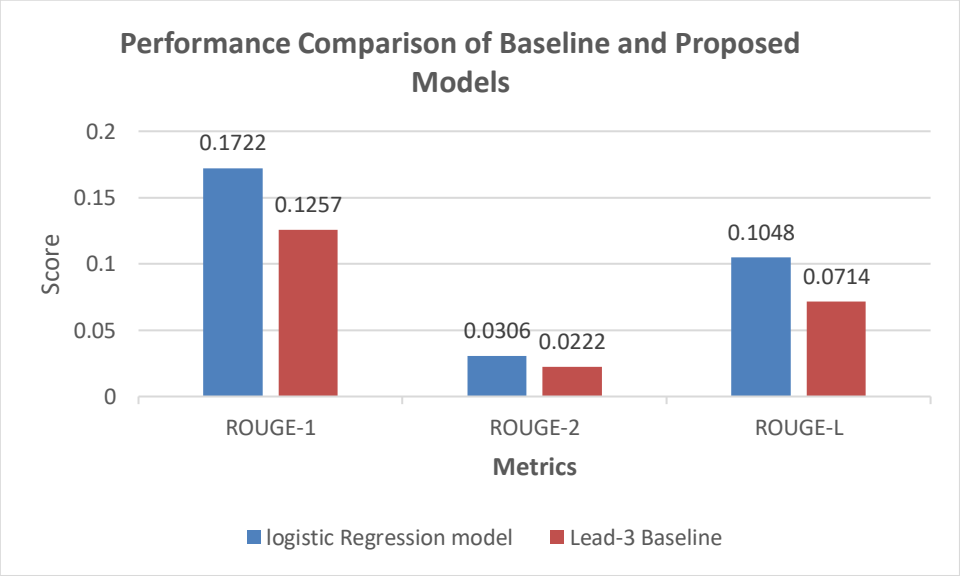

Fig.2**:** Comparison of ROUGE scores between the proposed Logistic Regression model and the Lead-3 baseline on the DUC 2003 dataset.

The comparison suggests that the proposed supervised learning approach provides competitive performance by leveraging sentence-level features and probability-based ranking. Unlike simple baselines, the proposed method incorporates both statistical and structural information for sentence selection.

TABLE IV

METHOD	ROUGE-1	ROUGE-2	ROUGE-L
Lead-3 Baseline	0.1257	0.0222	0.0714
Proposed Model	0.1722	0.0306	0.1048

D. Discussion

The experimental outcomes show that the proposed machine learning paradigm can generate multi-document extractive summaries with reasonable efficiency. The Logistic Regression classifier detects sentences with high informativeness based on TF-IDF and structural characteristics. The redundancy elimination module using cosine similarity enhances the summary quality by eliminating repeated content.

The biggest strength of the proposed method lies in its relatively low computational complexity in comparison with deep learning-based summarization approaches.

Although the obtained ROUGE-1 (0.1722), ROUGE-2 (0.0306), and ROUGE-L (0.1048) scores are lower than those reported by recent transformer-based summarization models, the proposed approach emphasizes computational efficiency, simplicity, and interpretability. Unlike deep learning methods such as BERT, PEGASUS, and Longformer, which require large-scale training data and substantial computational resources, the proposed Logistic Regression-based framework operates with significantly lower computational cost. The results demonstrate that sentence-level statistical and structural features can provide a practical baseline for multi-document extractive summarization in resource-constrained environments. Future work will investigate contextual embeddings such as Sentence-BERT and transformer-based representations to further improve summary quality and ROUGE performance.

A limitation of the present study is that the experimental evaluation was conducted exclusively on the DUC 2003 dataset. Although DUC 2003 is a widely used benchmark for multi-document summarization research, evaluating the proposed framework on a single dataset may limit the generalizability of the findings. Future work will include experiments on additional benchmark datasets such as DUC 2004, Multi-News, and CNN/DailyMail to assess the robustness and adaptability of the proposed approach across diverse document collections and domains.

Nevertheless, the ROUGE metrics reveal that improvements are still possible. First, the use of manually designed features might impede the model from capturing complex semantic relations between documents. Second, the static probability and word limit parameters might undermine the flexibility of the generated summaries.

Although the proposed model demonstrates improved ROUGE scores compared with the Lead-3 baseline, statistical significance testing was not performed in the current study. Future work will include significance analysis using bootstrap resampling or paired statistical tests to further validate the observed improvements.

Although the proposed Logistic Regression classifier achieved an accuracy of 93.09%, accuracy alone may not fully reflect classification performance in the presence of class imbalance. Precision, recall, and F1-score were not evaluated in the current study, which represents a limitation of the experimental analysis. Future work will include a comprehensive evaluation using additional classification metrics to provide a more balanced assessment of model performance.

Future research could consider using contextualized representations such as BERT or SBERT, adjusting probability limits adaptively, and applying better redundancy detection methods.

VI. CONCLUSION AND FUTURE WORK

A. Conclusion

The current work proposes a method of multi-document summarization using machine learning algorithms based on sentence-level features. The suggested framework involves appropriate data processing, such as CSV parsing, sentence tokenization, removing stopwords, and lemmatization, and feature selection through sentence length, position, numeric count, and TF-IDF vectors. Sentence classification is done with the help of a Logistic Regression classifier, while sentence redundancy is removed by comparing cosine similarities in the course of summary creation.

Evaluation on the DUC 2003 dataset shows that the developed algorithm provides an accuracy of 0.9309 when classifying sentences, along with ROUGE-1, ROUGE-2, and ROUGE-L metrics equaling 0.1722, 0.0306, and 0.1048, respectively [17]. In this way, it becomes possible to conclude that a relatively simple machine learning model can serve as an adequate basis for text

summarization. Nevertheless, it should be noted that the dataset under consideration demonstrates considerable class imbalance.

B. FutureWork

Extending the framework to domain-specific or multilingual multi-document summarization tasks.Evaluating the system using human judgment metrics in addition to ROUGE for more comprehensive assessment.

REFERENCES

[1] K. Shinde, T. Roy, and T. Ghosal, "An extractive-abstractive approach for multi-document summarization of scientific articles for literature review," in *Proc. Workshop on Scholarly Document Processing*, 2022.

[2] Y. Liu and M. Lapata, "Text Summarization with Pretrained Encoders," *Proceedings of EMNLP*, 2021.

[3] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization," *ICML*, 2021.

[4] H. Su, J. Jiang, and Y. Huang, "Multi-Granularity Adaptive Extractive Document Summarization with Heterogeneous Graph Neural Networks," *PeerJ Computer Science*, vol. 9, 2023.

[5] N. Moratanch and S. Chitrakala, "A Survey on Extractive Text Summarization," *International Journal of Computer Applications*, 2021.

[6] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," *EMNLP*, 2021.

[7] I. Beltagy, M. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," *ACL*, 2021.

[8] A. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *EMNLP*, 2021.

[9] C. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," *Text Summarization Branches Out*, updated usage widely cited, 2022.

[10] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., draft, 2023.

[11] K. Papineni et al., "BLEU: A Method for Automatic Evaluation of Machine Translation," adapted in summarization evaluation studies, 2022.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," updated applications in summarization, 2021.

[13] K. R. Varma and V. B. Balamurugan, "Machine learning approaches for extractive text summarization: A comparative study," Expert Systems with Applications, vol. 178, 2021.

[14] R. Bandaru and Y. Radhika, "Extractive multi-document text summarization leveraging hybrid semantic similarity measures," International Journal of Advanced Computer Science and Applications, 2022.

[15]Serasiya S, And U. Chauhan, "Abstractive Gujarati Text Summarization Using Sequence-To-Sequence Model and Attention Mechanism", Journal of Information Systems Engineering and Management, 2468-4376, 2025, PP: 754- 762

[16] M. Nasari and A. S. Girsang, "Automated multi-document summarization using extractive-abstractive approaches," *International Journal of Informatics and Communication Technology*, 2024.

[17] N. Gu, E. Ash, and R. Hahnloser, "MemSum: Extractive summarization of long documents using multi-step episodic Markov decision processes," *arXiv*, 2021.

[18] M. Chen et al., "SgSum: Transforming multi-document summarization into sub-graph selection," in *Proceedings of EMNLP*, 2021.

[19]Serasiya S, And U. Chauhan. "A Comprehensive Survey on Text Summarization For Indian Languages: Opportunities, Challenges And Future Prospects." South Eastern European Journal Of Public Health (2025): 1418-1434.

[20] R. Bandaru and Y. Radhika, "Extractive multi-document text summarization leveraging hybrid semantic similarity measures," *International Journal of Advanced Computer Science and Applications*, 2022.