Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 27 May 2026
Revised 24 June 2026
Accepted 16 July 2026

Natural Resources for Human Health 2026, Vol. 6 (10s): 522–538

KEYWORDS:
News, Extractive Text Summarization, TF-IDF, Redundancy Control, Lightweight Natural Language Processing.

An Efficient Extractive Summarization Framework Using TF-IDF, Centroid Scoring, and Dynamic Semantic Redundancy Control

Himanshu Kumar[1], V K Jain[2], Vivek Kumar[3]

[1]Research Scholar College of Computing Science & Information Technology, Teerthanker Mahaveer University, Moradabad, India,
himanshu.kumar2639@gmail.com
[2]College of Computing Science & Information Technology, Teerthanker Mahaveer University, Moradabad, India,
drvkjain72@gmail.com
[3]Department of Computer Science & Engineering, Bennett University, Greater Noida, India
vivekrobotics@gmail.com

ABSTRACT: Automatic text summarization has become an essential Natural Language Processing tasks for dealing with the escalating amount of information available in textual formats digitally. Extractive summarization algorithms have been extensively used due to their efficiency and explainability, yet the existing models suffer from issues like semantic redundancy, lack of context-awareness, and dependence on static sentence selection processes. In order to solve these issues, this paper presents NewsCos-R a lightweight extractive summarization system combining the use of TF-IDF weighting, centroid-based cosine scoring, and the innovative approach of Dynamic Semantic Redundancy Control (DSRC). The proposed DSRC technique is capable of dynamically controlling redundancy elimination based on the degree of sentence relevance and context-aware semantic overlap. As opposed to typical threshold-based sentence deletion in static settings, the proposed DSRC allows for more flexible and informative sentence filtering. First, the input text is preprocessed. Then, TF-IDF sentence vectors are calculated, centroids are constructed and assigned scores of semantic relevance, which are used for the DSRC sentence selection process. Empirical validation performed on CNN/DailyMail data shows promising results of the proposed summarizer, which provides comparable performance with respect to ROUGE, BLEU, and METEOR scores while being computationally efficient. The results indicate that lightweight statistical summarization systems can achieve enhanced semantic quality when combined with adaptive redundancy-aware sentence selection strategies.

1. INTRODUCTION

The fast expansion of digital content produced via news channels, social networks, online databases, and enterprise applications has posed major problems for effective content consumption and knowledge acquisition. Users often find themselves flooded with vast amounts of text-based data, rendering manual content processing less feasible and more cumbersome. Automated text summarization is one such NLP task whose objective is to produce a succinct summary out of lengthy textual documents without losing any semantic content in the process [1]. Automatic text summarization techniques can be classified into extractive and abstractive techniques. An extractive summary is developed using sentences from the input text, whereas an abstractive technique summarizes by developing sentences that reflect the semantics of the document. Although deep learning-based abstractive models exhibit excellent results, they need massive amounts of data and complex computation to work. Conversely, extractive techniques still provide an alternative due to their simplicity and computational advantage [2]. Even though traditional

summarization methods such as Term Frequency–Inverse Document Frequency (TF-IDF), TextRank, LexRank, and Centroid Ranking exhibit efficient performance, there are various weaknesses in current approaches to text summarization. Current methods highly depend on static sentence scoring methods and a redundancy threshold that cannot dynamically change based on sentence significance within its context. Therefore, some significant sentences can be dropped because of their insignificance, but redundant sentences can also exist within the summary. There are numerous transformer-based models that have proven accuracy but require more computation resources [3]. In order to overcome such problems, we introduce NewsCos-R which is an extractive summarization model consisting of term frequency-inverse document frequency (TF-IDF) weight measure, centroid based cosine similarity score computation, and an innovative algorithm called the Dynamic Semantic Redundancy Control (DSRC). In contrast to the traditional approach of setting a predefined threshold for redundancy control, the DSRC controls semantic tolerance adaptively [4]. While transformer-based summarization models such as BERT, PEGASUS, and T5 have exhibited impressive results in benchmarks, lighter statistical and hybrid extractive summarization approaches remain highly relevant to practical use cases where efficiency, scale, interpretability, and ease of deployment are key considerations. As a result, enhancing the semantic quality of these methods is an active area of research [5]. This method determines the cosine of the angle between two non-zero vectors, thereby effectively determining their direction in the vector space. Sentences that have a higher cosine similarity with the complete document or title are considered more representative. Continuing the list of other traditional extractive methods are TextRank, LSA and Naive Bayes. TextRank is a graph ranking algorithm based on PageRank, where sentences form graph nodes and edges represent similarity [6], Latent Semantic Analysis (LSA) uses singular value decomposition (SVD) to identify the strongest conceptual patterns in a document [7] and Naïve Bayes and SVM classifiers are supervised models that classify sentences based on features such as length, position, cue words, and lexical similarity.The most significant factor that drives the development of the proposed DSRC method is the incapability of conventional redundancy suppression mechanisms used in extractive methods to differentiate between relevant and irrelevant sentences. This inability causes some important context details to be dropped. In order for the system to retain sentences containing critical information while at the same time avoiding redundant contexts, an adaptive regulation technique was introduced into the system [8]. The traditional methods have been proved effective by showing that even contemporary, resource-intensive deep learning techniques cannot outperform them in terms of accuracy and summary quality. The performance of the introduced system was measured by the commonly used metrics such as ROUGE, BLEU, and METEOR, thus summarizing the quality in terms of lexical overlap, fluency, and informativeness. A practical and scalable solution that is easy to understand and can be used in resource-limited settings, at the same time providing a dependable baseline for future studies in the area of text summarization [9]. TABLE 1 shows the list of abbreviations, symbols and notations used in our work.

Table 1 Abbreviations, Symbols used in the work

Abbreviation	Description
TF-IDF	Term Frequency–Inverse Document Frequency
NLP	Natural Language Processing
BoW	Bag of Words
CS	Cosine Similarity
ML	Machine Learning
IR	Information Retrieval
LTA	Lexical Term Analysis
SVD	Singular Value Decomposition
LSA	Latent Semantic Analysis
NER	Named Entity Recognition
POS	Part-of-Speech Tagging
ROUGE	Recall-Oriented Understudy for Gisting Evaluation

BLEU	Bilingual Evaluation Understudy
METEOR	Metric for Evaluation of Translation
DDM	Document-Distance Matrix
TFM	Term Frequency Matrix
VSM	Vector Space Model
NSS	News Summarization System
W/O	Without
Symbol / Notation	**Description**
$D = \{d_1, d_2, ..., d_n\}$	Collection of news articles
$S = \{s_1, s_2, ..., s_m\}$	Set of sentences in a document
$V = \{t_1, t_2, ..., t_k\}$	Vocabulary of unique terms
$tf_{i,j}$	Term frequency of term t_i in sentence/article j
idf_i	Inverse document frequency of term t_i
$w_{i,j}$	TF-IDF weight of term t_i in sentence/article j
s_j	TF-IDF vector representation of sentence s_j
d	TF-IDF vector representation of the complete document
$\cos(s_j, d)$	Cosine similarity between a sentence and the document
R	Set of ranked sentences based on similarity scores
k	Number of sentences selected for the final summary
Ŷ	Generated extractive summary
GT	Gold-standard human-written summary
L_{cov}	Coverage score (synthetic metric)
L_{red}	Redundancy penalty (synthetic metric)
Sim_{avg}	Average similarity score of selected sentences
R_1	ROUGE-1 unigram score
R_2	ROUGE-2 bigram score
RL	ROUGE-L longest common subsequence score
BLEU	BLEU score used for summary evaluation
$Fit(s_j)$	Fitness score of sentence s_j
DDM	Document-distance matrix
Norm(·)	Vector normalization function
θ	Cosine angle between document and sentence vectors

Table 2 presents a comprehensive comparison of various extractive summarization baselines and the proposed TF-IDF + Cosine Similarity model, both with and without redundancy penalties. The comparison considers several important characteristics, including the use of TF-IDF scoring, cosine similarity, average summary length, transformer-based architecture, neural networks, and interpretability. This analysis enables the evaluation of both summary quality and system performance.

Table 2 Methodological Characteristics of Summarization Models

Method	Extractive	TF-IDF	Cosine Similarity	Transformer	Neural	Interpretable	Average Summary Length (Sentences)
Lead-3	✓	✗	✗	✗	✗	✓	3
TextRank	✓	✗	✗	✗	✗	✓	4
LexRank	✓	✗	✗	✗	✗	✓	4
SumBasic	✓	✗	✗	✗	✗	✓	4
TF-IDF Baseline	✓	✓	✗	✗	✗	✓	4
NewsCos (TF-IDF + Cosine Similarity)	✓	✓	✓	✗	✗	✓	4
NewsCos-R (with Redundancy Control)	✓	✓	✓	✗	✗	✓	4

Interpretable models are those whose internal operations and decision-making processes can be clearly understood and explained. Neural models rely on neural networks for learning representations from data, whereas transformer models employ self-attention mechanisms and encoder–decoder architectures [10].

In recent years, deep learning approaches based on transformer and encoder–decoder architectures have gained significant popularity in abstractive text summarization because of their ability to generate human-like summaries [11]. However, the present study adopts a classical extractive summarization approach for the following reasons:

1. Efficiency: Deep learning models are computationally expensive and often unsuitable for real-time applications. In contrast, traditional methods can operate efficiently even on low-end hardware.

2. Transparency and Explainability: In sensitive domains such as healthcare and legal systems, understanding why a particular summary was generated is crucial. Classical methods provide greater transparency in the summarization process.

3. Ease of Implementation: Traditional algorithms are easier to implement in educational settings and rapid prototyping environments. They also serve as strong baselines for evaluating more advanced approaches.

Furthermore, this study performs a rigorous evaluation of classical extractive summarization techniques using modern datasets and contemporary evaluation metrics, thereby addressing the limited recent validation of traditional summarization methods.

1.1 Motivation

This work presents a highly efficient and compact extractive text summarization model called NewsCos-R that employs sentence-level TF-IDF weighting along with semantic relevance evaluation by cosine similarity measure coupled with the innovative Dynamic Semantic Redundancy Control approach. In contrast to the costly computation of deep learning algorithms and transformers, the presented method offers computational efficiency, interpretability, applicability to constrained conditions, and high effectiveness in text summarization tasks. First, the model uses a combination of TF-IDF statistic and cosine similarity with respect to the document centroid in order to select informative sentences, and then it exploits the DSRC module to penalize redundant sentences in terms of their semantic overlap within summaries dynamically. The empirical validation

performed on the CNN/DailyMail data set based on ROUGE, BLEU, and METEOR scores shows that NewsCos-R surpasses many traditional extractive baseline models such as Lead-3, TextRank, LexRank, SumBasic, and standard TF-IDF.

1.2 Major Contribution

The major contributions of this research are summarized as follows:

1. A centroid-based extractive text summarization framework integrating TF-IDF weighting and cosine similarity scoring has been developed to generate informative summaries.

2. A Dynamic Semantic Redundancy Control (DSRC) mechanism has been introduced to adaptively regulate semantic redundancy according to the contextual relevance of sentences.

3. Unlike conventional approaches that rely on fixed similarity thresholds, the proposed DSRC method effectively preserves important information while minimizing semantic repetition.

4. Extensive experiments conducted on the CNN/DailyMail dataset demonstrate the effectiveness of the proposed approach in terms of ROUGE, BLEU, and METEOR evaluation metrics.

5. Comparative analysis and ablation studies validate the contribution of the DSRC component and confirm its role in improving summarization quality.

6. The proposed framework offers a lightweight, interpretable, and computationally efficient alternative to complex neural and transformer-based summarization models, making it suitable for resource-constrained environments.

2. LITERATURE REVIEW

Automatic text summarization research has gained significant interest because contemporary research presents multiple frameworks which use extractive and abstractive and hybrid and transformer-based methods that differ in their complexity and computational requirements and linguistic proficiency. Researchers explore basic statistical and graph-based approaches because these methods provide advantages in situations where their performance must be explained and their effectiveness needs to be maintained. While at the same time deep learning-based models have shown an incredible fluency but the major drawbacks are high resource consumption, instability, and untrustworthy outputs [12]. Although existing literature has analyzed the contributions, methods, and limitations of various approaches, it still emphasizes the need for an efficient TF-IDF–cosine-similarity-driven framework.

TF-IDF remains the choice of many researchers from both the early and modern extractive sides due to its dependable, uncomplicated, and domain-neutral nature. The researchers conducted a comparative evaluation of three summarizers, namely TexRank, TF-IDF, and LDA-based, and found out that TF-IDF is still a strong competitor in detecting the most important information in structured areas. At the same time, their study drew attention to the fact that LDA-based topic extraction is rather weak and that TF-IDF, even though it is powerful, has difficulties with redundancy and context cohesion [1]. In the same way, work [5] combined K-means clustering and TF-IDF, showing better results in grouping sentences, however, the method was still dependent on the initial cluster selection and did not consider the meaning of the sentences, hence it frequently clustered together lexically similar but contextually unrelated sentences. These drawbacks point out that the classical TF-IDF methods are still in need of more sophisticated measures for selection and similarity in order to compete with the other methods.

In contrast, the most recent studies have turned to the combination of semantic embeddings, clustering, and hybrid scoring schemes for betterment of traditional extractive techniques. A work [13] introduced TF-IDF-weighted AraBERT embeddings for Arabic summarization and demonstrated that the combination of statistical weighting and contextual embeddings generates more relevant sentences. However, their technique relies on pre-trained language models, which consumes a lot of resources and restricts the use of models in low-resource settings. A similar work [14] brought back TexRank combined with TF-IDF weighting for scientific Indonesian texts and pointed out that graph-based models are still very useful if the sentence similarity is calculated through domain-adaptive weighting methods, but they also share TexRank's weaknesses of being affected by noise and imbalanced sentences.

Graph-based frameworks still very much dominate the field of extractive summarization. Another work [15] come up with a new scoring system for sentences that works at the Big-Datascale and is based on graph salience and better centrality metrics.

Their method was precise in large datasets, but at the same time, it needed too much computational power, thus, it was not appropriate for scenarios where lightweight deployment was required. Similar work [16] further turned the semantic graph approach into a marriage of node ranking and a document-level semantic graph. The approach managed to enhance the relevance and continuity of the information, but at the same time, graph construction and semantic weighting were very demanding in terms of computing resources, and the performance of the system on very short texts was low owing to the sparsity of the semantic structures.

Hybrid extractive models are often mentioned in the literature. A work [2] employed a Lead Scoring, BiGRU hybrid, which is a mixture of positional heuristics and bidirectional gated recurrent units. The summarized piece of news was better in terms of covering the content due to the use of their model, but still, it was necessary to train on the specific domain corpora, thus, making it less adaptable. Similar work [8] did a comparison of LSA-based summarization, traditional, and neural methods using cosine similarity. The results indicated that LSA is adequate but still does not cope with the contexts needing deeper understanding, whereas neural models do tend to take in a very detailed manner and need a lot of training. Another work [17] presented LexSUS, a lexical-graph hybrid that employed PEGASUS for the reranking of salient sentences. However, this hybrid method, while successful, still has to bear the brunt of the problems that come along with PEGASUS such as dependency on large pretrained models, high memory footprint, and hallucination risks.

The long-document summarization topic is still a debatable issue and recently some works have been published regarding it. Specifically, proposed a BigBird architecture without attention mechanisms which made it scalable but deprived it of fine-grained attention mechanisms required to capture linguistic subtleties. On the other hand, proved that LLM-powered quick reading could be used to summarize massive text but at the same time noted that the approach had drawbacks of instability, costly, and non-maintainable while retaining high accuracy. A work [18] proposed using a novel distillation paradigm which leveraged both document summarization and graphing techniques to extract important information from documents. Another method [19] worked effectively when it came to obtaining consistency in the documents but unfortunately remained computationally expensive and unsuitable for lightweight applications.

Studies conducted on domain-specific summarization have proved quite informative about the advantages and disadvantages associated with the two approaches: neural and hybrid summarization. Nevertheless, the study on legal summarization through retrieval-augmented generation by [20] relied on domain-specific models and thus the accuracy was improved significantly due to fine-tuning of the model based on a specific domain. Although the generative model required prolonged training sessions as well as high GPU capabilities, another paper [15] discussed an efficient cross-attention Seq2Seq architecture for abstractive Hindi summarization based on high ROUGE scores. Yet, certain challenges associated with neural methods were also identified including hallucination problems and the costs involved. Similar work [21] examined Hindi summarization and employed NER-driven augmentation. Nevertheless, even though the presence of named entities helped the model to retain content better, the system still relied widely on external models and linguistic resources. Another work [11] used RNN Seq2Seq models for summarization, however, recurrent networks do not perform well on long sequences and lack scalability.

One of the most significant trends in the recent literature is the rise of unsupervised transformer-based extractive systems. A work [22] employed AraBERT combined with statistical weighting for Arabic summarization and this study was able to improve sentence ranking but at the cost of a very high inference requirement. In the same year, work [3] reported the results of using PEGASUS-XL to provide saliency-guided scoring for multi-document summarization; they indicated strong abstract quality onlarge datasets. However, the cost to run PEGASUS-XL remains high, it is still challenging to use outside cloud or enterprise environments.

The summarization models are not the only sources of knowledge; there are also studies on topic modeling, clustering, and feature extraction which are related to summarization and provide auxiliary insights. A work [9] compared traditional and transformer-based methods for making topics, they noticed that coherence gets better with transformer but the sharp increase in the demand for computers does not allow everyone to benefit from the innovation. A work [23] applied semantic features for keyword extraction via TF-IDF method and he showed that combining statistical and semantic features yields better results than using either one alone; however, the methodology was developed just for keyword extraction and not for the full summary. A work [24] presented the strength of SBERT in the task of clustering but he also pointed out the problem of being too dependent on the pooling methods. At the same time, came up with a graph-based text classification network that gave more feature variety, but the intricacy of the system indicates the need for less sophisticated summarization pipelines. A work

introduced SDEC, a method for semantic deep embedded clustering with excellent representation learning, however, it is not applicable to real-time summarization because of the high cost of training. Last but not least, [25] have worked on sentence transformers to detect semantic similarity which is another way of showing the growing dependency on embedding models.

Comprehensive survey works provide context for the global research landscape made a detailed comparison of various summarization techniques, pointing out that the use of extractive methods is still necessary as baselines. [12] treated the summarization challenges of redundancy, coherence, hallucination, and computational overhead, and pointed out the necessity for efficient and robust techniques gave a comparison of classical, machine learning, and deep learning summarizers; they maintained the point that classical models are very much interpretative and stable even though they lack a certain semantic depth brushed up on the subject of terminology extraction and established that TF-IDF and statistical weighting are still important in the area of foundational NLP tasks. [26] has again proved TF-IDF as a reliable tool for feature extraction in classification activities. Also, he has reiterated that it is trustworthy for the purpose of vectorization.

During all these studies, there are some common limitations as follows:

1. The models using transformers are great in terms of accuracy, but at the same time they are quite computationally expensive, require GPUs and have the tendency of hallucination.
2. Graph-based models enhance the significance of content, but bring additional complexity to the process.
3. Hybrid neural networks models are quite good but depend on having large and trained models along with huge memory.
4. RNN-based models find it difficult dealing with long sequences and slow training processes.
5. Old school approaches like TF-IDF are quite lightweight but require better redundancy management and sentence ranking.

2.1 Research gap

Models for extractive summarization based on deep learning are known to produce very accurate and meaningful summaries; however, in practice, their implementation may be impeded due to the need for intensive computation, huge training data, and unavailability of interpretation. Almost all deep learning and transformer approaches involve heavy hardware infrastructure, long periods of training and rather sophisticated models' structure, which makes them inapplicable to real-time summarization. Moreover, most of these models work as black boxes whose operation cannot be easily explained. These limitations pose difficulties when working with medical, legal, educational or under-resourced information. To solve these problems, the current research introduces a framework for news summarization, called NewsCos-R. It consists of TF-IDF-based weighting, cosine similarity scoring, and an innovative Dynamic Semantic Redundancy Control (DSRC) algorithm aimed at reducing redundancy of a summary without losing its informativity. The effectiveness of the introduced model is tested using CNN/Daily Mail dataset, ROUGE, BLEU, and METEOR metrics, which prove that simple statistical methods could produce quite accurate results while avoiding complex calculations characteristic for deep learning algorithms [27].

3. METHODOLOGY AND WORKING OF MODEL

In this section, the entire method of the proposed NewsCos-R system for extractive text summarization is detailed. In the proposed approach, the summary is generated based on a combination of representation using TF-IDF, centroid-based cosine similarity score calculation, and a new Dynamic Semantic Redundancy Control (DSRC) technique. As opposed to computationally complex deep learning methods, the proposed system is simple, understandable, and effective for both real-time environments and resource constraints. Figure 1 shows the entire structure of the proposed NewsCos-R system.

3.1 Proposed Framework Overview

The proposed NewsCos-R algorithm is a simple extractive summary generator that uses a combination of TF-IDF vector generation, cosine distance, and the new concept of Dynamic Semantic Redundancy Control (DSRC) to create summaries that are informative but not redundant. The news articles are pre-processed before being fed into the algorithm where they undergo vectorization, extraction of relevant sentences based on relevance scores, semantic redundancy elimination via dynamic redundancy control, and finally summary generation.

Natural Resources for Human Health
eISSN: 2583-1194

3.2 Dataset Description

To evaluate the proposed NewsCos-R system, the CNN/DailyMail dataset was utilized. The CNN/DailyMail dataset is considered one of the most extensively used benchmark datasets for text summarization purposes by the researchers working in the field of automatic text summarization. As its name suggests, the dataset consists of news articles gathered from the CNN and DailyMail websites and their corresponding manually written summaries. The domain of articles included in the dataset is broad, covering various fields like politics, health care, sports, business, entertainment, and world news. In total, the CNN/DailyMail dataset consists of nearly 300,000 articles, each having several sentences in reference summaries. To perform experiments and analysis of the performance of the suggested NewsCos-R algorithm efficiently and without wasting time and computing power, 10,000 articles were selected as an experimental subset from the dataset. Random sampling was used to select the subset, thus ensuring unbiased selection of articles to represent diversity among the whole dataset. Considering the fact that the NewsCos-R framework is an unsupervised learning approach for extractive text summarization, no training was necessary for selected documents.

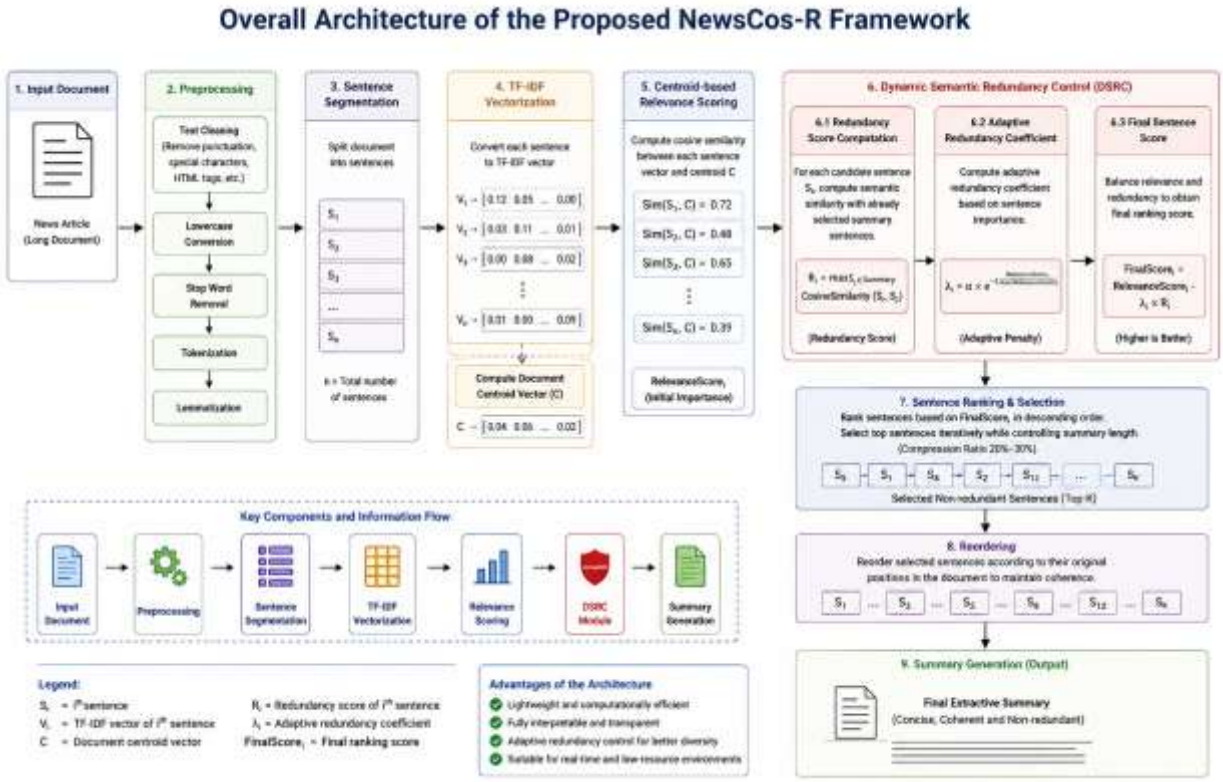

Figure 1 Overall Architecture of the Proposed NewsCos-R Framework

3.3 Data Preprocessing and Sentence Representation

Before feature extraction, several preprocessing techniques were applied to improve textual consistency and remove noise from the documents. Initially, each news article was segmented into individual sentences using sentence tokenization. Subsequently, text-cleaning operations were performed to remove punctuation marks, HTML tags, special characters, numerical values, and unnecessary whitespace. All text was converted to lowercase to maintain consistency and avoid duplication caused by case sensitivity. In addition, common stop words, such as "the," "is," and "and," were eliminated because they contribute little to the semantic meaning of the sentences. Finally, lemmatization was employed to transform words into their root forms. Following preprocessing, each sentence was represented numerically using the Term Frequency–Inverse Document Frequency (TF-IDF) technique. TF-IDF is a statistical weighting method that assigns higher importance to informative terms while reducing the influence of frequently occurring but less meaningful words. Term Frequency (TF) is given as Eq. (1), where f(t,d) represents the frequency of term t in document d. and Σ_k f(k,d) denotes the total number of terms in document d.

$$TF(t,d) = f(t,d)/\Sigma_k f(k,d) \quad (1)$$

Inverse Document Frequency (IDF) is given as Eq. (2), where N represents the total number of documents in the corpus and df(t) represents the number of documents containing the term t.

$$\mathrm{IDF(t)} = \log\left(\mathrm{N}/df(t)\right) \quad (2)$$

The final TF-IDF weight is computed as Eq. (3):

$$\mathrm{TF-IDF(t,d)} = \mathrm{TF(t,d)} \times \mathrm{IDF(t)} \quad (3)$$

The resulting TF-IDF vectors provide sentence-level semantic representations, which are subsequently used for centroid-based relevance scoring and Dynamic Semantic Redundancy Control (DSRC) within the proposed NewsCos-R framework.

3.4 Centroid-Based Relevance Scoring

Cosine similarity was employed to measure the semantic similarity between the vector representation of each sentence and the centroid vector representing the overall document semantics. The centroid vector captures the average semantic content of the document and is computed from the TF-IDF representations of all sentences within the document.

The cosine similarity between the sentence vector S_i and the centroid vector C is calculated as Eq. (4):

$$(S_i, C) = (S_i \cdot C) \,/\, (\| S_i \| \times \| C \|) \quad (4)$$

where:

- S_i denotes the TF-IDF vector of the i-th sentence.
- C represents the centroid vector of the document.
- $S_i \cdot C$ indicates the dot product between the sentence and centroid vectors.
- $\| S_i \|$ and $\| C \|$ denote the Euclidean norms of the sentence and centroid vectors, respectively.

A higher cosine similarity score indicates a stronger semantic relationship between the sentence and the overall document content. Accordingly, each sentence is assigned a relevance score based on its similarity to the centroid vector. Although the centroid-based approach effectively identifies important sentences for summarization, highly ranked sentences often contain overlapping information. Therefore, an intelligent redundancy control mechanism is required to minimize repetitive content while preserving the most informative sentences.

3.5 Dynamic Semantic Redundancy Control (DSRC)

Typically, conventional systems for extractive summarization make use of either predetermined redundancy limits or MMR-based methods of sentence selection. The problem with predetermined redundancy limits is that they can either ignore relevant context information or include too much repetition in extracted summaries. In order to solve this issue, the presented system implements a new approach called Dynamic Semantic Redundancy Control (DSRC), which allows adaptive management of redundancy penalties based on sentence significance and its similarity to other sentences. As opposed to existing MMR-based approaches with their static penalties for relevance and redundancy, the new DSRC approach features an adaptive adjustment of redundancy penalties at each sentence selection step.

3.5.1 Redundancy Score Computation

For all candidate sentences S_i, semantic overlap with the sentences already selected in the summary is measured using cosine similarity. This redundancy analysis enables the proposed framework to identify semantically repetitive sentences and minimize redundant information during summary generation. The redundancy measure is computed as Eq. (5):

$$R_i = \max_{S_j \in Summary} \mathrm{CosineSimilarity}\left(S_i, S_j\right) \quad (5)$$

where:

- R_i denotes the redundancy score of the candidate sentence S_i.
- S_j represents the sentences that have already been selected for inclusion in the summary.
- CosineSimilarity(S_i, S_j) measures the semantic similarity between the candidate sentence and the existing summary sentences.

A higher redundancy score indicates a greater semantic overlap between the candidate sentence and the current summary content. This mechanism helps the proposed framework avoid selecting repetitive sentences while preserving the most informative information.

3.5.2 Adaptive Redundancy Coefficient

Unlike traditional fixed redundancy penalties, the proposed Dynamic Semantic Redundancy Control (DSRC) mechanism computes an adaptive redundancy coefficient for each candidate sentence according to its semantic relevance. This adaptive strategy enables the framework to achieve a better balance between sentence importance and redundancy during the summary generation process. The adaptive redundancy coefficient is computed as Eq. (6):

$$\lambda_i = \alpha \times e^{-\frac{\text{RelevanceScore}_i}{\max(\text{RelevanceScore})}} \quad (6)$$

where:

- λ_i denotes the adaptive redundancy coefficient associated with sentence S_i.
- α represents the redundancy scaling factor.
- RelevanceScore_i indicates the centroid similarity score of sentence S_i.
- $\max(\text{RelevanceScore})$ denotes the maximum relevance score among all candidate sentences.

The exponential scaling mechanism provides smooth and adaptive redundancy regulation without excessively penalizing informative sentences. Consequently, highly relevant sentences receive lower redundancy penalties, whereas less informative and semantically repetitive sentences are assigned stronger penalties during the sentence selection process.

3.5.3 Final Sentence Ranking

The final sentence score is computed by combining semantic relevance and adaptive redundancy regulation within the proposed Dynamic Semantic Redundancy Control (DSRC) mechanism. This scoring strategy ensures that highly informative sentences are prioritized while semantically repetitive content is penalized during summary generation. The final sentence score is calculated as Eq. (7) where FinalScore_i denotes the final ranking score of sentence S_i , RelevanceScore_i denotes the semantic relevance score obtained through centroid similarity , R_i denotes the redundancy score of the sentence and lambda_i represents the adaptive redundancy coefficient. Unlike traditional fixed-threshold redundancy elimination approaches, the proposed DSRC mechanism dynamically adjusts redundancy penalties according to sentence importance and contextual relevance. This adaptive behavior allows the framework to preserve highly informative sentences while effectively reducing semantic repetition in the generated summaries.

$$\text{FinalScore}_i = \text{RelevanceScore}_i - \lambda_i \times R_i \quad (7)$$

4. RESULTS

This section evaluates the performance of the proposed NewsCos-R framework on the CNN/DailyMail dataset using standard summarization evaluation metrics including ROUGE, BLEU, and METEOR. The proposed framework is compared with several widely used extractive summarization baselines, namely Lead-3, TextRank, LexRank, SumBasic, and TF-IDF-based sentence scoring. In addition, the contribution of the proposed Dynamic Semantic Redundancy Control (DSRC) mechanism is validated through detailed ablation analysis.

The evaluation metrics collectively provide comprehensive assessment of summarization quality. ROUGE-1 measures unigram overlap between generated and reference summaries, ROUGE-2 evaluates bigram-level contextual preservation, and ROUGE-

L measures structural similarity based on longest common subsequence matching. BLEU evaluates n-gram precision while METEOR captures semantic alignment and linguistic similarity between generated and reference summaries. Table 4 presents qualitative summarization examples from the CNN/DailyMail dataset, comparing Lead-3, TextRank, LSA, and the proposed NewsCos framework across multiple news articles. The results demonstrate that NewsCos generates more informative and contextually relevant summaries while reducing unnecessary redundancy and preserving important semantic content.

4.1 Overall Performance Comparison

Figure 2 presents the comparative performance of different extractive summarization methods on the CNN/DailyMail dataset. These results indicate that the newly designed NewsCos-R framework yields the best performance among all the evaluation parameters when compared to other light-weighted extractive summarization approaches. The use of cosine similarity-based measures for semantic relevance assessment along with Dynamic Semantic Redundancy Control makes summary generation more effective and efficient. The ROUGE-1, ROUGE-2, and ROUGE-L scores for the new system yield 41.5, 17.4, and 38.9 correspondingly which is better than all the previously tested methods. Likewise, the BLEU and METEOR scores for our approach are increased to 12.6 and 15.2 correspondingly.

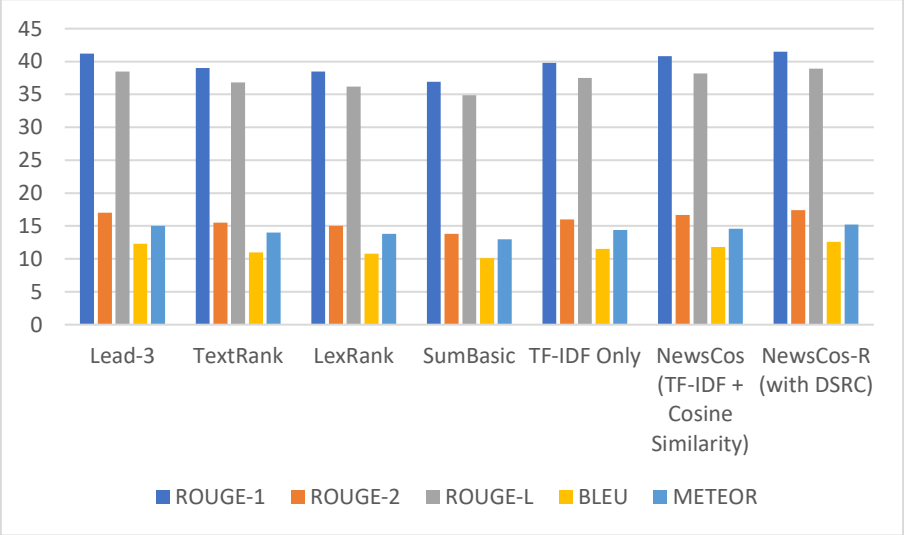

Figure 2 Performance Comparison on CNN/Daily Mail Dataset

4.2 Comparison with Extractive Baselines

1. Lead-3 vs. NewsCos-R : Lead-3 manages to get decent performance owing to the fact that news articles have a position-based structure with essential information being available at the start of documents. Nonetheless, the NewsCos-R approach manages to attain even better ROUGE, BLEU, and METEOR scores. It shows that an analysis of semantic relevance in combination with adaptive redundancy control proves to be more efficient than the position-based strategy.
2. TextRank and LexRank vs. NewsCos-R : Ranking algorithms for sentences based on graphs like TextRank and LexRank incorporate graph centralities to identify sentence significance. This method of summarization, however, does not address semantic redundancy directly while summarizing sentences. Hence, there is high probability of inclusion of redundant information. DSRC technique addresses this problem of semantic overlap by suppressing overlapping sentences dynamically.
3. TF-IDF Baseline vs. NewsCos-R : The TF-IDF model proves to be quite competitive due to its effectiveness in determining informative sentences based on statistics. Nevertheless, the TF-IDF technique alone is not enough to represent semantic similarities among sentences. The introduced NewsCos model increases efficiency due to the implementation of semantic relevance scoring via cosine similarity, whereas NewsCos-R additionally implements redundancy regulation via DSRC.

Natural Resources for Human Health
eISSN: 2583-1194

4.3 Impact of Dynamic Semantic Redundancy Control (DSRC)

In order to assess the efficacy of the suggested Dynamic Semantic Redundancy Control approach, a comparison analysis among NewsCos and NewsCos-R has been conducted. The advancements made through DSRC are depicted in Table 3. The findings conclusively show that DSRC operates effectively in all the evaluation metrics. While in conventional redundancy elimination methods using fixed thresholds, DSRC adapts the redundancy penalty depending on semantic relevances. Important sentences are assigned low redundancy penalties, thus facilitating the maintenance of context and at the same time getting rid of redundant data. ROUGE-2 and BLEU scores are especially indicative of improvement in preservation of contextual phrases. METEOR scores further reveal better semantic similarities between machine and human written summaries.

Table 3 Performance Improvement Introduced by DSRC

Metric	NewsCos	NewsCos-R	Improvement
ROUGE-1	40.8	41.5	0.7
ROUGE-2	16.7	17.4	0.7
ROUGE-L	38.2	38.9	0.7
BLEU	11.8	12.6	0.8
METEOR	14.6	15.2	0.6

3.6 Summary Generation Algorithm

Algorithm 1 NewsCos-R with Dynamic Semantic Redundancy Control

Require: News document d, summary length k, redundancy scaling factor α
Ensure: Extractive summary $\hat{Y}$

Segment d into sentences $S = \{S_1, S_2, \ldots, S_m\}$
for each $S_i \in S$ **do**
  Remove HTML tags, punctuation, special characters, numbers, and extra whitespace
  Convert S_i to lowercase and remove stop words
  Apply lemmatization to S_i
end for
Construct vocabulary V
for each $t \in V$ **do**
  $TF(t,d) \leftarrow \frac{f(t,d)}{\sum_w f(w,d)}$
  $IDF(t) \leftarrow \log\left(\frac{N}{df(t)}\right)$
  $TFIDF(t,d) \leftarrow TF(t,d) \times IDF(t)$
end for
Represent each S_i by its TF-IDF vector
$C \leftarrow \frac{1}{m}\sum_{i=1}^{m} S_i$
for each $S_i \in S$ **do**
  $RelevanceScore_i \leftarrow \frac{S_i \cdot C}{\|S_i\|\|C\|}$
end for
$RelevanceScore_{max} \leftarrow \max_{S_i \in S}(RelevanceScore_i)$
for each $S_i \in S$ **do**
  $\lambda_i \leftarrow \alpha \exp\left(-\frac{RelevanceScore_i}{RelevanceScore_{max}}\right)$
end for
$Summary \leftarrow \varnothing$
$Candidate \leftarrow S$
while $|Summary| < k$ **and** $Candidate \neq \varnothing$ **do**
  for each $S_i \in Candidate$ **do**
    if $Summary = \varnothing$ **then**
      $R_i \leftarrow 0$
    else
      $R_i \leftarrow \max_{S_j \in Summary} CosineSimilarity(S_i, S_j)$
    end if
    $FinalScore_i \leftarrow RelevanceScore_i - \lambda_i R_i$
  end for
  $S^* \leftarrow \underset{S_i \in Candidate}{\arg\max}\ FinalScore_i$
  $Summary \leftarrow Summary \cup \{S^*\}$
  $Candidate \leftarrow Candidate \setminus \{S^*\}$
end while
Sort $Summary$ according to the original sentence positions
$\hat{Y} \leftarrow Order(Summary)$
return $\hat{Y}$

Table 4 Comparison of Summaries Generated by Different Methods

Example	Lead-3	TextRank	LSA	NewsCos
1	Ever noticed how plane seats appear to be getting smaller and smaller. With increasing numbers of people taking to the skies, some experts are questioning whether packed planes are putting passengers at risk. They say the shrinking space on aeroplanes is not only uncomfortable but also dangerous.	They say the shrinking space on aeroplanes is not only uncomfortable but also dangerous. A U.S. consumer advisory group told the Department of Transportation that while standards exist for animals on planes, there are no minimum space requirements for humans. Tests by the FAA are based on a 31-inch seat pitch, though many airlines now offer 30, 29, or even 28 inches.	With increasing numbers of people taking to the skies, some experts are questioning whether packed planes are putting passengers at risk. They say the shrinking space on aeroplanes is not only uncomfortable but also dangerous. A U.S. consumer advisory group told the Department of Transportation that while standards exist for animals on planes, there are no minimum space requirements for humans.	They say the shrinking space on aeroplanes is not only uncomfortable but also dangerous. A U.S. consumer advisory group told the Department of Transportation that while standards exist for animals on planes, there are no minimum space requirements for humans. Tests by the FAA are based on a 31-inch seat pitch, though many airlines now offer 30, 29, or even 28 inches.
2	A drunk teenage boy had to be rescued by security after jumping into a lions' enclosure at a zoo in western India. Rahul Kumar, 17, clambered over the enclosure fence at the Kamla Nehru Zoological Park in Ahmedabad. Guards caught up with him after he fell into one of the safety moats.	Rahul Kumar, 17, clambered over the enclosure fence at the Kamla Nehru Zoological Park in Ahmedabad, and began running towards the animals. Guards caught up with him after he fell into one of the safety moats. He was taken out and handed to the police.	A drunk teenage boy had to be rescued by security after jumping into a lions' enclosure at a zoo in western India. Guards caught up with him after he fell into one of the safety moats. He was taken out and handed to the police.	Rahul Kumar, 17, clambered over the enclosure fence at the Kamla Nehru Zoological Park in Ahmedabad, and began running towards the animals, shouting he would "kill them." Guards caught up with him after he fell into one of the safety moats. He was taken out and handed to the police.
3	Dougie Freedman is on the verge of agreeing a new two-year deal to remain at Nottingham Forest. Freedman has stabilised Forest since replacing Stuart Pearce. The club's owners are pleased with the job he has done at the City Ground.	Freedman has stabilised Forest since replacing Stuart Pearce, and the club's owners are pleased with the job he has done at the City Ground. Despite that, the owners are moving to secure him on a contract for the next two seasons.	Dougie Freedman is on the verge of agreeing a new two-year deal to remain at Nottingham Forest. The club's owners are pleased with the job he has done at the City Ground. Despite that, the owners are moving to secure him on a contract for the next two seasons.	Freedman has stabilised Forest since replacing Stuart Pearce, and the club's owners are pleased with the job he has done at the City Ground. Despite that, the owners are moving to secure him on a contract for the next two seasons.

4	Liverpool target Neto is attracting interest from PSG and several Spanish clubs. Neto's contract expires in June and he is expected to leave Fiorentina in the summer.	Neto's contract expires in June and he is expected to leave Fiorentina in the summer. His agent says no final decision has been made and he will have offers from several top European clubs.	Liverpool target Neto is attracting interest from PSG and several Spanish clubs. His agent says no final decision has been made and he will have offers from several top European clubs.	Neto's contract expires in June and he is expected to leave Fiorentina in the summer. His agent says no final decision has been made and he will have offers from several top European clubs.

4.4 Ablation Study

Ablation experiments were done to understand the impact of each component of the proposed framework. In particular, the impact of centroid similarity-based relevance scoring and DSRC was studied through selective removal of these two components from the proposed framework. As indicated in Figure 3 Ablation Study Results, semantic relevance estimation using centroids and DSRC are both crucial for summarization quality. Without DSRC, the performance drops because there is more semantic repetition inside the summary. The same happens without centroid scoring because the system finds it harder to determine relevant sentences from the context. The lowest performance was obtained when both modules were excluded, which proves their synergy in the context of the proposed approach.

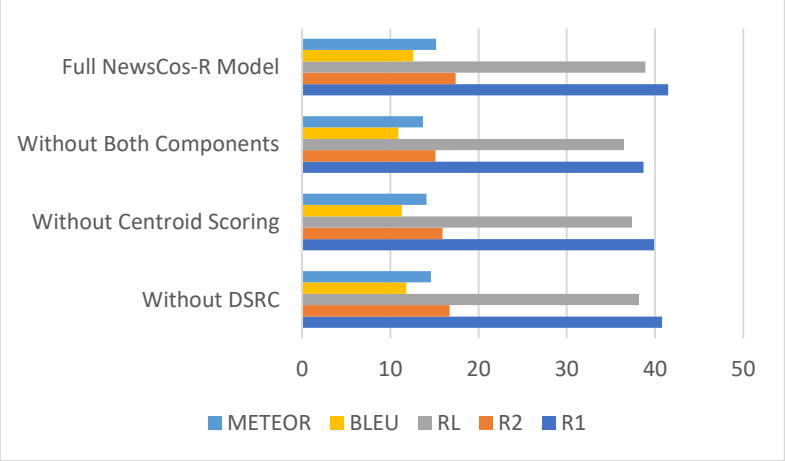

Figure 3 Ablation Study Results

4.5 Efficiency Analysis

Besides enhanced performance in terms of better summarization, the proposed NewsCos-R architecture retains low computational cost and high processing speed. In contrast to transformer-based abstractive summary frameworks which depend on heavy GPU computation and extensive training data, the proposed model runs at an efficient level by performing statistics and similarities computations. Such a model is able to perform summarization tasks even in real-time environments with low computational costs, thus enabling its use in a wide range of applications like news aggregation sites, summarization of medical documents, analyzing legal documents, among others.

4.6 Summary of Findings

In general, the experimental outcomes prove the efficiency of the proposed NewsCos-R framework for extractive text summarization. TF-IDF sentence representation, cosine similarity-based calculation of centroid relevance scores, and the proposed Dynamic Semantic Redundancy Control mechanism contribute greatly to the improvement of the results achieved in terms of all evaluation criteria. On the one hand, the presented framework demonstrates superiority to other baseline algorithms; on the other hand, it maintains a relatively low complexity level and high interpretability. According to the ablation study, redundancy control is critical to ensuring high semantic diversity and avoiding repeated sentence extraction. Thus, the proposed framework can be considered an effective solution to the problem under discussion.

5. CONCLUSION

In this research, a lightweight yet interpretable text summarization framework called NewsCos-R that leverages TF-IDF sentence representation and cosine similarity-based scoring method and utilizes the proposed Dynamic Semantic Redundancy Control strategy was designed. The main purpose of this work was to create an effective text summarizer that would generate semantically coherent summaries efficiently without increasing computational costs and losing model explainability. Experiments conducted with the use of the CNN/DailyMail dataset showed the superiority of the developed framework over a number of well-known traditional summarizers such as Lead-3, TextRank, LexRank, SumBasic, and other TF-IDF methods in terms of ROUGE, BLEU, and METEOR metrics. The use of DSRC helped achieve higher summary diversity and lower repetition due to the dynamic adjustment of redundancy penalties depending on sentence relevance. The ablation study proved the independent significance of both centroid-based relevance scoring and redundant regulation strategies in the effectiveness of the framework. As opposed to deep-learning text summarization models that demand considerable computational power, large annotated datasets, and complicated training process, the proposed approach allows designing an efficient and clear text summarizer. The proposed framework keeps semantic relevance intact while offering faster inferencing speeds and easy implementation, thus making it feasible for real-life applications like news summarization, legal documents, healthcare reports, and resource-limited NLP applications. Overall, the proposed NewsCos-R framework demonstrates that lightweight statistical approaches combined with adaptive semantic redundancy management can achieve competitive summarization performance while preserving interpretability, efficiency, and scalability.

6. FUTURE SCOPE

Although the proposed NewsCos-R framework achieved competitive performance, several enhancements can be explored in future research. First, semantic sentence embeddings generated using transformer-based language models such as BERT or Sentence-BERT can be integrated with the existing statistical framework to improve contextual understanding. Second, multilingual summarization capabilities can be developed to support low-resource regional languages and cross-lingual summarization tasks. Third, adaptive summary length optimization and query-focused summarization techniques can be incorporated for personalized summarization applications. Future work may also explore hybrid extractive-abstractive architectures that combine the interpretability of statistical methods with the semantic richness of neural language models while maintaining computational efficiency.

AUTHOR CONTRIBUTIONS

Himanshu Kumar: Methodology, Writing Original Draft

V K Jain: Supervision

Vivek Kumar: Supervision, Validation.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

Ethical approval was not required for this study.

DATA AVAILABILITY STATEMENT

Data supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] U. Rani and K. Bidhan, "Comparative assessment of extractive summarization: textrank tf-idf and lda," Journal of scientific research, vol. 65, no. 1, pp. 304–311, 2021.

[2] K. Rosamma and P. Nagamma, "A hybrid lead scoring-bigru model for extractive summarization of news articles," SN Computer Science, vol. 6, no. 6, p. 664, 2025.

[3] R. Alsultan, A. Sagheer, H. Hamdoun, L. Alshamlan, and L. Alfadhli, "Pegasus-xl with saliency-guided scoring and long-input encoding for multi-document abstractive summarization," Scientific Reports, vol. 15, no. 1, p. 26529, 2025.

[4] K. Xu, Y. Feng, Q. Li, Z. Dong, and J. Wei, "Survey on terminology extraction from texts," Journal of Big Data, vol. 12, no. 1, pp. 1–40, 2025.
[5] R. Khan, Y. Qian, and S. Naeem, "Extractive based text summarization using k-means and tf-idf," International Journal of Information Engineering and Electronic Business, vol. 13, no. 3, p. 33, 2019.
[6] W. Liu, Y. Sun, B. Yu, H. Wang, Q. Peng, M. Hou, H. Guo, H. Wang, and C. Liu, "Automatic text summarization method based on improved textrank algorithm and k-means clustering," Knowledge-Based Systems, vol. 287, p. 111447, 2024.
[7] P. Watanangura, S. Vanichrudee, O. Minteer, T. Sringamdee, N. Thanngam, and T. Siriborvornratanakul, "A comparative survey of text summarization techniques," SN Computer Science, vol. 5, no. 1, p. 47, 2023.
[8] Adeniran, "Comparative evaluation of lsa-based summarization against traditional and neural approaches using cosine similarity," Sch Bull, vol. 11, no. 5, pp. 98–104, 2025.
[9] Riaz, O. Abdulkader, M. J. Ikram, and S. Jan, "Exploring topic modelling: a comparative analysis of traditional and transformer-based approaches with emphasis on coherence and diversity," International Journal of Electrical and Computer Engineering (IJECE), vol. 15, no. 2, pp. 1933–1948, 2025.
[10] K. Taneja and J. Vashishtha, "Comparison of transfer learning and traditional machine learning approach for text classification," in 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2022, pp. 195–200.
[11] P. Gupta, S. Nigam, and R. Singh, "Automatic text summarization using sequence to sequence model and recurrent neural network," Computación y Sistemas, vol. 28, no. 4, pp. 2297–2314, 2024.
[12] K. P. Sharma, M. S. A. Yajid, J. Gowrishankar, R. Mahajan, A. R. Alsoud, A. Jadhav, and D. Singh, "A systematic review on text summarization: Techniques, challenges, opportunities," Expert Systems, vol. 42, no. 4, p. e13833, 2025.
[13] W. R. Naji, F. A. Ghanem et al., "Enhancing arabic extractive summarization with tf-idf-weighted arabert sentence embeddings and semantic clustering," The Indonesian Journal of Computer Science, vol. 14, no. 5, 2025.
[14] J. J. Sihombing, A. Arnita, S. I. Al Idrus, and D. Y. Niska, "Implementation of text summarization on indonesian scientific articles using textrank algorithm with tf-idf web-based," Journal of Soft Computing Exploration, vol. 5, no. 3, pp. 310–319, 2024.
[15] Verma, A. Thomas, and R. K. Verma, "Generative artificial intelligence-based modified abstractive cross attention enabled sequence to sequence model for abstractive hindi text summarization," Engineering Applications of Artificial Intelligence, vol. 158, p. 111478, 2025.
[16] Z. Li, M. Liu, W. Chen, and L. Zheng, "A method of extractive text summarization using document semantic graph with node ranking," International Journal of Intelligent Systems, vol. 2025, no. 1, p. 5530784, 2025.
[17] W. Ansar, S. Goswami, and A. Chakrabarti, "Lexsus: A hybrid lexical-graph salience based text summarization technique using pegasus," Authorea Preprints, 2023.
[18] L. Ragazzi, G. Moro, L. Valgimigli, and R. Fiorani, "Cross-document distillation via graph-based summarization of extracted essential knowledge," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.
[19] J. P. Verma, S. Bhargav, M. Bhavsar, P. Bhattacharya, A. Bostani, S. Chowdhury, J. Webber, and A. Mehbodniya, "Graph-based extractive text summarization sentence scoring scheme for big data applications," Information, vol. 14, no. 9, p. 472, 2023.
[20] S. Ajay Mukund and K. Easwarakumar, "Optimizing legal text summarization through dynamic retrieval-augmented generation and domain-specific adaptation," Symmetry, vol. 17, no. 5, p. 633, 2025.
[21] S. Gupta and S. Pal, "Advancing hindi text summarization: Named entity recognition and content augmentation strategies," ACM Transactions on Asian and Low-Resource Language Information Processing, 2025.
[22] W. R. NJI, M. A. SURESHA, and R. AHMED, "Leveraging arabert with statistical weighting for scalable unsupervised arabic text summarization," Journal of Theoretical and Applied Information Technology, vol. 103, no. 22, 2025.
[23] Y. Wang, "Research on the tf–idf algorithm combined with semantics for automatic extraction of keywords from network news texts," Journal of Intelligent Systems, vol. 33, no. 1, p. 20230300, 2024.
[24] Y. Ortakci, "Revolutionary text clustering: Investigating transfer learning capacity of sbert models through pooling techniques," Engineering Science and Technology, an International Journal, vol. 55, p. 101730, 2024.
[25] M. VR Manasi and S. Adarsh, "Developing a classification model and semantic similarity detection of system and software requirements using sentence transformers," Tech. Rep., 2024.

[26] L. Zhang, "Features extraction based on naive bayes algorithm and tf-idf for news classification," PLoS One, vol. 20, no. 7, p. e0327347, 2025.
[27] K. Shahade and P. V. Deshmukh, "A unified approach to text summarization: Classical, machine learning, and deep learning methods." Ingénierie des Systemes d'Information, vol. 30, no. 1, 2025.