Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Topic Driven Text Extraction for Kannada Document Summarization Using LDA

Veena R [1], Dr. D. Ramesh [2] , Dr. Hanumanthappa M [3]

[1]*Sri Siddhartha Academy of Higher Education*
[2]*Sri Sddhartha Academy of Higher Education*
[3]*Bangalore University*

ARTICLE INFO

Received: 11 Dec 2024

Revised: 05 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

Automatic Text Summarization (ATS) compacts source content into a concise format while preserving core information. While extensively studied for resource-rich languages, ATS remains challenging for low-resource languages like Kannada due to limited corpora and NLP tools. This work introduces an extractive, topic-driven method for summarizing Kannada news articles from multiple documents. We developed a custom dataset of 100 Kannada news story sets (3 articles per set) to address the lack of standardized benchmarks. The proposed approach leverages Latent Dirichlet Allocation (LDA) to identify latent themes across documents, followed by sentence selection using vector-space modeling. Sentences are scored based on their relevance to identified topics (via cosine similarity) and prioritized to maximize informational value while minimizing redundancy through Maximum Marginal Relevance (MMR). Evaluations using ROUGE metrics demonstrate that the LDA-based method outperforms existing summarization algorithms, producing summaries closer to human-generated references. The system achieves higher F-scores (e.g., 0.68 at 40% compression) compared to baseline models like TextRank and approaches for other Indian languages, validating its efficacy for low-resource linguistic contexts.

Keywords: Kannada NLP, multi-document summarization, topic modeling, LDA, redundancy reduction, extractive summarization.

INTRODUCTION

In the modern digital landscape, countless websites generate extensive volumes of news articles, often covering identical events with minor differences while leaving individual readers with fragmented details. Users browsing multiple articles on the same subject regularly face repetitive content. Condensing key insights from diverse sources into unified summaries would significantly enhance reader efficiency. To tackle this issue, multi-document summarization systems filter out redundant material and prioritize high-impact sentences to deliver cohesive overviews.

Automatic text summarization involves distilling vital information from documents to produce succinct synopses. Techniques vary depending on the number of source texts (single or multi-document) and the method of summary generation (extractive or abstractive). Extractive approaches identify and compile significant sentences verbatim from source materials, facing the challenge of pinpointing the most informative combinations due to uneven information distribution across sentences.

Abstractive methods, on the other hand, rephrase key content or generate entirely new sentences around central themes. While extractive systems often rely on machine learning—supervised or unsupervised—to rank sentence relevance, their effectiveness is constrained by the limited availability of annotated datasets. Manual labelling is labour-intensive, and even semi-supervised techniques struggle with data scarcity. As a result, unsupervised methods dominate extractive summarization.

Unsupervised systems commonly use feature-based ranking, assessing sentences via linguistic and statistical metrics. Topic-centric approaches analyse thematic word distributions in source texts, with methods like Latent Dirichlet Allocation (LDA) modelling latent themes to identify inter-document relationships. LDA's probabilistic framework has proven valuable in multi-document summarization by mapping topic structures within document collections.

Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Expanding on LDA's strengths, we propose an unsupervised extractive strategy that prioritizes sentences densely populated with topic-specific terms. This approach constructs summaries by capturing dominant themes through three stages: leveraging LDA to derive topic vectors from input documents, analysing sentence relevance, and compiling results. The method is tested using a custom dataset tailored for evaluation.

In the current digital era, numerous websites provide extensive news coverage. Many platforms reproduce similar content with minor adjustments, often failing to deliver comprehensive information per user. Readers accessing multiple articles on the same topic frequently encounter redundancy. Providing condensed summaries consolidating essential details from diverse sources would enhance efficiency. Multi-document summarization systems address this by selecting high-value sentences and eliminating unnecessary content [1].

Automatic text summarization extracts crucial information from documents to produce concise representations of source materials. Existing methods are categorized by document quantity (single or multi-document) and summarization technique. The latter includes extractive and abstractive approaches [1], distinguished by their summary-generation methods. Extractive summarization compiles key sentences verbatim from source documents. Since sentences vary in informational value, the core challenge involves identifying optimal combinations to form efficient summaries [2].

Abstractive summarization restructures critical phrases or generates new sentences based on identified themes. Most extractive multi-document systems use supervised or unsupervised learning to evaluate sentence relevance. Supervised/semi-supervised methods train classifiers on labeled data to assess importance [3]. While effective, these approaches face limitations due to scarce labeled datasets. Manual annotation is impractical given the volume of required training data, and even semi-supervised methods struggle with limited labeled examples. Consequently, unsupervised learning dominates extractive summarization [3].

Feature-based ranking techniques are widely adopted in unsupervised frameworks. These assess sentences using linguistic and statistical features to determine relevance [1]. Topic-based approaches analyze subject-related word distributions in source texts, leveraging thematic patterns for summary generation. Each sentence typically aligns with overarching document themes. Topic identification helps determine content suitability, while latent themes evaluate inter-document correlations [4]. Latent Dirichlet Allocation (LDA) [5,6], a probabilistic framework for document collections, has proven effective for multi-document summarization [7–11].

Building on LDA's success, we propose an unsupervised extractive strategy prioritizing sentences rich in topic-specific terms. The generated summary captures core ideas through three stages: deriving topic vectors via LDA, analyzing sentence relevance, and compiling results. Experiments used a custom Kannada ATS dataset developed collaboratively with language experts, addressing the absence of a benchmark equivalent to English DUC. This dataset was automatically evaluated to validate the approach.

This work's primary contributions include an unsupervised technique combining topic modeling and MMR for extractive multi-document summarization, a redundancy removal mechanism to enhance content diversity in summaries, and the creation of a dedicated Kannada dataset to bridge the gap in regional language resources. The proposed method was tested against the TextRank model [11] and compared with ATS systems for other Indian languages, demonstrating performance comparable to prior language-specific efforts.

The subsequent sections detail the research methodology and outcomes: Section 2 discusses LDA-based frameworks, Section 3 explains the proposed approach, Section 4 evaluates performance metrics, and Section 5 concludes with insights and implications.

RELATED WORKS

[12] Manju et al. (2021) developed a framework for extractive multi-document summarization of Malayalam news articles using topic modeling. Their method identifies latent themes via Latent Dirichlet Allocation (LDA), clusters content accordingly, and generates document-specific topic and sentence vectors. Sentences are ranked and prioritized based on semantic alignment between these vectors.

Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[13] Gunasundari et al. introduced an enhanced text summarization approach combining the TextRank algorithm with cosine similarity. Their unsupervised method leverages graph-based ranking to identify key sentences, improving summary quality through similarity-based sentence selection.

[14] Pokharkar et al. (2022) explored extractive and abstractive summarization techniques. Their extractive approach focuses on selecting key phrases and paragraphs using linguistic and statistical features, while the abstractive method analyzes text semantics to generate contextually accurate summaries. The study advocates NLP-driven solutions for automated summarization.

[15] Senthamizh and Arutchelvan (2022) proposed a multi-source text summarization system incorporating tokenization, lemmatization, and named entity recognition. Their technique employs document clustering to organize content thematically, bypassing the need for semantic structuring while preserving critical linguistic elements.

[16] A cosine similarity-based extractive summarization method was proposed to automatically condense texts while retaining essential information. The approach prioritizes sentences with high semantic relevance to the document's core themes.

[17] This work addresses multi-document summarization for Malayalam texts using computational methods. Despite limited training data for deep learning, the authors achieved promising accuracy, recall, and F-measure scores by prioritizing diverse sentences through statistical analysis.

[18] Topic-based methods, particularly LDA, remain foundational for multi-document summarization. LDA interprets documents as mixtures of latent topics, where each topic is defined by a unique distribution of words, enabling thematic clustering of content.

[19] The ROUGE metric evaluates summary quality in this study. Leveraging LDA's probabilistic modeling, the proposed strategy identifies sentence relevance by mapping text to latent topics. LDA treats documents as distributions of themes and sentences as representations of these themes, enabling topic-driven summarization.

METHODOLOGY

This section outlines the architecture's process flow and provides the dataset used in the experiments. Figure 1 depicts the system's architecture.

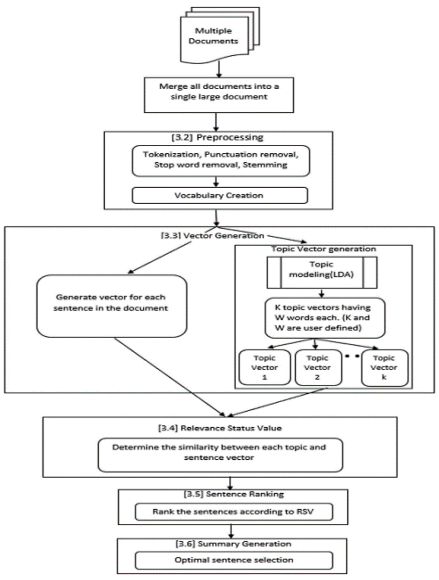

Fig 1-- Overview of Methodology

Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Dataset

While Kannada summarization systems remain in a nascent phase, standardized datasets for analytical frameworks are currently unavailable. To address this gap, we compiled 100 distinct collections, each comprising three articles sourced from three prominent Kannada e-newspapers.

Dataset parameters	
The total amount of record sets	100
Each set contains a certain number of documents.	3
The mean total of words in every record	21.7
Ideal number of clauses enable in a piece of paper	70
Minimal sentence count per record	10
(%) a brief length	40

Text preprocessing is a foundational stage in natural language processing, critically shaping the effectiveness of downstream algorithms. This phase standardizes raw text into a structured format suitable for computational analysis. Key tasks include sentence segmentation, tokenization, stopword removal, and stemming.

Documents are first decomposed into individual sentences using Python's Natural Language Toolkit (NLTK). Tokenization then splits sentences into discrete units (tokens). An 85-word stopword list filters out linguistically frequent but semantically low-value terms, improving cosine similarity calculations and streamlining sentence vectorization.

Stemming reduces words to their root forms using a method analogous to the Indic stemmer [37], which iteratively strips inflectional suffixes. For example, Malayalam words like "vanathil" and "vanathiloode" are normalized to the root "vanam." This standardization ensures related word forms receive identical treatment during vector generation, enhancing model accuracy.

The Gensim library facilitates vocabulary creation by mapping processed tokens to structured dictionary entries. These preprocessing steps collectively optimize the system's performance in subsequent analytical stages.

Vector generation

The process involves converting input topic words and textual content into numerical vectors. Each sentence within the input text is transformed into a vector representation. Among the prevalent methods for this conversion, binary encoding and TF-IDF (Term Frequency-Inverse Document Frequency) vectorization [38] are widely adopted. Following this, topic modeling is employed to generate topic vectors. The Latent Dirichlet Allocation (LDA) methodology is applied to analyze the distribution and representation of thematic content within the text.

LDA Implementation Steps:

(i) **Determine Topic Quantity:** Identify the number of topics NNN present in the document collection.

(ii) **Initial Topic Assignment:** LDA randomly allocates each word in every sentence to one of the NNN topics. This random assignment serves as a starting point but lacks precision.

(iii) **Topic Distribution Analysis:** To refine accuracy, LDA calculates the proportion of phrases assigned to each topic within the text.

The following probabilities guide the adjustment process:

$$P(T|S) = \text{Percentage of words in sentence S linked to topic T}$$

Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

P(T|S) =Percentage of words in sentence S linked to topic T
P(W|T) =Percentage of words currently assigned to topic T within sentence S

(iv) The product of P(T|S) and P(W|T) determines the likelihood of a word belonging to a specific topic. Each word is reassigned to the topic with the highest computed probability.

(v) **Iterative Optimization:** Steps (iii) and (iv) are repeated for every word across all sentences until convergence—a state where topic assignments stabilize.

The final output comprises K topics, each defined by a weighted combination of keywords that encapsulate its thematic essence.

The LDA topic vector generation process can be illustrated through an example. Consider two documents (as shown in Figure 2) discussing the appointment of a Chief of Defense Staff (CDS) to streamline military coordination. When processing this input text to derive three topics, LDA modeling identifies the following thematic clusters (Figure 2): the first theme concerns defense system modernization, the second addresses the necessity of the CDS role, and the third focuses on legislative terminology related to the CDS establishment.

Relevance Status Value (RSV)
RSV computation involves assessing the similarity between sentence vectors and topic vectors. This relevance metric is typically derived using established similarity measures such as Euclidean distance, Jaccard similarity, and Cosine similarity, which quantify the alignment between textual elements and thematic content.

$$C.S\ (T, S) = \frac{\sum_{k=1}^{n} T_i * S_i}{\sqrt{\sum_{k=1}^{n} T_i^2} * \sqrt{\sum_{k=1}^{n} S_i^2}}$$

Where, C.S is Cosine Similarities, the components of vectors T and S are Ti and Si, respectively.

Sentence Ranking
Sentences are arranged in descending order of their Relevance Status Value (RSV) and forwarded to the summary generation phase. This process aligns with the LDA-derived topic structure (Figure 2), which identifies three thematic clusters: defensive system updates, the necessity of a Chief of Defense Staff (CDS), and legislative terminology tied to the CDS establishment. The highest-ranked sentences correspond closely to the **K** topic vectors, ensuring thematic relevance in the output.

Summary Generation
The final stage involves synthesizing summaries by filtering redundant content from top-scoring sentences. Multi-document analysis often deals with large, diverse document sets, inherently containing more repetition than single-document contexts. Redundancy mitigation is critical here. Maximum Marginal Relevance (MMR) and clustering are primary strategies for this purpose. MMR evaluates textual overlap between candidate sentences and the evolving summary, excluding redundant phrases. The process initiates by incorporating the highest-ranked sentence into the summary, iteratively adding subsequent sentences while enforcing redundancy constraints.

Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

	Inputs	Outputs
1	ಸತ್ಯವನ್ನು ಅರಿತವರು ದುರಾಸಿಗಳಾಗುವುದಿಲ್ಲ, ದುರಹಂಕಾರಿಗಳಾಗುವುದಿಲ್ಲ, ಸ್ವಾರ್ಥಿಗಳಾಗುದಿಲ್ಲ, ಕ್ರೂರಿಗಳಾಗುವುದಿಲ್ಲ ಮತ್ತು ಕ್ರೋಧದಿಂದ ಯಾರನ್ನು ಯಾರೂ ದ್ವೇಷಿಸುವುದಿಲ್ಲ	ಸತ್ಯವನ್ನು ತಿಳಿದಿರುವ ಜನರು ದುರಾಶೆ, ದುರಹಂಕಾರ, ಸ್ವಾರ್ಥ ಅಥವಾ ಕ್ರೌರ್ಯವನ್ನು ಹೊಂದಿರುವುದಿಲ್ಲ ಮತ್ತು ಅವರು ಯಾರಿಗೂ ಇಷ್ಟವಾಗುವುದಿಲ್ಲ
2	ಮಾನವರಾದ ನಮಗೆ ಈ ಜೀವನ ಮತ್ತು ನಾವು ವಾಸಿಸುವ ಪರಿಸರವನ್ನು ಉಡುಗೊರೆಯಾಗಿ ನೀಡಿರುವುದು ಒಂದು ಸುಂದರವಾದ ಆಶೀರ್ವಾದವಾಗಿದೆ. ತಾಯಿಯ ಪ್ರೀತಿಯು ಅಪ್ರತಿಮವಾಗಿದೆ ಎಂದು ಪ್ರಕೃತಿಯನ್ನು "ತಾಯಿ" ಎಂದೂ ಕರೆಯಲಾಗುತ್ತದೆ.ಅವರು ನಮಗಾಗಿ ತಮ್ಮಲ್ಲಿರುವ ಎಲ್ಲವನ್ನೂ ನೀಡುತ್ತಾರೆ, ನಮ್ಮನ್ನು ರಕ್ಷಿಸುತ್ತಾರೆ, ನಮಗೆ ಆಹಾರವನ್ನು ನೀಡುತ್ತಾರೆ ಆದರೆ ಪ್ರತಿಯಾಗಿ ಏನನ್ನೂ ನಿರೀಕ್ಷಿಸುವುದಿಲ್ಲ. ಸಂಕ್ಷಿಪ್ತವಾಗಿ, ಪ್ರಕೃತಿಯು ಜೀವನದ ಅತ್ಯಂತ ಸ್ಫೂರ್ತಿಯಾಗಿದೆ.	ಪ್ರಕೃತಿಯು ಮಾನವರಿಗೆ ನೀಡಿದ ಅಮೂಲ್ಯ ಕೊಡುಗೆಯಾಗಿದೆ ಮತ್ತು ಇದನ್ನು ಸಾಮಾನ್ಯವಾಗಿ ಪ್ರೀತಿಯ ತಾಯಿಗೆ ಹೋಲಿಸಲಾಗುತ್ತದೆ. ಪ್ರಕೃತಿಯುಪ್ರತಿಯಾಗಿಏನನ್ನೂ ಕೇಳದೆ ನಮ್ಮನ್ನು ರಕ್ಷಿಸುತ್ತದೆ ಮತ್ತು ಅದನ್ನು ಜೀವನದ ಸಾರವನ್ನಾಗಿ ಮಾಡುತ್ತದೆ.

Fig 2. LDA example of topic word subsequent generations from input text.

Algorithm 1. Latent Dirichlet Allocation with Maximum Marginal Relevance -based Multi Document Summary.

Input
D: The combined manuscript thru n phrases where =... D SS S 1 2 ,, , n
K: The number of LDA modeling topics. (Specified by the user)
W: The number of terms that must be supplied in each topic. (Specified according to the user)
C: The number of statements that will be featured in the overview. (Specified depending on the user)

Output
S. D is the document number in the mining multi document summary.
i. Preprocess the following phases of segmenting the document D into statements as listed below:
(1) Tokenization, (2) Punctuation destruction, (3) Stopword elimination, (4) Stemming
ii. A vector is made for every sentence in document D.
iii. LDA is used to create a topic vector for document D (topic W words are assigned to K topics).
iv. Determine the RSV for each phrase through assessing the similarities of the wording and topic vectors.
v. Applying RSV, to arrange the sentences.
vi. To generate the ultimate general, run MMR. S′:
(a) Increase the initial phrase from the Ranklist to the overview
(b) Examine the resulting word to the present terms in the overview
(c) The sentence is added to S′ if the similarity between the new sentence and the other summary clauses falls less than 0.66.
Steps (b)-(c) should be repeated until the length () S′ C is reached.

PERFORMANCE EVALUATION METRICS

Evaluation Metric

This study employs the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) framework [34] for automated summary assessment. The implemented metrics include ROUGE-N (analyzing unigrams via ROUGE-1 and bigrams

Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

via ROUGE-2), ROUGE-L (measuring longest common subsequences), and ROUGE-SU4 (combining skip-bigrams and unigram co-occurrence patterns).
ROUGE-N evaluates semantic alignment between system-generated summaries and human-authored reference texts by calculating n-gram overlap. Specifically:

- ROUGE-1 quantifies unigram overlap.
- ROUGE-2 assesses bigram correspondence.

ROUGE-L identifies the longest sequence of matching terms between summaries, while ROUGE-SU4 incorporates flexible skip-bigram analysis with unigram frequency statistics. This multi-faceted approach ensures rigorous evaluation of content relevance and linguistic coherence in generated summaries. Here's how it's calculated as below:
ROUGE-N=
Where, N denotes the N-gram's length and (). The maximum number of N-grams that can be found in both the candidate and reference summaries is known as the count gram N match. ROUGE--1 unigrams and ROUGE--2 bigrams, which calculate the percentage of interleaved, respectively, are the most commonly used ROUGE measures values. Better ROUGE--S, ROUGE-SU4 [41] is a variant of ROUGE-SU. ROUGE-SU4 can skip up to four distances between bigrams. ROUGE-L assesses summary fluency using the lengthiest frequently encountered. Here's how it's calculated as below:

$$\text{ROUGE-N} = \frac{\sum_{\text{gram}_N \in \text{Ref}} \text{Count}_{\text{match}}(\text{gram}_N)}{\sum_{\text{gram}_N \in \text{Ref}} \text{Count}(\text{gram}_N)}$$

Where, N denotes the N-gram's length and (). The maximum number of N-grams that can be found in both the candidate and reference summaries is known as the count gram N match. ROUGE--1 unigrams and ROUGE--2 bigrams, which calculate the percentage of interleaved, respectively, are the most commonly used ROUGE measures values. Better ROUGE--S, ROUGE-SU4 [41] is a variant of ROUGE-SU. ROUGE-SU4 can skip up to four distances between bigrams. ROUGE-L assesses summary fluency using the lengthiest frequently encountered subsequence (LCS) method, which takes sentence-level structure comparison into consideration. Let S be the system summary and R denote the n-word reference of summary. The following formula is used to calculate

$$\text{ROUGE-L} = \frac{LCS(S,R)}{n}$$

4.2 Experiments and Results

All tests were executed on a system featuring an Intel Core i5-8250 CPU operating at 1.80 GHz with 16 GB RAM, using Python-based implementations. The experiments focused on Kannada document extraction and summarization using LDA (p. 401) to evaluate the proposed multi-document summarization (MDS) system.
A dedicated summarization dataset was utilized for comprehensive performance analysis. The model was tested across compression ratios (CRs) of 10%–40% and topic counts of 3, 5, and 9. Table 2 summarizes the precision, recall, and F-measure scores for the 10% CR configuration.
Observations reveal that summaries with narrower thematic focus (fewer topics) yield higher ROUGE scores, which decline as topic breadth increases. Table 3 details the ROUGE-1 and ROUGE-2 metrics for CRs spanning 10%–40% alongside topic counts of 3–9, demonstrating the interplay between compression granularity and thematic coverage.

Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table-2 shows the different ROUGE parameters for the suggested technique shown as the number of topics is varied with 10% CR.
Where, R=Recall, P=Precision, F-S= F-Score

Compression Raio = 10%					
# Topic	Measures .	ROUGE--L .	ROUGE--1 .	ROUGE--2.	ROUGE--SU4.
3	R	0.33182	0.29502	0.28632	0.29163
	P	0.82022	0.81915	0.78824	0.79737
	F-S	0 .47249	0.4338	0.42006	0.42706
5	R	0.29091	0.27586	0.23932	0.24543
	P	0.64	0.6729	0.59574	0.60714
	F-S	0.4	0.3913	0.34146	0.34955
9	R	0.28636	0.2567	0.23504	0.2435
	P	0.64948	0.64423	0.56122	0.56854
	F-S	0.39748	0.36712	0.33133	0.34097

Table 3. ROUGE-1 and ROUGE-2 values for the suggested model due to the large number of topics for various Compression Rations.

		Compression Ratios -10%		Compression Ratios -20%		Compression Ratios -30%		Compression Ratios -40%	
Topics	Measures	ROUGE--1	ROUGE--2	ROUGE-1	ROUGE-2	ROUGE-1	ROUGE-2	ROUGE-1	ROUGE--2
3	R	0.29502	0.28632	0.46743	0.42735	0.68199	0.65385	0.75862	0.7265
	P	0.81915	0.78824	0.58095	0.52083	0.55799	0.52577	0.54848	0.51672
	F-S	0.4338	0.42006	0.51805	0.46948	0.61379	0.58286	0.63666	0.60391
5	R	0.27586	0.23932	0.57471	0.54701	0.62835	0.59402	0.72414	0.68803
	P	0.6729	0.59574	0.65502	0.61244	0.57143	0.53053	0.5431	0.50789
	F-S	0.3913	0.34146	0.61224	0.57788	0.59854	0.56048	0.62069	0.58439
9	R	0.2567	0.23504	0.47127	0.44444	0.69349	0.66667	0.7931	0.77778
	P	0.64423	0.56122	0.62121	0.55026	0.58766	0.55516	0.59312	0.57233
	F-S	0.36712	0.33133	0.53595	0.49173	0.6362	0.60583	0.67869	0.65942

A comparative framework was designed to evaluate TextRank-based MDS against LDA-based MDS. Table 4 presents ROUGE-1 scores for both models across varying compression ratios (CRs) before redundancy removal. The results demonstrate that embedding sentences in topic space (LDA) enhances representation accuracy and information relevance for Kannada documents.

Analysis of 10% and 40% CRs in Table 4 reveals an inverse relationship between F-score and CR reduction. Lower CRs diminish co-occurrence alignment between system-generated summaries and reference texts, decreasing F-scores.

Redundancy elimination markedly improves summary quality. Integrating the MMR algorithm with LDA-derived rankings enhances informational diversity while minimizing repetition. Comparative analysis of Table 4 (pre-redundancy removal) and Table 5 (post-redundancy removal) confirms that filtering redundant content elevates multi-document summary coherence and precision.

Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table-4. ROUGE 1 values for the mathematical representations for various compression ratios prior to duplication elimination.

Model	compression ratios 10%	compression ratios 20%	compression ratios 30%	compression ratios 40%
Text Rank				
R-avg	0.27044	0.32143	0.50649	0.50649
P-avg	0.47253	0.61111	0.56727	0.51316
F-S-avg	0.344	0.42128	0.53516	0.5098
latent dirichlet allocation Model (Suggested system) # Topics: 9				
R_avg	0.27686	0.47659	0.71648	0.75
P-avg	0.6729	0.61353	0.60129	0.62738
F-S-avg	0.3913	0.54274	0.65385	0.68323

Table-5. ROUGE 1 values for various compression ratios after redundant elimination for the models

Model	compression ratios 10%	compression ratios 20%	compression ratios 30%	compression ratios 40%
Text Rank				
R-avg	0.2222	0.3295	0.40909	0.49573
P-avg	0.63736	0.53086	0.54878	0.42963
F-S-avg	0.32955	0.40662	0.46875	0.46032
latent dirichlet allocation Model (Suggested system) # Topics: 9				
R-avg	0.2467	0.46127	0.69349	0.7931
P-avg	0.64423	0.62121	0.58766	0.59312
F-S-avg	0.36712	0.53595	0.6362	0.67869

The procedure for summarizing documents in Kannada using MMR and topic model represented in Figure 3.

Fig 3: A sample of a summary produced by the suggested system**.**

INPUT TEXT	SUMMARISED TEXT
ಪರಿಸರವನ್ನು ಸ್ವಚ್ಛವಾಗಿ ಇಟ್ಟುಕೊಳ್ಳುವ ಕರ್ತವ್ಯ ನಮ್ಮದಾಗಿದೆ. ಇದಕ್ಕಾಗಿ ಕೆಲವು ಮಾರ್ಗಗಳನ್ನು ಅನುಸರಿಸಬೇಕು. ಗಾಳಿಯು ಮಲಿನವಾಗದಂತೆ ನೋಡಿಕೊಳ್ಳಬೇಕು. ಹೊಗೆ, ಧೂಳು, ಕೊಳೆತ ಪದಾರ್ಥಗಳಿಂದ ಗಾಳಿ ಕೆಡುತ್ತದೆ. ಆದ್ದರಿಂದ ಗಾಳಿಯನ್ನು ಸೂಕ್ತ ರೀತಿಯಲ್ಲಿ ಸಂರಕ್ಷಿಸಬೇಕು. ಜಲಮೂಲಗಳ ಬಳಿ ಮಲಮೂತ್ರ ವಿಸರ್ಜಿಸುವುದು, ದನಕರುಗಳ ಮೈ ತೊಳೆಯುವುದು, ಬಟ್ಟೆ ಮತ್ತು ಪಾತ್ರೆ ಸ್ವಚ್ಛಮಾಡುವುದು, ಶೌಚಗೃಹಗಳನ್ನು ನಿರ್ಮಿಸುವುದು, ಇವುಗಳಿಂದ ನೀರು ಅಶುದ್ಧವಾಗುತ್ತದೆ. ಆದ್ದರಿಂದ ಇವುಗಳನ್ನು ತಡೆಗಟ್ಟಬೇಕು. ಸಾಮಾನ್ಯವಾಗಿ ತಗ್ಗುಪ್ರದೇಶಗಳಲ್ಲಿ ನೀರು ನಿಂತು ರೋಗಾಣುಗಳ ಮೂಲ ಸ್ಥಾನವಾಗುತ್ತದೆ. ಇದರಿಂದ ಅನೇಕ ಕಾಯಿಲೆಗಳು ಹರಡುತ್ತವೆ. ಬಚ್ಚಲ ನೀರು, ಮೋರಿಯ ನೀರು, ಸುಗಮವಾಗಿ ಹರಿದುಹೋಗುವ ವ್ಯವಸ್ಥೆ ಮಾಡಬೇಕು. ಹೀರುಗುಂಡಿಗಳನ್ನು ನಿರ್ಮಿಸಿ ಕಲುಷಿತ ನೀರು ಭೂಮಿಗೆ ಸೇರುವಂತೆ ಮಾಡಬೇಕು.	ಜಲಮೂಲಗಳ ಬಳಿ ವಿಸರ್ಜನೆ, ಜಾನುವಾರುಗಳನ್ನು ತೊಳೆಯುವುದು, ಬಟ್ಟೆ ಮತ್ತು ಪಾತ್ರೆಗಳನ್ನು ತೊಳೆಯುವುದು, ಶೌಚಾಲಯಗಳ ನಿರ್ಮಾಣ ಎಲ್ಲವೂ ನೀರನ್ನು ಅಶುದ್ಧಗೊಳಿಸುತ್ತದೆ. ಶೌಚಾಲಯ ನೀರು ಮೋರಿ ನೀರು ಸರಾಗವಾಗಿ ಹರಿಯುವಂತೆ ವ್ಯವಸ್ಥೆ ಮಾಡಬೇಕು. ಸಾಮಾನ್ಯವಾಗಿ, ತಗ್ಗುಪ್ರದೇಶಗಳಲ್ಲಿ ನಿಂತಿರುವ ನೀರು ರೋಗಾಣುಗಳ ಸಂತಾನೋತ್ಪತ್ತಿಯ ಸ್ಥಳವಾಗಿದೆ.
Google translation from kannada to English	
It is our duty to keep the environment clean. For this some steps should be followed. The air should be kept free from pollution. Air is spoiled by smoke, dust, decaying matter. So the air should be properly conserved. Excretion near water bodies, washing of cattle, washing of clothes and utensils, construction of latrines all make the water impure. So these should be avoided. Usually, standing water in low-lying areas becomes a breeding ground for germs. It spreads many diseases. Toilet water, culvert water should be arranged to flow smoothly. Sewage tanks should be constructed and the polluted water should be drained into the ground.	Excretion near water bodies, washing cattle, washing clothes and dishes, All construction of toilets Purifies water. A toilet drain allows water to flow smoothly Must be arranged. Generally, Standing water in low-lying areas is a breeding ground for germs

Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Two news articles were randomly selected from a document corpus to illustrate phrase prioritization using LDA and the proposed algorithm's extractive methodology.
The framework was validated on the National Institute of Standards and Technology (NIST)'s DUC-2018 dataset for English-language multi-document summarization. Table 6 presents the ROUGE-1 unigram scores for the DUC2016 benchmark, demonstrating the model's robust performance in English text summarization tasks.

Table 6

Model	ROUGE-1 unigrams	ROUGE-2 bigrams	ROUGE-L longest matching sequence
Text rank	00.44703	00.20462	00.21490
Proposed model	00.48821	00.22471	0 0.24968

The proposed strategy's efficacy was benchmarked against existing Indian language summarization studies. Reference [42] tested their framework on 100 Hindi news articles, while other works addressed Tamil (semantic graph-based), Marathi (TextRank-based), Punjabi (hybrid model), and Hindi (graph-based) summarization.
As evidenced by Table 7, our method surpasses prior single-language approaches in cross-lingual performance, despite their focus on individual document analysis. The table confirms the model's superior accuracy across multiple Indian languages compared to current state-of-the-art techniques.

Table 7

Strategies.	Languages.	P.	R.	F-S.
Graph based [42]	Hindi	0.44	0.32	0.37
Hybrid [19]	Punjabi	0.45	0.21	0.29
Textrank [21]	Marathi	0.43	0.27	0.33
Semantic graph [22]	Tamil	0.42	0.31	0.35
Proposed model	Kannada	0.63	0.75	0.68

CONCLUSION

This study presents a methodology for automatic multi-document summarization (MDS) of Kannada texts using Latent Dirichlet Allocation (LDA)-based topic modeling. The framework extracts dominant thematic patterns from documents and maps sentences to a reduced-dimensional topic space, enabling efficient relevance scoring. A redundancy elimination module further refines summaries by enhancing content diversity.

The evaluation leveraged a custom Kannada MDS dataset, comprising curated news articles with thematic coherence. Additional validation on English and Kannada corpora confirmed the model's cross-linguistic adaptability. Comparative analysis against Indian language summarization systems (e.g., Hindi, Tamil, Marathi) demonstrated superior ROUGE scores, underscoring LDA's efficacy in capturing contextually rich summaries.

While the generated summaries exhibit non-redundant, topic-aligned content, cohesion across multi-source inputs remains a challenge. Nevertheless, the summaries retain human-readable clarity. Future work will integrate this topic-driven approach with graph-based algorithms and evolutionary optimization techniques to enhance coherence and scalability across languages.

REFERENCES

[1] Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., & Affandy, A. (2022). Review of automatic text summarization techniques & methods. Journal of King Saud University-Computer and Information Sciences, 34(4), 1029-1046.

Research Article

[2] Radev, D. R., Jing, H., Styś, M., & Tam, D. (2004). Centroid-based summarization of multiple documents. Information Processing & Management, 40(6), 919-938.
[3] Mao, X., Yang, H., Huang, S., Liu, Y., & Li, R. (2019). Extractive summarization using supervised and unsupervised learning. Expert systems with applications, 133, 173-181.
[4] Yau, C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. Scientometrics, 100, 767-786.
[5] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78, 15169-15211.
[6] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
[7] Arora, R., & Ravindran, B. (2008, July). Latent dirichlet allocation based multi-document summarization. In Proceedings of the second workshop on Analytics for noisy unstructured text data (pp. 91-97).
[8] Twinandilla, S., Adhy, S., Surarso, B., & Kusumaningrum, R. (2018). Multi-document summarization using k-means and latent dirichlet allocation (lda)–significance sentences. Procedia Computer Science, 135, 663-670.
[9] Yang, G., Wen, D., Chen, N. S., & Sutinen, E. (2015). A novel contextual topic model for multi-document summarization. Expert Systems with Applications, 42(3), 1340-1352.
[10] Rani, R., & Lobiyal, D. K. (2021). An extractive text summarization approach using tagged-LDA based topic modeling. Multimedia tools and applications, 80, 3275-3305.
[11] Rani, U., & Bidhan, K. (2021). Comparative assessment of extractive summarization: textrank tf-idf and lda. Journal of Scientific Research, 65(1), 304-311.
[12] Kondath, M., Suseelan, D. P., & Idicula, S. M. (2022). Extractive summarization of Malayalam documents using latent Dirichlet allocation: An experience. Journal of Intelligent Systems, 31(1), 393-406.
[13] Gunasundari, S., Shylaja, M. J., Rajalaksmi, S., & Aarthi, M. K. IMPROVED DRIVEN TEXT SUMMARIZATION USING PAGERANKING ALGORITHM AND COSINE SIMILARITY.
[14] Pokharkar, A., Dhumal, P., Singh, A., & Hadawale, H. (2022). Text Summarizer Using NLP. Available at SSRN 4097878.
[15] Senthamizh, S. R., & Arutchelvan, K. (2022). Automatic text summarization using document clustering named entity recognition. International Journal of Advanced Computer Science and Applications, 13(9).
[16] Jain, R. (2022). Automatic Text Summarization of Hindi Text Using Extractive Approach. ECS Transactions, 107(1), 4469