International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Extractive Summarization of Kannada Multi Documents using LDA

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

Submitted: 27/12/2023 **Revised**: 03/02/2024 **Accepted**: 11/02/2024

Abstract: With the main content of the source text intact, Automatic Text Summarization (ATS) condenses and presents the information to the user in a more manageable format. In the scientific literature, many strategies for summarizing texts have been studied for languages with substantial resources. However, ATS is a challenging system and difficult undertaking for languages with limited resources like Kannada. The absence of a reference corpus and Language processing presents challenges in terms of adequate processing tools. We prepared a dataset of news stories written in Kannada because there wasn't a standard collection available. The work demonstrates an extractive topic modelling approach to multi-document textual presentation for Kannada newspapers. To begin, we employ the latent Dirichlet allocation technique to identify latent themes on which the cluster contents modelling technique used. The vector space model is then used for creating the inputted document's sentence vector and dependent vector. Sentences are arranged in accordance with the topic and sentence vectors of the document, taking into account the appropriate status value. Non-redundancy is maximized in the resulting summary.

The assessment results for Kannada reports show that, in comparison to the existing text summarizing algorithms, the proposed technique produces a summary that is more similar to human-generated descriptions.

Keywords: *Kannada documents, relevant status value, topic modelling, LDA, and multi-document summarization.*

1 Introduction.

In the modern digital age, there are numerous websites that offer a vast amount of news. Several of these websites offer the same news with minor variations, and the majority don't offer comprehensive information for the per user. There's a good chance that the information in the news articles is redundant if readers read multiple articles on the same subject. A condensed version of all the information from the various sources should be provided to the reader as this will be beneficial. Systems for summarizing multiple documents can help achieve this. By choosing sentences with premium contents with removing information that isn't necessary.

The text summary automatically extracts key information from one or more text documents to create a brief summary of the original text documents. There have been several ways developed for summarizing documents. Based on the number of source documents used, text summarization can be classed as single or multi-document. The summarization techniques are largely classified as extractive and abstractive summarization techniques [1] based on how the ultimate outline is produced. The extractive summarising approach summarises significant sentences from a set of documents without modifying them. Generally speaking, each sentence does not contain the same amount of information. As a result, identifying the selected collection of sentences that serves as efficient document's summary [2].

The abstract summarising methodology creates a summary by dynamically restructuring phrases containing the most significant details or by producing new sentences based on themes detected in the content. Unsupervised or supervised learning techniques are employed by most extractive multi-document summary systems to ascertain the significance of sentences. On labelled data, a classifier that determines the importance of a sentence can be trained using supervised or semi-supervised machine learning [3]. Although supervised as well as semi-supervised learning can yield outstanding results, they are not applicable to

1Research Scholar, Sri Siddartha Academy of Higher Education, Tumakuru, India
2Professor and Head, Master of Computer Application, Sri Siddartha Academy of Higher Education, Tumakuru, India
3Senior Professor, Department of Computer Science and Applications, Bangalore University, Bengaluru, India
** Corresponding author's Email: veena.channig@gmail.com*

many datasets due to a lack of labels. It is likely to label the data manually, but this method is difficult due to the large amount of essential labeled training data. This is also true for a semi-supervised approach, which requires fewer data points to be labeled. As a result, unsupervised learning is frequently used in extractive summarization [3].

Feature-based ranking techniques have been widely applied in unsupervised learning. Sentences' relevance in summary is assessed using a variety of linguistic and statistical features by feature-based techniques [1]. The distribution of subject-related words in the text that is input serves as the foundation for topic-based approaches, and modeling strategies are used to determine the summary. In reality, every sentence in the provided text refers to a theme that runs throughout the document. Topic identification is one way for determining the acceptable content of text documents. Document correlations can be assessed using latent themes [4]. Latent dirichlet allocation refers to a procreant probabilistic structure for a set of documents. (LDA) [5,6]. LDA has ensued utilized successfully for multi-document summarization [7-11].

Unsupervised extractive summarization strategy is introduced that selects sentences with the maximum amount of embedded subject terms based on the LDA success. Sentences that summarize the major ideas are included in the summary that was generated. Three important steps are used in the suggested method to produce the summary. LDA is used to generate the topic vector for the supplied document. The dataset used in the experiments was specifically made for Kannada ATS. How effective. These data sets are automatically evaluated to illustrate this approach.

In summary, this work's primary contributions are as follows.

- Based on topic modelling and MMR, propose an unsupervised technique for extractive multi-document modelling.

-This work could remove redundant sentences and add more variety to the content in the document's final summary by utilizing the redundancy removal component.

-Kannada, unlike the English document understanding conference (DUC) dataset, lacks a benchmark an ATS dataset. In response, we collaborated with Kannada University language experts to develop a multi-document ATS information set for Kannada.

- Using the data set developed for Kannada ATS, this effort examined the proposed research to the Text rank [11] model. Additionally, it compares the results with those of ATS created for other Indian languages. The performance of the suggested strategy is shown in the results as a baseline, where it yields outcomes comparable to earlier attempts for other languages.

This is how the next part of the article is ordered. The part 2 presents the relevant LDA works.

Part 3 describes the approach used. Part 4 contains facts on performance evaluations. The inference is depicted in part 5.

Related works

[12] "A Framework for Generating Extractive Summary from Multiple Malayalam Documents" has been proposed by k Manju et al. (2021). Inside A multi-document text summarizing a way to summarize Malayalam newspapers that utilizes extractive topic modeling is proposed in this study. Employing the latent Dirichlet allocation topic modelling technique, they begin by finding latent themes on which to cluster the contents. The vector space model is then used to generate the provided document's subject vector alongside sentence vector. Sentences are prioritised among the subject and sentence vectors of the document based on the appropriate status value.

[13**]** Gunasundari et al. proposed "Improved driven text summarization using pageranking algorithm and cosine similarity" in their paper. Text summaries are extracted in this work using unsupervised learning approaches such as text rank. Text abstracts are often generated entirely using the text rank algorithm. To extract summaries, cosine similarity and the text rank algorithm are used in this work.

[14] In Pokharkar et al.'s (2022) the extractive summary method selects significant paragraphs, phrases, and others from the primary original text to combine into a shorter variant built on the linguistics and statistical features of the sentences. Whereas the abstractive summary approach analyses and evaluates the text using linguistic tools in order to find original concepts and expressions that successfully convey meaning. The authors suggested

developing a text summarizer with natural language processing (NLP).

[15] Senthamizh and Arutchelvan (2022) examine the most recently created text field summarization, allows the creation of abstracts from a number of different sources. Tokenizing, lemmatization, pronunciation, and the use of non-ASCII characters all help to create a summary in TS; this does not necessitate TS being semantically organised as an explanation for collecting accents from their specific location. The authors propose an automatic text summarization technique based on document clustering and named entity recognition.

[16] The author of the suggested work introduces an extractive technique for text summarization. This technique is centred on the cosine similarities technique, which automatically generates the summarised text while retaining essential details.

[17] This paper describes an approach for generating a summary from a collection of Malayalam documents. Furthermore, deep learning approaches are challenging to implement due to the limited quantity of the multi-document summarising data set. The study suggested demonstrates the success of current standard computational approaches in multi-document Malayalam. They offer a technique for sentence extraction that prioritises the most varied sentences. On a variety of input texts, the system performs satisfactorily in terms of accuracy, recollection, and F-measure.

[18] Topic-based approaches use word distribution patterns within documents to infer topics. Multi-document summarization was first solved using LDA, and researchers are continuously refining this topic model. According to LDA, every documents is a collection of various "topics," and every topic is a collections of distinct "words".

[19] ROUGE is the evaluation metric used to evaluate the strategy in relation to the dataset created for the distillation task. To extract briefs from the provided paperwork, we use topic modelling. The proposed approach uses the LDA modelling of subjects principle to identify the significance of the sentence by including its description. The LDA model, a generative probabilistic model, can be employed for analysing discrete data sets like text corpora. Every phrase in the text is interpreted by the model as a characteristic of one of the topics it covers. Every document is seen by LDA as an infinite collection of hidden themes, each of which can be distinguished by the way it is distributed throughout the text.

Methodology

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

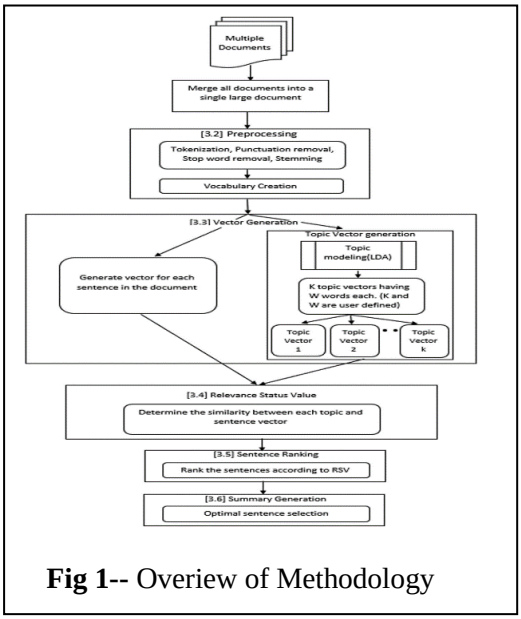

Fig 1-- Overiew of Methodology

3.1 Dataset

Although Kannada summary system is still in its early stages, there is no standard information set for analysis systems. As a result, we created a dataset of 100 capture sets, each of which contained three

articles from three widespread 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

3.2 Preprocessing

Natural language processing always involves text preprocessing. Moreover, the primary algorithm used in the processing phase's outcomes is greatly influenced by the precision of the preprocessing stage. The purpose of preprocessing is to standardize the source text and transform it to the format required for later processing. In the first step of input collection, the text of each document is collected. Preprocessing includes sentence categorization, tokenization, stop word elimination, and stemming. Before combining the texts from the collection of documents, the final document should be separated into its fundamental sentences. This is carried out using the Python Natural Language Toolkit module during the sentence segmentation stage. Tokenization divides each sentence into discrete pieces. Stop words are often the most commonly used terms in a language, and thus are ineffective for selecting relevant sentences in an input document. To weed out stop words, our system employs an 85-word stop word list. The removal of stop words enhances the cosine similarity score and makes vectorization of the sentence easier. Stemming is an important stage in input text preparation since it lowers a word's inductive and occasionally derivational forms to a single base form. The accuracy of the objective function is positively impacted by this task. The stemmer employed in our suggested system is comparable to Indic stemmer [37], which manages multiple levels of inflection using an iterative suffix stripping algorithm. For instance, the words vanathil and vanathiloode in Malayalam are changed to vanam where it consider it as root during the stemming process. As a result, during the vector generation stage of the processing phase, various word forms that share the same root are handled identically. Stemming is therefore essential to raising the suggested method's performance. To create the corpus's dictionary (vocabulary), we use the gensim Python library.

3.3 Vector generation

Vectorizing the input topic words and contained text is the second step. In the input text, every sentence is represented by a vector. Binary and TF-IDF vector representations are the most commonly utilized ones [38]. We then create the topic vector through topic modeling. To comprehend how topics are distributed and represented in a text, apply the LDA methodology. The user chooses W and K, and the top W words from K subjects that are selected based on the probability assigned to each phrase. LDA uses the following methods to identify topic representation:

(i) Firstly, identify the number of subjects in the documents (let it be N).

(ii) The LDA will then arbitrarily assign each word in each sentence to one of the N subjects. Because words are randomly assigned to themes, the resulting outcome is neither optimum nor accurate.

(iii) In order to improve representation, the LDA estimates the ratio of phrases in a text allotted to a specific topic.

$_P(T|S)$ = Percentage of words attributed to topic in sentence S

$_P(W|T)$ =Percentage of words presently allocated to a subject in sentence S

(4) Find the product of $_P(T|S)$ and $_P(W|T)$ to compute the controlled possibility that the term will appear on respective topic. Replace the word with the topic that has the highest controlled possibility.

(5) The preceding steps are repeated for each individual word in every phrase of the text till the confluence is achieved.

K topics are produced by the LDA, each of which is a keyword combination that gives the topic a specific weight.

The process of generating LDA topic vectors can be explained with an example. Assume, as indicated in Figure 2, that we have two documents to work with when creating the summary. The appointment of a Chief of Defense Staff (CDS) to enhance defense force coordination is covered in these two documents. Assume the provided text must be used to generate three topics. The topic along with associated terms are produced by applying LDA modelling to the input text, as shown in Figure 2. It indicates that the first item is about a defensive update, the second on the need for CDS, and the final one regarding language relevant to CDS creation.

3.4 Relevance status value (RSV)

The next step is to determine the sentence vector's relevance to the topic vector. Sentence-topic vector similarity can be used to approximate how relevant a sentence is to a given topic. It is common practice to use the Euclidean, Jaccard, and Cosine measures of similarity. To obtain the RSV, calculate the cosine of the link between each topic and sentence vectors pairs, as indicated in the equation (1)

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

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

3.5 Sentence ranking

After ranking the sentences in a decreasing sequence of RSVs, they are routed to the summary generation process. The overview shows the sentences that closely resemble the K subject vectors.

3.6 Summary generation

Our methodology concludes with the creation of summaries. It involves eliminating unnecessary phrases from the sentences with the highest scores. When using multi-document analysis, the variety of documents that need to be summarized can be fairly considerable. As a result, compared to single-document summarizing, multi-document distillation contains more redundant information. It's crucial to keep redundancy under control. The maximum marginal relevance (MMR) along with clustering are most often used ways to preventing redundancy in summarization. The textual overlap between the phrases in the final summary text and the phrases to be included in the output overview determines redundancy in MMR. The first sentence from the ranking list will be copied and pasted into the article's body to start.

	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 ಸೃಷ್ಟಿಯಾಗಿದೆ.	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.

4. Performance evaluation metrics.

4.1. Evaluation metric.

The evaluation method used in this experiment was the Recall--Oriented Understudy for Gisting Evaluation (ROUGE) [34], specifically ROUGE-N (ROUGE-1 **unigrams** and ROUGE-2 **bigrams**), ROUGE-L **longest matching sequence**, and ROUGE-SU4 Skip-bigram plus unigram-based co-occurrence statistics. ROUGE is a state-of-the-art fully automated method for evaluating text summaries. ROUGE-N assesses the degree of similarity between the related reference summary of documents and the system overview using n-gram comparison and overlap. 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 subsequence (LCS) method, which takes sentence-level structure comparison into consideration. $\frac{LCS(S,R)}{n}$ Let S be the system summary and R denote the n-word reference of summary. The following formula is used to calculate ROUGE-L:

ROUGE-L =

4.2 Experiments and results

All of the experiments were carried out on a computer equipped with an Intel Core i5-8250 CPU. operating at 1.80 GHz and 16 GB of RAM, using Python.

Kannada documents extraction and summarization using LDA 401

Experiments have been conducted on the dataset considered especially for summarization in order to acquire a comprehensive assessment of the proposed LDA-based MDS system. The proposed model was run with compression ratios (CRs) ranging from 10% to 40%. In addition, the proposed model's performance was tested with three, five, and nine topics instead of the original nine. Table 2 displays the F-measure, recall, and precision outcomes of the suggested MDS for 10% CR.

The table shows that compressed content covering fewer subjects has a good ROUGE score, which decreases with increasing topic coverage. The

ROUGE--1 and ROUGE--2 ratings for different CRs with topic values ranging from 3 to 9 are shown in the Table-3.

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

For contrasting the Textrank-based MDS alongside the LDA-based MDS, a test method was developed. The ROUGE-1 parameters of the models with varying CRs are assessed in Table 4 prior to the redundancy being removed. This shows that putting sentences into the subject space gives a more accurate representation of the input Kannada records along with more pertinent information. Based on a evaluation of the ROUGE-1 outcomes from the 10% compression ratios and the 40% compression ratios, Table-4 demonstrates that the F-score decreases as the compression ratios decreases because there is a decrease in the co-occurrence of the system overview and the reference summary in the documents.

The elimination of redundant text conents improves the quality of text summaries. The use of a redundancy elimination component significantly increases the final summary's accuracy. In order to improve information richness and eliminate redundancy, the MMR algorithm was applied. A summary with more variety and less repetition is obtained by applying MMR to the rankings generated by the LDA model. A comparison of Tables-4 and Table- 5 makes it clealy shows that removing redundant information improves the

quality of the final summary of the multo documents.

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 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**.**

Two news items are chosen at random from a collection of documents to demonstrate phrase ranking with LDA and our algorithm's extractive an overview.

Using the DUC-2018 dataset from the NIST, the recommended solution for the English multi-document summarizing problem was assessed. Table 6 hows the ROUGE-1 unigrams result for DUC2016. The results show that the suggested approach performs brilliantly when used with the English language.

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 efficiency of the proposed strategy was evaluated by comparing it to a few earlier studies on summaries in Indian languages. The authors of reference [42] experimented with 100 Hindi news articles.

Tamil, Marathi, and Punjabi are translated using the four Indian language methodologies. These are the Hindi graph-based strategy, the Marathi textrank-based approach, the Punjabi combination model, and the semantic Tamil graph-based technique. Table 7 shows that, even though the previous studies were limited to single papers, our proposed method performs better than earlier research in multiple langages. The table shows that our proposed model

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

outperformed every study currently conducted on Indian languages.

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

5. Conclusion

This paper aimed at providing topic that will demonstrate a technique for automatic extraction multi-document summarization of Kannada multi-documents. To mine dictating topic expressions from texts, LDA-based topic models are used. The proposed techniques provide value by giving a general mechanism for substituting the provided documents with the subject vector and measure that determines sentence relevance to topic in a reduced size dimension. Additionally, we try to remove redundancy component was employed to add to the final summary's content.

This study would not have been possible without the employment of specific MDS information set in Kannada, which comprises of organizing new terms and summarizing news events. Later, trials on English and Kannada MDS datasets, were conducted several masseurs separately to ensure the efficacy of the proposed model's generated summary. The technique yielded more positive and motivating findings than the baseline model, according to the testing data. This study conducted a comparison analysis with summary efforts carried out in Indian Languages, indicating that LDA provides more ROUGE value than other content summaries approaches utilized in other languages of India.

The proposed architecture model produced a non-redundant and appreciate relevant summary from multiple sources, but it was not as cohesive as a single summary. The users, on the other hand, might typically understand the summary. Although the model was developed and tested in Kannada, it is adaptable to any other language. It is proposed that in the future, topic modeling methodology be integrated into other extractive text summarization procedures, such as graph-based ways and algorithmic evolution.

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.

[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