Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7688572

DES: DOMAIN EXPERT SUMMARIZATION USING LDA

B. Lavanya
Associate Professor, Department of Computer Science, University of Madras, Chennai, India.
lavanmu@gmail.com

U.Vageeswari
Research Scholar, Department of Computer Science, University of Madras, Chennai, India.
uvageeswariphd@gmail.com

Abstract—The text data is unstructured. The amount of textual data available is excessive and continues to increase daily. The technique of shortening long documents into brief paragraphs or phrases is known as text summarization. The method ensures that the meaning of the paragraph is constant while also extracting crucial information. The two main goals of a text summarization are optimal topic inclusion and excellent readability. Extractive summarization methods emphasize identifying significant sentences from the document. The identification of important sentences is based on the sentence score. Most of the extractive summarization methods fall under one of the following categories. Graph-based methods, TF-IDF-based methods, Fuzzy Logic based methods, and Machine Learning methods. The key issue with all of these methods is that they only examine local knowledge or data that can only be found in a particular file. Document domain knowledge is not taken into consideration. The Domain Expert Summarization (DES) method is developed in this research to summarize a document like an SME, and the effectiveness of the DES method is assessed in comparison to other state-of-the-art works. To make a machine dexterously Domain Knowledge must be obtained. So that key points or keywords of the domain can be easily identified and a summary can be produced with all key points. The LDA topic modelling method is used to obtain domain knowledge. The experiment makes use of the BBC NEWS and NEWS Aggregator data set. Evaluation is done using ROUGE-1, ROUGE-2, and ROUGE- L measures. The experiment showed that, in terms of the ROUGE Score, the suggested DES method outperforms the current state-of-the-art methods. A statistical t-test is performed at a 5% significance level. The p values for ROUGE 1, ROUGE 2, and ROUGE L on the BBC NEWS dataset are 0.002174, 0.002174, and 0.001905. The p values for ROUGE 1, ROUGE 2, and ROUGE L on the NEWS Aggregator dataset are 0.032835, 0.014387, and 0.025124, respectively. It is clearly evident from the P values that the suggested DES Method is statistically significant.

Keywords —Text Summarization, LHUN, Lex Rank, Tex Rank, LSA, LDA

INTRODUCTION

The amount of textual data available is excessive and continues to increase daily. Take, for example, the internet, which is comprised of a wide assortment of content including blogs, websites, news articles and status updates. The text data is not structured. The key points of a lot of this text's content need to be emphasised in shorter, more focused summaries so that they

can be more thoroughly explored. The technique of shortening long documents into brief paragraphs or phrases is known as text summarization.

The method ensures that the meaning of the paragraph is constant while also extracting crucial information. The two main goals of a text summarization are optimal topic inclusion and excellent readability. The process of extractive summarization places a strong emphasis on locating significant sentences within the document. The sentence score provides the basis for determining which sentences are significant. Several approaches are developed in order to arrive at the sentence score. Most of the extractive summarization methods fall under one of the following categories. TF-IDF-based methods, Graph-based methods, Machine Learning methods, and Fuzzy Logic-based methods.

The sentences themselves are the "nodes" in a graph-based method, and the "edges" are the resemblance scores between each pair of sentences. Once this graph has been constructed, the well-known page rank algorithm is adopted to investigate the most important sentences are the most important. The most popular Lex-Rank and Tex-Rank extractive summarization methods are graph-based methods. In a TF-IDF-based method, TFIDF scores of each word are calculated. The TF-IDF score of each word in a sentence is used to determine the sentence's overall score. Following this step, the sentences are graded according to their score, and the top - k sentences are chosen for the summary. The popular LHUN method is the TF-IDF-based method.

In a method that uses fuzzy logic to score sentences, the location of the sentence, the quantity of nouns and verbs, and the quantity of keywords are all taken into account. Extractive Summarization using Machine learning methods needs a reference summary to train the model. Naive Bayes, Decision Tree, and SVM are some of the machine learning methods used in summarization.

In single-document summarising, only one source document is used to generate a summary. A multi-document summary is a condensed version of multiple papers that served as input.

The main issue with all these methods is that they only consider local information, which is the knowledge that can only be derived from a specific document. Therefore, document domain knowledge is not considered. This study takes domain knowledge into account. The LDA method is used to obtain domain knowledge.

Methods for abstractive summarization need a trained model for framing sentences. Except for NEWS, scientific publications, legal documents, and tweet data, trained models are not available for other domains. The primary objective of this research is to develop a framework for producing expert-level summaries of documents and to compare that framework's results to those of existing, cutting-edge solutions. A subject matter expert (SME) is an individual who has a general comprehension of a particular subject. The vital stage to turning into an SME is to acquire information regarding the specific domain. To make a machine like an SME, we must train a machine with more documents of the domain. So that key points or keywords of the domain can be easily identified and a summary can be produced with all key points.

The following are the primary goals of this research work.

Examine the various summarization methods

Propose Domain Expert Summarization (DES) method for document summarization as done by subject matter expert.

Evaluate the proposed method to cutting-edge techniques and demonstrate its effectiveness.

The rest of this paper is structured as follows: A comparison of several summarization techniques is presented in Section 2. The various techniques used in DES implementations are explained in Section 3. Section 4 provides a description of the suggested DES technique. The performance of the suggested approach and other cutting-edge methods is discussed in Section 5, which also describes the dataset utilised and tabulates the findings. The paper is concluded in Section 6.

RELATED WORK

Depending on the sentences that are chosen or the new sentences that are framed, summarization techniques are either extractive or abstractive. Sometimes more than one document is used to create a summary; other times, only one document is used. Table I shows recent summarization works. It displays the several summarising approaches employed, along with the datasets and methodologies used.

2.1. Single Document Summarization

The input for the summarization is one single document. After that, the texts are pre-processed to apply an algorithm or generate a sentence score. After pre-processing, a graph-based or sentence-scoring approach is used.

SciBERTSUM [1] makes BERTSUM work on long documents by incorporating a section embedding layer that adds section details to the sentence vector and creates section details to the sentence vector and puts in a sparse attention mechanism. According to this approach, each sentence will focus on the sentences immediately around it, while just a few sentences will analyze the entire language. The PS5K dataset is utilised for analysis.

A novel graph-based summarization method that [2], in addition to considering sentence similarity, also considers sentence similarity to the entire (input) document Two attributes are taken into account when distributing the weight among the graph's edges. In topic modelling, the similarities between nodes are one property, and the weight of an edge in relation to the document's subjects is another. Learning Free Integer Programming Summarizer (LFIPSUM) [3] is a methodology for extractive summarization that does not rely on human supervision. As the model does not require labelled data for training, this approach has the added benefit of removing the demand for parameter training. An integer programming issue is defined using trained sentence embedding vectors Principal component analysis can automatically decide how many sentences should be retrieved and how important each sentence is.

Domain Feature Miner [4], Three recently developed empirical observations were used to define the feature mining problem as a clustering problem: grouping semantics, frequency count, and distributional statistics of features. Asymmetric cluster extraction (ACE) and Symmetric cluster extraction (SCE) techniques are created to extract domain information from clusters.

This model employs semantic measure and topic modelling in a vector space framework [5]. The sentences in the provided document are expressed in an alternate dimension to generate the subject vector using a topic modelling and vector space model. Using a metric based on semantic similarity, we can evaluate the importance of the statement. The topic vector can be extracted from the provided document using either the Individual Topic Vector Model or the Combined Topic Vector Model.

Three techniques for extracting a single document's summary using supervised and unsupervised learning are suggested [6]. The statistical characteristics of sentences and their

relationship to one another are combined to determine a sentence's importance. To score sentences separately, the first method employs supervised models and graph models. The graph model is employed in the second technique. The third technique scores sentences using a biased graph model.

Both semantic and syntactic characteristics were used to assess the sentences' significance in the Candidate sentence selection model [7]. LSTM-NN was put out in this work as a method of summarization that manages the combination of syntactic and semantic information.

COSUM [8] presents a model for the identification of sentences that is comprised of two stages and is based on optimization and clustering strategies. In order to identify the themes included inside a document, it applies the k-means algorithm to the task of categorising the phrases into subject groups. The perfect summary was constructed with the help of an optimization algorithm that selected the most important sentences from each of the groups. In order to solve the optimization problem, a modified version of the differential evolution (DE) algorithm is developed.

An approach that is based on three parameters, including the rate of redundancy, the diversity rate, and the compression rate, is described [9] in order to produce diverse, least redundant, semantically feature-rich, and compressed summaries. The redundancy and variety of each sentence are determined by putting primary emphasis on minimising the amount of repetition and achieving the greatest possible level of overall variation in the summary.

The" EdgeSumm" framework [10], which is built on four proposed algorithms, is intended to improve ATS for single documents. The first method starts with the input content and creates a brand new representation of the text graph model from scratch. The second and third algorithms conduct a search for sentences that can be included in the candidate summary within the text graph that was just constructed. Once that final prospective summary again goes beyond the limit that the user requires, the fourth method is used to select the most important sentences to include in the summary. EdgeSumm combines a number of different extractive ATS techniques, such as statistical-based, graph-based, centrality-based techniques and semantic-based, in order to make use of each technique's individual benefits while minimising the negative effects of any drawbacks it may have.

An innovative review-to-text summary [11] is offered as a means of automatically extracting text summaries from reviews of various electronic devices. When determining the significance of a sentence, Both the review's content and the author's authority are considered. Both the content and the semantic similarity of each and every pair of phrases in a review are compared. Fuzzy c-means clustering is utilised to create the reviews' summary.

Making summaries of an organizations or brand's online reputation is a focused summarizing assignment with a unique feature: problems that could harm the entity's reputation are given precedence in the summary [12]. The banking and automotive industries' 31 organizations' tweet streams are compiled in a new test collection of manually compiled reputation reports. In the context of monitoring online reputation, a fresh methodology for evaluating summaries is proposed.

SUMMCODER [13] proposes a system for unsupervised text summarization involving deep neural networks. Recurrent neural networks employ sentence vector representations. Summary created utilising the three sentence qualities of position, novelty, and relevance. For

determining the relevance of a sentence's content, deep auto-encoders are used. A brand-new text summary dataset derived from darknet domains is presented.

This article [14] proposes an extractive linguistic information topic modelling text summarising approach for Hindi novels and short stories. There are four independent variations that are executed utilizing diverse sentence weighting algorithms. As there was no archive, a group of novels and short stories in Hindi were pulled together. For informative and diverse summaries, a smoothing technique is used, and then the effectiveness of the resulting summaries is assessed using three criteria gist diversity, retention ratio, and ROUGE score.

TABLE I: Recent Extractive summarization works

S.No	Paper ID	Year	Datasets Used	Single or Multi	Techniques
1	[1]	2019	CNN-DailyMail, PS5K	Single	BERT, Embedding info, Sparse Matrix
2	[2]	2020	Opinosis, CNN/ Daily Mail	Single	Graph with LDA
3	[3]	2021	CNN/Daily Mail, Wikihow, Cornell newsroom, Korean	Single	Principal component analysis (PCA), Integer Linear Programming (ILP)
4	[4]	2021	Amazon and TripAdvisor	Single	Review & Inverse Review-Feature mapping , Symmetric and Asymmet-ric Clustering
5	[5]	2021	CNN/ Daily Mail, Opinosis	Single	Topic modeling, Combined and Individual topic vector
6	[6]	2019	DUC2001 and 2002	Single	LEX Rank, statistical features
7	[7]	2021	125 research papers,	Single	GloVe and word2vec embeddings, semantic and syntactic features
8	[8]	2021	DUC2001 and 2002	Single	Clustering and optimization
9	[9]	2022	Novels in Hindi, DUC 2007 and news articles in English	Single	Gist diversity , Redundancy ratio
10	[10]	2020	DUC2001 and 2002	Single	graph-based, statistical-based, semanticbased, and centrality-based
11	[17]	2021	DUC 2001, 2002, 2006, and 2007	Multi	LDA, Classification, Silhouette
12	[22]	2020	DUC-2002 and 2004	Multi	Graph independent set, Textual graph
13	[19]	2020	DUC 2005 and 2006	Multi	Expansion of query, correct sense of a word
14	[23]	2021	BBCNEWS,DUC2002, 2006, and 2007	Multi	softmax regression, Spider Monkey Optimization
15	[11]	2019	Amazon dataset	Single	fuzzy c-means clustering
16	[12]	2019	New TWEET	Single	Fuzzy

17	[20]	2018	DUC 2002 and 2001	Multi	Contextual polarity, Sentiment dictionary
18	[21]	2021	DUC 2005, 2006, and 2007	Multi	Maximalmarginalrelevance, Greedy Search
19	[13]	2019	DUC 2002, Blog	Single	Deep auto-encoders, RNN
20	[14]	2020	Hindi Novels	Single	Tagged-LDA
21	[15]	2019	EASC Corpus	Single	PageRank Morphological analyzer Graph
22	[16]	2022	CNN/DailyMail, DUC 2002	Single	Siamese networks,graph-based

In [15] suggests a method that is based on graphs to accomplish the task of extractive Arabic text summarization. The sentences serve as the graph's vertices in this approach of representing the document. In the customised PageRank algorithm, the first ranking for every node is the amount of nouns within that sentence. The initial rank of a sentence is based on how many nouns it has since more nouns mean more information in the sentence. The cosine similarity among sentences is used to find the edges between sentences so that the last summary includes sentences with more details and that flow well together. Using the Modified PageRank method, different iterations were done to find the best number that gives the best summary results.
An approach to extractive text summarization of individual documents, Ranksum [16] is predicated on the rank fusion of four multi-dimensional sentence characteristics collected for each phrase: significant keywords, subject information, semantic content, and position. The fusion weights are learned using a tagged document collection, whereas the scores are created completely unsupervised. To determine the relative importance of various topics, probabilistic topic models are employed, while semantic information is gathered by means of phrase embeddings. Using a graph-based method, we may find the most relevant keywords and sentence ranks in the document. Each sentence in the document is ranked for each feature and then the aggregate scores are used to produce the final score for each sentence.

Multi-Document Summarization

In this method, Multiple documents serve as the summarization's input. The texts are then pre-processed in order to apply an algorithm or provide a sentence score. After pre-processing, a graph-based or sentence-scoring approach is used. The result is a single summary that incorporates information from each document.

In [17], a novel strategy for synthesising a corpus of articles into a unified summary by combining topic modelling and classification approaches is introduced. While analysing a collection of documents, it is possible to count the exact number of recurring themes by employing a technique that takes into account the randomness of latent Dirichlet allocation. By reducing the amount of sentences in a large body of papers without leaving out any relevant details, we are able to provide a concise and all-encompassing summary from the original material.
In Multiview Convolutional Neural Network [18], Multiview learning is applied to CNN in order to improve its capacity for learning. Using the consensus and the two-tiered

complimentary ideas to their maximum potential boosts the model's learning capacities. Pre-trained word embedding eliminates the requirement for human feature engineering. The proposed model employs sentence location embedding to boost its receptiveness to new information.

In Improved Query-based text summarization [19], a model is proposed based on word sense disambiguation and common sense. By extending the query words, common sense knowledge is merged. Query-based text summarization utilizes a semantic similarity measure between the input text content and the question to choose which sentences to extract. A semantic network is made up of a sizable collection of concepts, each of which is graphically connected to other concepts. When collocation and semantic relatedness are used in a linear equation, a measure of word similarity is also offered. By observing for common terms between two phrases, redundancy is checked.

Automated emergency sentiment-oriented summarising of multiple documents via soft computing (ASMUS) [20] takes into account content coverage and redundancy. Essentially, it combines the steps of sentiment classification with sentiment description. During the phase of sentiment classification, many methods are used to overcome the following restrictions: In order to enhance the ranking conclusion for sentences, the sentiment information is included in a graph-based ranking system that also takes into account statistical and linguistic techniques, contextual polarity, the word scope limit of a specific glossary, and sentence types.

The efficiency of various QF-MDS algorithms is influenced by the essential cornerstone of sentence and query representation. Unmonitored query-focused multi-document summarization [21] proposes a system that employs embedding vectors to depict sentences in manuscripts and users' questions by utilising transfer learning from pre-trained sentence embedding models. BM25 and the similarity measure function are linearly combined to identify sentences that are associated with the question. The selected phrases are then reordered using either the maximum possible marginal relevance criterion or which preserves query relevance while limiting redundancy. For text summarising, the proposed scheme is unmonitored and does not necessitate any labelled training data.

Figure (1) illustrates the three different summarising methods. Single-document summarization, multi-document summarization, and proposed pattern-identified document summarization are the first, second, and third respectively. In the single document summarising approach, the output is the document’s summary and the input is just one document. The multi-document summarizing approach produces a single summary for all the input documents, all of which are on the same theme. With more than one document of the same subject as input, the proposed pattern identified document summarization first finds the pattern in the document collection and then uses the pattern to provide a summary of each document. Figure (1) shows the input, the summary procedure, and the result for three different summarization methodologies.

METHODS

3.1. LHUN

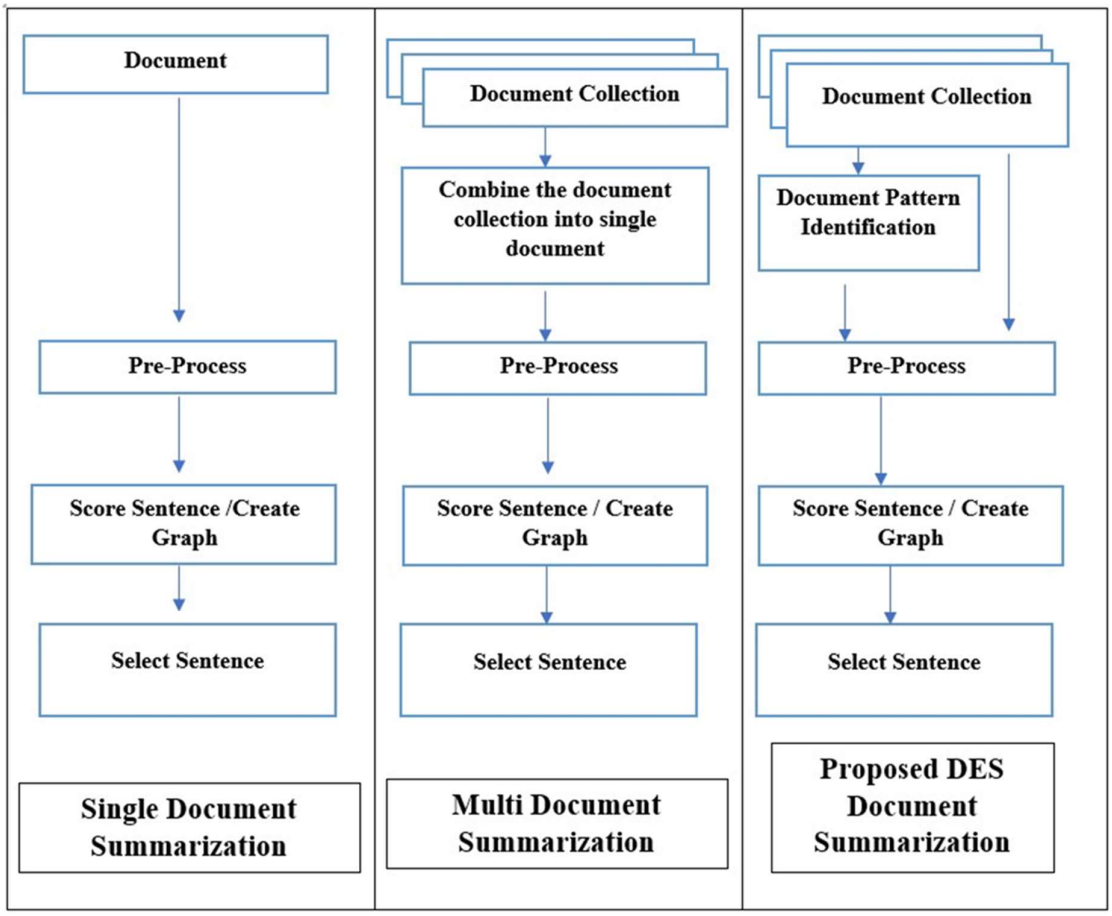


Fig. 1: Summarization Methods

Hans Peter Luhn, an IBM researcher, presented one of the earliest text summary algorithms. Luhn's method is a simple technique based on TF-IDF that analyses the" window size" of non-essential words between keywords terms. It also gives sentences that occur at the initial stage of a text more weight. The TF-IDF approach is employed in the Luhn Summarization algorithm (Term Frequency-Inverse Document Frequency). It is useful when both extremely low-frequency and very common words (stopwords) are absent. Sentence scoring is done in accordance with this, and the top-scoring sentences are included in the summary. Setting a minimum frequency threshold is a straightforward technique to eliminate low-frequency phrases. Using a maximum frequency criterion (statistically determined) in contrast to a common-word list is a smart idea.

$$Score = \frac{Number\ of\ Important\ words 2\ score}{Number\ of\ Unimportant\ words}$$

LDA

Latent Dirichlet Allocation is an algorithm that gives each document a strategy that focuses on what it is about. LDA considers each text as a collective of topics and every topic as a group of words. The LDA model needs a text that has been turned into a vector.

LDA is an unsupervised topic modelling algorithm. As a statistical approach, it depicts the conditional probabilities P (X, Y). The purpose of LDA is to discover topics a document

belongs to, based on the words in it. The very first and most important parameter to LDA is the corpus or document term matrix. The other two important parameters that must be set for LDA are max-iterations and the number of topics that must be identified using the LDA algorithm.

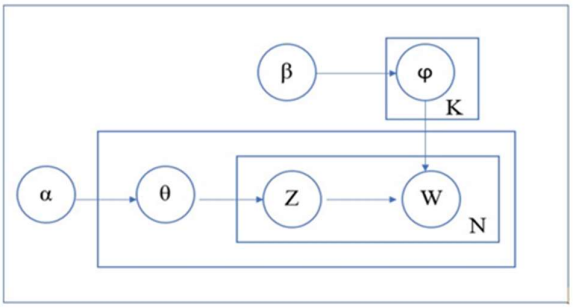


Fig. 2: LDA - Latent Dirichlet Allocation

Figure 2 shows the plate notation of the LDA Algorithm. In this diagram

M is the Total count of documents in the corpus.
N_i is the Total count of words in a particular document (Ni -Count of words in the ith document).
α is the hidden parameter. It is a density value for a document on a topic.
β is the hidden parameter. It is a density value for a word on a topic.
θ_i is the topic distribution value for the document i
φ_k is the word distribution value for topic k
z_{ij} is the topic for the j-th word in document i
W_{ij} is the specific word

Only W is directly measurable; all other variables are concealed latent factors. Long-term topic modelling with LDA is an example of an unsupervised algorithm. The conditional probability function P is a statistical approach (X, Y). The purpose of LDA is to discover topic a document belongs to, based on the words in it. The very first and most important parameter to LDA is the corpus or document term matrix. The other two important parameters that must be set for LDA are max-iterations and the number of topics that must be identified using the LDA algorithm. LDA algorithm has two hyperparameters. They are a and b. α - is a density value for a document on a topic β - is a density value for a word on a topic

If the document has a greater number of topics, then α must be assigned a higher value, else if the document has a smaller number of topics a must be assigned a lower value. If a topic has a greater number of words, then β must be assigned a higher value, else if the topic has a smaller number of words, then b must be assigned a lower value.

PROPOSED METHOD

In this Proposed Domain Expert Summarization (DES) method, a collection of documents is used, and each document covers a specific topic. The first step is to determine the document's topic. The main challenge in this technique is determining the topic of each document. The LDA topic modelling approach has been applied for topic identification and keyword selection. Both the document set and the total amount of themes in the set are fed into the LDA algorithm. The LDA algorithm lists the topic keywords for each article as well as the topic of the document. The document is summarised using this pattern of topic keywords.

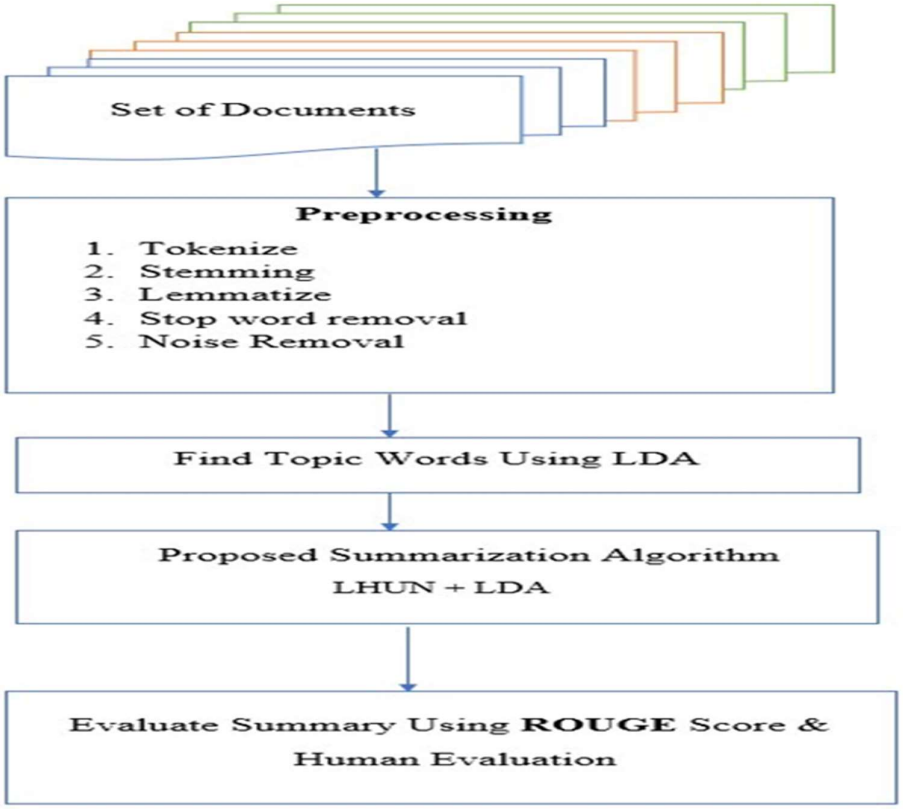


Fig. 3: Domain Expert Summarization

Algorithm 1 shows the proposed domain expert summarization (DES). The algorithm includes three methods. The main approach is the summarize method. This method takes three inputs: documents (D), domain words (DW), and summary count (SC). This method calls the Top Words Method first in order to choose the summary sentences. The Top Words Method uses the TF-IDF value to identify the top words that are specific to that document. When applied to a document, the term frequency-inverse document frequency (TF-IDF) provides a numerical value for each term. It examines the texts into sentences so that the terms can be quantified. And then sentences are split into words. The calculate score method is then used by the summarise method; it assigns scores to each sentence in a document. Top word and domain terms are considered while scoring the sentences. The LDA algorithm determines words as being domain words. The top N subject words also referred to as the domain words, are presented for each topic. After each sentence has been scored, a summary with a predetermined amount of summary counts is generated. Table II explains the symbols used in the algorithm.

Algorithm 1: Domain Expert Summarization (DES)

Input: Collection of text documents

```
Output: Document Summary
Function Top_Words(D)
    Lines = split(D)
    CW = stopword()
    Record = []
    for L in Lines do
        Words[] = splitWords(L)
        for W in Words do
            if W notin CW then
                if W in Record then
                    Record[W] += 1
                else
                    Record[W] = 1
    return TopWords
Function Calculate_Score(s, IW)
    Words = splitWords(s)
    ImportantWords = 0
    UnImportantWord = 0
    TotalWords = len(Words)
    for W in Words do
        if W in IW then
            IW += 1
    UnImportantWord = TotalWords − ImportantWords if UnImportantWord != 0 then
        SentenceScore = ImportantWords**2/UnImportantWord
    else
        SentenceScore = 1
    return SentenceScore
Function Summarized(D, DW, SC)
    Summary = []
    S = splitWords(D)
    IW = CALLTopWords(D) + DW
    Score = for S in S do
        Score[] = CALLCalculateScore(s, IW)
    SortedSentences = sort(Score)
    Summary = SortedSentences[SC]
    return
```

TABLE II: Symbols used in Algorithm

Symbol	Meaning
D	Document
DW	Domain Words
IW	Important words
S	Sentences
W	Words
CW	Common Words
SC	Summary sentence Count

EXPERIMENT

5.1. Evaluation Metric

The effectiveness of an automatic summarising strategy is evaluated using a variety of metrics, the most important of which are ROUGE scores and variants. ROUGE stands for "Recall-Oriented Understudy for Gisting Evaluation." Machine translation and automatic text summarization are analysed using this set of metrics. In order to function, it compares an

algorithmically created summary or translation to a database of summaries used as references (typically human-produced). ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S are the acronyms for the four different types of ROUGE.

ROUGE-N: It affects how much the summaries created by Machines and Humans coincide in terms of n-grams. A unigram is ROUGE-1. A bigram is ROUGE-2. ROUGE3 weighs three grammes, while ROUGE-4 weighs four grammes. The reference summary of the denominator increases as the n-gram number increases. As a result, these measures incorporate several sources. ROUGE-N gives more weight to the candidate summaries that have more common words with the reference summaries.

ROUGE-L: In order to compare the efficiency of the system summary to that of human summaries, the length of the Longest Common Subsequence (LCS) is used. When the LCS between the candidate and the reference summaries is large, the two summaries are quite close. The measure takes into account the restrictions imposed by the ROUGE-N metric, in particular the fact that ROUGE-L takes into account the LCS between the two text portions while **ROUGE-N** estimates similarity based on shorter text sequences. This metric is an improvement over ROUGE-N, but it still suffers from the same flaw—it requires continuous n-grams.

Precision, Recall and F-measure: Precision: When comparing a candidate summary to a reference summary, ROUGE Precision is the percentage of terms that appear in both.

P=(Number of overlapping words in summaries)/(Total Words on candidate summary)

$$P = \frac{Number\ of\ overlapping\ words\ in\ summaries}{Total\ Words\ on\ candidate\ summary} \quad (1)$$

Recall: ROUGE Recall is a ratio between the number of overlapping words in the candidate and reference summary to the total words in the reference summary.

R=(Number of overlapping words in summaries)/(Total Words in reference summary)

$$R = \frac{Number\ of\ overlapping\ words\ in\ summaries}{Total\ Words\ in\ reference\ summary} \quad (2)$$

F-measure: The F measure offers all of the data that recall and precision do independently.

$$F1 - Score = \frac{2*(Precision*Recall)}{(Precision+R\quad)} \quad (3)$$

5.2 Experiment on BBC NEWS Dataset

The BBC NEWS Dataset has been used for analysis. The data was gathered through Kaggle. Five main topics are covered by the news stories in the BBC NEWS dataset.

NEWS from 2004–2005 include business, technology, politics, entertainment, and sports. Since there are nearly equal amounts of papers on each topic, this dataset is balanced.

Figure 4 shows the document distribution in the BBC NEWS dataset. It is evident from the graph that the dataset is balanced. because there are roughly equal numbers of papers in each category. A summary of the BBC NEWS dataset's document distribution by category is provided in table (III).

Table (IV) lists the top forty domain words for each category in BBC NEWS dataset. LDA algorithm is used to identify these words. In the DES algorithm, these terms are employed as domain-specific words.

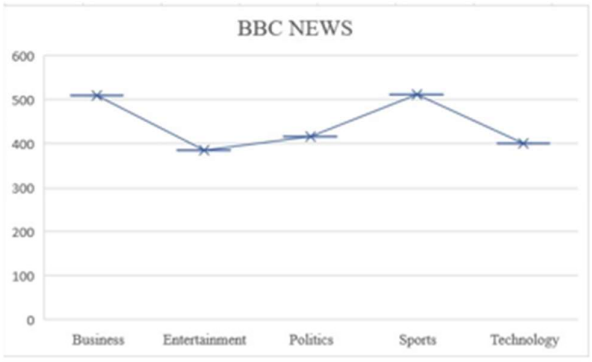


Fig. 4: Document distribution in BBC NEWS

TABLE III: Number of documents in BBC NEWS

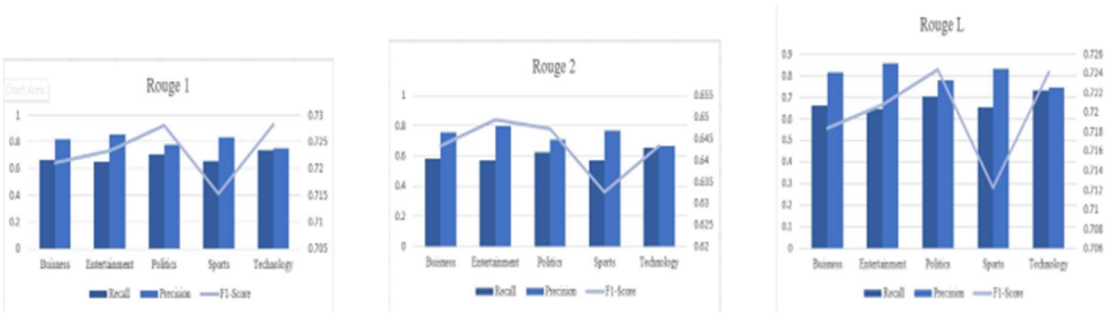
Category	Number of Documents
Business	510
Entertain	386
Politics	417
Sport	511
Technology	401

TABLE IV: Domain Words in each topic in BBC NEWS dataset

Domain	Top Words
Business	'company', 'firm', 'people', 'case', 'rule', 'bank', 'tell', 'group', 'right', 'unit', 'deal', 'club', 'legal', 'state', 'director', 'terror', 'home', 'trial', 'share', 'offer', 'support', 'come', 'work', 'hold', 'sale', 'time', 'foreign', 'judge', 'allow', 'claim', 'go', 'include', 'take', 'share hold', 'finance', 'human', 'plan', 'suspect', 'glazer', 'protect'
Sport	'play', 'time', 'best', 'player', 'come', 'go', 'music', 'half', 'match', 'think', 'wale', 'award', 'good', 'team', 'world', 'take', 'nation', 'minute', 'like', 'final', 'coach', 'injury', 'goal', 'second', 'great', 'franc', 'home', 'add', 'perform', 'start', 'leave', 'tell', 'break', 'right', 'open', 'people', 'look', 'know', 'victory', 'want', 'give', 'include', 'win', 'week', 'score', 'lose'
Politics	'government', 'people', 'elect', 'labour', 'party', 'minister', 'blair', 'plan', 'tori', 'tell', 'public', 'brown', 'work', 'service', 'campaign', 'lord', 'report', 'issue', 'time', 'claim', 'country', 'need', 'leader', 'come', 'want', 'secretary', 'polit', 'vote', 'prime', 'chancellor', 'britain', 'change', 'right', 'like', 'go', 'think', 'nation', 'support', 'call', 'council', 'propose', 'home', 'add', 'general', 'conservative', 'week', 'spokesman', 'increase', 'allow'
Entertain	'film', 'people', 'music', 'best', 'star', 'time', 'include', 'play', 'award', 'release', 'number', 'like', 'take', 'work', 'technology', 'million', 'director', 'video', 'digit', 'microsoft', 'go', 'record', 'movie', 'actor', 'online', 'sale', 'download', 'world', 'sell', 'industry', 'oscar', 'program', 'high', 'service', 'mobile', 'want', 'chart', 'company', 'product', 'week', 'search', 'screen', 'nominee', 'develop', 'sony', 'software', 'user', 'come', 'role', 'media'
Technology	'company', 'firm', 'phone', 'mobile', 'people', 'like', 'technology', 'market', 'secure', 'deal', 'user', 'service', 'time', 'bank', 'come', 'world', 'number', 'govern', 'work', 'want', 'china', 'group', 'state', 'think', 'product', 'operate', 'virus',

	'take', 'sale', 'player', 'sell', 'report', 'unit', 'month', 'look', 'open', 'need', 'help', 'gadget', 'develop', 'country', 'know', 'tell', 'apple', 'manage', 'consume', 'offer', 'business'

Figure (5) shows the Rouge Scores on BBC NEWS dataset on each topic wise.Table (V) shows the ROUGE 1 Score of each topic in BBC NEWS dataset. Table (VI) shows the ROUGE 2 score of each topic in BBC NEWS dataset. Table (VII) shows the ROUGE L Score of each topic in BBC NEWS dataset.



(a) ROUGE1 (b) ROUGE2 (c) ROUGEL

Fig. 5: Topic-wise ROUGE Score in BBC NEWS dataset

(a) ROUGE1 (b) ROUGE2 (c) ROUGEL

TABLE V: Topic Wise ROUGE 1 score on BBC NEWS

Rouge 1	Business	Entertain	Politics	Sports	Technology
Recall	0.6632	0.6481	0.7054	0.6550	0.7361
Precision	0.8192	0.8576	0.7818	0.8339	0.7492
F1-Score	0.7211	0.7233	0.7281	0.7153	0.7282

TABLE VI: Topic Wise ROUGE 2 score on BBC NEWS

Rouge 2	Business	Entertain	Politics	Sports	Technology
Recall	0.5835	0.5729	0.6260	0.5723	0.6566
Precision	0.7532	0.7991	0.7074	0.7654	0.6663
F1-Score	0.6432	0.6495	0.6473	0.6327	0.6434

Figure (6) shows the Rouge Scores achieved by DES method on BBC NEWS dataset along with other state of the art methods. Table (VIII) shows the ROUGE 1 score of all summarization methods on BBC NEWS dataset. Table (IX) shows the ROUGE 2 score of all summarization methods on BBC NEWS dataset. Table (X) shows the ROUGE L Score of all summarization methods on BBC NEWS dataset.

(a) ROUGE1 (b) ROUGE2 (c) ROUGEL

Fig. 6: ROUGE Score in BBC NEWS dataset

TABLE VII: Topic Wise ROUGE L Score on BBC NEWS

Rouge L	Business	Entertain	Politics	Sports	Technology
Recall	0.6605	0.6456	0.7016	0.6519	0.7318
Precision	0.8163	0.8549	0.7780	0.8307	0.7452
F1-Score	0.7184	0.7208	0.7244	0.7123	0.7241

TABLE VIII: ROUGE-1 values in BBC NEWS

Rouge 1	KLDIV	LEX	LHUN	LSA	TEX	DES
Recall	0.5241	0.7744	0.7972	0.6760	0.7706	0.68154
Precision	0.5236	0.5989	0.6643	0.5633	0.6084	0.80834
F1-Score	0.5038	0.6579	0.7098	0.5978	0.6655	0.7232

TABLE IX: ROUGE-2 values in BBC NEWS dataset

ROUGE-2	KL DIV	LEX	LHUN	LSA	TEX	DES
Recall	0.4041	0.6883	0.7301	0.5506	0.6899	0.6023
Precision	0.4023	0.5141	0.5801	0.4601	0.5232	0.73827
F1-Score	0.3820	0.5685	0.6290	0.4814	0.5785	0.64322

TABLE X: ROUGE-L values in BBC NEWS dataset

ROUGE-L	KL DIV	LEX	LHUN	LSA	TEX	DES
Recall	0.5134	0.7689	0.7933	0.6672	0.7646	0.6783
Precision	0.5119	0.5938	0.6605	0.5546	0.6029	0.805001
F1-Score	0.4929	0.6527	0.7060	0.5893	0.6599	0.7199

5.3 Experiment on News Aggregator Dataset

In order to evaluate the efficacy of the suggested DES method, it has been tested on a second dataset. Data for the News Aggregator was obtained from the UCI data archive. It is a NEWS dataset that has 4 categories of NEWS documents. The period of NEWS collected from 10 March 2014 to 10 August 2014. Business, Health, Technology and Entertain are 4 categories in the document collection.

Figure 7 shows the document distribution in the News Aggregator dataset. The amount of papers that fall into each category is detailed in Table (XI), which is part of the dataset for the News Aggregator.

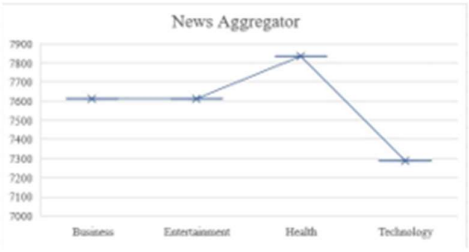


Fig. 7: Document distribution in NEWS Aggregator

TABLE XI: Number of documents in NEWS Aggregator

Category	Number of Documents
Business	7611
Entertain	7612
Health	7836
Technology	7288

Table (XII) lists the top fifty domain words to each to each category in the News Aggregator dataset. LDA algorithm is used to identify these words. Figure (8) shows the Rouge Scores on the NEWS Aggregator dataset on each topic wise. Table (XIII) shows the ROUGE 1 Score of each topic in NEWS Aggregator dataset. Table (XIV) shows the ROUGE 2 score of each topic in NEWS Aggregator dataset. The Rouge Scores that were obtained using the DES method on the NEWS Aggregator dataset are displayed in Figure (9) alongside those obtained using other state of the art methods. The ROUGE 1 Score of each summarization method that was used on the NEWS Aggregator dataset is displayed in Table (XVI). The ROUGE 2 Score for each method of summarization applied to the NEWS Aggregator dataset is displayed in Table (XVII). Table (XVIII) shows the ROUGE L Score of all summarization methods on NEWS Aggregator dataset.

TABLE XII: TopWords in each topic in NEWS Aggregator dataset

Domain	Top Words
Business	'say', 'company', 'year', 'market', 'per cent, 'month', 'report', 'million', 'bank', 'apple', 'google', 'price', 'time', 'sale', 'expect', 'share', 'user', 'billion', 'Microsoft', 'week', 'device', 'march', 'service', 'data', 'like', 'include', 'trade', 'govern', 'come', 'iPhone', 'business', 'stock', 'plan', 'product', 'sell', 'growth', 'state', 'rise', 'Mobil', 'window', 'release', 'accord', 'office', 'February', 'game', 'secure', 'custom', 'rate', 'increase', 'launch'
Health	'say', 'health', 'study', 'research', 'people', 'cancer', 'year', 'disease', 'report', 'patient', 'case', 'risk', 'state', 'virus', 'medic', 'drug', 'cigarette', 'care', 'test', 'Mer', 'women', 'recall', 'death', 'prevent', 'accord', 'include', 'hospital', 'per cent', cause', 'like', 'time', 'infect', 'effect', 'blood', 'treatment', 'increase', 'know', 'universe', 'develop', 'help', 'high', 'country', 'association', 'smoke', 'heart', 'cell', 'public', 'human', 'million', 'number'
Technology	'say', 'change', 'climate', 'report', 'music', 'year', 'world', 'like', 'country', 'Juan', 'Pablo', 'rise', 'food', 'water', 'people', 'risk', 'high', 'bachelor', 'time', 'impact', 'young', 'release', 'include', 'girl', 'come', 'American', 'season', 'global', 'level', 'increase', 'life', 'Nikki', 'miss', 'research', 'nature', 'work', 'album', 'know', 'women', 'life', 'million', 'dream', 'warm', 'area', 'scientist', 'boss', 'favourite', 'state', 'lead', 'group'
Entertain	'say', 'like', 'time', 'year', 'know', 'people', 'go', 'think', 'tell', 'star', 'look', 'want', 'come', 'love', 'film', 'take', 'right', 'thing', 'video', 'story', 'movie', 'life', 'work', 'Twitter', 'season', 'week', 'play', 'final', 'report', 'serial', 'photo', 'good', 'life', 'news', 'follow', 'leave', 'march', 'watch', 'episode', 'start', 'post', 'friend', 'best', 'whale', 'night', 'world', 'write', 'family', 'couple', 'call'

TABLE XIII: Topic Wise ROUGE 1 score on NEWS Aggregator

Rouge 1	Buisness	Health	Entertain	Technology
Recall	0.3717	0.3705	0.3647	0.3629
Precision	0.6841	0.7224	0.6500	0.7039
F1-Score	0.4533	0.4652	0.4436	0.4554

TABLE XIV: Topic Wise ROUGE 2 score on NEWS Aggregator

Rouge 2	Business	Health	Entertain	Technology
Recall	0.2605	0.2661	0.2511	0.2494
Precision	0.5654	0.6240	0.5237	0.5769
F1-Score	0.3280	0.3469	0.3162	0.3248

Table XV: Topic Wise ROUGE L score on NEWS Aggregator

Rouge L	Business	Health	Entertain	Technology
Recall	0.3717	0.3705	0.3717	0.3629
Precision	0.6840	0.7224	0.6500	0.7039
F1-Score	0.4533	0.4651	0.4436	0.4554

TABLE XVI: ROUGE-1 values in the News Aggregator

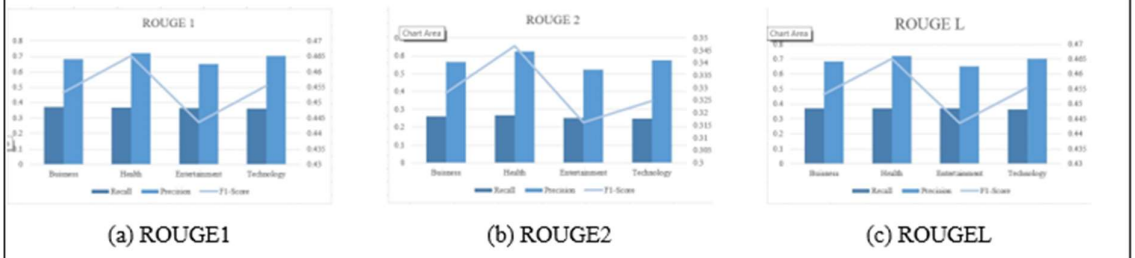
ROUGE 1	KL DIV	LEX	LHUN	LSA	TEX	DES
Recall	0.6048	0.7917	0.7765	0.7299	0.7376	0.8075
Precision	0.5236	0.4871	0.5024	0.4802	0.4689	0.6901
F1 Score	0.5513	0.5922	0.6006	0.5709	0.5642	0.6244

TABLE XVII: ROUGE-2 values in News Aggregator

ROUGE 2	KL DIV	LEX	Lhun	LSA	TEX	DES
Recall	0.5047	0.6883	0.68016	0.5952	0.6087	0.6968
Precision	0.4199	0.4014	0.4132	0.3843	0.3747	0.5725
F1 Score	0.4477	0.4943	0.5028	0.4585	0.4545	0.5790

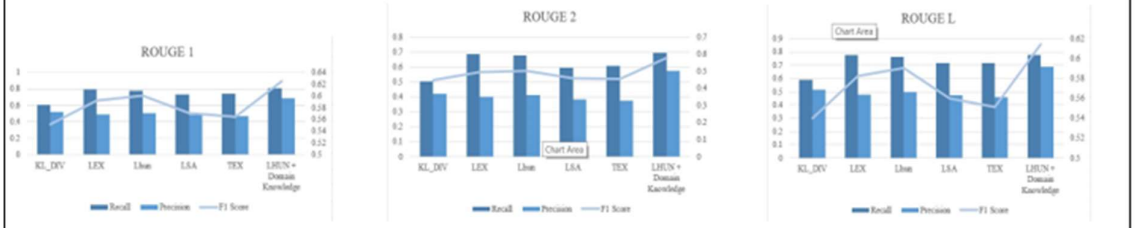
TABLE XVIII: ROUGE-L values in News Aggregator

ROUGE L	KL DIV	LEX	Lhun	LSA	TEX	DES
Recall	0.59225	0.77714	0.76141	0.71488	0.71729	0.77746
Precision	0.51291	0.47946	0.49471	0.47174	0.45903	0.6901
F1 Score	0.53975	0.58227	0.59051	0.56019	0.55097	0.61439



(a) ROUGE1 (b) ROUGE2 (c) ROUGEL

Fig. 8: Topic wise ROUGE Score in News Aggregator dataset



(a) ROUGE1 (b) ROUGE2 (c) ROUGEL

Fig. 9: ROUGE Score in News Aggregator dataset

STATISTICALLY SIGNIFICANT ANALYSIS TEST

TABLE XIX: P Values on Datasets

P Value	BBC NEWS	NEWS Aggregator
ROUGE 1	0.002174	0.032835
ROUGE 2	0.002174	0.014387
ROUGE L	0.001905	0.025124

A statistical t-test is performed at a 5% significance level to demonstrate that the findings obtained using the proposed DES method are statistically significant. A lower p-value indicates that the result is more significant.

NULL Hypothesis: The F1-Score values of the proposed DES method do not outperform significantly existing state-ofthe-art methods.

Alternate Hypothesis: According to the F1-Score, the suggested DES algorithm performs far better than the current top-tier approaches.

Here, we use two groups to determine the significance level (p-value). The ROUGE scores generated by the DES technique approach are in the first group, whereas the ROUGE scores generated by the current LHUN method are in the second group. In XIX, we can see a table with the p-values for both datasets when using the Rouge 1, Rouge 2, and Rouge L values. Both F1 Score-based significance levels are significantly lower than 0.05. So, the NULL hypothesis is rejected due to the strength of the evidence presented here. The alternative hypothesis is supported by the test findings, suggesting that the gains made using the suggested DES Method are not coincidental and that the method has statistical significance.

CONCLUSION AND FUTURE WORK

The primary drawback of current extractive summarising methods is that they only focus on local data that is unique to a given document. The document domain is not taken into

consideration. A DES technique is developed to address this problem, and domain expertise is established using the LDA topic modelling method. The suggested DES Summarization method achieves better results than the current best practices in terms of Rouge Score. Better results are achieved when summarising using domain knowledge. On two datasets, the proposed DES method was evaluated. Results from the suggested method are superior for both datasets. Additionally, statistical t-test analysis demonstrates the superiority of the DES method over the alternatives. Future abstractive summarization algorithms could leverage the learned domain knowledge more extensively, producing superior results.

ACKNOWLEDGEMENT

Funding for this study was provided by the University Grants Commission Junior Research Fellowship programme (JRF).

REFERENCES

- Sefid and C. L. Giles, "Scibertsum: Extractive summarization for scientific documents," in International Workshop on Document Analysis Systems, pp. 688–701, Springer, 2022.
- R. C. Belwal, S. Rai, and A. Gupta, "A new graph-based extractive text summarization using keywords or topic modeling," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 10, pp. 8975–8990, 2021.
- M. Jang and P. Kang, "Learning-free unsupervised extractive summarization model," IEEE Access, vol. 9, pp. 14358–14368, 2021.
- H. K. Thakkar, P. K. Sahoo, and P. Mohanty, "Dofm: domain feature miner for robust extractive summarization," Information Processing & Management, vol. 58, no. 3, p. 102474, 2021.
- R. C. Belwal, S. Rai, and A. Gupta, "Text summarization using topicbased vector space model and semantic measure," Information Processing & Management, vol. 58, no. 3, p. 102536, 2021.
- X. Mao, H. Yang, S. Huang, Y. Liu, and R. Li, "Extractive summarization using supervised and unsupervised learning," Expert systems with applications, vol. 133, pp. 173–181, 2019.
- Mutlu, E. A. Sezer, and M. A. Akcayol, "Candidate sentence selection for extractive text summarization," Information Processing & Management, vol. 57, no. 6, p. 102359, 2020.
- R. M. Alguliyev, R. M. Aliguliyev, N. R. Isazade, A. Abdi, and N. Idris, "Cosum: Text summarization based on clustering and optimization," Expert Systems, vol. 36, no. 1, p. e12340, 2019.
- R. Rani and D. Lobiyal, "Document vector embedding based extractive text summarization system for hindi and english text," Applied Intelligence, pp. 1–20, 2022.
- W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Edgesumm: Graph-based framework for automatic text summarization," Information Processing & Management, vol. 57, no. 6, p. 102264, 2020.
- R. K. Roul and K. Arora, "A nifty review to text summarizationbased recommendation system for electronic products," Soft Computing, vol. 23, no. 24, pp. 13183–13204, 2019.
- J. Rodr´ıguez-Vidal, J. Carrillo-de Albornoz, E. Amigo, L. Plaza, J. Gon-´ zalo, and F. Verdejo, "Automatic generation of entity-oriented summaries for reputation management,"

Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 4, pp. 1577–1591, 2020.

- Joshi, E. Fidalgo, E. Alegre, and L. Fernandez-Robles, "Summcoder:´ An unsupervised framework for extractive text summarization based on deep auto-encoders," Expert Systems with Applications, vol. 129, pp. 200–215, 2019.
- R. Rani and D. Lobiyal, "An extractive text summarization approach using tagged-lda based topic modeling," Multimedia tools and applications, vol. 80, no. 3, pp. 3275–3305, 2021.
- R. Elbarougy, G. Behery, and A. El Khatib, "Extractive arabic text summarization using modified pagerank algorithm," Egyptian informatics journal, vol. 21, no. 2, pp. 73–81, 2020.
- Joshi, E. Fidalgo, E. Alegre, and R. Alaiz-Rodriguez, "Ranksum—an unsupervised extractive text summarization based on rank fusion," Expert Systems with Applications, vol. 200, p. 116846, 2022.
- R. K. Roul, "Topic modeling combined with classification technique for extractive multi-document text summarization," Soft computing, vol. 25, no. 2, pp. 1113–1127, 2021.
- Y. Zhang, M. J. Er, R. Zhao, and M. Pratama, "Multiview convolutional neural networks for multidocument extractive summarization," IEEE transactions on cybernetics, vol. 47, no. 10, pp. 3230–3242, 2016.
- N. Rahman and B. Borah, "Improvement of query-based text summarization using word sense disambiguation," Complex & Intelligent Systems, vol. 6, no. 1, pp. 75–85, 2020.
- Abdi, S. M. Shamsuddin, S. Hasan, and J. Piran, "Automatic sentiment-oriented summarization of multi-documents using soft computing," Soft Computing, vol. 23, no. 20, pp. 10551–10568, 2019.
- S. Lamsiyah, A. El Mahdaouy, S. Ouatik El Alaoui, and B. Espinasse, "Unsupervised query-focused multi-document summarization based on transfer learning from sentence embedding models, bm25 model, and maximal marginal relevance criterion," Journal of Ambient Intelligence and Humanized Computing, pp. 1–18, 2021.
- T. Uc¸kan and A. Karcı, "Extractive multi-document text summarization based on graph independent sets," Egyptian Informatics Journal, vol. 21, no. 3, pp. 145–157, 2020.
- P. K. Wilson and J. Jeba, "A developed framework for multi-document summarization using softmax regression and spider monkey optimization methods," Soft Computing, vol. 26, no. 7, pp. 3313–3328, 2022.