Journal on Artificial Intelligence

Tech Science Press

DOI: 10.32604/jai.2024.052099

ARTICLE

A Hybrid Query-Based Extractive Text Summarization Based on K-Means and Latent Dirichlet Allocation Techniques

Sohail Muhammad[1], Muzammil Khan[2] and Sarwar Shah Khan[2,3,*]

[1]Department of Computer Science, City University of Science and IT, Peshawar, 24370, Pakistan

[2]Department of Computer and Software Technology, University of Swat, Swat, 19120, Pakistan

[3]Department of Computer Science, IQRA National University Swat, Swat, 19130, Pakistan

*Corresponding Author: Sarwar Shah Khan. Email: sskhan0092@gmail.com

Received: 22 March 2024 Accepted: 08 July 2024 Published: 07 August 2024

ABSTRACT

Retrieving information from evolving digital data collection using a user's query is always essential and needs efficient retrieval mechanisms that help reduce the required time from such massive collections. Large-scale time consumption is certain to scan and analyze to retrieve the most relevant textual data item from all the documents required a sophisticated technique for a query against the document collection. It is always challenging to retrieve a more accurate and fast retrieval from a large collection. Text summarization is a dominant research field in information retrieval and text processing to locate the most appropriate data object as single or multiple documents from the collection. Machine learning and knowledge-based techniques are the two query-based extractive text summarization techniques in Natural Language Processing (NLP) which can be used for precise retrieval and are considered to be the best option. NLP uses machine learning approaches for both supervised and unsupervised learning for calculating probabilistic features. The study aims to propose a hybrid approach for query-based extractive text summarization in the research study. Text-Rank Algorithm is used as a core algorithm for the flow of an implementation of the approach to gain the required goals. Query-based text summarization of multiple documents using a hybrid approach, combining the K-Means clustering technique with Latent Dirichlet Allocation (LDA) as topic modeling technique produces 0.288, 0.631, and 0.328 for precision, recall, and F-score, respectively. The results show that the proposed hybrid approach performs better than the graph-based independent approach and the sentences and word frequency-based approach.

KEYWORDS

Extractive text summarization; machine learning; natural language processing; K-Means; latent dirichlet allocation

1 Introduction

Natural Language Processing (NLP) is a subfield of computer science that helps computer systems to interact in natural language effectively. Both, supervised and unsupervised machine learning techniques are used in NLP to solve different problems. Digital data and information rapidly increasing, making it hard to get the associated information on demand using user queries. Data scientists work rigorously in the area of machine learning to automate the information retrieval process

from extensively large collections [1]. Text summarization helps to efficiently manage the digital collection. Text document summarization is a field of computer science involving Artificial Intelligence (AI), Data Mining, Statistics, and Psychology [2]. It has several applications, like summaries of newspapers, books, magazines, scientific papers, weather forecasting, stock marketing, and news [3]. Text Summarization is a procedure for key information extraction from the original textual contents. The information extracted is produced as a summarized report, which serves as a concise brief user summary. The huge textual content is difficult for humans to comprehend [4]. The use of automatic approaches for data processing became critical to the user. It is challenging to utilize all the related information accessible online without using those tools. Automatic text summarization systems are advantageous, allowing the user to obtain relevant knowledge quickly. In recent years, the NLP community has shown a high interest in the generation of automatic text summarization [2]. There are four different types of summarization approaches, as briefly discussed below:

1. **Abstractive text summarization** is the main idea of a document. It produces a summarized document which easy to read and grammatically accurate in a consistent form by modifying the original text [5].
2. **In extractive summarization approaches,** the related sentences are extracted and ranked by their importance. Then the sentences are grouped to compose the summary without making changes and modifications to the original text document [6].
3. **Generic summarization** is a stepwise procedure in which the first step ranks sentences from the given document and the second step extracts all those sentences from the documents [7]. The first step utilizes a conventional information retrieval approach to classify the most related phrases, while in the second step, the Latent Semantic analysis is followed to generate a summary [8].
4. **Query-based text summarization** is essential because it provides all the information according to the user's requirements. The user doesn't have to spend time browsing the information they need. Query-based summarization is key in the information-gaining process due to the speedy growth of web information and retrieving the related information by matching the user's query [6]. The query-based summarization is also called "user-focused summarization" or "query-focused summarization". In query-focus text summarization, the importance of every sentence and variable combination is determined [9].

With the growing data size, rapid improvement, and technology usage, query-based text summarization is vital in providing up-to-date information about a topic or query. Natural Language Processing is described as a stream of computer sciences and linguistics apprehensive about the connections between natural languages and computers. In supposition, natural language converting must be the neat technique of human-computer crossing point. Natural language is grateful is occasionally described as Artificial Intelligence's entire issue because natural language recognition appears to engage extensive information about the external world and the aptitude to handle it. People find much information daily in newspapers, and getting significant information from all the editorials is boring for individuals. There is a need for an automatic system that can extract only the related information from these sources. One needs to mine the text of the newspapers by which high-quality information is extracted in large amounts. Text mining uses some of the approaches of Natural Language Processing, such as parsing, Part-Of-Speech (POS) tagging, tokenization, etc., to execute the text analysis. Identification and analysis of machine learning methodologies for query-based extractive text summarization for multiple documents. The contributions of the paper are as follows:

- The study proposes a hybrid approach for query-based extractive text summarization.
- The Text-Rank Algorithm is utilized as the core algorithm for the implementation flow of the approach.
- The hybrid approach combines the K-Means clustering technique with Latent Dirichlet Allocation (LDA) for topic modeling.
- The results demonstrate that the proposed hybrid approach performs better than the graph-based independent approach and the sentences and word frequency-based approach.

Text summarization plays a crucial role in finding the most relevant information, whether it's a single document or multiple documents from this collection. In NLP, machine learning is employed for both supervised and unsupervised learning, leveraging probabilistic features to enhance the accuracy of retrieval. In this study, the aim is to propose a hybrid approach for query-based extractive text summarization. The core algorithm utilized in this approach is the Text-Rank Algorithm, chosen to help us achieve the research goals. The hybrid approach combines the K-Means clustering technique with LDA, which is a topic modeling technique. This combination enables us to perform query-based text summarization on multiple documents, resulting in the extraction of highly relevant information. The Text-Rank algorithm is used for query-based-text summarization, which works in hybrid combination, using the K-Means clustering technique with the LDA technique by extracting text from query-based text summarization for a document to a particular topic. It uses both "AND" and "OR" operations for retrieving data from multiple documents. The documents are news articles about different issues, i.e., politics, sports, science, etc.

2 Literature Review

Data and information are increasing daily and moving toward big data. The need for automatic text summarization is required to summarize text [5]. The information in the compressed and summarized form having semantic meaning is needed–the problems with multi-document summarization, i.e., redundancy, identifying differences among documents, and summary coherence. The main focus is on Abstractive Text Summarization. The intention behind abstractive summarization is to construct a generalized summary. It generally requires generation and language compression techniques to present the information concisely. The authors still regard the lack of a generalized framework for text parsing and alignment as the key challenge. A researcher proposes different techniques for extracting important sentences from the text and order. One such technique is the Support Vector Machine (SVM) proposed by [10]; SVM utilizes a regression model for query-based text summarization of multi-documents. Support Vector Regression (SVR) is used to assess the significance of a sentence in a document. The summarization is done through some pre-defined features in the paper. A fixed set feature-based technique is used as a learning model to search for the best possible composite scoring function. Sentences are ranked by their feature values. It is shown through experiments that regression models for estimating the importance of a sentence in a test document dominate learning-to-rank model and classification models.

Another study proposed regression-based ML techniques like SVR and gradient-boosted decision trees for selecting sentences, which is a crucial sub-task for query-based test summaries. A local topic identification-based technique that utilizes word frequency is proposed by [11] for single-document summarization. It uses logical structure features where documents are first clustered based on local topics after calculating sentence similarity and sorting. In order to calculate the word frequency, sentences from every local topic are chosen. For every local issue, the suggested technique reduces information redundancy more effectively. A proposed paradigm for local topic identification is the

vector space model. The sentence with the best score is selected from the best local topic to be included in the summary. The sum is obtained using the probabilistic model of document weight and Topic weight. Sentence location plays a vital part in text summarization.

A set of fixed feature-based techniques is used as a learning model to search for the best possible composite scoring function. Sentences are ranked by their feature values. It is shown through experiments that regression models for estimating the importance of a sentence in a test document dominate learning-to-rank and classification models. The importance of query-based text summarization is evident from the fact that it saves precious time for the user by providing the required information in the summarized form and requiring him to spend time searching and browsing. An author performed query-based text summarization in the Arabic language [12].

One Arabic document is given a query-based summary by the system. To extract the pertinent paragraph and produce its summary, it makes use of the conventional vector space model and cosine similarity. The redundancy problem is one of the major challenges in text summarization, which cannot be completely resolved and still requires further research to explore. A study describes a model or detail for determining the criteria that can retrieve sentences from multiple documents and uses a genetic algorithm to generate multiple summaries [13]. Then it evaluates and classifies them to produce the best summary. The generic algorithm produces some random solutions, then builds a set of solutions in each stage and evaluates them. The genetic algorithm produces, for each iteration, an extract's population while combining random sentences from the various documents. Then apply crossover and mutation operators. It assigns an objective value for each generated extract, which depends on criteria applied to extract as a whole unit. The importance of statistical criteria to determine the candidate sentences to form the extract, i.e., position of the sentence, Size of the document, TF $*$ IDF, Similarity to the title, Similarity to document keywords, Similarity to question keywords. The author of [14] used a graph-based ranking technique called query expansion algorithm for query-based summarization of multiple documents. This approach also resolves the information limit problem in the original query. To select the query-based informative words from the set of documents and to use them as query expansion for improving the ranking result of sentences, this method utilizes both sentence-to-sentence relationships and sentence-to-word relationships.

A provides a graph-based method for summarising Arabic text that includes text preprocessing, the use of the Firefly algorithm to select sentences for the summary, the creation of a graph of possible replies, and the calculation of similarity scores based on structural elements. Using the EASC corpus, the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics were used to evaluate the suggested approach. The study discusses other methods of text summarization and the challenges of finding an ideal summary ratio [15].

A novel weighted word embedding method for extractive text summarization. The suggested model is assessed using the ROUGE-N score metric and contrasted with three cutting-edge baselines. The results show that the proposed model outperforms the baselines in terms of precision, recall, and F-score. The model generates summaries of different lengths and compression rates, and the evaluation is conducted on a benchmark dataset [16].

The task of automatic text summarization and the different approaches used in the literature. It highlights the importance of summarization in compressing textual data and providing a cost-efficient solution for transferring information. The study also presents a performance study on extractive text summarization using Bidirectional Encoder Representations from Transformers (BERT) models, which achieved promising results [17].

RankSum is an extractive text summary framework that operates without supervision and integrates many structural and semantic aspects of a document. With the use of keywords, signature topics, sentence embeddings, and a phrase's position within the document, the method extracts multi-dimensional information from the document. The suggested approach performs better than a variety of approaches, including statistical, concept, optimisation, topic, graph, sentence centrality, semantic, and deep learning-based approaches [18].

An overview of the advancements made in Urdu linguistics and the challenges faced in generating automatic extractive summaries for Urdu text. The article discusses the two approaches used to compute the weight of words in the English language, the difficulties in generating automatic extractive summaries for Urdu text, and the publicly available dataset or framework for automatic Urdu extractive summary generation [19].

A study proposes a novel approach for semantic extractive text summarization that takes into account the meaning and context of the original content. The proposed approach surpasses current methods when it comes to precision, recall, and F-measure scores. It also allows for all features and excludes redundancy compared to other algorithms. The ANOVA technique can be used to tune the model for better accuracy with minimum variance. The proposed approach has potential applications in real-world scenarios such as news articles or scientific papers [20].

A new framework has been introduced for extractive text summarization, utilizing a neural network classifier enhanced with deep learning techniques. This algorithm employs entropy values to create concise and informative document summaries. According to experimental results, this new method outshines current top algorithms, achieving higher ROUGE and F1-scores. The proposed algorithm has potential applications in various domains, including business organizations and end-users [21].

A study proposes a deep learning-based approach for extractive text summarization. The authors compare sentence-level and paragraph-level summarization and find that their model performs better in the former. To evaluate their model using state-of-the-art metrics such as ROUGE and BLEU and compare it to other deep learning models. The proposed model achieves promising results in accurately summarizing large corpora of unstructured data. The authors also discuss future improvements such as using advanced data structures like wavelet trees. This study presents a new model that shows potential for improving the efficiency and accuracy of text summarization [22].

3 Framework

For this study, the information is extracted from the documents stored in secondary storage in a folder. These documents are from different categories from a folder that is collected. In the first step, the documents will be pre-processed, and the noise ID will be removed from both the text document and the user query [23,24]. This study focuses on the Text-Rank Algorithm, an entirely unsupervised method that operates through several steps. Initially, the text is segmented into marked parts of speech; this preprocessing step is crucial for applying syntactic filters [25]. To avoid an overly large graph from forming by including all possible combinations of multi-word sequences, only single words are considered for the graph, with multiple keywords being reconstructed later in the post-processing phase. The graph comprises lexemes that pass the syntactic filter, and edges are created between those lexical units that co-occur within a specific window of words. In the second step, to create clusters of documents using the K-Means clustering approach. The algorithm commonly assigns items by distance to the closest cluster. In the final step, the Latent Dirichlet Allocation topic Modeling Approach is used to extract the features of the query and rank the documents retrieved against a

user query and will be analyzed. LDA is a subject-modeling approach that works on classifying text in a document to a specific theme. The LDA approach works on subject patterning cantered on probabilistic path directions of words, which specify their relation to the text corpus. The framework is shown in Fig. 1.

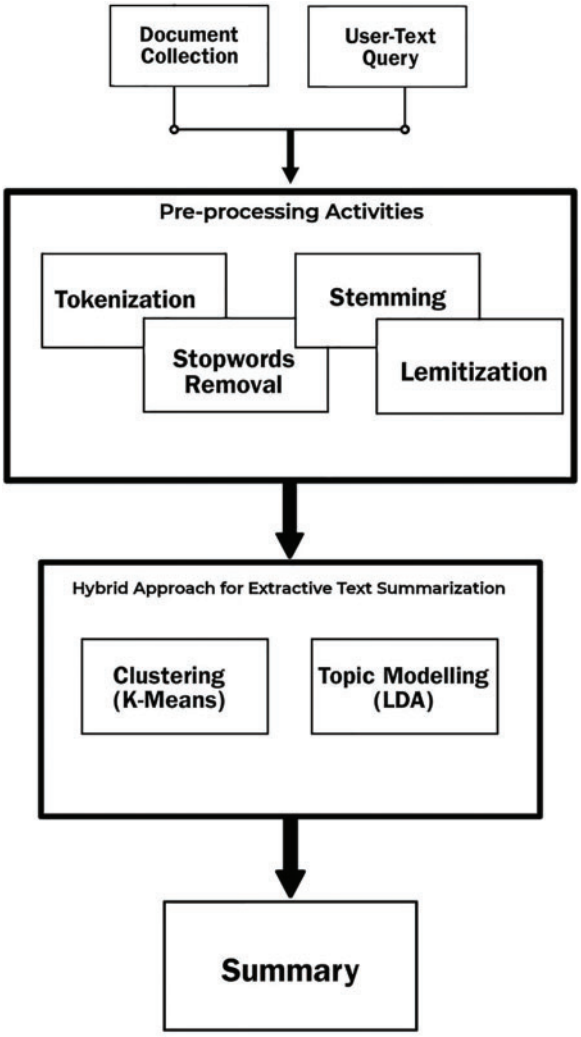

Figure 1: Framework for proposed hybrid approach

3.1 DUC-2004 Dataset

The DUC-2004 dataset is a dataset commonly used in the field of NLP and text summarization. It was created as part of the Document Understanding Conference (DUC) series, specifically for the DUC 2004 competition. This dataset consists of various documents and associated human-generated reference summaries. Researchers and participants in the competition use this dataset to develop and evaluate automatic text summarization systems. The goal is to create algorithms and models that can generate concise and coherent summaries of the given documents that are similar in content and quality to the reference summaries provided by human annotators [26]. The dataset used in the study is summarized as:

- The main purpose of DUC-2004 was to advance the research and development of automatic text summarization systems. Participants in the competition were tasked with creating software and algorithms capable of summarizing a set of documents effectively.
- DUC-2004 provided participants with a collection of documents, typically news articles or scientific papers, along with human-generated reference summaries. These documents covered a range of topics and domains.
- The main task in DUC-2004 was to develop systems that could automatically generate summaries of the provided documents. These summaries were expected to be concise, coherent, and informative, similar in quality to the reference summaries created by human annotators.
- The quality of the generated summaries was assessed using several metrics, such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation). ROUGE measures how much the generated summaries overlap with the reference summaries by comparing n-grams (word sequences).
- Challenges: DUC-2004, like other DUC competitions, presented several challenges to participants. These challenges included handling multi-document summarization, dealing with documents from various domains and genres, and addressing issues like redundancy and coherence in the generated summaries.
- The competition typically resulted in the identification of state-of-the-art techniques and systems in the field of automatic text summarization. It provided a benchmark for evaluating the performance of different summarization algorithms.
- The DUC-2004, along with other editions of the DUC series, contributed to the advancement of text summarization research and the development of summarization systems. It also served as a platform for researchers to compare their approaches and share their findings.

The proposed methodology works on "OR" and "AND" operations, which will extract all those sentences that will include the entire query individually like if the multiple words in a user's entered query. First, it will extract all those sentences, including the whole query, and then extract each word one by one. For example, if you enter "Asian Development Bank", then first it will extract all those sentences will, includes "Asian Development Bank", and then it will extract sentences having "Asian", "Development", and then "Bank", as shown in Table 1.

Table 1: Sample generated results against the query

Query	Generated result
Caribbean	• A hurricane warning was issued for the Caribbean coast of Guatemala. • Forecasters indicated that Mitch didn't pose an immediate threat to the United States but was expected to linger in the northwest Caribbean for five days. • The last major hurricane to hit Honduras was Fifi in 1974, which devastated the country's Caribbean coast and killed at least 2000 people. • At its height, Mitch was the fourth-strongest Caribbean hurricane of the century, following Gilbert in 1988, Allen in 1980, and the Labor Day hurricane of 1935. • Jerry Jarrell, director of the weather center, stated that Mitch was the most powerful hurricane to hit the Caribbean since Gilbert in 1988, which resulted in over 300 deaths.

(Continued)

Table 1 (continued)

Query	Generated result
	• As the storm remained off the coast of Honduras, officials in Mexico's Yucatan Peninsula eased emergency measures. However, hundreds of people stayed in shelters as a precaution on Wednesday night. • On Tuesday, Hondur as prepared for potential disaster as Hurricane Mitch swept through the northwest Caribbean, bringing high waves and heavy rain that forced coastal residents to seek safer ground.
Cambodia	• The prince fled Cambodia and returned only a few months before the July elections. • His assurances come just a week before the inaugural session of Cambodia's new parliament, the National Assembly. • On Wednesday, Ranariddh told reporters that he believed Cambodia was now safe for Sam Rainsy. • Sam Rainsy, a staunch critic of Hun Sen, and 14 other members of his party's parliament have stayed outside Cambodia since September due to security concerns. • Fearing for their safety, Sam Rainsy and his former ally, Prince Norodom Ranariddh, led an exodus of opposition lawmakers out of Cambodia after the ceremonial opening of parliament in late September. • In a letter to King Norodom Sihanouk, the prince's father and Cambodia's head of state, broadcast on television Tuesday, Hun Sen stated that the safety guarantees for Ranariddh extended to all politicians. • Concerned that his party colleagues might still face arrest due to their political stance, opposition leader Sam Rainsy sought further clarification on Friday regarding the security guarantees promised by Hun Sen. In a letter to King Norodom Sihanouk, he expressed eagerness to attend the first session of the new National Assembly on November 25 but noted that Hun Sen's assurances were insufficient to alleviate concerns about potential arrests of his party members upon their return to Cambodia. • After a three-month deadlock, an agreement was reached last week on a coalition deal that will make Hun Sen the sole prime minister and Ranariddh the president of the National Assembly.
Asian development bank	• "To advise the Asian Development Bank not to provide any new loans to the current regime in Cambodia," wrote the two-party leaders. • On Monday, Cambodia's two-party opposition asked the Asian Development Bank to halt loans to the sitting government, which they deem illegal. • This could also expedite Cambodia's admission into the Association of Southeast Asian Nations, which was postponed last year after Hun Sen ousted Ranariddh as co-prime minister in a violent coup. • Ranariddh and Sam Rainsy resumed their international campaign against the former government on Monday, sending a letter to ADB President Mitsuo Sato, urging the bank to cease lending to it.

(Continued)

Table 1 (continued)

Query	Generated result
	• No one should internationalize Cambodian affairs. • However, an agreement between Hun Sen and his main rival, Prince Norodom Ranariddh, to form a new government last week has paved the way for their return. • The men served as co-prime ministers until Hun Sen overthrew Ranariddh in a coup last year. • The agreement secures the two-thirds parliamentary vote necessary to approve a new government.
Capital	• Coastal Belize City was devastated by Hurricane Hattie in 1961, prompting the country to establish a new inland capital at Belmopan. • Rain squalls flooded the streets of Kingston, Jamaica, and government offices and schools were closed in the Cayman Islands, a British territory with a population of 28,000. • In El Progreso, 100 miles (160 kilometers) north of the Honduran capital, Tegucigalpa, the army evacuated over 5000 people living in low-lying banana plantations along the Ulua River, according to local resident Nolly Soliman. • Wind-driven waves nearly buried some houses near the shore. • Approximately 10,000 residents sought refuge in crowded shelters located in schools, churches, and firehouses. • The storm's intensity was diminishing, with sustained winds of 80 mph (130 kph) by 1200 GMT, down from 100 mph (160 kph) at midnight and significantly lower than its peak of 180 mph (290 kph) early Tuesday.
Diplomatic immunity	• Home Office Minister Alun Michael acknowledged on Sunday that Pinochet entered Britain with a diplomatic passport, but stated, "That does not necessarily convey diplomatic immunity." • On Monday, the magistrate expanded his charges to include the killings of both Chileans and Spaniards, and genocide, which does not grant diplomatic immunity. • Chile's ambassador formally protested to the Foreign Office on Monday, claiming that Britain violated Pinochet's diplomatic immunity. • Britain has justified the arrest of Gen. Augusto Pinochet, with one lawmaker dismissing Chile's claim of diplomatic immunity for the former dictator as absurd. • Chilean Ambassador Mario Artaza, who was exiled during Pinochet's regime, stated that Chile had a duty to protect a citizen with diplomatic immunity and senator status. • Trade Secretary Peter Mandelson expressed in a BBC interview on Sunday that most people in the country would find it deeply disturbing for such a brutal dictator as Pinochet to claim diplomatic immunity. • The Chilean government has protested Pinochet's arrest, asserting that as a senator traveling on a diplomatic passport, he was immune from arrest.

The algorithm gives output in the form of bullet sentences having that query that the user gave. It will provide the results according to the Text-Rank algorithm sentences having higher scores will

be on the top. Currently, it will give ten results, but it can increase or decrease the number of results from the code. In the end, it gives the precision, recall, and F1-score, which shows the performance of the software [27]. One of the main drawbacks of the deficiency of the low results is that it compares the result with the system-generated summary of that document, each sentence is the abstract of each document, and the system extracts those sentences having those words which were given in the query. The predicted answer against each query is mentioned in Table 2.

Table 2: Specific query, most relevant results, and generated results using the proposed model

Query	Most relevant answers	Generated result
Caribbean	• A hurricane warning was issued for the Caribbean coast of Guatemala. • Forecasters said Mitch did not pose an immediate threat to the United States, but it was expected to stay in the northwest Caribbean for five days.	• A hurricane warning was issued for Guatemala's Caribbean coast. • Forecasters reassured that Mitch did not immediately threaten the United States but was anticipated to linger in the northwest Caribbean for five days. • The most devastating recent hurricane to hit Honduras was Fifi in 1974, causing widespread destruction along the country's Caribbean coastline and claiming at least 2000 lives. • At its peak, Mitch ranked as the fourth-strongest Caribbean hurricane of the century, following Gilbert in 1988, Allen in 1980, and the Labor Day hurricane of 1935. • Jerry Jarrell, director of the weather center, noted that Mitch marked the most potent hurricane to hit the Caribbean since Gilbert in 1988, which resulted in significant casualties. • As the storm appeared stationary near Honduras, authorities in northern Mexico relaxed emergency measures along the Yucatan Peninsula's Caribbean coast, where hundreds remained in shelters as a precaution on Wednesday night.
Cambodia	• The prince left Cambodia and only returned a few months before the July elections. • He provided assurances a week before the inaugural session of Cambodia's new parliament, the National Assembly.	• The prince left Cambodia and returned only a few months before the July elections. • He expressed confidence a week before the first session of Cambodia's new parliament, the National Assembly. • On Wednesday, Ranariddh assured reporters that Cambodia was safe for Sam Rainsy. • Sam Rainsy, a fierce critic of Hun Sen, and 14 other members of his party's parliament have stayed outside Cambodia since September due to security concerns. • Concerned for their safety, Sam Rainsy and his former ally, Prince Norodom Ranariddh, led opposition lawmakers out of Cambodia after the ceremonial opening of parliament in late September.

(Continued)

Table 2 (continued)

Query	Most relevant answers	Generated result
		• In a televised letter to King Norodom Sihanouk, the prince's father and Cambodia's head of state, Hun Sen stated that safety assurances for Ranariddh extended to all politicians.
Asian development bank	• The two-party leaders wrote to advise the Asian Development Bank against issuing any new loans to Cambodia's current government. • On Monday, Cambodia's opposition from both parties urged the Asian Development Bank to cease lending to the existing government, which they deem illegitimate.	• The two-party leaders urged the Asian Development Bank to refrain from granting new loans to Cambodia's current government. • On Monday, Cambodia's opposition from both parties requested the Asian Development Bank to halt loans to the existing administration, which they describe as illegitimate. • This action could also expedite Cambodia's admission into the Association of Southeast Asian Nations, which was delayed last year after Hun Sen removed Ranariddh as co-prime minister in a violent overthrow. • Ranariddh and Sam Rainsy intensified their international lobbying efforts against the former government on Monday in a letter to ADB President Mitsuo Sato, urging the bank to cease financing it.
Capital	• Hurricane Hattie struck Coastal Belize City so severely in 1961 that the country decided to construct a new inland capital at Belmopan. • Rain showers flooded streets in the Jamaican capital of Kingston, leading to the closure of government offices and schools in the Cayman Islands, a British territory with a population of 28,000. • Nolly Soliman, a resident, reported that in El Progreso, located 100 miles (160 kilometers) north of Honduras' capital Tegucigalpa, the army evacuated over 5000 people residing in low-lying banana plantations along the Ulua River.	• Coastal Belize City suffered severe damage from Hurricane Hattie in 1961, prompting the country to establish a new inland capital at Belmopan. • Rain showers flooded streets in Kingston, Jamaica's capital, leading to the closure of government offices and schools in the Cayman Islands, a British territory with a population of 28,000. • Nolly Soliman, a resident of El Progreso located 100 miles (160 kilometers) north of Honduras' capital Tegucigalpa, reported that the army evacuated over 5000 people residing in low-lying banana plantations along the Ulua River. • Wind-driven waves nearly engulfed some houses near the shoreline. • Around 10,000 residents sought refuge in crowded shelters set up in schools, churches, and fire stations. • The storm's intensity was diminishing; by 1200 GMT, it sustained winds of 80 mph (130 kph), down from 100 mph (160 kph) earlier in the night and well below its peak of 180 mph (290 kph) early Tuesday morning.
Diplomatic immunity	• On Monday, Chile's ambassador formally protested to the Foreign Office, accusing Britain of breaching Pinochet's diplomatic immunity.	• On Sunday, Home Office Minister Alun Michael acknowledged that Pinochet used a diplomatic passport to enter Britain but clarified, "That does not automatically grant diplomatic immunity."

(Continued)

Table 2 (continued)

Query	Most relevant answers	Generated result
	• Britain has justified its arrest of Gen. Augusto Pinochet, with a lawmaker describing Chile's assertion of diplomatic immunity for the former dictator as absurd.	• The magistrate expanded his charges on Monday to include the killings of both Chileans and Spaniards, including genocide, for which diplomatic immunity does not apply. • Chile's ambassador formally protested to the Foreign Office on Monday, accusing Britain of violating Pinochet's diplomatic immunity. • Britain has defended its arrest of Gen. Augusto Pinochet, with one lawmaker criticizing Chile's assertion that the former dictator enjoys diplomatic immunity as absurd.

The observation is taken upon the following Criteria against each query:

Criteria 1: The query-related answer must be in the 1st position.

Criteria 2: The first five ranked sentences must relate to the user's query topic.

Criteria 3: The related answer for the user query must be at any top 5 positions.

Table 3 shows the criteria for each query.

Table 3: Criteria for each generated result

Query	C1	C2	C3
Caribbean	1	1	1
Cambodia	1	1	1
Asian development bank	1	1	1
Capital	1	1	1
Diplomatic immunity	0	1	1

4 Evaluation and Discussion

DUC 2004 corpus is used for descriptive purposes in this study. The DUC 2004 corpus has 30 clusters with ten (10) text documents about each. There is a subject in each cluster. The goal is to produce for each cluster a fluent bullet form summary. Three separate summaries were provided as a text reference summary, in which one file is generated in the form of bullets, one in the form of a paragraph, and the third in the form of a C file. These evaluators have provided a document cluster description of 250 words. The summary content is evaluated using these multiple reference summaries [9]. The results of the contrast between the methods implemented and the proposed method for ROUGE calculation. Comparisons in this table indicate that recall, accuracy, and F-measurement [28] have been enhanced by the proposed procedure, using the following formulas:

$$Precision = TP/(TP + FP) \text{ (where } TP \text{ is True Positive and } FP \text{ is False Positive)} \quad (1)$$

$$Recall = TP/(TP + FN) \text{ (where } FN \text{ is False Negative)} \quad (2)$$

$$F1\ Score = \frac{(precision * recall)}{(precision * recall)} \tag{3}$$

The value of each feature is obtained amongst the query and each phrase of the text file document. And then, the sentence is ranked upon the score from highest score to lowest score.

From Tables 4 and 5, it is obvious that the approach strengthens the outcome of the research performed in the past. DUC 2004corpus is used for descriptive purposes in this study. The DUC 2004 corpus has 45 clusters with 25 text documents of each. There is a subject in each cluster. The goal is to produce for each cluster a fluent 250-word summary. Four separate NIST assessors have provided each subject and its document cluster. These evaluators have provided a document cluster description of 250 words.

Table 4: Proposed approach comparison with graph-based feature extraction approach

Topics of my query	Proposed model			Graph-based independent set		
Query	Precision	Recall	F1-score	Precision	Recall	F1-score
Cambodia	0.251	0.541	0.343	0.233	0.198	0.198
Caribbean	0.283	0.636	0.308	0.174	0.215	0.192
Capital	0.23	0.597	0.332	0.184	0.203	0.193
Asian development bank	0.332	0.66	0.337	0.198	0.233	0.214
Diplomatic immunity	0.343	0.72	0.318	0.26	0.312	0.283
Average results	**0.288**	**0.631**	**0.328**	**0.211**	**0.232**	**0.216**

Table 5: Proposed approach comparison with sentence and word frequency approach

Topics of my query	Proposed model			Graph-based independent set		
Query	Precision	Recall	F1-score	Precision	Recall	F1-score
Cambodia	0.251	0.541	0.343	0.08	0.082	0.081
Caribbean	0.283	0.636	0.308	0.122	0.115	0.118
Capital	0.23	0.597	0.332	0.089	0.082	0.081
Asian development bank	0.332	0.66	0.337	0.226	0.184	0.203
Diplomatic immunity	0.343	0.72	0.318	0.204	0.233	0.198
Average results	**0.288**	**0.631**	**0.328**	**0.144**	**0.139**	**0.136**

The summary content is evaluated using multiple reference summaries. The results highlight the comparison between the implemented methods and the proposed method using ROUGE calculations. The comparisons in this table demonstrate that recall, precision, and F-measure have been improved by the proposed procedure. Specifically, the table presents both the outcomes of the analyses for ROUGE. The new approach shows higher performance compared to the form used, as illustrated in Figs. 2 and 3. This improvement is attributed to the enrichment of the feature set by incorporating query-oriented

functions into the main feature set. The features extracted are designed to be insightful and applicable to the application.

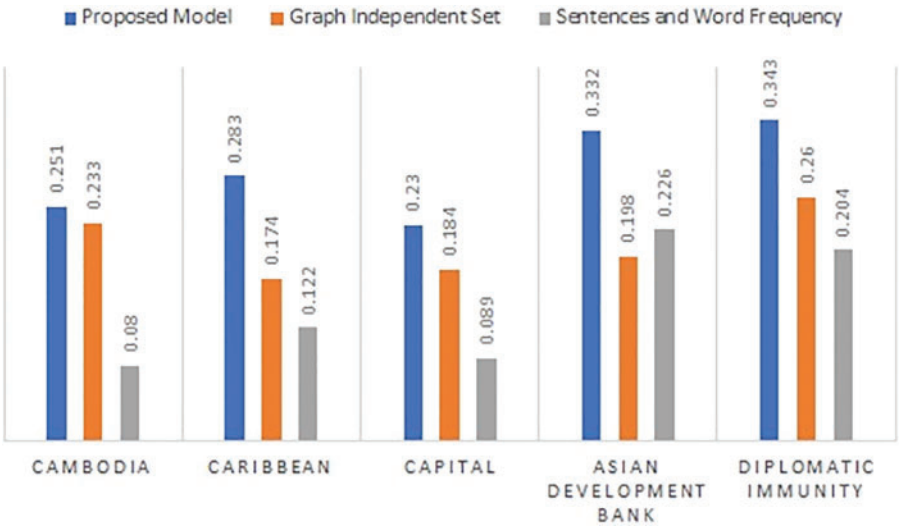

Figure 2: Precision comparison of proposed approach against graph independent approach and sentence and word frequency

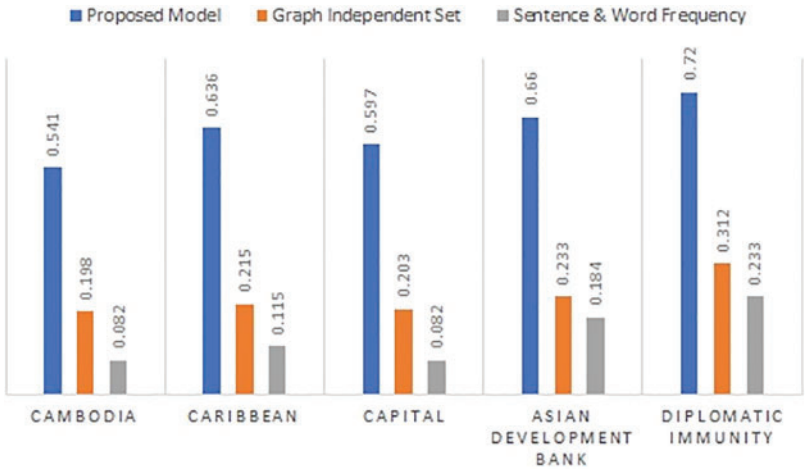

Figure 3: Recall comparison of proposed approach against graph independent approach and sentence and word frequency

The average results across different topics show that the proposed model significantly outperforms the sentences and word frequency method. Specifically, the proposed model achieves an average precision of 0.288, a recall of 0.631, and an F1-score of 0.328. In contrast, the sentences and word frequency method only achieves an average precision of 0.144, a recall of 0.139, and an F1-score of 0.136. This substantial improvement in performance metrics underscores the effectiveness of the proposed model in enhancing the accuracy and relevancy of text summarization. The superior performance of the proposed method can be attributed to its enriched feature set, which includes query-oriented functions that enhance the extraction of relevant information.

Fig. 4 shows that the approach used in this research gives better results than the previous one using graph-based and sentence extraction-based approaches. It shows a better average of 8% better results from the graph-based approach and an average of 9% better results from the sentence-based extraction approach.

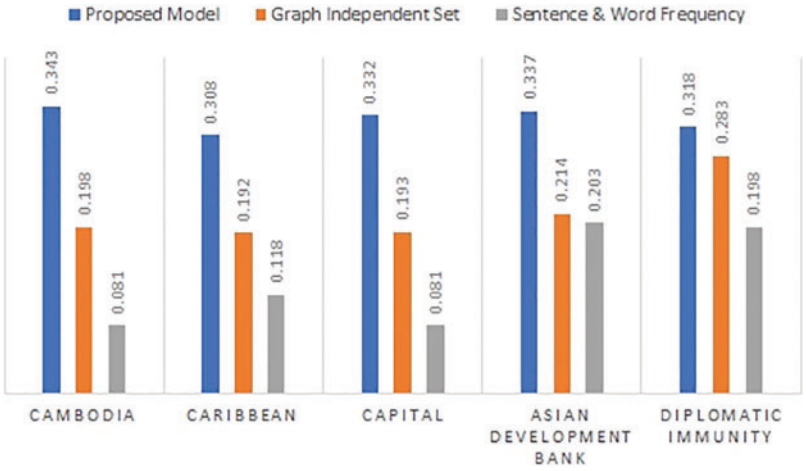

Figure 4: F-score comparison of proposed approach against graph-independent approach and sentence and word frequency

5 Conclusions

The amount of data is exponentially increasing, encompassing both web-based data and data on the local servers of various organizations. This provides data scientists with a substantial foundation to explore new techniques for meeting the market's requirements in information retrieval. This survey examines the machine learning techniques used to summarize query text. It has been observed that combining supervised and unsupervised learning algorithms to apply linguistic and statistical techniques yields better results than using a single stand-alone technique. The results show an average of 8% improvement from the graph-based approach and an average of 9% improvement from the sentence-based extraction approach. These improvements enhance the accuracy of the function, reduce calculation redundancy, and improve the identification of data through search.

6 Future Work

The study can be extended in various dimensions, such as:

- Continuing the refinement and optimization of the hybrid approach to enhance precision, recall, and F1-score metrics.
- Exploring alternative algorithms or machine learning techniques to further improve the accuracy of extractive text summarization.
- Investigating how the proposed approach can scale to manage even larger document collections, in response to the ongoing growth of big data.
- Adapting the hybrid approach to support multiple languages, addressing the challenge of summarizing texts in languages other than English.
- Conducting research on cross-lingual information retrieval and summarization techniques.

- Exploring ways to make the proposed approach more topic-sensitive, ensuring that generated summaries are tailored to specific domains or industries.
- Performing comprehensive evaluations and benchmarking studies to compare the proposed hybrid approach with other state-of-the-art summarization techniques.
- Investigating summarization techniques that can handle various types of data, such as images, audio, and video, to enable more comprehensive content summarization.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Sohail Muhammad, Muzammil Khan: Conceptualization, Methodology, Formal analysis, Investigation; Sohail Muhammad: Roles/Writing—original draft preparation. Muzammil Khan, Sarwar Shah Khan: Supervision, Validation, Writing—review and editing. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets analyzed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] G. Kaur, "Review on text classification by NLP approaches with machine learning and data mining approaches," *Int. J. Adv. Res., Ideas. Innov. Technol.*, vol. 3, no. 4, pp. 767–771, 2017.

[2] F. Ture and E. Boschee, "Learning to translate: A query-specific combination approach for cross-lingual information retrieval," in *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process.*, Doha, Qatar, 2014, pp. 589–599.

[3] A. Siva kumar, P. Premchand, and A. Govardhan, "Query-based summarizer based on similarity of sentences and word frequency," *Int. J. Data Min. Knowl. Manag. Process.*, vol. 1, no. 3, pp. 1–12, 2011. doi: 10.5121/ijdkp.2011.1301.

[4] L. Wang, H. Raghavan, C. Cardie, and V. Castelli, "Query-focused opinion summarization for user-generated content," arXiv preprint arXiv:1606.05702, 2012.

[5] D. Das and A. F. T. Martins, "A survey on automatic text summarization," in *2016 Int. Conf. Circuit, Power Comput. Technol.*, Nagercoil, India, 2016, pp. 1–31.

[6] N. Rahman and B. Borah, "A survey on existing extractive techniques for query-based text summarization," in *2015 Int. Symp. Adv. Comput. Commun. (ISACC)*, Silchar, Assam, India, Sep. 2015, pp. 98–102. doi: 10.1109/ISACC.2015.7377323.

[7] Y. Gong, "Generic text summarization using relevance measure and latent semantic analysis," in *Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, New Orleans, LA, USA, Sep. 9–13, 2001.

[8] H. Jin, T. Wang, and X. Wan, "Semantic dependency guided neural abstractive summarization," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, pp. 8026–8033, 2020. doi: 10.1609/aaai.v34i05.6312.

[9] M. Afsharizadeh, "Query-oriented text summarization using sentence extraction technique," in *2018 4th Int Conf. Web Res.*, Philadelphia, PA, USA, 2018, pp. 128–132.

[10] A. Sharaff, M. Jain, and G. Modugula, "Feature based cluster ranking approach for single document summarization," *Int. J. Inf. Technol.*, vol. 14, no. 4, pp. 2057–2065, 2022. doi: 10.1007/s41870-021-00853-1.

[11] D. Metzler and T. Kanungo, "Machine learned sentence selection strategies for query-biased summarization," in *Sigir Learning to Rank Workshop*, 2008, pp. 40–47.

[12] M. O. El-Haj and B. H. Hammo, "Evaluation of query-based Arabic text summarization system," in *2008 Int. Conf. Nat. Lang. Process. Knowl. Eng.*, Beijing, China, 2008, pp. 1–7.
[13] F. K. Jaoua and A. Ben Hamadou, "A learning technique to determine criteria for multiple document summarization," *Population*, vol. 1, no. 1, pp. 121–126, 2008.
[14] L. Zhao, L. Wu, and X. Huang, "Using query expansion in graph-based approach for query-focused multi-document summarization," *Inf. Process. Manag.*, vol. 45, no. 1, pp. 35–41, 2009. doi: 10.1016/j.ipm.2008.07.001.
[15] Y. A. AL-Khassawneh and E. S. Hanandeh, "Extractive Arabic text summarization-graph-based approach," *Electronics*, vol. 12, no. 2, pp. 437, 2023. doi: 10.3390/electronics12020437.
[16] R. Rani and D. K. Lobiyal, "A weighted word embedding based approach for extractive text summarization," *Expert Syst. Appl.*, vol. 186, no. 3, pp. 115867, 2021.
[17] S. Abdel-Salam and A. Rafea, "Performance study on extractive text summarization using BERT models," *Information*, vol. 13, no. 2, pp. 67, 2022. doi: 10.3390/info13020067.
[18] A. Joshi, E. Fidalgo, E. Alegre, and R. Alaiz-Rodriguez, "RankSum—An unsupervised extractive text summarization based on rank fusion," *Expert Syst. Appl.*, vol. 200, no. 6, pp. 116846, 2022.
[19] A. Nawaz, M. Bakhtyar, J. Baber, I. Ullah, W. Noor and A. Basit, "Extractive text summarization models for Urdu language," *Inf. Process. Manag.*, vol. 57, no. 6, pp. 102383, 2020. doi: 10.1016/j.ipm.2020.102383.
[20] Z. Fatima *et al.*, "A novel approach for semantic extractive text summarization," *Appl. Sci.*, vol. 12, no. 9, pp. 4479, 2022. doi: 10.3390/app12094479.
[21] B. Muthu *et al.*, "A framework for extractive text summarization based on deep learning modified neural network classifier," *Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 20, no. 3, pp. 1–20, 2021. doi: 10.1145/3392048.
[22] A. K. Yadav *et al.*, "Extractive text summarization using deep learning approach," *Int. J. Inf. Technol.*, vol. 14, no. 5, pp. 2407–2415, 2022. doi: 10.1007/s41870-022-00863-7.
[23] M. M. Danyal, M. Haseeb, S. S. Khan, B. Khan, and S. Ullah, "Opinion mining onmovie reviews based on deep learning models," *J. Artif. Intell.*, vol. 6, no. 2, pp. 23–42, 2024.
[24] M. Khan, M. S. Khan, and Y. Alharbi, "Text mining challenges and applications—A comprehensive review," *IJCSNS*, vol. 20, no. 12, pp. 138, 2020.
[25] S. S. Khan, M. Khan, Q. Ran, and R. Naseem, "Challenges in opinion mining, comprehensive review," *A Sci. Technol. J.*, vol. 33, no. 11, pp. 123–135, 2018.
[26] M. M. Danyal, S. S. Khan, M. Khan, S. Ullah, M. B. Ghaffar and W. Khan, "Sentiment analysis of movie reviews based on NB approaches using TF-IDF and count vectorizer," *Soc Netw. Anal. Min.*, vol. 14, no. 1, pp. 1–15, 2024. doi: 10.1007/s13278-024-01250-9.
[27] M. M. Danyal, S. S. Khan, M. Khan, S. Ullah, F. Mehmood and I. Ali, "Proposing sentiment analysis model based on BERT and XLNet for movie reviews," *Multimed. Tools Appl.*, vol. 83, no. 24, pp. 1–25, 2024. doi: 10.1007/s11042-024-18156-5.
[28] M. M. Danyal, S. S. Khan, M. Khan, M. B. Ghaffar, B. Khan and M. Arshad, "Sentiment analysis based on performance of linear support vector machine and multinomial naïve bayes using movie reviews with baseline techniques," *J. Big Data*, vol. 5, pp. 1–18, 2023. doi: 10.32604/jbd.2023.041319.