mathematics

MDPI

Article

Multi-Document News Web Page Summarization Using Content Extraction and Lexical Chain Based Key Phrase Extraction

Chandrakala Arya [1], Manoj Diwakar [2], Prabhishek Singh [3], Vijendra Singh [4], Seifedine Kadry [5,6,7] and Jungeun Kim [8,*]

1 School of Computing, Graphic Era Hill University, Dehradun 248002, India
2 CSE Department, Graphic Era Deemed to be University, Dehradun 248002, India
3 School of Computer Science Engineering and Technology, Bennett University, Greater Noida 201009, India
4 School of Computer Science, University of Petroleum and Energy Studies, Dehradun 248007, India
5 Department of Applied Data Science, Noroff University College, 4608 Kristiansand, Norway
6 Artificial Intelligence Research Center (AIRC), Ajman University, Ajman 346, United Arab Emirates
7 Department of Electrical and Computer Engineering, Lebanese American University, Byblos 13-5053, Lebanon
8 Department of Software and CMPSI, Kongju National University, Cheonan 31080, Republic of Korea
* Correspondence: jekim@kongju.ac.kr

Abstract: In the area of text summarization, there have been significant advances recently. In the meantime, the current trend in text summarization is focused more on news summarization. Therefore, developing a synthesis approach capable of extracting, comparing, and ranking sentences is vital to create a summary of various news articles in the context of erroneous online data. It is necessary, however, for the news summarization system to be able to deal with multi-document summaries due to content redundancy. This paper presents a method for summarizing multi-document news web pages based on similarity models and sentence ranking, where relevant sentences are extracted from the original article. English-language articles are collected from five news websites that cover the same topic and event. According to our experimental results, our approach provides better results than other recent methods for summarizing news.

Keywords: news web page summarization; extractive summarization; multi-document summarization; keyphrase extraction; sentence length; ROUGE; sentence ranking; similarity measure

MSC: 68T50

Citation: Arya, C.; Diwakar, M.; Singh, P.; Singh, V.; Kadry, S.; Kim, J. Multi-Document News Web Page Summarization Using Content Extraction and Lexical Chain Based Key Phrase Extraction. *Mathematics* **2023**, *11*, 1762. https://doi.org/10.3390/math11081762

Academic Editor: Florin Leon

Received: 23 February 2023
Revised: 28 March 2023
Accepted: 1 April 2023
Published: 7 April 2023

1. Introduction

With the widespread availability of the internet, more and more users are turning towards the internet for fast and easy access to various services. One prominent among these is reading news available through multiple online platforms and e-papers. Online news reading offers many advantages over conventional media. As many online news sources can be accessed freely, many people can now easily access these sites. The additional benefit of reading online news could be that the reader is often given categorized news articles and a comprehensive summarization of news articles, which is possible due to the gaining momentum of automatic text summarization research. The goal of text summarization is to present the essential information of the original text in a concise form while keeping its main content [1,2]. It helps the user to understand the large volume of information quickly, determine what the document is about, and avoid reading the document itself.

In earlier research [3,4], automatic text summarization techniques are classified into two classes: extraction and abstraction. Abstractive approaches are domain-specific and may contain words that may not be present in the original document, whereas an extractive approach usually identifies the main concepts in the document and is, therefore, more robust and practically easier to implement [5]. The extractive summarization is the main

emphasis of this paper. Classification of news articles is an important phase of this problem. Numerous news websites published articles by categorizing them into national, International, and sports classes. News articles from different sources often describe the same event from other aspects, and users often compare these articles. Therefore, a news summarization system is required to collect news articles that describe the same topic from various sources. This paper uses online news articles about the same topic for multi-document summarization. One of the major problems in the multi-document summarization task is the identification of similarities and differences across documents. We assume that a user can access various news stories that belong to a similar subject, but nobody has the time to look at every story. Therefore, to be kept up to date on the subject, a user goes through the details only when the reported story is interesting.

News summarization is considered a document summarization method that extracts multiple significant news events and summarizes them for readers [6]. The key purpose of summarization is to present the central concepts of a document in no time. Currently, summarization comprises single-document summarization and a variety of multi-document summarization approaches [7–11]. This paper summarizes news from multiple sources related to the same topic. Correlated content across news articles shows a high degree of content redundancy. Therefore, identifying and utilizing the correlation among the documents is important for news summarization. A couple of news documents take numerous forms of correlation. For example, they could describe a similar topic or related events from a different perspective.

We recommend a generic news summarization technique based on keyphrases-based extractive summarization. First, automatically extracted keyphrases from news articles are used to assess the importance of each sentence in summary. Sentences are extracted to produce extraction summaries by assigning some score to a sentence for the summary, called sentence weighting, and then choosing the high scorer sentences to form the documents summary. Third, we use the cosine similarity measure to reduce the redundancy of sentences in summary. The purpose of the study is to show that the result of the summarization is based not only on the core content or sentence extraction and similarity model but also on keyphrases, sentence position, and sentence length.

Research Contribution:

i. We have tried to develop a news web page summarization model for Indian news websites (English) to summarize news articles from different sources.
ii. Classification is the important phase in this work, and the news web page classification approach correctly classifies the news web pages from non-news web pages; correct classification of news web pages is important for news summarization. Important content has been extracted based on the Tag tree, efficiently extracting meaningful information from news articles.
iii. Keyphrases have been extracted in this work, giving brief and precise information about the article. Accurately extracted keyphrases play an essential part in news summarization. Therefore, our approach used lexical chain-based keyphrase extraction and showed better results.
iv. For precise summarization, sentence selection and ranking are important. The results reflect the suitability and appropriateness of the approach.

This paper also creates a dataset based on the same event containing five articles from different sources. Our dataset collects all subsequent articles that discuss the same event used for news summarization and will eventually help in the performance evaluation.

2. Related Work

Research on document summarization begins very early by Luhn in 1958 and Edmundson in 1967 and becomes one of the traditional topics in the natural language processing research. Previously various papers about document summarization have been recommended [3,12,13].

This literature mainly tries to address extractive multi-document news summarization. A considerable number of studies have been carried out in this field. Some of the important work has been illustrated in this section.

Early work on news summarization can be dated back to 1990s when the SUMMONS summarizer was created [14]. SUMMONS was designed for summarizing news articles on a single event such as terrorist events. After that, numerous studies were performed on this field.

The authors in ref. [15] proposed an extraction-based multi-document summarization which used maximal marginal relevance multi-document (MMR-MD) metric for reducing redundancy and achieving high compression ratios in generated summaries. Their approach is different from other approaches as it is completely domain-independent and depends upon fast statistical processing to maximize the novelty of the information that had been selected. In ref. [16], researchers used diverse summarization approaches dependent on the type of documents in the input set to develop a multi-document summarization system. In their system, for the automatic identification of the input set of documents, a router is used, which is also invokes the appropriate summarization subcomponents. Their system performs well on summary content as compared to other systems; it is ranked third or fourth with different systems ranked ahead of it for each analysis. The authors in ref. [17] used an ontology-based fuzzy event extraction agent for Chinese news summarization. In their work, for testing the performance of their summarization agent, they constructed an experimental website at Chang Jung University. Experimental results show that their approach can effectively summarize the Chinese weather e-news retrieved from the China Times website. In ref. [18], authors used keyphrase extraction methodology to develop the LAKE System. They used linguistic features for identifying relevant terms in the document. The generated summaries considered both the relevance and the coverage of keyphrases for a certain topic. Their experimental results show an average responsiveness and high linguistic quality of the summaries. However, their obtained results are very competitive to the pyramid metric. The researchers in ref. [19] used extractive methods for document summarization; they designed their system based on keyphrase extraction from the documents and select the sentences in the resultant summary. Their system gives a high-quality compressed summary. In ref. [20], authors proposed an optimized generic extractive Arabic and English multi-document summarization technique for summarization, which used a translation summary machine. Their approach uses cluster size and selection model as parameters in extractive summarization process. The experimental results show that performance of their summarization system is good in comparison with other top performing systems at DUC (Document Understanding Conference)—2002. In ref. [4], researchers used an Integer linear programming for multi-document summarization technique that jointly maximizes the significance of the sentences in the summary and their diversity beyond a maximum permissible length of the summary. Their findings show that the approach can attain better results. In ref. [21], researchers proposed a bigram-based supervised method for extractive document summarization and used the ILP (Integer Linear Programming) method as a core component. Their experimental results show that the improvement in system performance depends on the supervised bigram estimation module that successfully gathers the important bigram and gives them appropriate weights. Authors in ref. [22] proposed the SRRank algorithm for multi-document summarization. They used a graph-based ranking algorithm based on semantic role information. Their algorithm used a heterogonous ranking process to rank sentences, semantic roles, and words. They used DUC (Document Understanding Conference) datasets for the experiment and show that SRRank outperforms a few baselines approaches. Authors in ref. [23] developed a ranking framework to rank sentences for a multi-document summarization system based on recursive neural network [R2N2]. It transforms the sentence ranking task into a hierarchical regression process by using the recursive neural networks model. They designed an optimized sentence selection method based on the words and sentences ranking scores. They conduct experiments on the DUC benchmark; experimental results show that their

model achieves a higher ROUGE score than the previous summarization approaches and makes much more accurate predictions than traditional support vector regression. Ref in [24] presented mover's distance metric (WMD), in conjunction with semantic-aware continuous space representation of words, and has been proposed to accurately estimate the similarity degree between a pair of documents for effective use of the summarization process. They investigate their approach to other state-of-the-art approaches and show the effectiveness of their approach over other summarization frameworks. In ref. [25], authors proposed a cat swarm optimization (CSO)-based multi-document summarization model to create a standard extractive summary. The performance of their summarizer shows a better ROUGE score, F score, sensitivity, positive predictive value, and summary accuracy on a DUC dataset. Researchers in ref. [26] proposed a multi-objective artificial bee colony (MOABC) algorithm for extractive multi-document summarization. Their approach shows improved ROUGE-2 and ROUGE-L scores and less dispersion of around 13 and 6 times more robust than the other comparable approaches. In ref. [27], researchers used KUSH text processing tool to realm the semantic cohesion between sentences from the text document. They used the concept of maximum independent set for extractive, generic text document summarization. Researchers in ref. [28] used the MIRANEWS dataset for single-document news summarization. They introduced a new job called multi-resource-assisted news summarization to produce a summary that includes main article events. Their evaluation metrics confirm that introduced assisted documents provide better grounding than the reference summaries. Authors in ref. [29] presents an OntoRealSumm for real time tweet summarization. They used a three-phase approach for challenge handling such as classification quality improvement and found out the importance of each class. They also ensured that the final summary included information diversity and coverage for each class. Their results show the increase of 6–42% in Rouge-N1-score than the existing work. Chao Zhao et al. in ref. [30] proposed an approach where an extractive multi-document news summarization problem is reformulated by concatenating all documents as a single meta-document. They reorder the documents according to the order of meta-document salience. Their approach outperforms previous state-of-the-art methods. In ref. [31], the author proposed a textual graph-based model for comprehensible summary generation for Arabic. They used Essex Arabic Summary Corpus (EASC) dataset and achieved an F-score of 0.617 using a ROUGE-2 performance metric. Tianyi Zhang et al. in ref. [32], used two popular news summarization benchmarks for the evaluation of ten large language models (LLMs). They perform human evaluation over high-quality summaries collected from the freelance writers. Their findings highlight the role of good reference summaries in both summarization model development and evaluation. Andrea Pozzi et al. in ref. [33] proposed a methodology for news summarization related to cryptocurrencies that helps in the financial sector. They perform their experiment on 22,282 news articles, and their findings show that 86.8% of the examined summaries were considered as coherent and 95.7% of the articles were summarized correctly.

Conclusive Findings

In the past, the major research was stressed upon single-document summarization [34,35]. Recently, effort transferred to multi-document summarization [4,14,36]. Our literature review indicates that many significant studies have been performed on multi-document news web page summarization. However, more in-depth analysis is required to improve the performance of such systems to match end-user expectations.

In this research, extraction-based multi-document summarization is used for news summarization. Unlike previous approaches, this work combines different phases such as web page classification, content extraction, keyphrase extraction, and finally an extraction-based method, which is used to calculate the saliency score of each sentence and then rank the sentences in the document.

3. Methodology

News web summarization is an ideal solution to provide condensed, informative document reorganization for faster and better representation of news evolution. There have been few existing systems developed for news summarization, but little effort has been made on the combination of supervised algorithms-based classification, content extraction, and keyphrase extraction for summarization. In this work, we demonstrate the outcome of this combination for the summarization of Indian news web pages based on the similar event. This section describes the framework of the proposed system in detail. It comprises four features, i.e., news web page classification, content extraction, keyphrase extraction, and sentence selection as shown in Figure 1. The input is a collection of documents, which are classified into news and non-news web pages in the news web page classification phase. The system extracts the news article content from the news web pages in the next phase. Each document covers one or more keyphrases and tries to pick sentences that cover keyphrases with respect to summary length. Then, it extracts significant and non-redundant keyphrases in order to select sentences from the news document. This phase generates a set of sentences containing keyphrases. Now, the weight of each sentence (discussed in Section 3.4.4) is computed, which is used for sentence ranking. In the next step, redundancy is reduced by similarity computation using cosine similarity discussed in Section 3.5. After eliminating the redundant sentences, we select the final sentences for the summary.

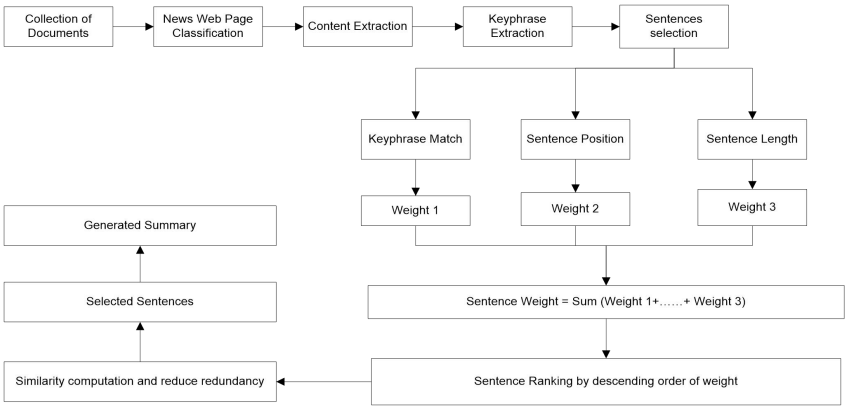

Figure 1. Architecture of news summarization system.

3.1. News Web Page Classification

In news web page filtering and summarization, the task of news web page classification has remained in sharp focus for a long time. In the work of news web page classification, an automatic recognition method has been used based on classification rules for web news based on a combination of content, structure, and uniform resource locator (URL) attributes using a Naïve–Bayes algorithm discussed in our earlier work [37]. This phase classifies a news web page from a non-news web page. Correctly classified news web pages are used for the content extraction task.

3.2. Extracting Article Content

For news summarization, it is important to extract the actual news content from the news web pages because it not only contains the actual news content but also some noisy content such as advertisement, comments, and branding banners, etc. In this paper for extracting article content, we use the content extraction approach from our prior work [38]. Our previous approach to extracting content from news web pages used the concept of tokenization of HTML pages; web pages are parsed into tag tree and corresponding template is produced to determine matching patterns and multiple sequence alignment. The relevant information is extracted from the web pages by finding and removing shared tokens.

3.3. Keyphrase Extraction

Keyphrases are extracted from the sentences of the news articles. The set of keyphrases is essential to analyze because human summarizers have put their effort collectively but independently to read the news articles and extract its keyphrases that contribute to the summarization. In this work, we introduce a technique to summarize the news documents by extracting keyphrases that cover the major contribution of the target documents. The proposed methodology has been to find the set of keyphrases that signify the main units of data and find the best set of sentences that cover more relevant information. Here, the keyphrases are extracted using the lexical chain as we discussed in our previous paper [39]. According to our previous work, lexical chain is created by taking a new phrase and finding a related chain for it according to lexical cohesion. These keyphrases are the key elements in our summarization.

3.4. Sentence Selection and Ranking

Generally, web pages contain diverse content, so to summarize the entire web page as one unit is not a good idea. Rather, we believe it is best to select the sentences from the articles that are more significant. To generate a summary, highly ranked sentences are selected, which are different from each other and cover the article's main content with less redundancy.

Our goal was to find the sentence rank when making summaries of news articles. We use three kinds of features for sentence ranking including the direct keyphrase match, sentence position, and sentence length, and by using a cosine similarity model, we reduce redundancy and select sentences.

3.4.1. Direct Keyphrases Match

Keyphrases are used to evaluate the sentence importance. After extracting the set of keyphrases for each document, the main task is to pick sentences for each document that cover most significant and non-redundant keyphrases. Basically, keyphrases that have been repeated in more sentences are more important and could represent a more important keyphrase. Therefore, sentences that comprise more recurrent keyphrases are more important. The approach of keyphrase extraction is discussed in the previous Section 3.3.

We score the sentence by direct keyphrase match. Those keyphrases that occurs in two or more sentences are more important than others. We calculate the direct keyphrase match by the following formula shown in Equation (1)

$$\text{Direct Keyphrase match} = \text{When two or more sentences were containing same keyphrases.} \quad (1)$$

$$\text{Kmatch} = \text{KN}/\text{T (S)} \quad (2)$$

where Kmatch denotes the direct keyphrase match in the document set. KN denotes the number of times a keyphrase occurs in the document set. T (S) denotes the total number of sentences in the document set. Direct keyphrase score of the six keyphrases are shown in Table 1.

Table 1. Keyphrase score.

Keyphrases	Score of Direct Keyphrase Match
K. Srikant	38
Australian open	14
Chen Long	6
Shuttler	8
Olympic	6
Badminton	3

3.4.2. Sentence Position

Sentence position [5] is a simple and effective feature for the news summarization. The perception is that leading sentences in the news article contain summarizing information. We used the positional information; occurrence of a sentence in the document whether the sentence 's' occurs very early or very late in a document, boosts the top sentences of an article.

$$\mathrm{SP(s)} = 1 - \frac{\mathrm{P}}{\mathrm{N}} \tag{3}$$

where N is the total number of sentences in the articles and P is the position of the sentence 's' in the article.

3.4.3. Sentence Length

Sentence length [5] is a binary feature that helps in reducing the noisy short text in the summary. It checks if the sentence contains at least 10 words. The sentences below the given limit of 10 words will be ignored to generate the summary.

$$\mathrm{SL(s)} = \begin{cases} 1 & \text{if len(s)} \geq 10 \\ 0 & \text{Otherwise} \end{cases} \tag{4}$$

3.4.4. Sentence Weight

According to the above four features, we compute the final significant score of a sentence by specifying a certain weight for each kind of feature, as shown in Equation (5)

$$SentenceWeight = \lambda_1 keyphrasematch + \lambda_2 SentencePosition + \lambda_3 SentenceLength \tag{5}$$

where λ_1, λ_2, λ_3 represent the weight parameters of the three kinds of features. The value of these parameters is lying between 0 and 1 according to the features. If the keyphrase match is maximum or an exact match, then we set the value at '1' for λ_1. If the sentence position is higher or lies in the top five, then λ_2 is set to '1'; otherwise, it is set to '0.5'. If the sentence length is greater than 10, then λ_3 is set to '1'; otherwise, it is set to '0'.

After weight computation, we use Algorithm 1 for sentence ranking. It accepts sentence set (SS) as input and produces ranked list of sentences in descending order.

Algorithm 1: Sentence Weight Selection Algorithm

Input: Sentences in Sentence Set (SS)
Output: A list of sentences sorted by descending order of their weight (LS).

```
1.  begin
2.      for each sentences S in SS
3.          if (length (S) < 10)
4.              Remove S into SS
5.          else
6.              Add S in SS
7.          ranking (SS, SenWeight)
8.      end for
9.  Output LS
10. end
```

In this algorithm, sentences are ranked in main loop run from (step 2 to 9). We have fixed the length of the sentence by 10 words, and only those sentences that satisfy this condition are included in the sentence set (SS) to calculate the score; otherwise, it has been removed from the sentence set.

3.5. Similarity Model for Reduce Redundancy

News summarization tasks face the major problem of identifying the similarity and differences across news articles. The reduction of similarity or redundancy is a difficult task

because the characteristics of discrete sentences are reliant on other sentences included in the summary. The main idea in our research is to find out the similarity of the news articles belonging to the same event where several sentences may have a substantial information overlap. Some researchers used clustering to obtain groups of similar sentences [8], while in this work, finding similarity among sentences depends on keyphrases in the news articles that report about the same event and link the similar keyphrases together. Summarizers can identify similarities and differences among documents that can eliminate redundant information across documents to provide a concise summary.

The similarity measure is usually based on matching keyphrases only. We used keyphrase-based similarity because different articles use different styles for writing the same event, and articles' sentences are not same line by line. In this step, similarity among keyphrases in the news articles are found that report about the same events and link the similar keyphrases together. Our approach uses TF-IDF, phrase distance, and lexical chains to identify the several keyphrases that convey approximately the same information discussed in Section 3.3 and find the similarity among the same events. We analyze the news stories written in English. We have limited our focus to the textual contents of the articles. Thus, pictures and other multimedia are rejected. News articles from different news websites described the same event in different aspects; users often compare articles from different sources. Therefore, news articles are gathered from different news websites and link articles describing the same event. News articles are updated frequently, and their descriptions are overlapped in series of news articles. Therefore, by removing duplicate descriptions, the user can obtain efficient information. We have to summarize different news articles on the same event in a single extract. It is far from clear that sentence scores from different news articles should be comparable.

If the two sentences contain similar keyphrases, their approximate feature values may probably be the same, so the approximate scores of the sentences also may probably be the same. Therefore, the summary extracted by this method may include similar high-score sentences, and this will lead to redundant information in the summary.

Therefore, we need to remove sentences that are redundant to others in articles on an event. Minimizing redundancy between passages and the selection of most significant sentences are the main concerns in summarization. For redundancy identification of news articles, we use cosine similarity to measure the similarity between documents [40]. First, sentences are selected based on the keyphrase extraction. After that, similarity in sentences is calculated with the help of a cosine similarity formula. Our task is to assign a score to every sentence that indicates the importance of that sentence in the summary. We use Algorithm 1 to select the sentences. The formula of cosine similarity is shown in Equation (6).

$$\text{Cos}\theta = \frac{\text{d1.d2}}{|\text{d1}||\text{d2}|} \tag{6}$$

Our system collects news articles that need to be summarized. First of all, article collections are spit into sentences in such a way that they are indexed by a letter and a number combination. The letter shows the corresponding document, and the number indicates the sentence position within its respective document [41].

We collect the five news articles from different sources that belong to the same event. We take one sentence from each document. We consider one sentence until the first full stop occurs. Figure 2 shows the example sentences from five documents.

D1 s1= India shuttler **Kidambi Srikanth** defeated Olympic champion Chen Long in a dominating fashion to win the Australia Open Super Series men's singles title on Sunday.

D2 s1= The **Badminton** Association of India announced a cash award of Rs. 5 lakh to **Srikanth** for clinching the **Australian Open Super Series** title in Sydney on Sunday.

D3 s1= **Kidambi Srikanth** has resurfaced on Indian **badminton**'s horizon since the surprise of a win over Lin Dan in the China Open final in November 2014.

D4 s1= Indian **shuttler Kidambi Srikanth** notched up his second successive Super Series title with a stunning straight-game triumph over reigning Olympic champion **Chen Long** in the **Australian Open summit** clash in Sydney on Sunday.

D5 s1= **Kidambi Srikanth** is enjoying the best phase of his career.

Figure 2. Example sentences from five documents.

Table 2 shows the five news documents first sentence matching according to keyphrases.

Table 2. The five news documents' first sentence-matching according to keyphrases.

Documents	Sentence	Keyphrases
D1	India shuttler **Kidambi Srikanth** defeated Olympic champion Chen Long in a dominating fashion to win the Australia Open Super Series men's singles title on Sunday.	India, shutttler, Kidambi Srikant, Olympic champion, Chen Long, Australia open super series, title, sunday
D2	The **Badminton** Association of India announced a cash award of Rs. 5 lakh to **Kidambi Srikanth** for clinching the **Australian Open Super Series** title in Sydney on Sunday.	Badminton, India, Kidambi Srikant, Australian open super series, title, Sydney, Sunday
D3	**Kidambi Srikanth** has resurfaced on Indian **badminton**'s horizon since the surprise of a win over Lin Dan in the China Open final in November 2014.	Kidambi Srikanth, Indian, Badminton
D4	**Indian shuttler Kidambi Srikanth** notched up his second successive Super Series title with a stunning straight-game triumph over reigning Olympic champion **Chen Long** in the **Australian Open summit** clash in Sydney on Sunday.	Indain, shuttler, Kidambi Srikanth, Olympic, Chen Long, Australian Open, Sydney, Sunday
D5	**Kidambi Srikanth** is enjoying the best phase of his career.	Kidambi Srikanth, career

Similarities of documents are shown in Table 3.

Table 3. Similarity measure.

	D1 S1	**D2 S1**	**D3 S1**	**D4 S1**	**D5 S1**
D1 S1	1	0.71	0.43	0.93	0.38
D2 S1	0.71	1	0.65	0.8	0.38
D3 S1	0.43	0.65	1	0.41	0.59
D4 S1	0.93	0.8	0.41	1	0.36
D5 S1	0.38	0.38	0.59	0.36	1

The above similarity measure table is calculated by the cosine similarity formula mentioned in Equation(6). All the extracted sentences are scored from highest to lowest, and the sentences are arranged according to the score; high-scorer sentences are ranked high, then the lower scorer. For the summary generation, sentences are selected iteratively. Every time the selected sentence is compared with the existing sentences in the summary, so whenever candidate sentence is not much similar to the existing summary sentences or

if the similarity measure of both the sentences is lower than a given threshold value, the sentence is considered as selected for the summary.

Table 3 shows the similarity among the first sentence of five documents. In cosine similarity, '1' denotes the exactly similar sentence; these sentences can cause redundancy in the summary, and therefore are removed from the final summary. Exactly different sentences are denoted by the '0' in the cosine similarity; such types of sentences also do not contribute to the summary and are excluded from the final summary. Based on the example document similarity measure computed in Table 3, we set the threshold value of cosine similarity is 0.65. We explicitly choose this threshold value because in the summary, non-matching sentences are not relevant in the summary and matching sentences produce redundancy, so we choose the value that is not so far and not so close to 1. When two sentences are approximately similar (that is close to 1), the one with the higher weight is selected for the summary.

3.6. Generate Summary

The summary is generated by extracting top-ranking sentences according to scores assigned to the sentences. However, to reduce redundancy, we use a cosine similarity model [40]. A sentence is selected for summary generation if it obtains the highest rank and is not too similar to any sentences existing in the summary. To determine similarity between sentences, we use cosine similarity at threshold t = 0.6.

The following Algorithm 2 describes the summary generation strategy in our system; in this algorithm, the output of Algorithm 1 is considered as the input for this algorithm.

Algorithm 2: Summary Generation Algorithm

Input: Multiple news articles on the same topic.
Output: Final summary of news articles.

```
Begin
    Main news content extracted from the document
    Keyphrase extracted from the main news content
    sentence weight computed as per the Equation (5)
    Sort sentences in descending order of weight using algorithm 1
    redundant sentences removed based on cosine similarity
    Finally selected sentences are used in summary generation.
    Output summary
END
```

4. Experimental Dataset

In order to evaluate our work, we collect the news articles from five different news websites, named as *The Hindu*, *The Times of India*, *Hindustan Times*, *Indian Express*, and *Deccan Harald*. Ten events that occurred between 15 May 2017 to 16 June 2017 were manually selected by these five news websites. Each event contained more than five interlinked articles. The news topics are collected from different categories. Keyphrases from all the related documents are extracted by the annotator. It consists of ten different categories of news articles, namely Market, Business, India, Technology, National, Science and Environment, Politics, World, Entertainment, Sports. Each category has two sets, and each set contains five news articles on the same topic related to the individual category, i.e., a total of 10 articles per news category, so a total of 100 news articles have been included in the dataset. All news articles are collected from five famous Indian news websites (English) such as *The Economic Times*, *The Hindu*, *The Times of India*, *Hindustan Times*, and *Indian Express*. The number of sentences contained in these news articles ranges from 10 to 60. We manually constructed the 150-word summary as a reference summary. Details of this dataset are given in Table 4.

Table 4. Analysis of dataset.

S. N.	Topic	Number of Articles Set	Number of Articles in Both Sets	Average Number of Sentences per Article Set	Average Number of Words per Article Set
1	Market	2	10	42	464
2	Business	2	10	36	514
3	Sports	2	10	49	593
4	India	2	10	34	554
5	National	2	10	39	667
6	Technology	2	10	45	385
7	World	2	10	53	753
8	Politics	2	10	47	581
9	Entertainment	2	10	25	497
10	Science & Environment	2	10	37	209
11	Total	20	100	40	632

5. Experimental Results and Evaluation

We performed our experiments on a machine that was equipped with an Intel Core i7 processor that ran at 1.80 GHz, had 8 GB RAM, Windows 10 pro 64 bit. We used the ROUGE tool to analyze the performance of our approach. ROUGE has been tested for extraction-based summaries with a focus on content overlap [19,42,43]. In this work, the extraction method is used for the summarization, which is why we used this tool for the analyses.

ROUGE (recall-oriented understudy for gisting evaluation) [44] has been used as an automatic evaluation method and it based on the similarity of n-gram. Automated machine summaries can be compared with reference summaries (human summaries) using the ROUGE summarization evaluation tool. It is one of the standard ways to compute the effectiveness of auto-generated summaries by comparing it to a set of reference summaries that is typically produced by the human. In this work, we also compare the reference summary sentences with system summary sentences using ROUGE metrics.

There are several metrics within the ROUGE, and the most widely used are ROUGE-1, ROUGE-2, and ROUGE-SU4, and these three metrics are used in this work. ROUGE-1 and ROUGE-2 calculates the unigram and bigram overlap among the computers generated and reference summaries, whereas ROUGE-SU4 computes the intersection of the skip bigram with up to four superseding terms.

The collection of two sets of documents belonging to the same event has been considered as the system input. The generated output contained the condensed and concise summaries of the input documents. A good evaluation measure should assign a good score to a good summary and poor score to a bad summary.

Computation of ROUGE-1, ROUGE-2, and ROUGE-SU4 values for the system and reference summary sentences has been described in Table 5. Example sentences have been taken from the documents (D1, D2, D3, D4, and D5) of our dataset discussed in Section 4.

Table 5. Example of system and reference summary sentences.

System Summary Sentence	Srikanth defeated Olympic champion Chen Long in a dominating fashion to win the Australia Open Super Series men's singles title on Sunday.
Reference Summary Sentence	India shuttler Kidambi Srikanth defeated Olympic champion from China Chen Long in a dominating fashion to win the Australia Open Super Series men's singles title on Sunday.

According to the above example shown in Table 5, the values of ROUGE -1, ROUGE-2, and ROUGE-SU4 are computed as:

ROUGE-1 denotes the intersection of unigrams between the system summary and reference summary. In the above example, there are 22 words in the system summary, which matched with words of the reference summary. The formula for determining the ROUGE-1 value can be demonstrated as follows:

$$\text{ROUGE} - 1 = \frac{\text{Matching unigrams in system and reference summary}}{\text{Total number of unigrams in reference summary}} \quad (7)$$

ROUGE-1 = 22/27 = 0.815= 81.5%.

ROUGE-2 denotes the overlap of bigrams between the system summary and reference summary.

$$\text{ROUGE} - 2 = \frac{\text{Matching bigrams in system and reference summary}}{\text{Total number of bigrams in reference summary}} \quad (8)$$

ROUGE-2 = 11/13 = 0.846 = 84.6%

ROUGE-SU4 is considered as a comprehensive version of ROUGE-2 that allows maximum 4-length word-level gaps between the bigram [39].

$$\text{ROUGE} - \text{SU4} = \frac{\text{SKIP 4 (System Summary, Reference Summary)}}{\text{Reference Summary, 4}} \quad (9)$$

ROUGE-SU4 = 5/6 = 0.833 = 83.3%

Further, to have testing of our proposed approach, we take the articles from different categories to obtain the ROUGE values for each of these categories separately. Results are shown in Table 6.

Table 6. All three ROUGE values for different type of categories.

Categories		Sentences	Words	ROUGE-1	ROUGE-2	ROUGE-SU4
Market	Set 1	28	382	73.09%	72.03%	70.23%
	Set 2	42	537	84.21%	82.29%	85.53%
Business	Set 1	37	502	77.77%	79.94%	76.24%
	Set 2	39	522	83.13%	81.32%	79.67%
Sports	Set 1	34	457	73.27%	76.55%	72.54%
	Set 2	37	489	81.05%	80.78%	81.58%
India	Set 1	49	695	95.99%	92.67%	94.04%
	Set 2	43	549	90.34%	89.21%	91.55%
Technology	Set 1	18	277	69.11%	64.28%	65.79%
	Set 2	32	435	78.98%	74.69%	77.98%
National	Set 1	41	529	89.98%	88.57%	89.48%
	Set 2	36	472	77.26%	78.14%	76.57%
Politics	Set 1	43	553	91.89%	87.85%	91.09%
	Set 2	42	542	89.20%	85.34%	88.66%
World	Set 1	27	384	72.76%	71.53%	71.66%
	Set 2	35	463	86.72%	83.31%	85.98%

Table 6. *Cont.*

Categories		Sentences	Words	ROUGE-1	ROUGE-2	ROUGE-SU4
Entertainment	Set 1	46	569	94.32%	91.92%	93.38%
	Set 2	38	511	88.86	86.02%	87.66%
Science & Environment	Set 1	32	399	76.01%	79.23%	77.36%
	Set 2	21	192	70.41%	69.85%	66.76%

In the above table, the first column contains the ten different categories of news articles (as already discussed). Second column contains the two set of news articles for each particular category and their corresponding sentences and words are given in the third and fourth columns, respectively. The last three columns show the ROUGE-1, ROUGE-2, and ROUGE-SU4 values, respectively.

From the above results, we can say that better results were observed in the categories having a large number of sentences and words size, and poor results are found for those categories where sentences and words are small in size. Like in the India category, set 1 contains the highest value of sentences and words at 49 and 695, respectively, and therefore shows the highest ROUGE-1, ROUGE-2, and ROUGE-SU4 values as 95.55%, 92.67%, and 94.04%, respectively. Meanwhile, in the Technology category, set 1 contains the lowest value of sentences (18) and words (277), and hence shows the lowest ROUGE-1, ROUGE-2, and ROUGE-SU4 values as 69.11%, 64.28%, and 65.79%, respectively.

6. Comparative Evaluation of Proposed Approach with Other Baseline Approaches

Out of various approaches available, we choose the three baseline approaches—LAKE, TSES, and SRRank—for comparison with our proposed approach. To know the accuracy of our proposed approach, we used the same dataset for all the approaches. The brief description of baseline approaches is:

LAKE (linguistic analysis-based keyphrase extractor) [18]: LAKE is a multi-document summarization approach for DUC-2005. This approach used the concept of keyphrase extraction as an important calculation for summarization. The selection of significant keyphrases from the documents has been performed by the machine learning framework. Generated summaries contained relevant information and important keyphrases of the document.

TSES (text summarization extraction system) [19]: This approach extracts important keyphrases to select the important sentences. Each sentence is ranked according to the specified features and extracts the highest-ranking sentence to generate the final summary. TSES generates summaries in four steps; firstly, it removes stop words and assigns a POS tag for each word in the document. In the second step, it extracts important keyphrases from the document and ranks them by implementing a new algorithm. In the next step, sentences are ranked according to the extracted keyphrases, and in the final step, the amount of the candidate sentences in the summary is reduced in order to produce a qualitative summary using KFIDF measurement.

SRRank (semantic role rank) [22]: An extractive multi-document summarization system. It uses semantic role information to develop multi-document summarization, and a saliency score of all sentences are obtained by greedy algorithms for sentence selection.

The reason for using these approaches as our baseline is that both LAKE and TSES used the keyphrase-based approach for their experiments, and we also have keyphrases as an important feature in our approach. The SRRank incorporates the semantic role information into the graph-based ranking algorithm, and we also used semantic role information for lexical chain construction.

In the proposed approach, keyphrase is identified as an important feature for sentence ranking. For sentence ranking other than keyphrase extraction and semantic role information, we used additional features besides the baseline approaches such as direct keyphase match, matching terms, sentence position, and sentence length. We also used redundancy

reduction, which helps in sentence ranking and minimization of redundancy for the news summarization. Experimental results show that these combinations of features give better results than other baseline approaches.

Further, a comparative analysis of our results with other baseline approaches would help us to understand the overall performance of the proposed approach with the other popular approaches, i.e., SRRank, TSES, and LAKE. In this work for comparison, we re-implement all three baseline approaches on our dataset, having ten different categories containing two set each.

Table 7 shows the ROUGE-1, ROUGE-2, and ROUGE-SU4 score for each set from ten different news websites. From the table, results show that, on average, the proposed approach performs better than the other three baseline approaches for all ROUGE values. Further, in Table 8 the actual improvement in performance of our approach compares with each of the three baseline approaches for ROUGE-1, ROUGE-2, and ROUGE-SU4 values. Overall, the proposed approach performs better than other baseline approaches. Among the three approaches, LAKE is the strongest, and it can outperform the other two TSES and SRRank approaches. By the analysis of the results, we can also say that the performance of the proposed approach is affected by the size of the number of sentences and number of words. That category that contains a large number of sentences and words shows good results; otherwise, it shows poor results like in Market, Technology, and World.

To evaluate the effectiveness of the proposed approach, a survey was conducted. The size of the dataset is 370, collected from the Google form. The Google form was circulated to the experts of the related fields. Experts rate our approach with the other baseline approaches based on criteria relevance, coherence, and informativeness. Experts rate the results from 1 to 5, where 1 represents poor, 2 represents average, 3 represents good, 4 represents very good, and 5 represents excellent. According to Table 9, experts also rate our proposed approach at a higher ranking than the other baseline approaches.

Table 7. Experimental evaluation of dataset.

		ROUGE-1				ROUGE-2				ROUGE-SU4			
		Proposed Approach	**SRRank**	**TSES**	**LAKE**	**Proposed Approach**	**SRRank**	**TSES**	**LAKE**	**Proposed Approach**	**SRRank**	**TSES**	**LAKE**
Market	Set 1	73.09%	69.28%	71.29%	74.34%	72.03%	68.77%	69.11%	72.98%	70.23%	68.27%	69.93%	71.04%
	Set2	84.21%	79.95%	80.67%	82.47%	82.29%	79.86%	80.05%	81.03%	85.53%	79.98%	80.87%	83.18
Business	Set 1	77.77%	74.86%	76.24%	75.65%	79.94%	74.83%	75.89%	78.44%	76.24%	72.59%	73.99%	75.37%
	Set 2	83.13%	78.02%	79.78%	81.99%	81.32%	77.04%	77.94%	79.82%	79.67%	74.78%	75.67%	78.02%
Sports	Set 1	73.27%	71.64%	69.97%	72.99%	76.55%	73.86%	74.65%	75.23%	72.54%	70.75%	71.78%	71.98%
	Set 2	81.05%	77.87%	78.06%	80.56%	80.78%	76.01%	77.21%	78.35%	81.58%	76.24%	79.67%	80.99%
India	Set 1	95.55%	89.47%	88.88%	92.87%	92.67%	88.59%	89.12%	90.43%	94.04%	89.48%	88.97%	91.79%
	Set 2	90.34%	87.09%	89.17%	89.88%	89.21%	83.25%	85.87%	86.98%	91.55%	88.78%	86.08%	89.23%
Technology	Set 1	69.95%	65.52%	67.48%	70.15%	64.28%	62.82%	63.89%	65.21%	65.79%	63.15%	62.24%	66.03%
	Set 2	78.98%	74.03%	75.44%	77.89%	74.69%	72.02%	73.15%	74.06%	77.98%	73.36%	73.01%	76.97%
National	Set 1	89.98%	84.09%	85.32%	87.23%	88.57%	82.39%	83.66%	86.76%	89.48%	84.54%	85.39%	88.56%
	Set 2	77.26%	74.78%	75.89%	76.99%	78.14%	73.02%	74.88%	76.98%	76.57%	71.93%	72.78%	74.87%
Politics	Set 1	91.89%	85.45%	88.67%	90.54%	87.85%	82.51%	83.41%	85.98%	91.09%	87.99%	87.15%	89.79%
	Set 2	89.20%	84.77%	87.01%	87.96%	85.34%	81.82%	81.09%	83.75%	88.66%	83.86%	84.45%	87.77%
World	Set 1	72.76%	71.87%	71.98%	73.98%	71.53%	70.98%	69.02%	72.89%	71.66%	68.64%	69.54%	72.11%
	Set 2	86.72%	80.75%	81.83%	84.55%	83.31%	79.24%	78.58%	80.69%	85.98%	80.79%	82.65%	84.57%
Entertainment	Set 1	95.99%	89.09%	92.80%	94.34%	92.67%	87.45%	88.56%	90.05%	94.04%	89.11%	90.04%	92.29%
	Set 2	88.86	83.53%	83.79%	85.39%	86.02%	80.05%	81.88%	84.99%	87.66%	81.93%	83.21%	86.16%
Science & Environment	Set 1	76.01%	73.71%	74.04%	74.97%	79.23%	75.77%	74.78%	78.65%	77.36%	73.24%	73.96%	75.02%
	Set 2	70.41%	68.54%	69.94%	71.25%	69.85%	67.62%	68.12%	69.95%	66.76%	63.79%	64.52%	67.32%

Table 8. Performance improvement of proposed approach over the baseline approaches.

		ROUGE-1			**ROUGE-2**			**ROUGE-SU4**		
		Improvement over SRRank	**Improvement over TSES**	**Improvement over LAKE**	**Improvement over SRRank**	**Improvement over TSES**	**Improvement over LAKE**	**Improvement over SRRank**	**Improvement over TSES**	**Improvement over LAKE**
Market	Set 1	5.5%	2.5%	−1.7%	4.7%	4.2%	−1.3%	2.8%	0.42%	−1.1%
	Set2	5.3%	4.3%	2.1%	3.0%	2.8%	1.6%	6.9%	5.7%	2.8%
Business	Set 1	3.9%	2.0%	2.8%	6.8%	5.3%	1.9%	5.0%	3.0%	1.1%
	Set 2	6.5%	4.2%	1.3%	5.6%	4.3%	1.8%	6.4%	5.2%	2.1%
Sports	Set 1	2.3%	4.7%	0.38%	3.6%	2.5%	1.7%	2.5%	1.0%	0.77%
	Set 2	4.1%	3.8%	0.61%	6.2%	4.6%	3.1%	7.0%	2.3%	0.72%
India	Set 1	5.4%	6.1%	1.6%	3.7%	3.1%	1.6%	4.3%	4.9%	1.7%
	Set 2	3.7%	1.3%	0.51%	7.2%	3.9%	2.6%	3.1%	6.3%	2.6%
Technology	Set 1	6.7%	3.6%	−0.28%	2.3%	0.61%	−1.4%	4.1%	5.7%	−0.4%
	Set 2	6.6%	4.7%	1.3%	3.7%	2.1%	0.85%	6.2%	6.8%	1.3%
National	Set 1	7.0%	5.4%	3.2%	7.5%	5.8%	2.1%	5.8%	4.7%	1.0%
	Set 2	3.3%	1.8%	0.35%	7.0%	4.3%	1.5%	6.4%	5.2%	2.2%
Politics	Set 1	7.5%	3.6%	1.5%	6.4%	5.3%	2.2%	3.5%	4.5%	1.4%
	Set 2	5.2%	2.5%	1.4%	4.3%	5.2%	1.8%	5.7%	4.9%	1.0%
World	Set 1	1.2%	1.1%	−1.6%	0.77%	3.6%	−1.9%	4.3%	3.0%	−0.60%
	Set 2	7.3%	5.9%	2.6%	5.1%	6.0%	3.2%	6.4%	4.0%	1.6%
Entertainment	Set 1	7.7%	3.4%	1.7%	5.9%	4.6%	2.9%	5.5%	4.4%	1.8%
	Set 2	6.3%	6.1%	4.0%	7.4%	5.1%	1.2%	6.9%	5.3%	1.0%
Science & Environment	Set 1	3.1%	2.6%	1.3%	4.5%	5.9%	0.73%	5.6%	4.5%	3.1%
	Set 2	2.7%	0.67%	−1.1%	3.2%	2.5%	−0.1%	4.6%	3.4%	−0.8%

Table 9. Result evaluation based on survey.

		Relevance				**Coherence**				**Informativeness**			
		Proposed Approach	**SRRank**	**TSES**	**LAKE**	**Proposed Approach**	**SRRank**	**TSES**	**LAKE**	**Proposed Approach**	**SRRank**	**TSES**	**LAKE**
Market	Set 1	3	2	2	3	4	3	2	3	4	2	2	4
	Set2	4	2	3	3	4	3	3	4	3	2	2	3
Business	Set 1	4	3	3	3	3	2	2	2	3	1	2	3
	Set 2	3	2	2	3	3	2	1	3	4	2	3	3
Sports	Set 1	4	1	3	3	4	2	2	3	4	2	2	3
	Set 2	4	2	2	2	3	2	2	3	3	2	1	3
India	Set 1	3	2	1	3	4	1	3	3	4	2	2	3
	Set 2	4	2	2	3	4	2	2	2	3	1	1	2
Technology	Set 1	3	2	2	3	4	3	2	2	4	4	3	2
	Set 2	4	2	3	3	4	3	2	2	3	4	3	3
National	Set 1	3	1	2	3	4	1	3	3	4	3	2	3
	Set 2	3	2	1	3	4	2	2	2	4	1	2	2
Politics	Set 1	4	2	2	3	4	3	2	3	3	2	2	3
	Set 2	3	2	2	3	4	3	3	4	4	2	3	3
World	Set 1	4	3	2	2	3	2	2	3	3	2	1	3
	Set 2	4	3	2	2	4	2	3	3	4	2	2	3
Entertainment	Set 1	3	2	1	3	4	1	3	3	3	2	1	2
	Set 2	4	3	2	3	3	2	2	3	4	1	3	3
Science & Environment	Set 1	4	3	3	4	4	2	3	3	4	2	2	2
	Set 2	3	2	2	3	4	3	3	4	3	2	1	3

7. Conclusions and Future Work

In this paper, we present a method to generate an extractive summary of multiple news articles based on keyphrase-based sentence weight and use cosine similarity to reduce redundancy. We firstly filter out the main content from the news article by content extraction approach. After extracting the main content from the news articles, we identify and extract the keyphrases. It encloses significant information about the document content and compromises a brief and precise description of the document content, which is important for news articles summarization. To calculate the weight of the sentence, we combine three features—direct keyphrase match, sentence position, and sentence length—and to reduce redundancy, we used cosine similarity. We compared our proposed approach with other approaches on English news documents dataset. The experimental results indicate that our approach performs well on several multi-document summarization approaches for English news documents. This study did not carry out any post-processing of the sentences, such as compression of sentences and information fusion.

In our future work, we plan to summarize multi-lingual news articles, which are not covered in this paper. On the other hand, the dataset we used in this paper contains 100 news articles; it is considered small compared to the other standard summarization datasets. However, we will try to build a larger database for the more confident results. We will also apply our approach to some more existing datasets to test its robustness. Furthermore, we focus on utilizing more evaluation methods to evaluate the proposed summarization approach.

Author Contributions: Conceptualization, C.A., M.D. and P.S.; data curation, C.A., M.D. and P.S.; formal analysis, V.S. and S.K.; investigation, C.A., M.D. and P.S.; project administration, V.S., S.K., and J.K.; resources, C.A., M.D. and P.S.; software, C.A., M.D. and P.S.; supervision, M.D.; validation, C.A., M.D. and P.S.; visualization, C.A., M.D. and P.S.; writing—original draft, C.A., M.D. and P.S.; writing—review and editing, V.S, S.K. and J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly supported by the Technology Development Program of MSS [No. S3033853] and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A4A1031509).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mitchell, C.C.; West, M.D. *The News Formula: A Concise Guide to News Writing and Reporting*; St. Martin's Press: New York, NY, USA, 1996.
2. Radev, D.R.; Blair-Goldensohn, S.; Zhang, Z.; Raghavan, R.S. Interactive, domain-independent identification and summarization of topically related news articles. In *International Conference on Theory and Practice of Digital Libraries*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 225–238.
3. Kupiec, J.; Pedersen, J.; Chen, F. A trainable document summarizer. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, USA, 9–13 July 1995; pp. 68–73.
4. Galanis, D.; Lampouras, G.; Androutsopoulos, I. Extractive multi-document summarization with integer linear programming and support vector regression. In Proceedings of the 24th International Conference on Computational Linguistics, COLING 2012, Mumbai, India, 8–15 December 2012; pp. 911–926.
5. Wong, K.F.; Wu, M.; Li, W. Extractive summarization using supervised and semi-supervised learning. In Proceedings of the 22nd International Conference on Computational Linguistics, Manchester, UK, 18–22 August 2008; pp. 985–992.
6. Chen, H.H.; Kuo, J.J.; Huang, S.J.; Lin, C.J.; Wung, H.C. A summarization system for Chinese news from multiple sources. *J. Assoc. Inf. Sci. Technol.* **2003**, *54*, 1224–1236. [CrossRef]
7. Mani, I.; Bloedorn, E. Multi-document summarization by graph search and matching. *arXiv* **1997**, arXiv:cmp-lg/9712004.
8. McKeown, K.R.; Klavans, J.L.; Hatzivassiloglou, V.; Barzilay, R.; Eskin, E. Towards multidocument summarization by reformulation: Progress and prospects. In Proceedings of the AAAI-99, Orlando, FL, USA, 18–22 July 1999.
9. Radev, D.R.; McKeown, K.R. Generating natural language summaries from multiple on-line sources. *Comput. Linguist.* **1998**, *24*, 470–500.

10. Radev, D.R.; Blair-Goldensohn, S.; Zhang, Z. Experiments in single and multi-document summarization using MEAD. *Ann. Arbor.* **2001**, *1001*, 1–8.
11. Lin, C.Y.; Hovy, E. From single to multi-document summarization: A prototype system and its evaluation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; Association for Computational Linguistics: Toronto, ON, Canada, 2002; pp. 457–464.
12. Carbonell, J.; Goldstein, J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, 24–28 August 1998; pp. 335–336.
13. Hovy, E.H.; Marcu, D. Automated Text Summarization. In *Pre-Conference Tutorial of the COLING/ACL*; ACL: Berkeley, CA, USA, 2000; Volume 98.
14. McKeown, K.; Radev, D.R. Generating summaries of multiple news articles. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, USA, 9–13 July 1995; pp. 74–82.
15. Goldstein, J.; Mittal, V.; Carbonell, J.; Kantrowitz, M. Multi-document summarization by sentence extraction. In Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization; Association for Computational Linguistics: Toronto, ON, Canada, 2000; pp. 40–48.
16. McKeown, K.; Hatzivassiloglou, V.; Barzilay, R.; Schiffman, B.; Evans, D.; Teufel, S. *Columbia Multi Document Summarization: Approach and Evaluation*; Columbia University: New York, NY, USA, 2001.
17. Lee, C.S.; Chen, Y.J.; Jian, Z.W. Ontology-based fuzzy event extraction agent for Chinese e-news summarization. *Expert Syst. Appl.* **2003**, *25*, 431–447. [CrossRef]
18. D'Avanzo, E.; Magnini, B. A keyphrase-based approach to summarization: The lake system at duc-2005. October. In Proceedings of the DUC 2005, Sydney, Australia, 24 June 2005.
19. Al-Hashemi, R. Text Summarization Extraction System (TSES) Using Extracted Keywords. *Int. Arab J. e-Technol.* **2010**, *1*, 164–168.
20. El-Haj, M.; Kruschwitz, U.; Fox, C. Exploring clustering for multi-document arabic summarization. In *Asia Information Retrieval Symposium, December*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 550–561.
21. Li, C.; Qian, X.; Liu, Y. Using supervised bigram-based ilp for extractive summarization. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Sofia, Bulgaria, 4–9 August 2013; Volume 1, pp. 1004–1013.
22. Yan, S.; Wan, X. SRRank: Leveraging semantic roles for extractive multi-document summarization. *IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP)* **2014**, *22*, 2048–2058. [CrossRef]
23. Cao, Z.; Wei, F.; Dong, L.; Li, S.; Zhou, M. Ranking with Recursive Neural Networks and Its Application to Multi-Document Summarization. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 29.
24. Liu, S.-H.; Chen, K.-Y.; Hsieh, Y.-L.; Chen, B.; Wang, H.-M.; Yen, H.-C.; Hsu, W.-L. Exploring Word Mover's Distance and Semantic-Aware Embedding Techniques for Extractive Broadcast News Summarization. *Interspeech* **2016**, *2016*, 670–674.
25. Rautray, R.; Balabantaray, R.C. Cat swarm optimization based evolutionary framework for multi document summarization. *Phys. A Stat. Mech. Its Appl.* **2017**, *477*, 174–186. [CrossRef]
26. Sanchez-Gomez, J.M.; Vega-Rodríguez, M.A.; Pérez, C.J. Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach. *Knowl.-Based Syst.* **2018**, *159*, 1–8. [CrossRef]
27. Uçkan, T.; Karcı, A. Extractive multi-document text summarization based on graph independent sets. *Egypt. Inform. J.* **2020**, *21*, 145–157. [CrossRef]
28. Xu, X.; Dušek, O.; Narayan, S.; Rieser, V.; Konstas, I. MiRANews: Dataset and Benchmarks for Multi-Resource-Assisted News Summarization. *arXiv* **2021**, arXiv:2109.10650.
29. Garg, P.K.; Chakraborty, R.; Dandapat, S.K. OntoRealSumm: Ontology based Real-Time Tweet Summarization. *arXiv* **2022**, arXiv:2201.06545.
30. Zhao, C.; Huang, T.; Chowdhury, S.B.R.; Chandrasekaran, M.K.; McKeown, K.; Chaturvedi, S. Read Top News First: A Document Reordering Approach for Multi-Document News Summarization. *arXiv* **2022**, arXiv:2203.10254.
31. AL-Khassawneh, Y.A.; Hanandeh, E.S. Extractive Arabic Text Summarization-Graph-Based Approach. *Electronics* **2023**, *12*, 437. [CrossRef]
32. Zhang, T.; Ladhak, F.; Durmus, E.; Liang, P.; McKeown, K.; Hashimoto, T.B. Benchmarking Large Language Models for News Summarization. *arXiv* **2023**, arXiv:2301.13848.
33. Pozzi, A.; Barbierato, E.; Toti, D. Cryptoblend: An AI-Powered Tool for Aggregation and Summarization of Cryptocurrency News. In *Informatics*; Multidisciplinary Digital Publishing Institute: Basel, Switzerland, 2023; Volume 10, p. 5.
34. Vore, K.; Vanderwende, L.; Burges, C. Enhancing single-document summarization by combining RankNet and third-party sources. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Prague, Czech Republic, 28–30 June 2007.
35. Litvak, M.; Last, M. Graph-based keyword extraction for single-document summarization. In Proceedings of the Workshop on Multi-Source Multilingual Information Extraction and Summarization, Association for Computational Linguistics, Toronto, ON, Canada, 23 August 2008; pp. 17–24.

36. Gonçalves, P.N.; Rino, L.; Vieira, R. Summarizing and referring: Towards cohesive extracts. In *Proceedings of the Eighth ACM Symposium on Document Engineering—DocEng 2008, Sao Paulo, Brazil, 16–19 September 2008*; Association for Computing Machinery: New York, NY, USA, 2008; pp. 253–256.
37. Arya, C.; Dwivedi, S.K. News web page classification using url content and structure attributes. In Next Generation Computing Technologies (NGCT), Proceedings of the 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun, India, 14–16 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 317–322.
38. Arya, C.; Dwivedi, S.K. Content extraction from news web pages using tag tree. *Int. J. Auton. Comput.* **2018**, *3*, 34–51. [CrossRef]
39. Arya, C.; Dwivedi, S.K. Keyphrase Extraction of News Web Pages. *Int. J. Educ. Manag. Eng. (IJME)* **2018**, *8*, 48–58. [CrossRef]
40. Ding, C.H. A similarity-based probability model for latent semantic indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, 15–19 August 1999; pp. 58–65.
41. Qazvinian, V.; Radev, D.R.; Özgür, A. Citation summarization through keyphrase extraction. In Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, 23–27 August 2010; Association for Computational Linguistics: Toranto, ON, Canada, 2010; pp. 895–903.
42. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004.
43. Chin-Yew, L.; Och, F.J. Looking for a few good metrics: Rouge and its evaluation. In Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization, NTCIR-4, National Center of Sciences, Tokyo, Japan, 2–4 June 2004; National Institute of Informatics (NII): Tokyo, Japan, 2005; ISBN 4-86049-030-4.
44. Nenkova, A.; Passonneau, R. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Hlt-naacl 2004*; Association for Computational Linguistics: Boston, MA, USA, 2004.