VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Helmholtz-Based Automatic Document Summarization

Ahmet TOPRAK[1]*, Metin TURAN[2]

[1]Istanbul Commerce University, Computer Engineering, Istanbul, TURKEY
[2]Istanbul Commerce University, Computer Engineering, Istanbul, TURKEY

Makale Bilgisi

Başvuru:
01/07/2022
Kabul:
06/09/2022

Abstract

Nowadays, internet networks and social media have allowed people to express and interpret their opinions about other people or institutions easily and clearly. With the increasing prevalence of this opportunity, a growing body of rich content emerges. As a result, analyzing big data obtained from the internet, transforming it into meaningful information, and using it is a subject that has been studied intensively in recent years. In this process, automatic text summarization has become an important task. This study presents the Helmholtz-based extractive summarization method to create an automatic text summarization system. The BBC News data set was used to test the proposed method. In this data set, there are both original full-text documents and summary documents of these original documents produced by human summarizers. The similarity of the summary document produced by the proposed Helmholtz-based extractive text summarization method with the original summary in the BBC News data set was calculated using the Simhash text similarity algorithm. When the results are examined, summary documents can be produced with a 38.9% Simhash text similarity rate with the proposed Helmholtz-based extractive summarization method. The Experiments section also shares the results obtained with other third-party extractive summarization algorithms.

Keywords: *Automatic Document Summarization, Helmholtz Principle, Kullback-Leibler Divergence, Latent Semantic Analysis, LexRank, Summarization Algorithms, TextRank, TF-IDF*

***** İletişim e-posta: ahmetoprak190363@gmail.com
** Bu çalışmanın bir kısmı V. International Conference on Data Science and Applications 2022'de sözlü olarak sunulmuştur.

1 Introduction

Automatic document summarization is a system that tries to give the user basic information about the document without reading the entire text. It reduces the size of the document while keeping the important information in the document [1]. In this system process, the parts other than the summary are unimportant in terms of the document and do not reflect the main subject. Therefore, it is not a problem that these parts are not included in the summary. However, there are some difficulties faced by this system. The most obvious of these is the selection of parts containing imports study [2] presents a Helmholtz-based extractive summarization method have been proposed to overcome these difficulties.

Automatic summarization systems are divided into two extractive and abstractive in terms of analyzing the text and creating a summary. Summarizing by selecting important sentences from the text document is called extractive summarization. A level of importance is determined for sentence fragments whose score is assigned, and sentences included in the summary are obtained with a determined threshold upper limit. In abstractive summarization systems, the process is completely different from extractive summarization. In this approach, sentences are interpreted and re-expressed with certain steps. Parts of the text, such as paragraphs or sentences, are obtained from the main text using linguistic methods. This method is based on understanding the real text, interpreting it, and obtaining it concisely using fewer words [3, 4]. Abstractive summarization systems consist of more complex steps than extractive summarization systems.

Summarization systems can also be classified as single-document summarization and multi-document summarization, depending on the number of input documents provided. The data set used to test the proposed method in this study belongs to the multi-document summarization type. This study presents a Helmholtz-based extractive summarization method to create an automatic text summarization system. BBC News data set was used to evaluate the proposed method. In this data set, there are both original full-text documents and summary documents of these original documents produced by human summarizers. First, the original documents in the data set that the user wanted to summarize were taken into the data preprocessing process. After the data pre-processing process, the summary of the original document in the data set was provided to each of the third-party and proposed Helmholtz-based extractive algorithms known in the literature. The ratio of the summary to be extracted was obtained from the division of the document's word count in summary in the data set to the word count of the original document. Using this ratio, the similarity of the summary document obtained from each algorithm and the original summary document in the BBC News data set was calculated using the Simhash text similarity algorithm. When the summary documents obtained were examined in terms of average Simhash text similarity rates, it was seen that the proposed Helmholtz-based summarization algorithm achieved a higher similarity rate than most third-party extractive summarization algorithms. The obtained results are shared in detail in the Experiments section. This study aims to create an automatic document summarization system without requiring human intervention on a technical problem.

The organization of the remaining paper is as follows: In the second section, document summarization research in literature or studies covering methods applied in this study are mentioned. In the third section, the used methods in the document summarization study are explained in detail. The results are commented on, and future studies are discussed.

2 Related Works

Although text summarization studies were started about 50 years ago, studies in this field continue to be popular in light of recent technological and linguistic studies. Luhn made the first study on automatic text summarization in 1959 [5]. In this study, Luhn made use of the frequency of using words in sentences to summarize, and he intuited that the words with the highest frequency of use give the most important views about that study.

In 2014, in the study by Pal et al. [6], most of the text summarization approaches in the literature were based on the spelling of the sentence, its position in the text, the frequency of a particular word in a sentence, etc. It was stated that he performed the summarization based on some manually labeled rules as it has been emphasized that such predefined constraints greatly affect the quality of text summarization. On the other hand, the proposed approach performs the summarization task with the unsupervised learning methodology.

The importance of a sentence in the original text is evaluated with the help of the Simplified Lesk algorithm, and WordNet is used as a semantic dictionary. First of all, with this approach, the weights of the sentences of an original text are evaluated separately using the Simplified Lesk algorithm and arranged in descending order according to their weights. Then, a certain number of sentences are selected from this ordered list, according to the given summarization percentage. It has been stated that the proposed approach gives the best results in summarizing up to 50% of the original full-text.

In 2019, in the study by Dani Gunawan et al. [7], a method for summarizing multi-document data was proposed. Two or more important sentences will likely share similar information in the multi-document summary. Including these sentences in the summary result will cause unnecessary information to be created in the summary document. Therefore, this study aims to reduce similar sentences in multiple documents that share similar information to obtain a more concise text summary. The data set used in the study is a combination of several online news articles divided into six groups. The merged articles are pre-processed to produce clean text. After obtaining the clean text, the TextRank summarization algorithm extracted important sentences using the similarity measure. As a result of this process, the summarized text was obtained. However, it is stated that the summarized text still contains similar sentences. Therefore, to further reduce similar sentences in the resulting summary document, calculate the Maximal Marginal Fit (MMR). Finally, the summary text was obtained. Rouge (Recall-Oriented Understudy for Gisting Evaluation)-1 and Rouge-2 were used to evaluate the summary text. When the results were examined, the mean F score was obtained as 0.5103 and 0.4257, respectively.

In 2018, in the study by Kamal Al-Sabari and her team [8], another way of using the attention mechanism to construct a sentence and document embedding was proposed. In the study, it is stated that recent developments in neural network architecture and training algorithms increase the efficiency of representational learning. Therefore, neural network-based models have better representation ability than traditional models. They are capable of automatically learning distributed representation for sentences and documents. For this purpose, it is stated that a new model is proposed in this study, which is not adequately modeled by the previously proposed models, which addresses various problems, including memory problems and document structure information. The proposed model uses a hierarchical structured self-attention mechanism to create a sentence and document embedding. This architecture mirrors the document's hierarchical structure and, in turn, enables us to obtain better feature representation. The attention mechanism provides an additional source of information to guide the summary extraction. The new model treated the summarization task as a classification problem in which the model computes the respective probabilities of sentence summary membership. The model predictions are broken up by features such as information content, salience, novelty, and positional representation. The performance of the proposed model was evaluated on two well-known data sets, CNN/Daily Mail and DUC 2002. Rouge-1 and Rouge-2 were used for the model evaluation metrics. When the results were examined, the Rouge-1 and Rouge-2 values were obtained as 0.423 and 0.416, respectively.

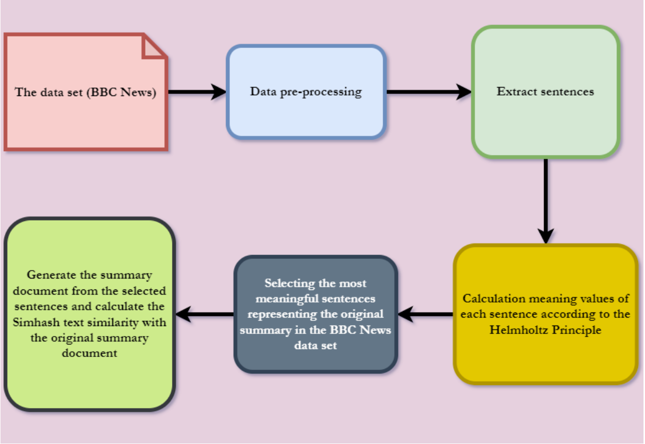
In 2019, the study by Aneesh Vartakavi et al. [9] proposed PodSumm. This new system automatically generates podcast audio summaries using Automatic Speech Recognition (ASR) and extractive text summarization, allowing listeners to preview podcast episodes quickly. The proposed system first copies the audio from a podcast using ASR, then summarizes the transcript using the Transformer-based summarization model, one of the extractive text summarization methods, and finally returns the audio associated with the text summary. Since no data set can be used for inferential summarization of podcasts, a data set was produced with manual methods to support the development and evaluation of the proposed system. For the data set, 19 unique podcast series from different genres were selected, with an average of 16.3 episodes per series. The full data set includes 309 different podcast episodes with an average duration of 36.5 minutes per episode for 188 hours of audio. Rouge-1 and Rouge-2 metrics were used to evaluate the system. When the results were examined, the Rouge-1 and Rouge-2 values were obtained as 0.63 and 0.53, respectively.

3 Material and Method

In this study, the Helmholtz-based extractive automatic document summarization model is proposed to extract appropriate summaries from the full texts that the user intends to summarize. The proposed model consists of four main stages. In the first stage, the original full-text documents in the BBC News data set were taken to the data pre-processing process and cleaned and incorrect entries were corrected. In the second stage, the sentences belonging to the documents subjected to the data preprocessing process were determined. In the third stage, summary extractions were made with both third-party. They suggested Helmholtz-based summary algorithms for the original documents in the BBC News data set, based on the summary size parameter. In the last stage, the similarities of the summaries generated by the algorithms with the original summary documents in the BBC News data set were calculated using the Simhash text similarity algorithm.

The working topology of the proposed Helmholtz-based extractive summarization approach is given in Figure 1. All steps in topology will be covered under separate headings.

Figure 1. The topology of the proposed Helmholtz-based extractive summarization approach


The data set (BBC News)
Data pre-processing
Extract sentences
Calculation meaning values of each sentence according to the Helmholtz Principle
Selecting the most meaningful sentences representing the original summary in the BBC News data set
Generate the summary document from the selected sentences and calculate the Simhash text similarity with the original summary document

3.1 *The Data Set*

In this study, the BBC News data set [10] was used to test the accuracy of the proposed Helmholtz-based summarization method. This data set was created using a data set used for data categorization that consists of 2225 documents from the BBC news website corresponding to stories in five topical areas from 2004-2005 used in Greene and Cunningham's [19] study. The number of documents in the BBC News data set according to the topics is given in Table 1.

Table 1. The number of documents in the BBC News data set according to the topics

Document Topic	Number of Document
Business	510
Entertainment	386
Politics	417
Sport	511
Technology	401

3.2 *Data Pre-processing*

Data pre-processing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and used for machine learning. It is also an important step in data mining, as we cannot work with raw data. The quality of the data should be checked before being applied to machine learning or data mining algorithms. Data pre-processing steps applied in the research were data cleaning, data integration, data transformation, data reduction, and data discretization in sequence as given in Figure 2 [18].

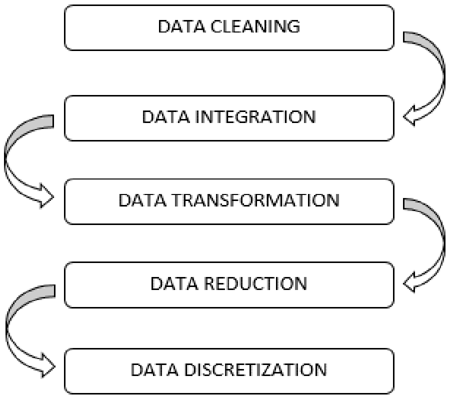


Figure 2. Data pre-processing steps applied in the research

Before processing the data set used, it was subjected to the data pre-processing. The steps given in Figure 2 were carried out. Initially, the documents entered into the system should be separated into the units to which they can be processed later. In this study, the documents were divided into sentences because they are processed sentence-based. The document contains many words that are meaningless in the text and removing them from this sentence would not cause a semantic loss. They, called stop words in the language, were removed from the document using the Natural Language Toolkit (NLTK) library. In addition, incorrect entries in both full-text and summary were corrected with the spell checker function in the NLTK library. Then all the words in the document were changed to lowercase format to ensure similar letter compatibility (regularization). After this process, the inflectional suffixes and derivational affixes of the words were removed and the root form of the words were obtained.

3.3 Helmholtz-Based Extractive Summarization

In the proposed Helmholtz-based extractive summarization method, the original full-text documents in the BBC News data set were evaluated on a sentence basis. The meaning values of the words in each sentence were calculated. Helmholtz's principle [11, 12, 13] was used to calculate the meaning values of words. Then, the words whose meaning values were calculated were collected on a sentence basis. The sentences with the highest meaning value were selected according to the determined summary length ratio and added to the summary document.

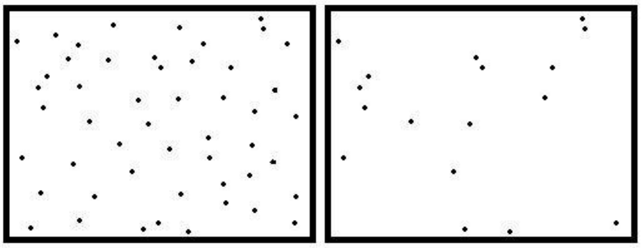


Figure 3. The Helmholtz principle in human perception [14]

According to the Helmholtz principle, "Humans immediately perceive whatever could not happen by chance" [15]. This means a structure is easily recognized when it exhibits a large deviation from randomness. For example, in Figure 3, a group of five aligned dots appears in both images. These dots can be easily perceived in the right-hand side image because it exhibits a large deviation from randomness, which is unlikely to happen by chance. By contrast, the same aligned dots are difficult to identify in the left-hand side image, submerged by a large quantity of random information [16].

Helmholtz principle, for each word; was used in the sentence of the document to determine whether it is probable that the word occurs m times. The meaning value based on this theory was calculated with the following formulas [25].

NFA (Number of False Alarm) is inversely proportional to the meaning value. In other words, the lower the NFA value; the higher the meaning value of the selected word (k) for documents of that class (D). A high meaning value indicates that the words are effective and efficient. In the study, the following approach was used to select the best words.

$$\mathrm{NFA}(k,P,D)=\binom{K}{m}\frac{1}{N^{m-1}} \quad (1)$$

$$\mathrm{MV}\ (k,P,D)=\frac{1}{m}\log \mathrm{NFA}(k,P,D) \quad (2)$$

$$\mathrm{N}=\frac{|\mathrm{D}|}{|\mathrm{P}|} \quad (3)$$

$$\log\ (\mathrm{NFA}(k,P,D))=\log(\ \binom{K}{m}\frac{1}{N^{m-1}}\ ) \quad (4)$$

The formula for calculating the Helmholtz principle includes the following abbreviations:

MV: Meaning value.
D: Document to be processed.
k: Word to be calculated.
P: The total number of sentences in the document.
m: The total number of sentences in which the word to be calculated occurs.
K: Total number of words to be calculated in the document.
N: The total number of words in the document divided by the total number of words in the sentence to be processed.

The meaning value was made for each word in the sentence. Then, these meaning values were collected and the meaning value of the sentence was calculated by dividing it by the total number of words in the sentence. After determining the meaning values for all sentences, the summary document was created by selecting the sentences with high meaning values according to the summary length ratio.

If this process is considered through an example scenario;

The word to calculate is "team" and

Let D = 100, P =20, m= 3, K=10.

First, the N value is calculated.

N= 100/20=5

NFA ("team", P, D) = $\binom{10}{3} * \frac{1}{5^2}$ = 120* $\frac{1}{25}$= 4.8

MV ("team", P, D) = 1/3∗ log (4.8)

log (4.8) = 0.68

MV ("team", P, D) = 0.68 * (1/3) = 0.22

The meaning value of the word "team" was calculated as 0.22.

3.4 Extractive Summarization Algorithms Used in Experiments

To compare and evaluate the proposed Helmholtz-based extractive summarization algorithm in this study, algorithms that are widely known in the literature and produce successful results (LexRank, TextRank, Kullback-Leibler (KL) Divergence, Latent Semantic Analysis (LSA), and Term Frequency–Inverse Document Frequency-based summarization (TF-IDF)) are used. These algorithms perform extractive summarization.

LexRank summarization algorithm [21] is an unsupervised graph-based approach for automatic text summarization. The scoring of sentences is done using the graph method. LexRank is used for computing sentence importance based on the concept of eigenvector centrality in a graph representation of sentences. The main idea is that sentences "recommend" other similar sentences to the reader. Thus, if one sentence is very similar to many others, it will likely be a sentence of great importance. The importance of this sentence also stems from the importance of the sentence "recommending" it. Thus, to get ranked highly and placed in a summary, a sentence must be similar to many sentences that are in turn also similar to many other sentences. This makes intuitive sense and allows the algorithms to be applied to any arbitrary new text.

TextRank summarization algorithm [20] is an unsupervised algorithm, used for automated summarization of texts written in natural languages. It is an algorithm based on PageRank, which is often used in keyword extraction and extractive text summarization. It is also a graph-based ranking model for text processing which can be used to find the most relevant sentences in the text and to find keywords. The basic steps involved in the TextRank algorithm are as follows.

- Extract all the sentences from the text document.
- A Graph is created out of the sentences extracted in step 1. Bag-of-Words (BOW), TF-IDF, or one of the ready word vectors Global Vectors for Word Representation (Glove) can be used.
- The similarity values in the sentence vectors are stored in the similarity matrix.
- The similarity matrix is converted to a graph.
- Sentence selection is made according to the similarity scores ranked in the graph.

KL Divergence algorithm [24] is an information-based measure of disparity among probability distributions. In mathematical statistics, the KL Divergence is a measure of how one probability distribution is different from another. Less the divergence, more the summary and the document are similar to each other in terms of understandability and meaning conveyed.

LSA algorithm [23] is an NLP technic that analyzes relationships between a set of documents and the terms contained within. It uses singular value decomposition, a mathematical technique, to scan unstructured data to find hidden relationships between terms and concepts.

The basic steps involved in the LSA algorithm are as follows.

- The text is converted into matrices to represent sentences. Each cell in the matrix contains the number of times a certain word appears in a certain sentence.
- The matrix is factorized so that every sentence is represented as a vector. The value for each vector is the sum of vectors representing its component words.

- Dot products, cosines, or similar metrics are used to represent similarities between words and sentences.

TF-IDF [22] is the weighting factor calculated by statistical methodology, which shows the importance of a word in a document. It is used for statistical analysis of the texts that are the common application of Natural Language Processing (NLP) and text mining. It is also often referred to as a ranking algorithm under topics such as information retrieval.

Term Frequency (TF): A method used to calculate word weights in a document. There are several ways of calculating this frequency, with the simplest being a raw count of instances a word appears in a document.

Inverse Document Frequency (IDF): Measures the rank of the specific word for its relevancy within the text. Stop words that contain unnecessary information such as “the”, “a” and “with” carry less importance despite their occurrence.

In this study, formulas (5) and (6) were used for TF and IDF calculations.

t = Word to be calculated.
NT = Number of times the word “t” appears in a document.
TT = Total number of words in that document.
TD = Total number of documents.
ND = Number of documents these includes the word “t” in it.

$$\mathrm{TF}\ (\mathrm{t}) = \frac{\mathrm{NT}}{\mathrm{TT}} \tag{5}$$

$$\mathrm{IDF}\ (\mathrm{t}) = \frac{\mathrm{TD}}{\mathrm{ND}} \tag{6}$$

Finally, the TF-IDF for the “t” word is calculated as the multiplication of TF (t) and IDF (t) values given in formula (7).

$$\mathrm{TF} - \mathrm{IDF}\ (\mathrm{t}) = \mathrm{TF}\ (\mathrm{t}) * \mathrm{IDF}\ (\mathrm{t}) \tag{7}$$

The success rates of the summaries made by these algorithms on the BBC News data set are discussed in detail in the Experiments section based on the document topic.

3.5 *Determining Summary Length Ratio*

The length of the summary document generated with the extractive summarization algorithms was calculated concerning the BBC News data set. The summary length ratio was determined by dividing the total number of words of the original full-text document to be summarized by the number of words in the original summary document.

When the summary length ratio is calculated for the document belonging to the BBC News data set in Table 2;

Table 2. A sample document of the BBC News data set

Original Full-text Document	Original Summary Document
Maternity pay for new mothers is to rise by £1,400 as part of new proposals announced by the Trade and Industry Secretary Patricia Hewitt. It would mean paid leave would be increased to nine months by 2007, Ms. Hewitt told GMTV's Sunday program me. Other plans include letting maternity pay be given to fathers and extending rights to parents of older children. The Tories dismissed the maternity pay plan as "desperate", while the Liberal Democrats said it was misdirected. Ms. Hewitt said: "We have already doubled the length of maternity pay, it was 13 weeks when we were elected, we have already taken it up to 26 weeks. “We are going to extend the pay to nine months by 2007 and the aim is to get it right up to the full 12 months by the end of the next Parliament." She said new mothers were already entitled to 12 months leave, but that many women could not take it as only six of those months were paid. "We have made a firm commitment. We will definitely extend the maternity pay, from the six	She said her party would boost maternity pay in the first six months to allow more women to stay at home in that time. She said new mothers were already entitled to 12 months leave, but that many women could not take it as only six of those months were paid. The Tories dismissed the maternity pay plan as "desperate,” while the Liberal Democrats said it was misdirected. She said ministers would consult on other proposals that could see fathers being allowed to take some of their partner's maternity pay or leave period, or extending the rights of flexible

months where it now is to nine months, that's the extra £1,400." She said ministers would consult on other proposals that could see fathers being allowed to take some of their partner's maternity pay or leave period, or extending the rights of flexible working to careers or parents of older children. The Shadow Secretary of State for the Family, Theresa May, said: "These plans were announced by Gordon Brown in his pre-budget review in December and Tony Blair is now recycling it in his desperate bid to win back women voters." She said the Conservatives would announce their proposals closer to the General Election. Liberal Democrat spokeswoman for women Sandra Gidley said: "While mothers would welcome any extra maternity pay the Liberal Democrats feel this money is being misdirected." She said her party would boost maternity pay in the first six months to allow more women to stay at home in that time. Ms. Hewitt also stressed the plans would be paid for by taxpayers, not employers. But David Frost, director general of the British Chambers of Commerce, warned that many small firms could be "crippled" by the move. "While the majority of any salary costs may be covered by the government's statutory pay, recruitment costs, advertising costs, retraining costs and the strain on the company will not be," he said. Further details of the government's plans will be outlined on Monday. New mothers are currently entitled to 90% of average earnings for the first six weeks after giving birth, followed by £102.80 a week until the baby is six months old.	working to careers or parents of older children. Liberal Democrat spokeswoman for women Sandra Gidley said: "While mothers would welcome any extra maternity pay the Liberal Democrats feel this money is being misdirected. "We will definitely extend the maternity pay, from the six months where it now is to nine months, that's the extra £1,400." Ms. Hewitt said: "We have already doubled the length of maternity pay; it was 13 weeks when we were elected, we have already taken it up to 26 weeks. Other plans include letting maternity pay be given to fathers and extending rights to parents of older children.

When the summary length ratio was calculated for the document belonging to the BBC News data set in Table 2,

F (t) = Total word count of the original full-text document (BBC News).

F (t) = 446

S (t) = Total number of words in the original summary document.

S (t) = 200

G (t) = Summary length ratio = (S (t) / F (t))*100

G (t) = (200 / 446) * 100 = 44.8%.

The summary length ratio value (44.8%) determined for the document given in Table 2 was used to generate the summary of this original full-text document with LexRank, TextRank, KL Divergence, Lsa, TF-IDF, and suggested Helmholtz principle algorithms. The meaning value of each sentence of the original full-text document in Table 2 was calculated with the proposed Helmholtz principle approach (Table 3). Then, sentences with high meaning were added to the summary document. The summary document produced with the Helmholtz principle was given in Table 4.

Table 3. The meaning values of the sentences of the document in Table 2 calculated according to the Helmholtz principle

Sentence	Meaning Value
Ms. Hewitt said: "We have already doubled the length of maternity pay, it was 13 weeks when we were elected, we have already taken it up to 26 weeks. "We are going to extend the pay to nine months by 2007 and the aim is to get it right up to the full 12 months by the end of the next Parliament."	0.35

Other plans include letting maternity pay be given to fathers and extending rights to parents of older children.	0.39
New mothers are currently entitled to 90% of average earnings for the first six weeks after giving birth, followed by £102.80 a week until the baby is six months old.	0.34
Maternity pay for new mothers is to rise by £1,400 as part of new proposals announced by the Trade and Industry Secretary Patricia Hewitt.	0.36
But David Frost, director general of the British Chambers of Commerce, warned that many small firms could be "crippled" by the move.	0.37
She said ministers would consult on other proposals that could see fathers being allowed to take some of their partner's maternity pay or leave period, or extending the rights of flexible working to careers or parents of older children.	0.42

Table 4. The generated summary document by the Helmholtz principle

Generated Summary Document By Helmholtz Principle
Ms. Hewitt said: "We have already doubled the length of maternity pay, it was 13 weeks when we were elected, we have already taken it up to 26 weeks. “ We are going to extend the pay to nine months by 2007 and the aim is to get it right up to the full 12 months by the end of the next Parliament. “Other plans include letting maternity pay be given to fathers and extending rights to parents of older children. New mothers are currently entitled to 90% of average earnings for the first six weeks after giving birth, followed by £102.80 a week until the baby is six months old. Maternity pay for new mothers is to rise by £1,400 as part of new proposals announced by the Trade and Industry Secretary Patricia Hewitt. But David Frost, director general of the British Chambers of Commerce, warned that many small firms could be "crippled" by the move. She said ministers would consult on other proposals that could see fathers being allowed to take some of their partner's maternity pay or leave period, or extending the rights of flexible working to careers or parents of older children.

When the summary document generated by the Helmholtz principle was examined, it is seen that the number of document words is 195.

$G(t) = (195 / 446) * 100 = 43.7\% < 44.8\%$

Since the summary length ratio calculated for the document given in Table 2 is 44.8%, the number of summary words generated by the Helmholtz principle remained at this level. Otherwise, if a new sentence had been added to the document, it would have exceeded this rate.

3.6 Calculation of Similarity Rate

In this study, the similarities of the summary documents generated by both third-party and proposed Helmholtz-based extractive summarization algorithms with the original full-text documents found in the BBC News data set were calculated using the Simhash text similarity algorithm.

Simhash is a hashing function and its property is that the more similar the text inputs are, the smaller the Hamming distance of their hashes is (Hamming distance – the number of positions at which the corresponding symbols are different). The algorithm works by splitting the text into chunks and hashing each chunk with a hashing function of your choice. Each hashed chunk is represented as a binary vector and the bit values are transformed into +1 or -1 depending on whether the bit value is 1 or 0. To get the Simhash, we add up all the bit vectors bitwise. Finally, the resulting bits are set to 0 if the sum is negative; otherwise, to 1 [17].

4 Experiments

In this part of the study, summaries of original full-text documents of different document topics (Business, Tech, Politics, Sport, Entertainment) in the BBC News data set were generated with both third-party and proposed Helmholtz-based extractive summarization algorithms. Then, the similarities of these summary documents generated by the algorithms and the original summary documents in the BBC News data set were calculated with the Simhash text similarity algorithm. The average operation times of each extractive summarization algorithm were also calculated. In each experiment, the results obtained by the algorithms in a different document topic in the BBC News data set are shared.

4.1 Experiment I

In Experiment I, summaries of the original full-text documents of the business document topic in the BBC News data set were generated with each extractive summarization algorithm. Then, the similarities between the summary document generated by each summarization algorithm and the original summary document were calculated. The average Simhash text similarity rates obtained by the algorithms are given in Table 5.

Table 5. Results obtained by summarization algorithms with business document topic

Summarization Algorithm	Document Topic	Average Similarity Rate (%)	Average Operation Time(sec)
LexRank	Business	41.3	1.6
TextRank	Business	33.1	1.6
KL Divergence	Business	31.4	2.2
Lsa	Business	32.0	2.3
TF-IDF Based	Business	29.2	1.8
Helmholtz-Based	Business	37.0	1.7

4.2 Experiment II

In Experiment II, summaries of the original full-text documents of the politics document topic in the BBC News data set were generated with each extractive summarization algorithm. Then, the similarities between the summary document generated by each summarization algorithm and the original summary document were calculated. The average Simhash text similarity rates obtained by the algorithms are given in Table 6.

Table 6. Results obtained by summarization algorithms with politics document topic

Summarization Algorithm	Document Topic	Average Similarity Rate (%)	Average Operation Time(sec)
LexRank	Politics	40.9	1.6
TextRank	Politics	35.4	1.7
KL Divergence	Politics	33.8	2.1
Lsa	Politics	32.2	2.4
TF-IDF Based	Politics	30.1	1.9
Helmholtz-Based	Politics	38.9	1.8

4.3 Experiment III

In Experiment III, summaries of the original full-text documents of the sport document topic in the BBC News data set were generated with each extractive summarization algorithm. Then, the similarities between the summary document generated by each summarization algorithm and the original summary document were calculated. The average Simhash text similarity rates obtained by the algorithms are given in Table 7.

Table 7. Results obtained by summarization algorithms with sport document topic

Summarization Algorithm	Document Topic	Average Similarity Rate (%)	Average Operation Time(sec)
LexRank	Sport	39.5	2.0
TextRank	Sport	33.4	1.9
KL Divergence	Sport	34.6	2.6
Lsa	Sport	31.3	2.4
TF-IDF Based	Sport	30.5	1.9
Helmholtz-Based	Sport	37.8	2.1

4.4 Experiment IV

In Experiment IV, summaries of the original full-text documents of the entertainment document topic in the BBC News data set were generated with each extractive summarization algorithm. Then, the similarities between the summary document generated by each summarization algorithm and the original summary document were calculated. The average Simhash text similarity rates obtained by the algorithms are given in Table 8.

Table 8. Results obtained by summarization algorithms with entertainment document topic

Summarization Algorithm	Document Topic	Average Similarity Rate (%)	Average Operation Time(sec)
LexRank	Entertainment	40.1	1.7
TextRank	Entertainment	33.4	1.7
KL Divergence	Entertainment	33.2	2.1
Lsa	Entertainment	33.1	2.1
TF-IDF Based	Entertainment	29.7	1.9
Helmholtz-Based	Entertainment	36.7	1.7

4.5 *Experiment V*

In Experiment V, summaries of the original full-text documents of the tech document topic in the BBC News data set were generated with each extractive summarization algorithm. Then, the similarities between the summary document generated by each summarization algorithm and the original summary document were calculated. The average Simhash text similarity rates obtained by the algorithms are given in Table 9.

Table 9. Results obtained by summarization algorithms with tech document topic

Summarization Algorithm	Document Topic	Average Similarity Rate (%)	Average Operation Time(sec)
LexRank	Tech	41.0	1.6
TextRank	Tech	35.2	1.7
KL Divergence	Tech	32.6	2.1
Lsa	Tech	33.9	2.3
TF-IDF Based	Tech	30.1	2.0
Helmholtz-Based	Tech	37.7	1.8

5 Conclusion and Future Studies

This study presents the Helmholtz-based extractive summarization method to create an automatic text summarization system. The study used the BBC News data set to validate the proposed automatic document summarization model. In this data set, there are both original full-text documents and summary documents of these original documents generated by human summarizers. The Helmholtz-based extractive summarization model proposed in this study was compared with LexRank, TextRank, KL Divergence, Lsa, and TF-IDF based summarization algorithms known in the literature. Summaries of original full-text documents for five different document topics (Business, Tech, Politics, Sport, Entertainment) in the BBC News data set were generated with extractive summarization algorithms. The similarities of the generated summaries with the original summary documents in the BBC News data set were determined by the Simhash text similarity algorithm.

When the results obtained in the experiments were examined, the extractive summarization algorithm that achieves the highest average Simhash similarity rate in all different document topics in the BBC News data set is LexRank with 40.6%. The proposed Helmholtz-based summarization method achieved a Simhash similarity rate of 37.6%, close to the LexRank algorithm. The average Simhash similarity rates obtained by the algorithms in different document topics in the BBC News data set were given in Table 10.

Table 10. Similarity rates of the extractive summarization algorithms in average

Summarization Algorithm	Average Similarity Rate (%)
LexRank	40.6
TextRank	34.1
KL Divergence	33.1
Lsa	32.5
TF-IDF Based	30.1
Helmholtz-Based	37.6

When the results obtained from extractive summarization algorithms are evaluated in terms of document topics, the LexRank algorithm in business document topic obtained the highest Simhash text similarity rate. However, when examined in general, the content of the documents and the meaningful word selection methods of these algorithms directly affected the success rates rather than the document topic.

The following factors are likely to be effective in the LexRank algorithm's higher Simhash text similarity rate than other extractive summarization algorithms discussed in this study.

- Like TextRank, LexRank uses not only the PageRank approach but also similarity metrics.

- Considers position and length of sentences.

As a continuation of this study, the following studies that are thought to contribute to the literature can be discussed.

- The results obtained using a different data set instead of the BBC News data set used in this study can be evaluated comparatively.
- The proposed Helmholtz-based extractive summarization method can also be compared with other third-party algorithms known in the literature.

References

[1] Lee J. H, Park S, Ahn C. M, Kim D. “Automatic generic document summarization based on non-negative matrix factorization”. *Information Processing & Management,* 45(1), 20-34, 2009.

[2] Torres-Moreno J. M. “Automatic text summarization”. *John Wiley & Sons*, 2004.

[3] Joshi A, Fidalgo E, Alegre E, Fernández-Robles L. “SummCoder: An Unsupervised Framework for Extractive Text Summarization Based on Deep Auto-encoders”. *Expert Syst Appl.,* 129(1), 200-215, 2019.

[4] Cigir C, Kutlu M, Cicekli I. “Generic text summarization for Turkish”. *The Computer Journal*, 53(8), 1315-1323, 2010.

[5] Luhn H. P, “The automatic creation of literature abstracts”. *IBM Journal of research and development*, 2(2), 159-165, 1958.

[6] Pal A. R, Saha D. “An approach to automatic text summarization using WordNet”. *2014 IEEE International Advance Computing Conference (IACC)*, India, 1169-1173, 2014.

[7] Gunawan D, Harahap S.H, Rahmat R.F, “Multi-document Summarization by using TextRank and Maximal Marginal Relevance for Text in Bahasa Indonesia”. *2019 International Conference on ICT for Smart Society (ICISS)*, Indonesia, 1-5, 2019.

[8] Kamal A, Zhang Z, Mohammed N. “A Hierarchical Structured Self-Attentive Model for Extractive Document Summarization (HSSAS)”. *IEEE Access*. 1(1) 1-10, 2018.

[9] Vartakavi A, Garg A, Rafii Z. "Audio Summarization for Podcasts." *2021 29th European Signal Processing Conference (EUSIPCO)*, 431-435, 2021.

[10] Kaggle. “BBC News Summary”. https://www.kaggle.com/pariza/bbc-news-summary (29.08.2022).

[11] Dadachev B, Balinsky A, Balinsky H, Simske S. “On the Helmholtz Principle for Data Mining”. *Third International Conference on Emerging Security Technologies (EST), Lisbon,* Portekiz, 2012.

[12] Dadachev B, Balinsky A, Balinsky H, Simske S. “On Helmholtz’s Principle for Documents Processing,” *Proceedings of the 10th ACM Symposium on Document Engineering,* Manchester, England, 283-286, 2010.

[13] Toprak A, Turan M. "English Automatic Dictionary Creation with Natural Language Processing". *2019 Innovations in Intelligent Systems and Applications Conference (ASYU),* 1-6, Izmir, Turkiye, 2019.

[14] Raphaël K, Lei S, Abdelwahab H. “Key Elements Extraction and Traces Comprehension Using Gestalt Theory and the Helmholtz Principle”. *2016 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 478-482, NC, USA, 2016.

[15] Sarkar K, Saraf K, Ghosh A. “Improving graph based multidocument text summarization using an enhanced sentence similarity measure”. *2015 IEEE 2nd International Conference on Recent Trends in Information Systems,* 359-365, 2015.

[16] Desolneux J, Jean-Michel L. “From Gestalt Theory to Image Analysis". 34(1), 2006.

[17] Toprak A, Turan M. "The Positive Effect of PMI on the Selection of Meaningful Words". *2019 11th International Conference on Electrical and Electronics Engineering (ELECO),* 911-915, Bursa, Turkiye, 2019.

[18] Nadzurah Z.A, Ismail A.R, Emran N. “Performance Analysis of Machine Learning Algorithms for Missing Value Imputation”. *International Journal of Advanced Computer Science and Applications,* 9(6), 442-447, 2018.

[19] Derek G, Padraig C. “Practical solutions to the problem of diagonal dominance in kernel document clustering”. *ACM International Conference Proceeding,* 377-384. 2006.

[20] Balcerzak B, Jaworski W, Wierzbicki A. "Application of TextRank Algorithm for Credibility Assessment". *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT),* 451-454, 2014.

[21] Erkan G, Radev D.R. “LexRank: Graph-based Centrality as Salience in Text Summarization”. *Journal of Artificial Intelligence Research,* 22(1), 457-479, 2004.

[22] Shahzad Q, Ramsha A. “Text Mining: Use Sof TF-IDF to Examine the Relevance of Words to Documents”. *International Journal of Computer Applications,* 181(1), 25-29, 2018.

[23] Kherwa P, Bansal P. "Latent Semantic Analysis: An Approach to Understand Semantic of Text*". 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC),* 870-874, Mysore, India, 2017.

[24] Ji S, Zhang S, Ying S, Wang L, Zhao X, Gao Y. "Kullback–Leibler Divergence Metric Learning". *in IEEE Transactions on Cybernetics,* 52(4), 2047-2058, 2022.

[25] Turan M, Ögtelik S. "İngilizce Dokümanlarda Tema ve Alt Kavramlar Tespit Modeli". *Düzce Üniversitesi Bilim ve Teknoloji Dergisi,* 6(4), 754-764, 2018.