Journal of Soft Computing Exploration

Homepage: https://shmpublisher.com/index.php/joscex

p-ISSN: 2746-7686
e-ISSN: 2746-0991

Implementation of text summarization on indonesian scientific articles using textrank algorithm with TF-IDF web-based

Jeremia Jordan Sihombing[1], Arnita[2], Said Iskandar Al Idrus[3], Debi Yandra Niska[4]
[1,2,3,4]Department of Mathematics, Universitas Negeri Medan, Indonesia

Article Info

Article history:

Received Sep 6, 2024
Revised Sep 13, 2024
Accepted Nov 6, 2024

Keywords:

Indonesian scientific articles
Text summarization
Textrank
TF-IDF
Web-based system

ABSTRACT

The development of information technology has significantly changed how information is accessed, necessitating readers to absorb content efficiently and make quick decisions. To address this challenge, this research developed a text summarization system specifically for Indonesian scientific articles using a web-based implementation of the TextRank and TF-IDF algorithms. TextRank was selected for its capability to identify key sentences without requiring training data, while TF-IDF was employed to weight words based on their frequency within the document. The dataset comprised 100 scientific articles in Indonesian from the Unimed Kode Journal, covering the years 2022-2024. The summarization process included several critical stages: text preprocessing, TF-IDF weighting, cosine similarity calculation, and sentence ranking. The resulting summaries were rigorously evaluated by language experts and website specialists using a Likert scale to assess both the quality of the summaries and the usability of the system. The findings demonstrated that the system effectively generated summaries that retained essential information from the original articles, with the highest accuracy observed at a 50% compression rate (88.533%). Additionally, the system achieved good performance at 40% compression (85.133%) and 30% compression (81.26%). The web-based system allows users to input article text and quickly obtain a summary, offering a practical tool for researchers and readers to efficiently comprehend academic content.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Jeremia Jordan Sihombing,
Department of Mathematics,
Universitas Negeri Medan,
Jl. Willem Iskandar Pasar V Medan Estate, Kota Medan, Indonesia.
Email: jeremasihombing@mhs.unimed.ac.id
https://doi.org/10.52465/joscex.v5i3.475

1. INTRODUCTION

The development of information technology has brought significant changes in access to information, especially through online media [1]. Scientific articles have become an important source of information for researchers and students in various scientific fields [2]. According to data from GARUDA [3], there is a significant increasing trend in the number of scientific articles registered yearly. From the data on the number of scientific articles registered in GARUDA per year, there is a significant increasing trend from 2010 to 2022. In 2010, the number of articles registered was 27.426, and this figure increased consistently every year until it peaked in 2022 with a total of 453.690 scientific articles. The most significant increase occurred after 2017, when the number of scientific articles experienced a large spike, especially in 2021 and 2022. This increase reflects the dynamics and relevance of the role of scientific articles as a source of information in the development of research and academia. However, with the increasing number of publications, readers face challenges in absorbing information efficiently due to time constraints [4] . Therefore, readers need to absorb information efficiently and make quick decisions considering their limited time, which necessitates quick and efficient article summaries [5]. To overcome this challenge, text summarization techniques have been the focus of research in the field of natural language processing.

Text summarization is a technique that automatically generates a summary containing important sentences and includes all relevant information from the original document [6]. Text summarization consists of two main approaches: abstractive summarization and extractive summarization. Abstractive summarization involves creating a summary using new sentences [7], while extractive summarization involves selecting important sentences from the source text and combining them to form a summary [8]. Responding to the need for efficient text summarization, several web-based applications have been developed and are available today, such as paraphraser.io. These applications generally use an abstractive summarization approach to generate summaries. However, this approach has significant drawbacks. The method used tends to create new sentences that differ from the original source text, potentially leading to loss of important information and distortion of meaning [9]. Given the importance of maintaining information integrity in the summarization of scientific articles, it is necessary to develop a text summarization system that can retain important information from the source text without changing the original meaning. Therefore, this research focuses on the extractive summarization approach, which aims to develop a text summarization system that applies sentence selection methods to generate summaries of long Indonesian texts. This research chose to use a combination of the TextRank algorithm with Term Frequency-Inverse Document Frequency (TF-IDF) weighting in the extractive summarization approach. TextRank is a graph-based ranking algorithm that is effective in finding relevant sentences and keywords, and one of the advantages of the TextRank algorithm is that it does not require training using training data and has been applied in various studies for text summarization in various languages [10], [11]. TF-IDF is a weighting method that can identify important words in a document by giving each word a weight based on the frequency of its occurrence in the document and the frequency of documents containing the word [12], [13]. The combination of these two methods in the context of extractive summarization is expected to produce a more accurate and informative summary while maintaining the original sentences that are most representative of the source text. Some previous research has been done in the field of text summarization.

The research conducted by Kodicherla [14] is a comparative analysis between the TextRank and Latent Semantic Analysis (LSA) algorithms for extractive news summarization. This study used a large dataset consisting of 2225 news articles from five different categories. The performance of both algorithms was evaluated using ROUGE-N and ROUGE-L metrics. The results show that TextRank consistently outperforms LSA across all news categories, with higher precision, recall, and F-score values. Research conducted by Barrios [15] discusses the development of variations of the TextRank algorithm for automatic summarization of text. The results show that the use of the original TextRank has some limitations in capturing semantic relationships between sentences. To overcome this limitation, Barrios proposed new variations that use similarity functions such as BM25, Longest Common Substring, and others. These variations are proven to improve the algorithm's performance in producing more accurate and relevant summaries compared to the original version of TextRank. In particular, the TextRank algorithm variation using the BM25 similarity function showed the best performance. This variation achieved a ROUGE score improvement of up to 2.92% compared to the original TextRank, making it the most effective method in the study. This finding indicates that the original TextRank is less effective without modification for text summarization. Research by Eris [16] applied the TextRank algorithm for automatic summarization of Indonesian-language documents. This web-based system uses a sentence extraction method with stages including preprocessing, calculation of content overlap similarity, calculation of TextRank value, and sentence ranking. Tests on 50 technology news documents with 50% and 75% compression using the Q&A Evaluation method showed promising results, where the 50% summary was able to present 82.48% of important information, while the 75% summary reached 93.76%. This research proves the effectiveness of the TextRank algorithm in producing informative extractive summaries for Indonesian documents without the need for training data. For further development, the researcher suggests the use of other similarity methods such as cosine similarity, the application of stopword removal to improve accuracy, and the use of evaluation methods that involve the agreement of several experts. Research by Kresna [17] developed an automatic text summarization system for COVID-19 news articles using the Maximum Marginal Relevance (MMR) method. The system aims to produce a compact summary of long news articles, with the main stages including text preprocessing, TF-IDF weighting, cosine similarity calculation, and application of the MMR algorithm. Tests were conducted on 30 samples of COVID-19 news articles from Kompas.com. The results show that the best adjustment coefficient (α) is 0.5 with f-measure 0.7, the system with stemming produces a better summary (f-measure 0.7) than without stemming (f-measure 0.532), and the optimal number of words for the summary is under 300 words. This research demonstrates the effectiveness of the MMR method in generating relevant automatic summaries of COVID-19 news articles, with the best accuracy reaching 70% (f-measure 0.7). Previous research by Apriani [18] compared online news classification using the Term Frequency-Inverse Document Frequency (TF-IDF) and Term Frequency Absolute (TF-ABS) weighting methods with the Support Vector Machine (SVM) algorithm. In a study involving 2225 online news articles from the BBC, it was found that the TF-IDF method performed better than TF-ABS when combined with SVM, with an average accuracy of 96.63% and f1-score of 97.06%. Meanwhile, TF-ABS only achieved 89.66% accuracy and 96.63% f1-score. It can be concluded that TF-IDF weighting is more suitable and effective in this study. Based on these studies, we identified a gap in the development of a text summarization system specifically for Indonesian articles using a combination of TextRank and TF-IDF. Therefore, this research aims to develop and implement a text summarization system for Indonesian articles using the TextRank algorithm with web-based TF-IDF weighting. In addition, this research will also evaluate the resulting text summarization results based on expert judgment to ensure its effectiveness and accuracy. It is expected that the results of this research can make a significant contribution to helping students and researchers to get good and accurate article summaries efficiently. The web-based implementation will also facilitate the access and use of this system by readers and researchers.

2. METHOD

This study adopts a Research and Development (R&D) approach to develop a web-based text summarization application. The study will be conducted at the Mathematics Laboratory of FMIPA UNIMED, North Sumatra, Indonesia, from May to June 2024. The research process includes several main stages: data input, text pre-processing (including sentence segmentation, case folding, tokenization, punctuation removal, and stopword removal), weighting using TF-IDF, cosine similarity calculation, ranking using TextRank algorithm, and finally generating summaries based on the specified compression levels. The developed application will be validated by two experts: a linguist who will assess the quality of the summary based on relevance, grammar, non-redundancy, clarity, and coherence, and a web development expert who will evaluate the learning ability, recall ability, efficiency, errors, and user satisfaction of the application. Validation will be conducted using a Likert scale questionnaire, providing a comprehensive assessment of the quality of the summaries and the usability of the web application. The flow of this research can be seen in Figure 1.

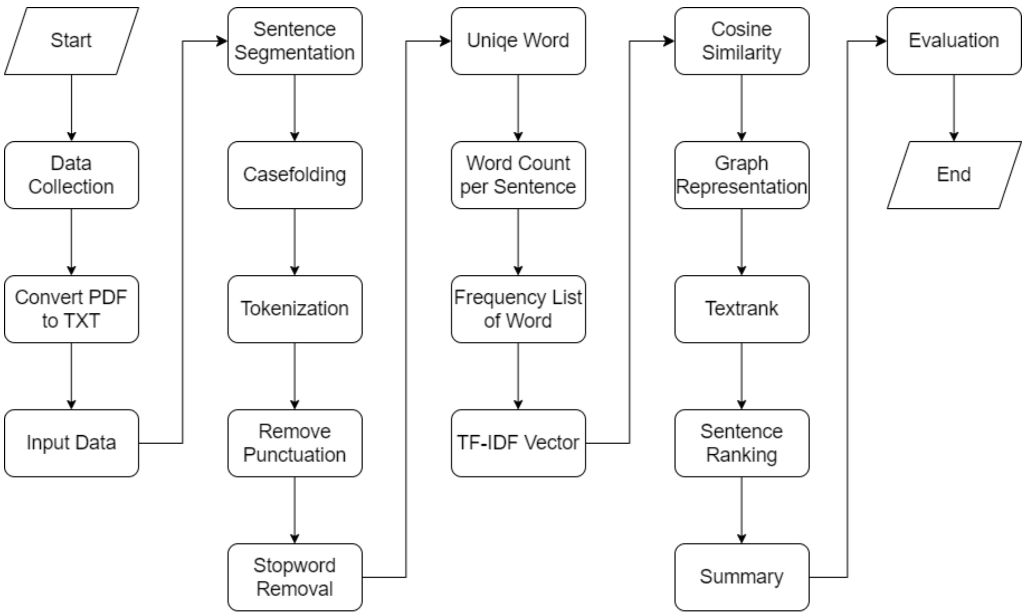


Figure 1. Research method flow

.

Data Collection

The data used in this research consists of PDF files of articles in Indonesian from the Kode Unimed journal, totaling 100 journals from the years 2022 to 2024. The Kode journal is managed by the Indonesian Language and Literature Education Program, Faculty of Language and Arts, Universitas Negeri Medan.

Input Data

The data that will be used in this research is articles in PDF format that have been converted into plain text documents with a txt extension using online conversion tools such as pdf to txt. The abstract section of the scientific articles has been removed.

Text Preprocessing

Text preprocessing is an important stage in text processing for automatic summarization. The main goal of text preprocessing is to clean and tidy up the raw text into a cleaner and more structured representation, making it easier for the next stage of analysis [19]–[21]. The process starts with sentence segmentation, which breaks the article text into individual sentences. Next, case folding converts all letters to lowercase to standardize the format. Tokenization then divides each sentence into individual words or tokens. The remove punctuation stage removes punctuation marks that are deemed unimportant for content analysis. Finally, stopword removal removes common words that do not carry significant meaning in the context of the analysis. This whole process aims to produce clean and standardized text, so that it is ready for further processing stages such as TF-IDF weighting calculation.

TF-IDF Weighting

TF-IDF (Term Frequency-Inverse Document Frequency) weighting is a key stage in the text summarization process. It begins by calculating Term Frequency (TF), which represents the frequency of occurrence of a word in each sentence. TF is calculated by dividing the number of occurrences of a word in a sentence by the total number of words in the sentence. Next, Inverse Document Frequency (IDF) is calculated to measure the importance of the word in the whole document. IDF is obtained by calculating the logarithm of the total number of sentences divided by the number of sentences containing the word. The TF and IDF values are then multiplied for each word in each sentence, resulting in a TF-IDF score. This score reflects the importance of each word in the context of the sentence and the whole document. Words with higher TF-IDF scores are considered more significant in the representation of the document content. Here are the formulas for the three main components in TF-IDF weighting [22]:

$$TF(t_k, d_j) = f(t_k, d_j) \tag{1}$$

$$IDF(t_k) = log\frac{N}{dj(t)} \tag{2}$$

$$TF - IDF(t_k, d_j) = TF(t_k, d_j) \times IDF(t_k) \tag{3}$$

With F being the frequency of occurrence of a term, N the total number of documents, and $dj(t)$ the number of documents containing that term. The TF-IDF method [22] allows accurate weighting of the importance of words in documents. TF measures the frequency of a word in a single document, while IDF measures the importance of a word across the entire set of documents. The combination of the two results in a score that reflects the relative importance of each word. A word with high frequency in one document but rare in other documents gets a high TF-IDF score. These weighting results become the basis for the next stage in automatic text summarization, such as cosine similarity calculation and sentence ranking.

Cosine Similarity

Cosine similarity is used to measure the similarity between sentences in the text summarization process [23]. The formula used is:

$$\cos\theta = \frac{A \cdot B}{||A||B||} \tag{4}$$

Explanation:

$A \cdot B$ = The dot product of vector A and vector B
$|A|$ = The magnitude of vector A
$|B|$ = The magnitude of vector B

$||A||B||$ = The magnitudes of vector A and vector B

The values of A and B represent the TF-IDF vectors of the two sentences being compared [24]. The calculation is done by dividing the dot product of the two vectors by the product of their vector lengths. The result of the calculation ranges from 0 to 1, where 0 indicates no similarity and 1 signifies identical sentences. This similarity value plays an important role in the graph formation for the TextRank algorithm. In the graph, the sentences act as nodes, while the cosine similarity value determines the weight of the edges connecting the nodes. This method is effective in capturing semantic relationships between sentences, thus enabling accurate representation of text structure for summarization purposes.

Sentence Ranking

The sentence ranking method in this study adopts the TextRank algorithm, which is an adaptation of PageRank for text analysis [25]. The process starts by forming an undirected graph, where sentences act as nodes and cosine similarity values as edge weights. The TextRank algorithm is then run iteratively, using the formula:

$$WS(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} w_{ji} \times WS(V_j) \quad (5)$$

Explanation:

$WS(V_i)$ = Weight of sentence V_i
d = Damping factor, a value between 0 and 1
V_j = Set of neighboring nodes of V_i
$In(V_i)$ = Represents the edges connected to node V_i
w_{ji} = The weight of the edge between node j and node i
$WS(V_j)$ = Weight of node V_j

with d as the damping factor which is generally 0.85. Iteration continues until the score reaches convergence or the maximum number of iterations is reached. Next, the sentences are sorted based on their final score, with the highest score being the most important. The final summary is created by selecting the top sentences according to the desired compression level, which is 30%, 40%, or 50% of the original text length. This method allows the identification of key sentences based on structural and semantic relationships within the text, resulting in a summary that reflects the core of the original document.

Evaluation

This study employs a Likert scale with five categories (Excellent, Good, Fair, Poor, and Very Poor) to assess the quality of text summaries and the implementation of a web-based system [26]. This study employs expert validation from two domains: language and web. The questionnaire for language validators evaluates aspects such as Relevance, Grammaticality, Non-redundancy, Clarity, and Coherency of the journal text summaries [27]. Conversely, the questionnaire for web validators covers Learnability, Memorability, Efficiency, Errors, and Satisfaction [28], focusing on user interface, ease of use, and overall system performance. Average scores for each criterion are calculated using the formula:

$$P = \frac{F}{N} \times 100\% \quad (6)$$

where P represents the percentage score, F is the obtained score, and N is the maximum score. The percentage scores and eligibility levels are divided into categories based on the criteria shown in Table 2.

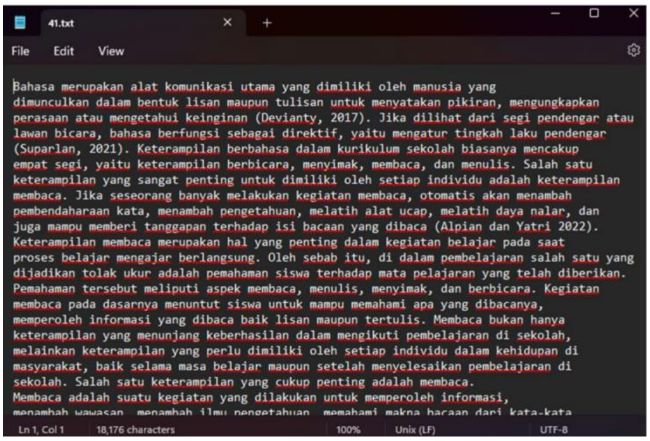
Table 1. Eligibility score rating

Eligibility	Score	Percentage %
Excellent	5	80% < total score ≤ 100%
Good	4	60% < total score ≤ 80%
Fair	3	40% < total score ≤ 60%
Poor	2	20% < total score ≤ 40%
Very Poor	1	0% < total score ≤ 20%

3. RESULTS AND DISCUSSIONS

Data Collection

This study utilizes secondary data consisting of Indonesian language articles in PDF format, sourced from Jurnal Kode Unimed from 2022 to 2024, with a total of 100 article files. The PDF files were then converted into plain text (.txt) documents using online conversion tools, such as PDF to TXT converters. Prior to their use in the study, the abstracts of each scientific article were removed. Figure 2 shows an example of the data converted into TXT format.


41.txt
File Edit View

Bahasa merupakan alat komunikasi utama yang dimiliki oleh manusia yang
dimunculkan dalam bentuk lisan maupun tulisan untuk menyatakan pikiran, mengungkapkan
perasaan atau mengetahui keinginan (Devianty, 2017). Jika dilihat dari segi pendengar atau
lawan bicara, bahasa berfungsi sebagai direktif, yaitu mengatur tingkah laku pendengar
(Suparlan, 2021). Keterampilan berbahasa dalam kurikulum sekolah biasanya mencakup
empat segi, yaitu keterampilan berbicara, menyimak, membaca, dan menulis. Salah satu
keterampilan yang sangat penting untuk dimiliki oleh setiap individu adalah keterampilan
membaca. Jika seseorang banyak melakukan kegiatan membaca, otomatis akan menambah
pembendaharaan kata, menambah pengetahuan, melatih alat ucap, melatih daya nalar, dan
juga mampu memberi tanggapan terhadap isi bacaan yang dibaca (Alpian dan Yatri 2022).
Keterampilan membaca merupakan hal yang penting dalam kegiatan belajar pada saat
proses belajar mengajar berlangsung. Oleh sebab itu, di dalam pembelajaran salah satu yang
dijadikan tolak ukur adalah pemahaman siswa terhadap mata pelajaran yang telah diberikan.
Pemahaman tersebut meliputi aspek membaca, menulis, menyimak, dan berbicara. Kegiatan
membaca pada dasarnya menuntut siswa untuk mampu memahami apa yang dibacanya,
memperoleh informasi yang dibaca baik lisan maupun tertulis. Membaca bukan hanya
keterampilan yang menunjang keberhasilan dalam mengikuti pembelajaran di sekolah,
melainkan keterampilan yang perlu dimiliki oleh setiap individu dalam kehidupan di
masyarakat, baik selama masa belajar maupun setelah menyelesaikan pembelajaran di
sekolah. Salah satu keterampilan yang cukup penting adalah membaca.
Membaca adalah suatu kegiatan yang dilakukan untuk memperoleh informasi,
menambah wawasan, menambah ilmu pengetahuan, memahami makna bacaan dari kata-kata

Ln 1, Col 1 18,176 characters 100% Unix (LF) UTF-8

Figure 2. Example of an article

Text Preprocessing

An example of an article before text preprocessing can be seen in Table 2. This table illustrates the initial content of the article before any preprocessing steps were applied.

Table 2. Article prior to text preprocessing

Article	Content
1	*Bahasa merupakan alat komunikasi utama yang dimiliki oleh manusia yang dimunculkan dalam bentuk lisan maupun tulisan untuk menyatakan pikiran, mengungkapkan perasaan atau mengetahui keinginan (Devianty, 2017). Jika dilihat dari segi pendengar atau lawan bicara, bahasa berfungsi sebagai direktif, yaitu mengatur tingkah laku pendengar (Suparlan, 2021). Keterampilan berbahasa dalam kurikulum sekolah biasanya mencakup empat segi, yaitu keterampilan berbicara, menyimak, membaca, dan menulis. Salah satu keterampilan yang sangat penting untuk dimiliki oleh setiap individu adalah keterampilan membaca. Jika seseorang banyak melakukan kegiatan membaca, otomatis akan menambah pembendaharaan kata, menambah pengetahuan, melatih alat ucap, melatih daya nalar, dan juga mampu memberi tanggapan terhadap isi bacaan yang dibaca (Alpian dan Yatri 2022).*

Next, sentence segmentation is performed, where the article text is divided into individual sentences. For this step, the researcher uses the `sent_tokenize` function from the `nltk.tokenize` module in the NLTK library. This function splits long texts into separate sentences. The results of the sentence segmentation are shown in Table 3.

Table 3. Article after sentence segmentation

Article	Content
1	*Bahasa merupakan alat komunikasi utama yang dimiliki oleh manusia yang dimunculkan dalam bentuk lisan maupun tulisan untuk menyatakan pikiran, mengungkapkan perasaan atau mengetahui keinginan (Devianty, 2017).* *Jika dilihat dari segi pendengar atau lawan bicara, bahasa berfungsi sebagai direktif, yaitu mengatur tingkah laku pendengar (Suparlan, 2021).* *Keterampilan berbahasa dalam kurikulum sekolah biasanya mencakup empat segi, yaitu keterampilan berbicara, menyimak, membaca, dan menulis.* *Salah satu keterampilan yang sangat penting untuk dimiliki oleh setiap individu adalah keterampilan membaca.* *Jika seseorang banyak melakukan kegiatan membaca, otomatis akan menambah pembendaharaan kata, menambah pengetahuan, melatih alat ucap, melatih daya nalar, dan juga mampu memberi tanggapan terhadap isi bacaan yang dibaca (Alpian dan Yatri 2022).*

After performing the sentence segmentation step, the next step is case folding. The researcher uses the lower() function because lower() is a string method used to convert all letters in a string to lowercase. The results of the case folding are shown in Table 4.

Table 4. Article after case folding

Article	Content
1	*bahasa merupakan alat komunikasi utama yang dimiliki oleh manusia yang dimunculkan dalam bentuk lisan maupun tulisan untuk menyatakan pikiran, mengungkapkan perasaan atau mengetahui keinginan (Devianty, 2017).* *jika dilihat dari segi pendengar atau lawan bicara, bahasa berfungsi sebagai direktif, yaitu mengatur tingkah laku pendengar (Suparlan, 2021).*

After performing the case folding step, the next step is tokenization. Tokenization is the process of dividing text, which can be sentences, into tokens or specific parts. The results of the tokenization are shown in Table 5.

Table 5. Article after tokenization

Article	Content
1	*['bahasa', 'merupakan', 'alat', 'komunikasi', 'utama', 'yang', 'dimiliki', 'oleh', 'manusia', 'yang', 'dimunculkan', 'dalam', 'bentuk', 'lisan', 'maupun', 'tulisan', 'untuk', 'menyatakan', 'pikiran', ',', 'mengungkapkan', 'perasaan', 'atau', 'mengetahui', 'keinginan', '(', 'devianty', ',', '2017', ')', '.']*

After performing the tokenization step, the next step is removing punctuation. Removing punctuation is the process of eliminating punctuation marks from the text because they are considered unnecessary. For example, punctuation marks such as periods, commas, question marks, and exclamation points are removed to reduce the load during text processing. The results of the removal of punctuation are shown in Table 6.

Table 6. Article after removing punctuation

Article	Content
1	*['bahasa', 'merupakan', 'alat', 'komunikasi', 'utama', 'yang', 'dimiliki', 'oleh', 'manusia', 'yang', 'dimunculkan', 'dalam', 'bentuk', 'lisan', 'maupun', 'tulisan', 'untuk', 'menyatakan', 'pikiran', 'mengungkapkan', 'perasaan', 'atau', 'mengetahui', 'keinginan', 'devianty', '2017']*

After performing the removal of punctuation step, the next step is stopword removal. Stopword removal is the process of eliminating words that are considered unimportant in a sentence. Examples include "dan," "itu," and "tapi." The results of the stopword removal are shown in Table 7.

Table 7. Article after stopword removal

Article	Content
1	*bahasa merupakan alat komunikasi utama dimiliki manusia dimunculkan bentuk lisan maupun tulisan menyatakan pikiran mengungkapkan perasaan mengetahui keinginan devianty 2017*

TF-IDF Weighting

The weighted content of the training program, calculated using TF-IDF, is presented in Table 8. This table displays the TF-IDF weights for all terms across each document or training program.

Table 8. TF-IDF weighting results

Article	*keinginan*	*laku*	*menulis*	*bahasa*	…
1	0.2472041	0.0	0.0	0.1797112	…
2	0.0	0.26089694	0.0	0.1896655	…
3	0.0	0.0	0.2725653	0.0	…
…	…	…	…	…	…

Cosine Similarity

After obtaining the TF-IDF vectors, the next step is to calculate cosine similarity. Cosine similarity measures the similarity between two vectors, in this case, TF-IDF vectors for each sentence. The calculation involves multiplying the two vectors and dividing by the product of their magnitudes. The researcher uses the `cosine_similarity` function from the `sklearn.metrics.pairwise` module in scikit-learn. The table 9 below shows the results of the cosine similarity for the article.

Table 9. Cosine similarity results

Article	1	2	3	…	115	116	117
1	1.0000	0.0341	0.0000	…	0.0000	0.0000	0.0375
2	0.0341	1.0000	0.0614	…	0.0000	0.0000	0.0000
3	0.0000	0.0614	1.0000	…	0.0000	0.0092	0.0428
…	…	…	…	…	…	…	…
115	0.0000	0.0000	0.0000	…	1.0000	0.2868	0.0534
116	0.0000	0.0000	0.0092	…	0.2868	1.0000	0.1625
117	0.0375	0.0000	0.0428	…	0.0534	0.1625	1.0000

Ranking

The cosine similarity results are used for ranking with the TextRank algorithm, which constructs a graph where each sentence is a node, and the similarity between sentences forms edges with weights based on cosine similarity. The process begins by constructing the graph with sentences as nodes and edges based on

.

cosine similarity. Ranking is performed over 100 iterations, using a damping factor of 0.85 and a convergence tolerance of 0.000001 to ensure stable results. After the process is complete, sentences are ranked based on their final scores, resulting in a list from most important to least important. The sentence rankings are shown in Table 10.

Table 10. Ranking results

Rank	Score	Sentence Index	Sentence
1	0.01454	91	*Kemampuan membaca pemahaman siswa diperoleh melalui observasi dan tes kemampuan membaca pemahaman dengan jumlah 10 butir soal, yang terdiri atas 3 tingkat pemahaman sesuai teori Ruddell (Zuchdi, 2008) yaitu pemahaman faktual, Interpretif dan aplikatif.*
2	0.01375	116	*Berdasarkan data tersebut, dapat disimpulkan bahwa membaca pemahaman siswa SMP 14 Kota Jambi dikategorikan kurang mampu karena pemahaman faktual dan aplikatifnya belum sepenuhnya dapat dipahami oleh siswa di kelas VIIIB.*
…	…	…	…
116	0.00454	67	*Ketujuh subketerampilan yang dikategorikan oleh Ruddell adalah sebagai berikut.*
117	0.00448	2	*Jika dilihat dari segi pendengar atau lawan bicara, bahasa berfungsi sebagai direktif, yaitu mengatur tingkah laku pendengar (Suparlan, 2021).*

Table 8 shows that the highest-scoring sentence is at index 91, while the lowest-scoring sentence is at index 2. After ranking, the number of sentences to be used for summarization (compression rate) needs to be determined. In this study, three compression rates are used: 30%, 40%, and 50%. The calculations for each rate are as follows: for a 30% compression rate, the summary includes 35 sentences; for a 40% rate, it includes 46 sentences; and for a 50% rate, it includes 58 sentences. The selected sentences are then combined to create the summary, preserving the original order of the text.

Evaluation

The evaluation of the text summarization system was conducted by two groups of validators: language experts and website experts. Language experts reviewed 20 journal summary samples using a validity questionnaire. The evaluation was conducted using a Likert scale of 1-5, as explained in Table 1**.** This scale measures the quality and adequacy of the summaries based on aspects such as information completeness, readability, and accuracy. The Table 11 below shows the results of the evaluation based on the scores and adequacy percentages calculated by both groups of validators.

Table 11. Ranking results

Evaluation Type	Compression Rate	Score Obtained	Maximum Score	Adequacy Percentage	Rating
Language Experts Evaluation	30%	1219	1500	81.27%	Excellent
	40%	1276	1500	85.13%	Excellent
	50%	1328	1500	88.53%	Excellent
Website Comparison (paraprasher.io)	-	1153	1500	73.4%	Good
Website Experts Evaluation	-	53	75	70.67%	Good

Table 11 presents the evaluation results of the text summarization system conducted by language and website experts. The system performed excellently in summarizing Indonesian articles, with adequacy percentages improving as the compression rate decreased, reaching up to 88.53%. In comparison, another website's summaries received a "Good" rating with a score of 73.4%. Website experts also rated the system as "Good" with a score of 70.67%. The Score Obtained reflects the points earned by the system based on expert evaluations, while the Maximum Score represents the highest possible score. The Adequacy Percentage is the ratio of the score obtained to the maximum score, expressed as a percentage, indicating the system’s overall effectiveness in summarizing texts.

System Implementation

The next step is the website development phase. The website is built with HTML, CSS, and JavaScript for the front-end, while text summarization is implemented on the back-end using Flask. The process works as follows: when an article is input, it is processed in Python and then the results are returned to the front-end. The website is a single-page application. Figure 3 is a screenshot of the website.

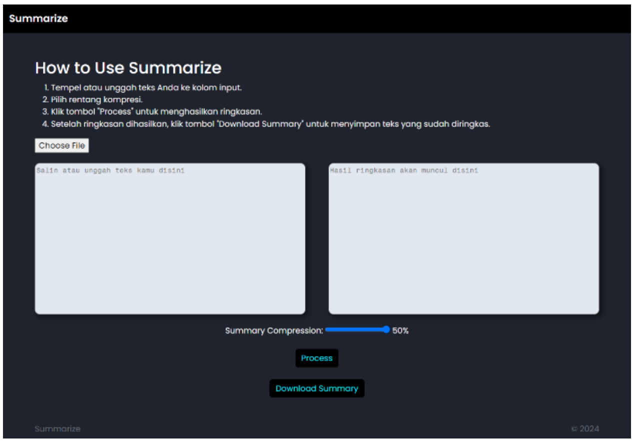


Figure 3. Main page of the website

The front-end interface consists of several elements: a header displaying the platform name "Summarize," usage guidelines, a text area for input on the left, an output text area on the right, a "Choose File" button for uploading .txt files, a slider to adjust the compression rate (30-50%, default 50%), and buttons for "Process" and "Download Summary." The text summarization workflow is designed to offer a seamless user experience. Users input text into the left text area or upload a file, adjust the compression level using the slider, and press "Process" to begin summarization.

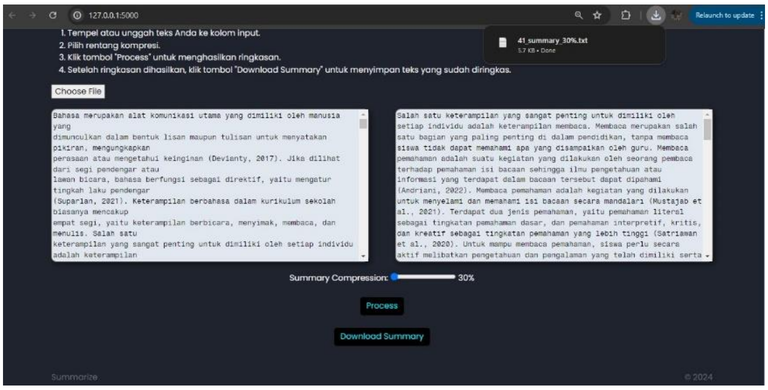


Figure 4. Summary results

Once the process is triggered, the input text and compression rate are sent to the back-end, where the summarization algorithm processes the text and generates a summary. The summary is displayed in the output area on the right, which initially reads "Summary will appear here" until the process is complete. Users can review the summary directly on the screen, and if satisfied, download it using the "Download Summary" button located beneath the input and output areas. Figure 4 is the summary result created by the website.

4. CONCLUSION

The conclusions from this study indicate that the developed system is capable of effectively generating summaries of Indonesian articles while retaining key sentences, with the best performance at a 50% compression rate, achieving 88.533%, followed by 40% at 85.133%, and 30% at 81.26%. When compared to another website, this system's summary results were superior, scoring 73.4%. Evaluations by language and website experts also showed that the system performs excellently in preserving important information, even at higher compression rates, and is suitable for summarizing Indonesian text.

As recommendations, future research is encouraged to explore other automatic text summarization methods, such as BERT, and consider using abstractive summarization. Additionally, further development could involve using a larger and more diverse dataset, including other languages, to enhance the system's performance and flexibility.

.

REFERENCES

[1] M. Ngafifi, “Kemajuan Teknologi Dan Pola Hidup Manusia Dalam Perspektif Sosial Budaya,” *TUTURAN J. Ilmu Komunikasi, Sos. dan Hum.*, vol. 1, no. 3, pp. 33–47, 2023, doi: 10.47861/tuturan.v1i3.272.

[2] S. Suherman, “Online Research Skills (ORS): Solusi Literasi Mahasiswa Perguruan Tinggi di Masa Pandemi Covid-19,” *IQRA` J. Ilmu Perpust. dan Inf.*, vol. 16, no. 1, p. 70, 2022, doi: 10.30829/iqra.v16i1.10153.

[3] Kemendikbud, “Garuda - Garba Rujukan Digital.” Accessed: Feb. 20, 2024. [Online]. Available: https://garuda.kemdikbud.go.id/home/graphic

[4] D. Fatmalasari and F. R. Lumbanraja, “Peringkasan Teks Artikel Ilmiah Berbahasa Indonesia dengan Metode Pembobotan Kalimat,” *J. Pepadun*, vol. 3, no. 3, pp. 314–322, 2022, doi: 10.23960/pepadun.v3i3.127.

[5] D. Fitrianah and R. N. Jauhari, “Extractive text summarization for scientific journal articles using long short-term memory and gated recurrent units,” *Bull. Electr. Eng. Informatics*, vol. 11, no. 1, pp. 150–157, 2022, doi: 10.11591/eei.v11i1.3278.

[6] A. P. Widyassari *et al.*, “Review of automatic text summarization techniques & methods,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1029–1046, 2022, doi: 10.1016/j.jksuci.2020.05.006.

[7] Q. A. Itsnaini, M. Hayaty, A. D. Putra, and N. A. . Jabari, “Abstractive Text Summarization using Pre-Trained Language Model ‘Text-to-Text Transfer Transformer (T5),’” *Ilk. J. Ilm.*, vol. 15, no. 1, pp. 124–131, 2023, doi: 10.33096/ilkom.v15i1.1532.124-131.

[8] V. Gupta and G. S. Lehal, “A Survey of Text Summarization Extractive techniques,” *J. Emerg. Technol. Web Intell.*, vol. 2, no. 3, pp. 258–268, 2010, doi: 10.4304/jetwi.2.3.258-268.

[9] R. Paulus, C. Xiong, and R. Socher, “A deep reinforced model for abstractive summarization,” *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.*, no. i, pp. 1–12, 2018.

[10] V. K. Manojkumar, S. Mathi, and X. Z. Gao, “An Experimental Investigation on Unsupervised Text Summarization for Customer Reviews,” *Procedia Comput. Sci.*, vol. 218, pp. 1692–1701, 2022, doi: 10.1016/j.procs.2023.01.147.

[11] M. F. Hidayattullah and A. Azizi, “Peringkas Otomatis Teks Berbahasa Arab Menggunakan Algoritma TextRank,” *J. Ilm. Inform.*, vol. 6, no. 1, pp. 33–42, 2021, doi: 10.35316/jimi.v6i1.1231.

[12] N. K. Widyasanti, I. K. G. Darma Putra, and N. K. Dwi Rusjayanthi, “Seleksi Fitur Bobot Kata dengan Metode TFIDF untuk Ringkasan Bahasa Indonesia,” *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 6, no. 2, p. 119, 2018, doi: 10.24843/jim.2018.v06.i02.p06.

[13] R. Ramadhan, Y. A. Sari, and P. P. Adikara, “Perbandingan Pembobotan Term Frequency-Inverse Document Frequency dan Term Frequency-Relevance Frequency terhadap Fitur N-Gram pada Analisis Sentimen,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 11, pp. 5075–5079, 2021, [Online]. Available: https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/10173

[14] P. Kodicherla, S. R. Sathineni, and J. S. Sai, “Comparative Analysis of TextRank and Latent Semantic Analysis Algorithms for Extractive News Summarization,” *2023 3rd Asian Conf. Innov. Technol. ASIANCON 2023*, pp. 1–6, 2023, doi: 10.1109/ASIANCON58793.2023.10270050.

[15] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, “Variations of the Similarity Function of TextRank for Automated Summarization,” 2016, [Online]. Available: http://arxiv.org/abs/1602.03606

[16] Eris, V. C. M, and J. Pragantha, “PENERAPAN ALGORITMA TEXTRANK UNTUK AUTOMATIC SUMMARIZATION PADA DOKUMEN BERBAHASA INDONESIA,” *J. Ilmu Tek. dan Komput.*, vol. 1, no. 1, pp. 71–78, 2017.

[17] Y. Ananda Kresna, I. Cholissodin, and Indriati, “Peringkasan Teks Menggunakan Metode Maximum Marginal Relevance terhadap Artikel Berita terkait COVID-19,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 9, pp. 3901–3907, 2021, [Online]. Available: http://j-ptiik.ub.ac.id

[18] I. Apriani, Y. Sibaroni, and I. Palupi, “Perbandingan Pembobotan Fitur TF-IDF dan TF-ABS Dalam Klasifikasi Berita Online Menggunakan Support Vector Machine (SVM),” *e-Proceeding Eng.*, vol. 10, no. 3, pp. 3652–3663, 2023.

[19] M. A. Palomino and F. Aider, “Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis,” 2022.

[20] D. W. Septiana and P. B. Pastika, “Classification of travel class with k-nearest neighbors algorithm using rapidminer,” *J. Student Res. Explor.*, vol. 2, no. 2, pp. 88–100, Jul. 2024, doi: 10.52465/josre.v2i2.357.

[21] V. A. Savitri, M. Sa’id, H. Husni, and A. Muntasa, “A sentiment analysis of madura island tourism news using C4.5 algorithm,” *J. Soft Comput. Explor.*, vol. 5, no. 1, pp. 9–17, 2024, doi: 10.52465/joscex.v5i1.258.

[22] E. T. Wijaya, “Perancangan Information Retrieval (IR) Berbasis Term Frequency-Inverse Document Frequency (TF-IDF) Untuk Peringkasan Teks Tugas Khusus Berbahasa Indonesia,” *J. Ilm. Teknol. Inf. Asia*, vol. 7, no. 1, pp. 79–93, 2013.

[23] R. Samuel, R. Natan, and U. Syafiqoh, “Application of Cosine Similarity and K-Nearest Neighbor (K-NN) in Classification and Book Search,” *J. Big Data Anal. Artif. Intell.*, vol. 1, no. 1, pp. 9–14, 2018.

[24] M. F. Abdurrafi and D. H. U. Ningsih, “Content-based filtering using cosine similarity algorithm for alternative selection on training programs,” *J. Soft Comput. Explor.*, vol. 4, no. 4, pp. 204–212, 2023, doi: 10.52465/joscex.v4i4.232.

[25] R. Mihalcea and P. Tarau, “TextRank: Bringing order into texts,” *Proc. 2004 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2004 - A Meet. SIGDAT, a Spec. Interes. Gr. ACL held conjunction with ACL 2004*, vol. 85, pp. 404–411, 2004.

[26] E. T. ARYANI and H. KURNIANINGSIH, “Pengaruh Inovasi Produk, Kepercayaan Merek, Dan Keragaman Produk Terhadap Keputusan Pembelian Honda Beat Di Surakarta,” *J. Maneksi*, vol. 12, no. 2, pp. 368–377, 2023, doi: 10.31959/jm.v12i2.1488.

[27] J. Steinberger and K. Ježek, “Evaluation measures for text summarization,” *Comput. Informatics*, vol. 28, no. 2, pp. 251–275, 2009.

[28] W. Handiwidjojo and L. Ernawati, “Pengukuran Tingkat Ketergunaan (Usability) Sistem Informasi Keuangan,” *Juisi*, vol. 02, no. 01, pp. 49–55, 2016.