JoCoSiR
Journal of Computer Science Research
ISSN 2986-2337 (Online)

Text Summarization of Online News Articles Using the Text Rank Algorithm

Indra Marto Silaban

Program Studi Magister Teknologi Informasi, Universitas Pembangunan Pancabudi, Medan, Sumatera Utara

email:indra.marto.silaban@gmail.com

ARTICLE INFO

Keywords:
Summarization,
Online News Media,
Summarization, Extraction,
TextRank, PageRank

IEEE style in citing this article:
I. M. Silaban, " Aplikasi Text Summarization Dengan Metode TextRank Pada Artikel Berita Online *JoCoSiR: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 2, no. 4, pp. 10-14, 2024.

ABSTRACT

Text Summarization in online news media is useful for helping readers to get the essence of a news story. Summarization will be less effective if it is done manually by humans, so we need an application that can do summaries quickly and precisely. By utilizing preprocessing technics with sastrawi python library and the implementation of TextRank algorithm which is part of the extraction method, news that was previously long can be presented in a very concise form. This application is developed using the Python programming language with sastrawi libraries, nltk and StemmerFactory. While the framework used is Django as the backend and bootstrap as the frontend framework.

1. Introduction

The rapid development of information technology today has produced various products that significantly assist humans in their daily lives. One such product is online news media, which can deliver news to readers with great speed. Events occurring across the globe can be read almost instantly by anyone with internet access. At present, traditional news media are also competing to keep up with the trend of publishing their news online. Printed newspapers are gradually declining in use, as paper-based news presentation is far less effective and slower compared to online news media.

In addition to its speed, online news media also offer several advantages that conventional print media lack. For instance, online media managers can instantly obtain statistics about their readership. Moreover, most online news media platforms provide a comment section that allows readers to express their opinions, enabling media managers to receive direct feedback from their audience [1].

However, as human activities become increasingly busy, the time available for reading news is limited. This requires news articles to be concise, clear, and dense in information. Advances in information technology, particularly in intelligent systems for data processing, can help address this issue. Through intelligent system approaches, lengthy articles can be summarized into shorter forms without losing their meaning. Summarization itself is a technique for condensing long articles into concise summaries composed of fewer words or sentences. This technique provides the most important information from an article in a shorter version

State of the Art

Text summarization has evolved into a crucial area of Natural Language Processing (NLP), especially in the era of digital information overload. Early summarization techniques were primarily based on statistical and heuristic approaches, such as *Term Frequency–Inverse Document Frequency (TF-IDF)* and *Latent Semantic Analysis (LSA)*, which focus on word occurrence and co-occurrence patterns to extract key sentences.

In recent years, graph-based algorithms, particularly TextRank (Mihalcea & Tarau, 2004), have become one of the most influential methods in extractive summarization. TextRank applies the concept of PageRank, initially designed for web page ranking, to identify the most relevant sentences within a document based on their connectivity and importance within a sentence graph. Several studies have demonstrated the robustness of TextRank for multilingual summarization and domain-specific tasks such as news and scientific abstracts [4].

Further advancements have led to hybrid methods that combine statistical and semantic analysis using word embeddings (e.g., Word2Vec, GloVe) to capture contextual meaning rather than relying solely on word frequency [5]. More recently, neural-based and transformer-based models, such as BERTSUM (Liu & Lapata, 2019) and PEGASUS (Zhang et al., 2020), have revolutionized the field by enabling abstractive summarization, where the system generates new sentences that preserve the original meaning.

Despite these advancements, the application of sophisticated models in Indonesian-language news summarization remains relatively limited due to resource constraints such as dataset availability, computational cost, and linguistic variation. Therefore, applying and optimizing TextRank-based extractive summarization for

ISSN 2502-3357 (Online)

Indonesian online news media (e.g., *Kompas.com*, *Detik.com*) provides a practical and explainable solution that balances performance, interpretability, and computational efficiency. [2]

2. Research Methodology

Research Stage

The initial stage of this research involved data collection through observation of several Indonesian online news media, specifically kompas.com and detik.com [3]. The observations revealed that news categorized as breaking news tends to be presented in a series of short articles, while regular news is generally written in longer articles that are divided into several pages. Some news items are also equipped with multimedia elements such as image slides and videos. After identifying these patterns, the researcher selected text-based news articles as the primary data for processing.

Extraction Method

From the perspective of output, summarization can be performed using two main approaches: extraction and abstraction. In simple terms, the extraction method works by selecting words, phrases, or sentences directly from the source document to form a summary. In contrast, the abstraction method generates summaries that may use different wording from the original text but still convey the same essential meaning. In fact, the words in an abstractive summary may not appear at all in the original article, yet they maintain equivalent meaning [4].

As explained above, the extraction method in summarization involves selecting words, phrases, or sentences, calculating their ranking, and then choosing those with the highest scores to form the summary. The parameters used in the extraction process are as follows [5]

a. Frequency
   Words that appear more frequently are considered more important within a document. The more often a word appears, the higher its score will be. The most common method for calculating word frequency is TF-IDF (Term Frequency–Inverse Document Frequency).
b. Title/Headline
   Words appearing in the title or main headline are usually closely related to the summary, as they often indicate the main topic of a document.
c. Sentence Length
   Sentences used in summaries are generally not too long and not too short.
d. Similarity
   Similarity can be measured using linguistic knowledge. It reflects how similar the sentences in the body of the document are to those in the title.
e. Proximity
   The distance between words in an entity serves as a factor in determining the relationships among entities.

TextRank Method

The **T**extRank method used in this study is one of the extraction-based techniques. TextRank is an algorithm adapted from Google's PageRank algorithm. As illustrated in Figure 1, A, B, C, and D represent vertices or pages, while the arrows indicate edges that show the relationships or links between vertices. Figure 1 shows an example of four linked web pages:

a. Page A links to (B)
b. Page B links to (A, C, D)
c. Page C links to (B, D)
d. Page D has no outgoing links (dangling page).

The ranking of each page is determined based on the following principles:

a. The probability of a link between *j* and *i* is 1 / (number of unique links on the page)**.**
b. If there is no link between *j* and *i*, the probability is 0**.**
c. If a user is on a dangling page, the probability is 1 / (total number of pages)**.**

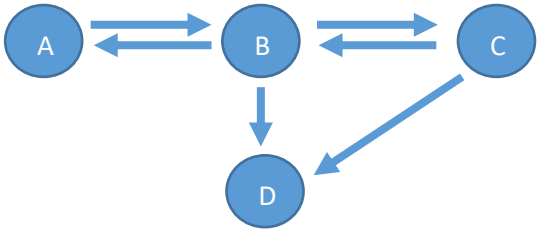

Figure 1. Four pages on a website

Based on the rules described above, the following calculation results are obtained:

Table 1. PageRank Calculation Results

	A	B	C	D
A	0	1	0	0
B	0,25	0,25	0,25	0,25

ISSN 2502-3357 (Online)

	A	B	C	D
C	0	0,5	0	0,5
D	0,25	0,25	0,25	0,25

The final ranking results, arranged according to their PageRank scores, are B, D, (A, C). TextRank is a graph-based ranking algorithm that can be applied to process human language text from a single document [6]. There are two main types of linguistic processing in TextRank, namely:

1. TextRank for Keyword Extraction – used to identify the most significant words or phrases representing the core topics of a document.
2. TextRank for Sentence Extraction – used to select and rank sentences according to their importance, which are then combined to form a concise summary of the text.

Steps of Text Summarization

As previously explained, the TextRank algorithm is an adaptation of the PageRank algorithm [7]. The general process of text summarization using TextRank can be illustrated as shown in Figure 2.

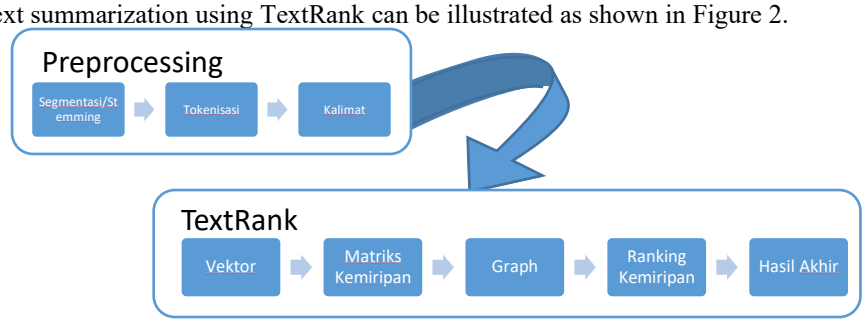

Figure 2. Steps of Text Summarization

a. Preprocessing

The preprocessing stage consists of two main steps: segmentation and tokenization. In the segmentation process, sentences within the document are divided into individual sentences. The first step in segmentation is stemming, which refers to the process of removing affixes (prefixes, infixes, and suffixes) to obtain the root form of each word [8]. The stemming process typically relies on a predefined dictionary (lexical library) that contains words and their stemmed forms. After stemming, the text is divided into single sentences based on punctuation marks such as a period (“.”), question mark (“?”), exclamation mark (“!”), and newline characters [6].

Segmentation also includes a stop-word removal procedure, which removes less meaningful or commonly used words that do not contribute significantly to the meaning of the sentence. Examples of such stop words in Indonesian include “yang”, “di”, “dari”, and other similar terms. The tokenization process follows segmentation. Tokenization involves collecting all separated sentences into a sequence (array) of tokens that will serve as input for the next phase of processing.

b. TextRank

The text processing type discussed in this study is TextRank for sentence extraction, which constructs relationships among sentences in the form of a graph. In the TextRank algorithm, each single sentence obtained from segmentation acts as a vertex (node), analogous to the concept used in PageRank, while the edges represent the degree of similarity between pairs of sentences [9].

The similarity between two sentences can be measured using methods such as cosine similarity, Jaccard similarity, or semantic similarity based on word embeddings. The stronger the similarity between two sentences, the higher the weight of the edge connecting them. Sentences with higher cumulative edge weights will have higher TextRank scores, indicating greater importance in the context of the overall document.

3. Results and Discussion

System Design

The system developed in this study is a text summarization application designed to summarize news articles from the online media platform *detik.com*, implemented using the Python programming language. The system is web-based and developed using the Django framework.

For the preprocessing stage, several libraries are utilized to handle linguistic processing in the Indonesian language, namely Sastrawi, NLTK, and Stemmer Factory [10]. The Sastrawi library, in particular, is used for stemming and stop-word removal to obtain root words from the original text. The first interface of the system displays a list of news articles retrieved from the URL https://news.detik.com/indeks, which shows daily news updates at the time of access. This interface can be seen in Figure 3.

After selecting one of the news articles from the list, the system automatically performs text summarization using the TextRank algorithm. The resulting summary is then displayed on the next page, as shown in Figure 4.

ISSN 2502-3357 (Online)

Figure 3. News list page retrieved from the detik.com index

Figure 4. Summary display page of the selected news

4. Conclusions

Based on the research conducted, several conclusions can be drawn as follows: The TextRank algorithm has proven to be effective in generating summaries of online news articles, where approximately 50% of the entire article content can be represented as a concise and meaningful summary. This significantly reduces reading time while preserving the core information of the original article. For Indonesian language preprocessing, the Sastrawi Python library has shown to be highly capable and practical. It serves as a reliable tool for stemming and stop-word removal, making it particularly suitable for applications in Natural Language Processing (NLP) involving Indonesian-language texts.

5. References

[1] A. Romadhony, F. Z.R, N. Yusliani, and L. Abednego, "Text Summarization untuk Dokumen Berita Berbahasa Indonesia," *Konferensi Nasional ICT-M Politeknik Telkom*, 2017.

[2] M. A. Zamzam, "SISTEM AUTOMATIC TEXT SUMMARIZATION MENGGUNAKAN ALGORITMA TEXTRANK," *MATICS*, vol. 12, no. 2, pp. 111–116, Sep. 2020, doi: 10.18860/mat.v12i2.8372.

[3] S. Tuhpatussania, "Automatic Text Summarization Artikel Berita Menggunakan Metode Maximum Marginal Relevance," 2022. Accessed: Nov. 19, 2022. [Online]. Available: http://download.garuda.kemdikbud.go.id/

[4] R. Adelia, S. Suyanto, and U. N. Wisesty, "Indonesian abstractive text summarization using bidirectional gated recurrent unit," in *Procedia Computer Science*, 2019. doi: 10.1016/j.procs.2019.09.017.

[5] Yulyardo, Okta Purnama Rahadian, Martin Sujono, and S. Kom. , Ph. D. Amalia Zahra, "Peringkas Teks Otomatis (Automatic Text Summarization)," https://mti.binus.ac.id/2018/12/26/peringkas-teks-otomatis-automatic-text-summarization/#:~:text=Text%20Summarization%20atau%20ringkasan%20teks,suatu%20dokumen%20teks%20yang%20panjang.

[6] V. M. Christanti and J. Pragantha, "PENERAPAN ALGORITMA TEXTRANK UNTUK AUTOMATIC SUMMARIZATION PADA DOKUMEN BERBAHASA INDONESIA," 2017.

ISSN 2502-3357 (Online)
[7] Y. Yuliska and K. U. Syaliman, “Literatur Review Terhadap Metode, Aplikasi dan Dataset Peringkasan Dokumen Teks Otomatis untuk Teks Berbahasa Indonesia,” *IT Journal Research and Development*, vol. 5, no. 1, 2020, doi: 10.25299/itjrd.2020.vol5(1).4688.
[8] A. Romadhony, F. Z. R, N. Yusliani, and L. Abednego, “Text Summarization untukDokumenBeritaBerbahasa Indonesia”.
[9] G. S. Budhi, R. Intan, and S. R. R, “Indonesian Automated Text Summarization.”
[10] Prateek Joshi, “An Introduction to Text Summarization using the TextRank Algorithm (with Python implementation),” https://www.analyticsvidhya.com/blog/2018/11/introduction-text-summarization-textrank-python/.