Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinfsormatic.org/

15th October 2025. Vol. 5. No. 1; e-ISSN: 2808-4519

Automatic Criminal News Summarization System with Extractive Method Based on Latent

Christ Chandra[1*], Robet[2], Kelvin Leonardi Kohsasih[3]

[1,2]*STMIK TIME, Medan,Indonesia*
Cangchrist@gmail.com

Abstract

The rapid growth of digital information demands automatic systems to help users efficiently extract the core of information, especially in criminal news which often attracts significant public attention. This study aims to design and develop an automatic summarization system for criminal news using an extractive method based on *Latent Semantic Analysis* (LSA). In the process, textual features are first extracted using the *Term Frequency-Inverse Document Frequency* (TF-IDF) method to weigh the importance of each word in the document. The resulting TF-IDF matrix is then used as input for LSA to model semantic relationships between sentences and identify those most representative of the document content. The dataset consists of Indonesian-language criminal news articles collected from various online news portals. The system is evaluated by comparing the automatically generated summaries with human-written summaries using the ROUGE metric. The experimental results show that the combination of TF-IDF and LSA can generate informative and relevant summaries, achieving a ROUGE-1 score of 0.72. This system is expected to help users understand news content quickly and efficiently.

Keywords: *Summarization System, TF-IDF, LSA, ROUGE*

1. Introduction

The rapid proliferation of digital information, particularly in the form of online news [1], presents significant challenges in efficiently extracting relevant and essential content. Criminal news, due to its societal impact, requires concise and accurate presentation to facilitate quick comprehension by readers [2]. However, the vast volume and diversity of news articles often hinder the identification of core information.To address this issue, automatic text summarization has emerged as a promising solution. By employing Natural Language Processing (NLP) techniques, such systems can automatically generate summaries that capture the main content of a document [3]. One widely adopted approach is extractive summarization [4], which selects and compiles key sentences from the source text.

Among extractive methods, Latent Semantic Analysis (LSA) has gained prominence for its ability to identify semantic relationships between words and sentences through Singular Value Decomposition (SVD) [5]. LSA has been effectively utilized in various studies to improve the quality of summaries, including those in the domain of criminal news [6][7][8].Complementary to LSA, Term Frequency-Inverse Document Frequency (TF-IDF) is frequently employed to assign weights to terms based on their importance within a document. TF-IDF helps emphasize contextually significant terms, which can enhance the relevance of the generated summaries [9][10].

Several studies have explored the integration of LSA and TF-IDF to enhance summarization effectiveness. For example, Al-Sabahi et al. demonstrated that combining these methods yields more informative and relevant summaries [11]. Nonetheless, their application to Indonesian-language criminal news remains limited.Despite existing research on text summarization using LSA and TF-IDF, a notable gap exists in their implementation for Bahasa Indonesia criminal news. Most prior works focus on English texts or other domains such as general news or financial reports [12][13]. Moreover, evaluation of summarization systems in the context of criminal news often lacks robust and standardized metrics like ROUGE**,** which can offer an objective measure of summary quality [14].

This study aims to develop an automatic summarization system for Indonesian criminal news articles, utilizing an extractive approach based on LSA and TF-IDF. The system's performance is evaluated using ROUGE metrics to assess the informativeness and relevance of the generated summaries.

2. Method

This study adopts a systematic approach to develop an automatic summarization system for Indonesian-language criminal news articles. Figure 1 illustrates the system workflow, which consists of several main stages: starting from data input (articles), preprocessing, term

weighting using Term Frequency-Inverse Document Frequency (TF-IDF), semantic feature extraction using Latent Semantic Analysis (LSA), and finally generating the summary as output.

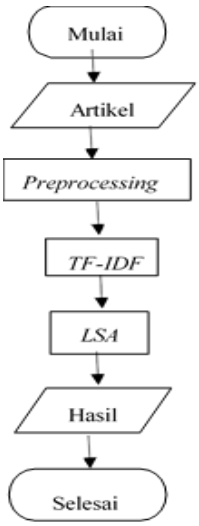

Fig. 1: Research Workflow

2.1. Data Collection

The data used in this study consist of Indonesian-language criminal news articles. The articles were manually collected from various popular Indonesian online news platforms, such as TEMPO.CO, CNN Indonesia, Kompas, Detik, and Tribun News, all of which feature dedicated sections on crime and criminality. The data were stored in plain text format (.txt) to facilitate further processing.
The criteria for data and source selection are presented in Table 1 below.

Table 1: Data Selection Criteria

Aspect	Description	Example
Data Type	The type of data used for conducting the research	Criminal news articles
Data Source	Credible sources of data	Online news portals such as TEMPO.CO, CNN Indonesia, KompasTV, Detik.com, Tribun News
Relevance Criteria	Criteria used to assess the relevance of an online news article	Incident details, time of occurrence, location, and date of the event

2.2. Preprocessing Text

The preprocessing stage aims to clean and standardize textual data to enable effective processing in subsequent stages [15]. The preprocessing steps include the following:

1. Case folding
   All text is converted to lowercase to avoid distinctions between uppercase and lowercase letters. For example, “Polisi” and “polisi” are treated as the same word.s
2. Tokenization
   This step breaks paragraphs into tokens. Tokenization allows the system to recognize individual words.
3. Stopward removal
   Common words such as “yang”, “dan”, and “di” that do not contribute significantly to semantic analysis are removed.
4. Stemming
   Words are reduced to their root forms. For instance, “pembunuhan” is reduced to “bunuh”.
5. Text normalization
   This step converts informal or inconsistent word forms into standard forms. For example, “gk” is normalized to “tidak”, and repetitive characters such as “baguuusss” are simplified to “bagus”.

2.3. Text Representation Using TF-IDF

TF-IDF is a statistical method used to evaluate the importance of a word in a document relative to a corpus. Term Frequency (TF) measures how frequently a word appears in a document, while Inverse Document Frequency (IDF) reduces the weight of words that appear frequently across many documents, as they are considered less informative.

The TF-IDF formula is as follows:

$$TF - IDF(t,d) = TF(t,d)\ x \log\left(\frac{N}{DF(t)}\right) \quad (1)$$

As shown in the formula above, the first step is to calculate the Term Frequency (TF), which refers to the frequency of a term appearing within a specific document.

$$TF(t,d)\frac{Number\ of\ times\ term\ t\ appears\ in\ document\ d}{Total\ numbers\ of\ terms\ in\ document\ d} \quad (2)$$

The formula above indicates that the more frequently a word appears in a news article, the higher its raw frequency; however, this value is normalized by the total number of terms in the document. The result of this normalization is then multiplied by the Inverse Document Frequency (IDF). This requires a further calculation using the following formula:

$$IDF(t) = \log\left(\frac{N}{DF(t)}\right) \quad (3)$$

The IDF calculation reflects the idea that the more frequently a word appears across multiple documents, the closer its IDF value approaches zero, indicating that the word carries little informational value. Once both the TF and IDF values are obtained, they are multiplied to compute the TF-IDF score. If the resulting value is close to zero, the term is considered irrelevant to the document.
While IDF alone focuses on the distribution of terms across documents, TF-IDF combines both term frequency and inverse document frequency, providing a more comprehensive measure of term importance. In the context of this study, TF-IDF plays a crucial role in evaluating the significance of words throughout the entire news article corpus targeted for summarization.

2.4. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a semantic analysis method used to identify latent relationships between words and sentences within a document. This process is carried out by applying Singular Value Decomposition (SVD), a matrix decomposition technique that transforms the TF-IDF matrix into a simpler, more analyzable form. Fundamentally, SVD decomposes a matrix A into three component matrices, expressed as:

$$A=U\Sigma VT \quad (4)$$

In this context, LSA leverages the SVD method to uncover hidden semantic representations within the text. LSA can be used to analyze the relationships between words and extract deeper meanings embedded in the textual content.

2.4. Result

After all sentences in the document are scored using the results of LSA, the system selects a number of sentences with the highest scores. The number of selected sentences is adjusted based on the desired summary length, typically ranging from 30% to 50% of the total number of original sentences.
Sentence selection also considers the original order of appearance in the article to maintain the summary's readability and logical flow. To evaluate the quality of the generated summary, the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric is used—specifically ROUGE-1, which measures unigram (single-word) overlap between the system-generated summary and the reference summary.
The ROUGE-1 formula is as follows:

$$Rouge-1 = \frac{Number\ of\ overlapping\ unigrams}{Total\ unigrams\ in\ the\ refrence\ summary} \quad (5)$$

The final output of the designed system, including the text summarization result and its evaluation score, is presented through a web-based interface, as shown in Figure 2.

Fig. 2: Text Summarization System Interface

3. Result and Discussion

3.1. Preprocessing and Keyword Extraction Results

After preprocessing the criminal news articles, the textual representation was cleaned from irrelevant characters and common words, as well as standardized through normalization and stemming. For example, in an article titled “Shooting of Police Officers in South Jakarta,” the preprocessing step identified important keywords such as *penembak* (shooter), *anggota* (member), *polisi* (police), *Jakarta*, and *selatan* (south). The TF-IDF process then assigned the highest weights to words that frequently appeared in a single document but were rare in others. This highlights the most relevant keywords for each article. For instance, the word *penembak* received a TF-IDF weight of 0.87 in one document, indicating a dominant semantic contribution.

3.2. Dimensionality Reduction and Sentence Selection Using LSA

Using Singular Value Decomposition (SVD) within LSA, dimensionality reduction was applied to the term-document matrix. From this, sentences with the highest contribution to the main semantic components were selected as candidate summary sentences.
As an example, the system summarized an article from which originally contained 531 words. The article was summarized to 50% of its length, producing a summary of 290 words with a ROUGE-1 score of 70.62%.

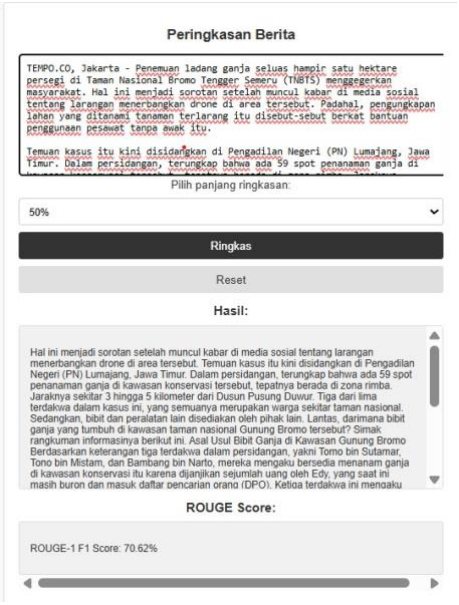

Fig. 3: Results and Evaluation with 50% Summary Length

The second experiment, using a summary length of 40%, produced 236 words with a ROUGE-1 score of 61.58%.

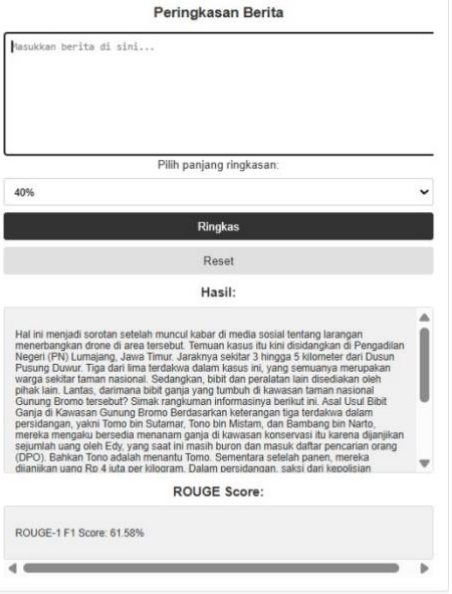

Fig. 4: Results and Evaluation with 40% Sumarry Length

Meanwhile, the final experiment with a summary length of 30% produced 158 words with a ROUGE-1 score of 46.04%.

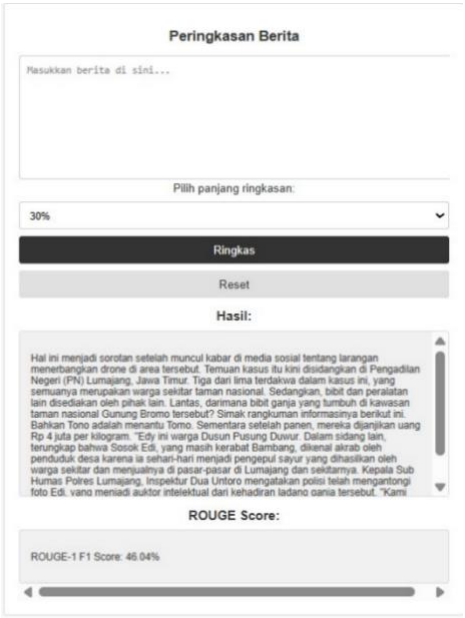

Fig. 5: Results and Evaluation with 30% Summary Length

Based on the results above, a graph illustrating the relationship between summary length and ROUGE-1 score can be observed.

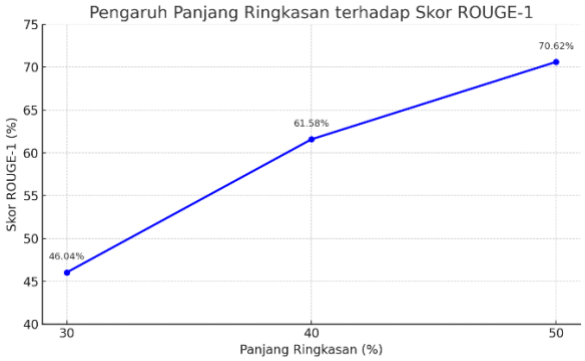

Fig. 6: Graph of the Relationship Between Summary Length and ROUGE-1 Score

From the graph, it can be observed that as the summary length decreases, the ROUGE-1 score tends to decline. The summary with a length of 50% achieved the highest score of 70.62%, indicating that the system is capable of preserving most of the important infomation from the original text. This is because at the 50% length, the system has more room to select sentences with high semantic weights based on the extraction results from TF-IDF and dimensionality reduction using LSA.

Conversely, at the 30% summary length, the ROUGE-1 score dropped sharply to 46.04%. This indicates that although the summary becomes more concise, the amount of important information captured by the system decreases significantly. This is due to the limitation in the number of sentences that can be included in the summary, resulting in many key terms being underrepresented. Therefore, to produce informative and representative summaries in the context of online criminal news summarization, a summary length of 50% is considered the most optimal choice.

4. Conclusion

This study successfully developed an automatic summarization system for Indonesian-language criminal news articles using an extractive approach that combines Term Frequency–Inverse Document Frequency (TF-IDF) and Latent Semantic Analysis (LSA). The proposed system is capable of identifying key sentences within a document by considering both the statistical importance of words and the latent semantic relationships among sentences. Evaluation using the ROUGE-1 metric demonstrates that the system can generate high-quality summaries. A summary length of 50% of the original text yielded the highest ROUGE-1 score of 0.72, indicating that most of the essential

information from the source text was preserved. Meanwhile, summaries of 40% and 30% length resulted in ROUGE-1 scores of 61.58% and 46.04%, respectively, reflecting a decline in quality as the number of summarized sentences was reduced.

These results highlight a trade-off between summary length and information completeness. A 50% summary length appears to be the most optimal choice for preserving informational content without losing critical context. This system can serve as a supportive solution for generating quick summaries of criminal news, especially on digital platforms that require processing of large volumes of text.

For future development, the system could be enhanced using abstractive summarization approaches powered by deep learning models, enabling the generation of more coherent, flexible, and semantically enriched summaries that capture implicit meanings in criminal news texts.

References

[1] A. Apriansyah, H. Fithriansyah, and T. Rahadian, "Eksistensi Surat Kabar Media Indonesia di Era Digital," *Popul. J. Sos. dan Hum.*, vol. 8, no. 1, pp. 74–81, 2023, doi: 10.47313/pjsh.v8i1.2351.

[2] A. Nurfahmi, D. Suherdiana, and A. H. Sumadiria, "Penggunaan Bahasa Jurnalistik pada Rubrik Lifestyle di Situs Prfmnews . id," vol. 8, no. September 2023, pp. 245–264, 2024.

[3] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural Language Processing : State of The Art , Current Trends and Challenges Department of Computer Science and Engineering Accendere Knowledge Management Services Pvt . Ltd ., India Abstract," *Sentim. Anal. has become one Most profound Res. areas with increasing growth Soc. media web. Nowadays, millions users Exch. their views, ideas, expressions, Feel. Opin. Soc. media like twitter Facebook. Se*, no. Figure 1, 2017.

[4] V. Reji, "Information Extraction Using Natural Language Processing," *Interantional J. Sci. Res. Eng. Manag.*, vol. 06, no. 05, 2022, doi: 10.55041/ijsrem13271.

[5] M. G. Ozsoy, F. N. Alpaslan, and I. Cicekli, "Text summarization using latent semantic analysis," *J. Inf. Sci.*, vol. 37, no. 4, pp. 405–417, 2011, doi: 10.1177/0165551511408848.

[6] N. Evangelopoulos, T. Ashton, K. Winson-Geideman, and S. Roulac, "Latent semantic analysis and real estate research: Methods and applications," *J. Real Estate Lit.*, vol. 23, no. 2, pp. 355–380, 2015, doi: 10.1080/10835547.2015.12090411.

[7] O. M. Foong, S. P. Yong, and F. A. Jaid, "Text Summarization Using Latent Semantic Analysis Model in Mobile Android Platform," *Proc. - AMS 2015 Asia Model. Symp. 2015 - Asia 9th Int. Conf. Math. Model. Comput. Simul.*, pp. 35–39, 2016, doi: 10.1109/AMS.2015.15.

[8] D. B. P. D. BP, K. Wilis, and ..., "Summarization of Speech to Text from Reporter in Police Office with Latent Semantic Analysis (LSA) Method," *Int. J. ...*, vol. 13, no. 2, pp. 933–943, 2020, [Online]. Available: http://eprints.upnyk.ac.id/23091/%0Ahttp://eprints.upnyk.ac.id/23091/1/5. summarization of speech.pdf

[9] H. Christian, M. P. Agus, and D. Suhartono, "Single Document Automatic Text Summarization using Term Frequency-Inverse Document Frequency (TF-IDF)," *ComTech Comput. Math. Eng. Appl.*, vol. 7, no. 4, p. 285, 2016, doi: 10.21512/comtech.v7i4.3746.

[10] W. Yulita, M. C. Untoro, M. Praseptiawan, I. F. Ashari, A. Afriansyah, and A. N. Bin Che Pee, "Automatic Scoring Using Term Frequency Inverse Document Frequency Document Frequency and Cosine Similarity," *Sci. J. Informatics*, vol. 10, no. 2, pp. 93–104, 2023, doi: 10.15294/sji.v10i2.42209.

[11] K. Al-Sabahi, Z. Zuping, and K. Yang, "Latent Semantic Analysis Approach for Document Summarization Based on Word Embeddings," 2018, [Online]. Available: http://arxiv.org/abs/1811.06567

[12] Z. Jiang, M. Srivastava, S. Krishna, D. Akodes, and R. Schwartz, "Combining Word Embeddings and N-grams for Unsupervised Document Summarization," 2020, [Online]. Available: http://arxiv.org/abs/2004.14119

[13] X. Zhang, M. Lapata, F. Wei, and M. Zhou, "Neural latent extractive document summarization," *Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018*, pp. 779–784, 2018, doi: 10.18653/v1/d18-1088.

[14] K. Ganesan, "ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks," pp. 1–8, 2018, [Online]. Available: http://arxiv.org/abs/1803.01937

[15] R. Robiyanto, N. Nugraha, and I. Apriatna, "Peringkasan Teks Otomatis Berita Menggunakan Metode Maximum Marginal Relevance," *JEJARING J. Teknol. dan Manaj. Inform.*, vol. 4, no. 1, pp. 23–32, 2019, doi: 10.25134/jejaring.v4i1.6712.