STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **12**, November 2024, pp 1973–1983.
Published online in International Academic Press (www.IAPress.org)

Indonesian News Extractive Summarization using Lexrank and YAKE Algorithm

Julyanto Wijaya*, Abba Suganda Girsang

Computer Science Department, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Abstract The surge in global technological advancements has led to an unprecedented volume of information sharing across diverse platforms. This information, easily accessible through browsers, has created an overload, making it challenging for individuals to efficiently extract essential content. In response, this paper proposes a hybrid Automatic Text Summarization (ATS) method, combining LexRank and YAKE algorithms. LexRank determines sentence scores, while YAKE calculates individual word scores, collectively enhancing summarization accuracy. Leveraging an unsupervised learning approach, the hybrid model demonstrates a 2% improvement over its base model. To validate the effectiveness of the proposed method, the paper utilizes 5000 Indonesian news articles from the Indosum dataset. Ground-truth summaries are employed, with the objective of condensing each article to 30% of its content. The algorithmic approach and experimental results are presented, offering a promising solution to information overload. Notably, the results reveal a two percent improvement in the Rouge-1 and Rouge-2 scores, along with a one percent enhancement in the Rouge-L score. These findings underscore the potential of incorporating a keyword score to enhance the overall accuracy of the summaries generated by LexRank. Despite the absence of a machine learning model in this experiment, the unsupervised learning and heuristic approach suggest broader applications on a global scale. A comparative analysis with other state-of-the-art text summarization methods or hybrid approaches will be essential to gauge its overall effectiveness.

Keywords automatic text summarization, unsupervised learning, extractive text summarization, sentence extraction, term weight

DOI: 10.19139/soic-2310-5070-1976

1. Introduction

Recent advancements in technology that happen around the world made many drastic changes regarding how information can be shared, especially with many platforms to share information around the world [1] [2]. This information can be accessed through a browser anywhere and at any time the user wants to see it [3]. As an illustration, Berisha & Mëziu in 2021 [4] conducted research on Big Data Analytics and presented a graph estimating a data volume of 149 Zettabytes in 2024. This illustrates the vast amount of information that can be cloned or created in the world of digitalization. The ease of accessing various information in this world then makes us feel as if we are in an ocean of information. This is because of how effortlessly we can access the information we seek [3].

The overload of information can significantly increase the time it takes for someone to extract the summary or essence of the article they are reading [4] [5]. This paper aims to implement automatic text summarization (ATS) to enable users to quickly grasp the essence of an article without the need to read the entire document, while still preserving the main ideas or essence of the content [6] [7]. Furthermore, this approach employs an

*Correspondence to: Julyanto Wijaya (Email: julyanto.wijaya@binus.ac.id). Computer Science Department, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia (11480)

ISSN 2310-5070 (online) ISSN 2311-004X (print)

unsupervised method due to its benefits, including faster summarization results, reduced processing requirements, and the elimination of the need for dataset training [8].

In short, the Automatic Text Summarization (ATS) took a sentence with the highest score (illustrated in figure 1). It was then grouped with other high-score sentences creating a summary of the article based on the percentage of the summary we want to create [6] [9]. ATS itself has been the focus of research by many scholars. Below, we aim to summarize previous ATS research that shares similarities with the current study, serving as inspiration for our experiment.

The first study, conducted by J.N. Madhuri in 2019, focused on extractive text summarization using a sentence-ranking approach. Madhuri extracted high-scoring sentences from a corpus, achieving average F1 scores of 0.62 for Rouge-1, 0.44 for Rouge-2, and 0.55 for Rouge-L. The dataset for this research comprised five different corpora [10]. In another study of the same year, Miller, D., explored using the BERT model for text summarization, especially text embedding. Miller then applied K-Means clustering to group sentences, determining the percentage of summarization needed to produce an output with sufficiently important sentences for reader comprehension [11]. Lastly, in the same year, another study was conducted by Mahmoud R. Alfarra, implementing an unsupervised approach to ATS by incorporating Fuzzy logic to optimize the generated summaries. This study introduced Graph-based Fuzzy Logic Extractive Text Summarization (GFLES), aiming to enhance the results of extractive text summarization from ATS. The dataset for this research comprised 500 datasets from DUC 2004, with an F1 measure of 0.48 in a single-document corpus and an F1 measure of 0.14 in a multiple-document corpus for the proposed method [2].

In 2020, Agrawal, K., conducted an intriguing study on extractive text summarization. The experiment utilized IT Law cases as its dataset topic, comprising a total of 10 data points. In this research, Agrawal, K. employed multiple methods for text summarization and used Rouge for accuracy testing. Based on the experiment, Agrawal, K. found that hybridizing LUHN & LSA, yielded the best F1 measure for this type of dataset, achieving a score of 0.64 [12]. Another study about text summarization came from Zamzam, M. A., where he tried to summarize the text of Indonesian news articles at tito.id. The dataset total they used in the study is 10 with an average score of 0.416 [13].

In 2021, Gheata, M., conducted research on text summarization, introducing QuBART. QuBART focused on summarization based on user-defined topics. The study utilized datasets from CNN/DailyMail, WikiHOW, and XSUM. The F1 scores for each dataset were as follows: for CNN/DailyMail, Rouge-1: 0.389, Rouge-2: 0.163, and Rouge-L: 0.328; for WikiHOW, Rouge-1: 0.236, Rouge-2: 0.052, and Rouge-L: 0.206; and for XSUM, Rouge-1: 0.186, Rouge-2: 0.022, and Rouge-L: 0.123 [14].

In 2022, Khotimah N. and Girsang AS conducted research on text summarization utilizing genetic algorithms. The dataset employed was sourced from Indosum, with the study achieving extractive summarization results. The highest Rouge-1 score noted was 0.489 at a 10% summarization percentage, followed by 0.387 at 20%, and 0.330 at 30% [7].

Following this, in 2023, Girsang AS and Amadeus FJ undertook a study on text summarization, this time employing the ant system algorithm and again utilizing a dataset from Indosum. This research resulted in extractive summarization with Rouge-1 scores of 0.424 at a 10% summarization percentage, 0.511 at 20%, and an impressive 0.602 at 30%. Rouge-2 scores were 0.272 at 10%, 0.378 at 20%, and 0.446 at 30%, while Rouge-L scores showed a steady increase from 0.336 at 10%, to 0.436 at 20%, and finally 0.585 at a 30% summarization percentage [8].

In this paper, we will implement lexrank as the main method to do the sentence score in the article while YAKE algorithm will calculate every word score individually and will be merged with the lexrank score. Both of the methods have the same advantage where it didn't need any training data or any fine model tuning because both algorithms were unsupervised learning method algorithms so both algorithms will only use a statistical approach based on the current document it read [15] [16] [17]. This paper will use 5000 news articles from indosum [18] as a dataset where every Indonesian news article will contain a ground-truth summary and will be summarized to 30% of its content.

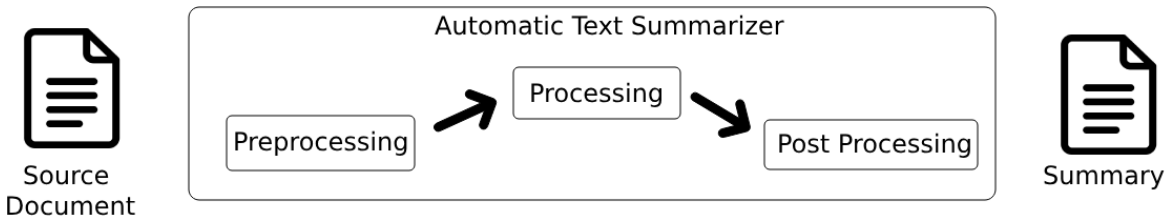

Figure 1. An illustration of how ATS works

1.1. Motivation

The evolution of information technology has exponentially increased the availability of textual data, necessitating efficient and effective summarization techniques to aid comprehension and decision-making. Existing methods, while proficient within their specific domains, often fall short in addressing the multifaceted challenges of automatic text summarization, such as balancing between content relevance and brevity, and adapting to diverse contexts and document types. LexRank, with its strength in understanding sentence importance through a graph-based centrality approach, and YAKE, renowned for its capability to identify key phrases efficiently, independently offer partial solutions. However, each method possesses inherent limitations when applied in isolation, such as LexRank's potential oversimplification of complex texts and YAKE's occasional overemphasis on keyword density over semantic coherence. This research is motivated by the pressing need to transcend these limitations, proposing a hybrid method that leverages the complementary strengths of LexRank and YAKE to enhance the accuracy and adaptability of text summarization.

1.2. Main Contributions

This paper introduces a novel hybrid approach to automatic text summarization, integrating LexRank and YAKE to overcome the noted limitations of existing methods. The key contributions of this study are *Hybrid Methodology* We present a novel algorithm that synergizes the graph-based sentence importance analysis of LexRank with the keyword-centric focus of YAKE, thus ensuring a more balanced and comprehensive summary generation process, *Enhanced Summarization Accuracy* Through empirical studies, we demonstrate that our hybrid approach significantly outperforms existing methods in terms of summarization accuracy, as measured by standard metrics such as Rouge-1, Rouge-2, and Rouge-L. *Theoretical and Algorithmic Analysis* We provide a rigorous mathematical modeling and algorithmic analysis of the hybrid method, detailing the integration process and illustrating how the combined strengths of LexRank and YAKE contribute to overcoming the limitations of singular approaches. *Adaptability and Scalability* Our method shows superior adaptability to diverse document types and contexts, evidenced by consistent performance across various datasets. Additionally, the scalability of the approach is discussed, highlighting its applicability to large-scale text corpora.*Insights into Unsupervised Summarization* By adopting an unsupervised learning approach, this research contributes valuable insights into the potential of unsupervised methods in automatic text summarization, opening avenues for further exploration and development in the field.

The paper will be organized as follows, Section 2 presents the algorithms and methods employed in this research. Section 3 discusses the results of the experiment. Section 4 concludes the paper, and Section 5 outlines areas for future research and improvement

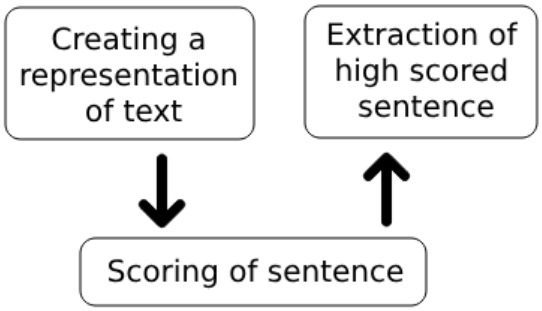

Figure 2. An illustration of how ATS will choose what sentence will be chosen as a summary of the corpus

2. Methodology

2.1. Preprocessing

Preprocessing is crucial in any natural language processing task, particularly text summarization. This process involves eliminating noise from raw data, which includes irrelevant characters, punctuation, or special characters that might impede the effectiveness of score calculations [1]. Typically, every preprocessing step begins with ***sentence segmentation/word tokenization***. During this step, each sentence in the document is converted into smaller units, allowing each sentence or word to stand alone independently and make sense independently. In the Indonesian language, sentences usually end with dots, question marks, exclamation marks, etc. [19] [20]. This segmentation step is skipped as this research employs the Indosum dataset, which already includes segmented articles.

Following segmentation, the next step is ***punctuation removal & case folding***, where punctuation marks are removed from each sentence, and all words are converted to lowercase [19]. Simultaneously or subsequently, ***stop words removal*** can be performed. Stop words removal eliminates common words that lack meaning or are considered irrelevant for the Automatic Text Summarization (ATS) process. Stop words are characterized by their low discrimination power, serving only to connect other words with high discrimination power in sentence formation. These words cannot stand alone, lack individual meaning, and have a high frequency in a document [19].

After punctuation and stop words removal, sentences containing fewer than two words are also eliminated in this research. This step reduces the time required for calculating sentence scores while enhancing ATS accuracy [21]. With clean sentences, the final preprocessing step involves ***stemming***, a process to obtain the root word of each word in the corpus. Stemming removes prefixes, infixes, and suffixes, allowing the system to recognize the meanings of words. The implementation of stemming varies for each language, as different languages have distinct stemming techniques [22].

2.2. Lexrank Algorithm

After completing the preprocessing tasks, the next step involves extracting the highest sentence scores from the corpus and ranking them from the highest to the lowest score (illustrated in Figure 2). To obtain the required score for each sentence, we employ LexRank as our primary model. LexRank is a summarization heuristic method used for both single and multi-document summarization, based on centroids that combine scores known as LexRank scores [15] [23]. The algorithm evaluates the importance of each sentence by calculating similarity using the cosine similarity algorithm [24] for every sentence in the corpus [12]. To calculate the LexRank score, eq (1) is applied, where p(v) represents the LexRank score results for sentence v, with d as the damping factor, N as the total number of sentences in the corpus, idf-modified-cosine(u,v) as the cosine similarity score of the sentence pair (u, v), and $\sum_{z \epsilon adj[v]}$ idf-modified-cosine(z,v) as the sum of all cosine similarity scores for sentence v [25].

$$p(v) = \frac{d}{N} + (1-d) \sum_{v \epsilon adj[u]} \frac{idf - modified - cosine(u,v)}{\sum_{z \epsilon adj[v]} idf - modified - cosine(z,v)} p(v) \tag{1}$$

2.3. YAKE Algorithm

An algorithm for extracting keywords from a corpus has been developed, utilizing a heuristic approach that enables quick implementation across different cases, domains, and languages. This algorithm operates by extracting various features from the corpus and consolidating them into a final score [16].

The first feature considered in this algorithm is called ***casing*** (T_{case}). This feature extraction assigns an additional score to a word if it has uppercase letters at the beginning, provided it is not the beginning of a sentence or is entirely in uppercase. The result of this feature will be called T_{case}, where TF(U(t)) is the frequency of appearance of the term candidate or word with uppercase at the beginning of a sentence, TF(A(t)) is the frequency of appearance of the term candidate or word that has all uppercase letters, and ln(TF(t)) is the natural logarithm of the term frequency appearance 2.

$$T_{case} = \frac{max(TF(U(t)), TF(A(t)))}{1 + ln(TF(t))} \tag{2}$$

The second feature is called ***term position*** ($T_{position}$), where it assigns an additional score to the word for calculation if the word has uppercase letters at the beginning of it and it's not the beginning of a sentence or is entirely in uppercase. This feature calculation can be done by applying eq (3), where Sen_t is the position where the term appears.

$$T_{position} = ln(ln(3 + median(Sen_t))) \tag{3}$$

The third feature is called ***term frequency normalization*** (TF_{norm}), which indicates that the frequency of appearance of a word in a document can provide a useful measure of the word's significance. However, it is worth noting that the relationship between the frequency of appearance of a word and its level of importance is not always linear. Therefore, to avoid bias in the calculation of TF_{norm}, we will use $TF(t)$ to represent the frequency of appearance of the term candidate or word, and we will divide it by the average frequency appearance of all term candidates or words ($MeanTF$) plus one times the standard deviation (σ) of every term candidate or word, excluding stopwords 4.

$$TF_{norm} = \frac{TF(t)}{MeanTF + 1 * \sigma} \tag{4}$$

The fourth feature is called ***term relatedness to context*** (T_{rel}), which will be used to determine the distribution of words in relation to their context. Before calculating the relatedness score, we first need to determine the local score of the term for every term candidate or word. There are two local score representatives that we will need to calculate, represented by either DL or DR 5, based on the local placement of the term. To get the final result of either local representative, we will have to use $|A_t, W|$ as the frequency of the unique appearance of the term candidate or word (in this research, we will use a window size of 1), and $\sum k \epsilon A_{t,w} CoOccur_{t,k}$ as the sum total frequency both to the left and right of the term candidate or word based on the window size used in this research.

$$DL[DR] = \frac{|A_t, W|}{\sum k \epsilon A_{t,w} CoOccur_{t,k}} \tag{5}$$

After getting the local term relatedness score, we can now calculate the global score to be used as the T_{rel} final score. We will calculate it using the eq (6), where $TF(t)$ represents the frequency of appearance of the term candidate or word in the corpus, $MaxTF$ as the highest frequency of appearance of all term candidates or words in the corpus, and DL & DR as the local score representative of the term that we calculated using the eq (5).

$$T_{rel} = 1 + (DL + DR) * \frac{TF(t)}{MaxTF} \tag{6}$$

The fifth and final feature is called ***term different sentence*** ($T_{sentence}$), where it calculates how often a word appears in the sentences within a corpus. This reflects the assumption that a word that frequently appears in many sentences has a high probability of being important. The calculation is done using the following eq (7), where $SF(t)$ will be the total number of sentences where the term candidate or word appeared, and #Sentence is the total number of all sentences in the corpus.

$$T_{sentence} = \frac{SF(t)}{\#Sentence} \tag{7}$$

Now that every feature has been calculated, the next step will be combining them to get the combined score of the term based on every feature described before. We can obtain the combined score with eq (8) and use the equation for $S(kw)$ 9 to get the final score of the term. Here, $\prod_{t \epsilon kw} S(t)$ represents the combined feature score of the term, and $KF(kw)$ is the keyword frequency of the term.

$$S(t) = \frac{T_{rel} * T_{posittion}}{T_{case} + \frac{TF_{norm}}{T_{rel}} + \frac{T_{Sentence}}{T_{rel}}} \tag{8}$$

$$S(kw) = \frac{\prod_{t \epsilon kw} S(t)}{KF(kw) * (1 + \sum_{t \epsilon kw} S(t))} \tag{9}$$

2.4. Proposed hybrid Algorithm

After establishing the foundational methodologies of LexRank and YAKE for evaluating sentence importance and keyword significance respectively, we proceed to the crux of our research. The intricate process of fusing these distinct but complementary insights. This fusion is not merely additive but a sophisticated integration that leverages the unique strengths of each approach to yield a more nuanced and effective summarization algorithm. The forthcoming sections detail the mathematical underpinnings and algorithmic considerations pivotal to realizing this hybrid method, elucidating how the interplay between sentence coherence and keyword relevance is meticulously balanced to enhance summarization accuracy.

The initial step in calculating the combined score involves normalizing the LexRank scores, as shown in eq (10). This normalization ensures comparability with YAKE scores and mitigates potential bias towards longer sentences within the corpus. To achieve this, we divide each sentence's LexRank score by the maximum LexRank score observed across all sentences.

$$L_{norm}(i) = \frac{L_i}{Max(L)} \tag{10}$$

Following the normalization of LexRank scores, the next step focuses on the calculation of YAKE keyword scores, as detailed in eq (11). This process underscores the significance of individual words within sentences, leveraging YAKE's ability to highlight key terms based on their features. For each sentence, we compute the product of scores for all keywords identified, thereby emphasizing the impact of each keyword's relevance to the sentence's overall importance. This keyword scoring mechanism is pivotal for integrating the local textual significance into our hybrid summarization framework

$$Y_{score}(i) = \prod_{k \in kw_i} S(kw)_k \tag{11}$$

Having established the individual scores through the normalization of LexRank and the calculation of YAKE keyword scores, we arrive at the culmination of our methodology which is the calculation of the combined score, articulated in eq (12). This critical step harmonizes the previously derived LexRank and YAKE scores into a singular, comprehensive metric for each sentence. By weighting and summing these normalized LexRank scores and keyword significance scores, we effectively balance global sentence relevance with local keyword importance, thus optimizing our hybrid summarization algorithm's ability to identify and prioritize the most pivotal sentences in

a text. This combined score embodies the core innovation of our approach, facilitating a nuanced balance between sentence coherence and keyword significance for enhanced summarization accuracy.

$$C_{score}(i) = ((W_k * Y_{score}(i)) + (W_s * L_{norm}(i))) \tag{12}$$

3. Experimental results

To present the experimental results, we will utilize a corpus from the Indosum dataset as an example. The chosen corpus is from the article available at https://dailysocial.id/post/pooka-magic-and-mischief. The content of the article will be displayed in Table 1.

Table 1. Indosum dataset

Anda mendambakan bermain game RPG dengan grafis colorful yang cocok untuk dimainkan bersama ponakan Anda? Pooka: Magic and Mischief garapan KISS Limited bisa menjadi jawabannya. Pooka: Magic and Mischief adalah sebuah game petualangan open-world, di mana Anda dapat merancang dan membuat karakter Pooka sendiri. Ada lebih dari 190 bagian yang bisa dikustomisasi, jadi memungkinkan Anda membuat jutaan kombinasi dan gaya yang berbeda-beda untuk membuat Pooka yang benar-benar unik. Selain merancang sendiri, Anda juga bisa secara acak membuat Pooka dari sistem. Setelah karakter menjadi, Anda dapat memotret Pooka dan bisa Anda langsung kirim ke media sosial. Tujuan Anda dalam game ini ialah untuk menyelamatkan dunia dari invasi yang dilakukan Gloom. Ia hadir membawa kegelapan yang menyebar seperti virus dan merusak semua yang disentuhnya. Dengan kekuatan sihir yang Anda miliki, Anda pun bisa membantu memulihkan dunia yang sekarat. Ada beragam quest menarik yang bisa Anda jalankan sambil menikmati dunia yang tidak akan Anda temui di dunia nyata. Berkat adanya sistem multiplayer asynchronous, Anda tidak harus melakukan petualangan sendiri. Pasalnya, Anda dapat mengajak teman-teman untuk memerangi Gloom bersama-sama dan mengirim hadiah satu sama lain. Tertarik? Langsung saja Anda unduh Pooka: Magic and Mischief melalui Play Store. Sumber: Toucharcade. DailySocial.id adalah portal berita startup dan inovasi teknologi. Kamu bisa menjadi member komunitas startup dan inovasi DailySocial.id, mengunduh laporan riset dan statistik seputar teknologi secara cuma-cuma, dan mengikuti berita startup Indonesia dan gadget terbaru.

The segmentation of the table in 1 resulted in a nested list containing chunks of sentences, resembling Table 2. Using the segmented data, we proceeded with punctuation removal, case folding, stopword removal, stemming, and the removal of sentences containing fewer than two words. This process resulted in a clean corpus ready for processing and calculation. The final result before the calculation is displayed in Table 3, which also indicates that sentence 13 will not be included in the calculation as the word count in the sentence is less than two, even before preprocessing began. Now, to obtain the sentence scores, we implemented LexRank with a damping factor of 0.85, a convergence threshold of 0.0001, and a maximum iteration of 100. Simultaneously, we calculated the individual term scores for the same corpus using the original non-preprocessed data to derive every term score. The LexRank scores are presented in Table 4, while the results of every term score are shown in Table 5. As shown in section 2 before, We first normalized the sentence score with the maximum LexRank score to merge the term score and sentence score. Next, we assigned a 50:50 weight to the term and sentence scores. Every term that appeared in a sentence was then added to the normalized sentence score, resulting in the new score for each sentence, combining both the sentence and term scores, as shown in the table 6. To evaluate our results, we used the ROUGE method to calculate the accuracy of the summarization. We averaged all of the results to obtain the average accuracy of our hybrid method. Figure 3 displays the average results for the 5000 datasets used in this research, compared with the base model that does not implement YAKE.

Table 2. Segmentation result

Sentence 1	[”Anda”, ”mendambakan”, ”bermain”, ”game”, ”RPG”, ”dengan”, ”grafis”, ”colorful”, ”yang”, ”cocok”, ”untuk”, ”dimainkan”, ”bersama”, ”ponakan”, ”Anda”, ”?”]
Sentence 2	[”Pooka”, ”:”, ”Magic”, ”and”, ”Mischief”, ”garapan”, ”KISS”, ”Limited”, ”bisa”, ”jadi”, ”jawabannya”, ”.”]
.....	.....
Sentence 16	[”DailySocial.id”, ”adalah”, ”portal”, ”berita”, ”startup”, ”dan”, ”inovasi”, ”teknologi”, ”.”]
Sentence 17	[”Kamu”, ”bisa”, ”menjadi”, ”member”, ”komunitas”, ”startup”, ”dan”, ”inovasi”, ”DailySocial.id”, ”,”, ”mengunduh”, ”laporan”, ”riset”, ”dan”, ”statistik”, ”seputar”, ”teknologi”, ”secara”, ”cuma-cuma”, ”,”, ”dan”, ”mengikuti”, ”berita”, ”startup”, ”Indonesia”, ”dan”, ”gadget”, ”terbaru”, ”.”]

Table 3. Preprocessed result

Sentence 1	[”damba”, ”main”, ”game”, ”rpg”, ”grafis”, ”colorful”, ”cocok”, ”main”, ”ponakan”]
Sentence 2	[”pooka”, ”magic”, ”and”, ”mischief”, ”garap”, ”kiss”, ”limited”, ”jawab”]
.....	.....
Sentence 16	[”dailysocialid”, ”portal”, ”berita”, ”startup”, ”inovasi”, ”teknologi”]
Sentence 17	[”member”, ”komunitas”, ”startup”, ”inovasi”, ”dailysocialid”, ”unduh”, ”lapor”, ”riset”, ”statistik”, ”putar”, ”teknologi”, ”cumacuma”, ”ikut”, ”berita”, ”startup”, ”indonesia”, ”gadget”, ”baru”]

Table 4. Lexrank score

Sentence 1	0.06162344
Sentence 2	0.06347439
...	...
Sentence 16	0.06238682
Sentence 17	0.06256823

Table 5. YAKE score

Pooka	0.00516
Magic	0.00899
Mischief	0.01171
...	...
gadget	0.07718

Table 6. Final sentence score

Sentence 1	0.9471
Sentence 2	0.9919
...	...
Sentence 16	0.9588
Sentence 17	0.9616

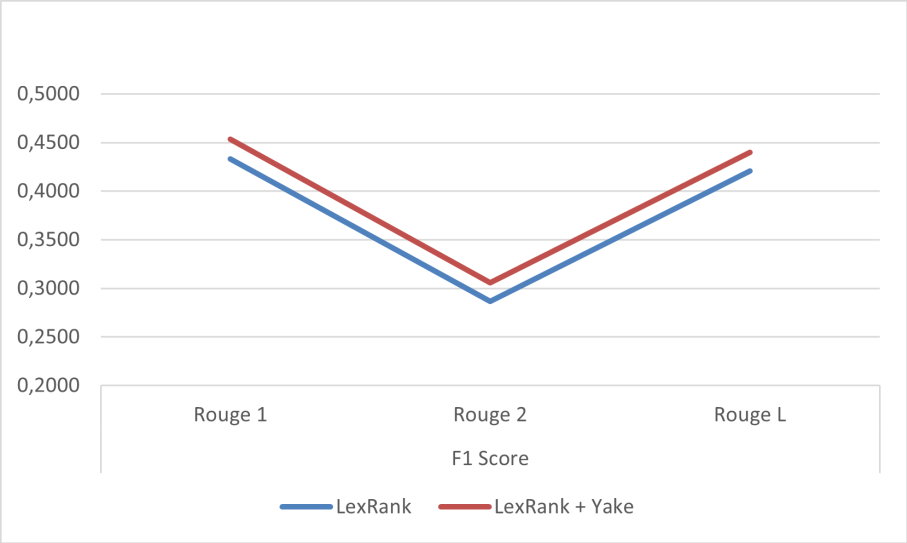

Figure 3. Chart of F1 score average

Following the initial comparative analysis that evaluates the performance of our base model devoid of YAKE hybrid integration as shown in Figure 3, our research endeavors to extend this comparison further. We aim to juxtapose the outcomes obtained from our model against those achieved by various other text summarization methodologies. This comprehensive comparison, ensuring all methods adhere to an identical compression ratio of 30% , is designed to offer a balanced and fair assessment of the relative efficacy and nuances of each summarization strategy within a consistent framework of content reduction. The forthcoming Table 7 will detail how the proposed hybrid method stacks up against other methods, providing a clear visual representation of our findings.

Table 7. Comparison table

Method	Precision	Recall	F1 Measure
SumBasic	0.350	0.469	0.388
LSA	0.283	0.637	0.381
TextRank	0.281	0.698	0.392
KLSum	0.311	0.423	0.344
LexRank	0.355	0.589	0.433
LexRank + YAKE	0.386	0.586	0.453
TextRank + GA	0.330	0.640	0.421

4. Conclusion

In our extensive comparative analysis, various text summarization methods were assessed against our innovative hybrid method, LexRank + YAKE, with the outcomes detailed in Table 7. SumBasic, displaying moderate precision, achieved a relatively modest F1 measure of 0.388. This suggests a balanced yet improvable precision-recall trade-off. LSA and TextRank, praised for their simplicity and efficacy, showed commendable recall rates of 0.637 and 0.698 respectively. However, their F1 scores—0.381 for LSA and 0.392 for TextRank—indicated a tendency to favor comprehensive content over succinctness, reflecting in their lower precision. KLSum, with the lowest F1 measure at 0.344, highlighted its cautious approach, potentially missing out on capturing all key text elements. Conversely, LexRank demonstrated a more balanced precision and recall, culminating in an F1 measure of 0.433, illustrating its efficiency in pinpointing relevant sentences.

Elevating this performance, our proposed LexRank + YAKE approach achieved the highest F1 measure at 0.453. This underscores its superior capability to identify and summarize critical content effectively, enhancing both relevance and conciseness. Similarly, integrating a genetic algorithm with TextRank (TextRank + GA) yielded a promising F1 measure of 0.421, showcasing the potential of optimization algorithms to significantly uplift summarization quality.

These findings, corroborated by the enhancements observed in the Rouge-1 and Rouge-2 scores (an increase of two percent) and the Rouge-L score (a one percent increase) as depicted in figure 3, affirm the merit of incorporating keyword scoring within the LexRank framework to amplify summary accuracy. It's particularly noteworthy that this experiment harnessed an unsupervised and heuristic approach, entirely foregoing the implementation of machine learning models for summary generation. This denotes the method's adaptability and potential for widespread application. However, a comprehensive evaluation against other state-of-the-art text summarization techniques and hybrid models remains imperative. Such comparative analysis is essential to validate the method's effectiveness fully and to explore its possible superiority within the domain of automatic text summarization.

To illustrate the practical application of LexRank + YAKE, consider its use in media monitoring. Media firms often need to process vast amounts of news quickly to keep up with current events. A media monitoring tool using LexRank + YAKE was deployed to summarize daily news efficiently, providing executives and analysts concise yet comprehensive updates. This tool significantly reduced reading time while ensuring that critical information was not omitted, thereby enhancing decision-making processes.

It's particularly noteworthy that this experiment harnessed an unsupervised and heuristic approach, entirely foregoing the implementation of machine learning models for summary generation. This denotes the method's adaptability and potential for widespread application. However, a comprehensive evaluation against other state-of-the-art text summarization techniques and hybrid models remains imperative. Such comparative analysis is essential to validate the method's effectiveness fully and to explore its possible superiority within the domain of automatic text summarization.

5. Future work

In response to the valuable insights provided during the peer-review process, we identify several critical areas for future development: optimization control, advanced information computing, and enhanced model generalization. The integration of optimization control principles, particularly through the development of a dynamic feedback mechanism adjusting the LexRank and YAKE parameters based on summary quality, promises to significantly enhance accuracy and adaptability. Such an approach would allow our method to dynamically respond to the diverse characteristics of different documents, potentially leading to more precise and contextually rich summaries.

Additionally, we aim to test the robustness and effectiveness of our model across a broader spectrum of languages and text genres. This initiative will assess the model's generalizability and identify necessary adaptations to maintain performance across various linguistic and contextual environments. Exploring this dimension is crucial, as it would confirm the scalability and applicability of our method beyond Indonesian news articles.

Moreover, enriching our summarization process with advanced NLP techniques, including Named Entity Recognition, sentiment analysis, and topic modeling, opens up exciting possibilities. This could lead to summaries that not only capture the text's essence more effectively but also provide deeper insights into its semantic and contextual layers. By exploring these advanced techniques, our aim is to extend the capabilities of text summarization beyond its current limits, offering a tool that is both technologically sophisticated and highly adaptable to the nuanced demands of information processing across various domains.

REFERENCES

1. Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679, 2021.
2. M R Alfarra, A M Alfarra, and J M Alattar. Graph-based fuzzy logic for extractive text summarization (gfles). pages 96–101, 2019.

3. Andrzej Szymkowiak, Boban Melović, Marina Dabić, Kishokanth Jeganathan, and Gagandeep Singh Kundi. Information technology and gen z: The role of teachers, the internet, and technology in the education of young people. *Technology in Society*, 65, 5 2021.
4. Blend Berisha and Endrit Mëziu. *Big Data Analytics in Cloud Computing: An overview*. 2 2021.
5. Adhika Pramita Widyassari, Supriadi Rustad, Guruh Fajar Shidik, Edi Noersasongko, Abdul Syukur, Affandy Affandy, and De Rosal Ignatius Moses Setiadi. Review of automatic text summarization techniques & methods. *Journal of King Saud University - Computer and Information Sciences*, 34:1029–1046, 2022.
6. Divakar Yadav, Jalpa Desai, and Arun Kumar Yadav. Automatic text summarization methods: A comprehensive review, 7 2022.
7. Nurul Khotimah and Abba Suganda Girsang. Indonesian news articles summarization using genetic algorithm. *Engineering Letters*, 30, 2022.
8. Abba Suganda Girsang and Fransisco Junius Amadeus. Extractive text summarization for indonesian news article using ant system algorithm. *Journal of Advances in Information Technology*, 14:295–301, 2023.
9. H P Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2:159–165, 4 1958.
10. J N Madhuri and R Ganesh Kumar. Extractive text summarization using sentence ranking. pages 1–3, 2019.
11. Derek Miller. Leveraging bert for extractive text summarization on lectures. 6 2019.
12. K Agrawal. Legal case summarization: An application for text summarization. pages 1–6, 2020.
13. Muhammad Adib Zamzam. Sistem automatic text summarization menggunakan algoritma textrank. *MATICS*, 12:111–116, 9 2020.
14. Miruna Gheata and Javier Varona. Automatic text summarization using a filter-based approach, 10 2021.
15. Ahmad Fauzi. Penerapan algoritma text mining dan lexrank dalam meringkas teks secara otomatis. *Bulletin of Data Science*, 1:65–72, 2022.
16. Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020.
17. Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. A text feature based automatic keyword extraction method for single documents. pages 684–691. Springer International Publishing, 2018.
18. Kemal Kurniawan and Samuel Louvan. Indosum: A new benchmark dataset for indonesian text summarization. pages 215–220. IEEE, 11 2018.
19. Halimah, Surya Agustian, and Siti Ramadhani. Peringkasan teks otomatis (automated text summarization) pada artikel berbahasa indonesia menggunakan algoritma lexrank. *Jurnal CoSciTech (Computer Science and Information Technology)*, 3:371–381, 12 2022.
20. C Slamet, A R Atmadja, D S Maylawati, R S Lestari, W Darmalaksana, and M A Ramdhani. Automated text summarization for indonesian article using vector space model. *IOP Conference Series: Materials Science and Engineering*, 288:12037, 1 2018.
21. D J Ladani and N P Desai. Stopword identification and removal techniques on tc and ir applications: A survey. pages 466–472, 2020.
22. Hari Dwiharyono and Suyanto Suyanto. Stemming for better indonesian text-to-phoneme. *Ampersand*, 9, 1 2022.
23. Pradeepika Verma and Anshul Verma. A review on text summarization techniques. *Journal of scientific research*, 64:251–257, 2020.
24. Ken H. Guo. Testing and validating the cosine similarity measure for textual analysis in accounting, 2023.
25. Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization, 2004.