Enhancing Indonesian Text Summarization with Latent Dirichlet Allocation and Maximum Marginal Relevance

Muhammad Faisal[1*], Bima Hamdani Mawaridi[2], Ashri Shabrina Afrah[3], Supriyono[4],
Yunifa Miftachul Arif[5], Abdul Aziz[6], Linda Wijayanti[7], Melisa Mulyadi[8]
Department of Informatics Engineering, Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia[1, 2, 3, 4, 5]
Faculty of Humanities, Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia[6]
Profesional Engineer Program, Universitas Katolik Indonesia Atma Jaya, Jakarta, Indonesia[7, 8]

***Abstract*—Maximum Marginal Relevance (MMR) Summarization of text is very important in grasping quickly long articles particularly for people who are very busy. In this paper, we use LDA to give topic queries for news articles, which then become inputs to the MMR method. According to this paper's summarization system, the ROUGE metric is employed to evaluate the summaries of news articles with 30 percent compression and 50 percent compression. Experimental findings show that the LDA-MMR combination outperforms MMR on its own in all our tests across all query lengths or number of sentences used and gives highest average ROUGE value of 0.570 for a 50% compression rate; 0.547 at 30% This implies that our system efficiently produces meaningful summaries using content-based keywords rather than click bait titles, which should not lead to complaints about misleading advertisements. This summarizer can convey the main points of a piece of news coverage in a concise form, thus offering people useful new tools for quickly digesting information.**

***Keywords*—Indonesian summarization; LDA; MMR; ROUGE evaluation**

I. Introduction

Natural Language Processing (NLP) is a field that combines computer methods and cognitive, seeking to make computers understand, process, or produce human language. This field entails such tasks as sentence analysis, terminology analysis, and decision-making. The usefulness of NLP is many including machine translation for instant translation between languages, electronic mail spam recognition and rejection of unwanted messages, information mining to recover relevant information from large text repositories, and chatbots as a kind of automatic customer service. An important application of NLP is the generation of automatic text summaries to produce shorter, more understandable summaries of long texts, while retaining all their essential meaning. This is particularly important with the explosion of textual data on the Internet and in digital archives [1].

By exploring and applying different methods or algorithms, automatic text summary aims to produce shorter versions of texts. These methods can be divided according to the input type (single document and multiple documents) and output type (extractive or abstractive summaries) [2]. Extractive summarization means picking sentences, phrases or sections out of the original text, while abstractive summarization involves constructing new sentences in one's own voice which interpret or compress the essence of the original text. Extractive methods are often preferred for their simplicity and lower computational requirements vs. the more sophisticated natural language understanding and generation capabilities that are needed in abstractive techniques.

An important technique in extractive summarization is called Maximum Marginal Relevance (MMR). MMR checks sentence for their relevance to a given query and removes redundancy in a dataset containing similar content [3] while it uses the cosine similarity matrix plus Vector Space Model (VSM) to assess sentence significance. It is well suited to making summaries from both single documents and multiple. However, text queries must be made by hand, taking up a lot of time. And given the arrival of large-scale editorial systems that are now reaching their limits on efficiency through human interaction alone, automated query generation methods will therefore be needed to raise productivity and meet higher quality levels. Latent Dirichlet Allocation (LDA), a popular topic modeling technique, can uncover topics from a text corpus without any human intervention. Efficiently raising queries, the second use for LDA is to find topics hidden in a data set and model them [4]. The utilization of LDA can facilitate the streamlining of manual query generation, thereby enhancing the efficacy of the summarization process. Empirical evidence has demonstrated that the integration of LDA with other summarization techniques can markedly enhance the quality of the resulting summaries. For instance, the conjunction of LDA with MMR has been observed to yield outcomes that are superior to those obtained by either method in isolation [5].

The method proposed in this study serves an inventive approach to use LDA in conjunction with MMR constructed along the lines of an algorithm for efficient summary-making Indonesian news articles. It begins by using LDA to reveal the most important themes present in each article and then builds queries for MMR onto these constituent word distributions This no-nonsense approach is designed to make the summaries both brief and germane to the essence of the articles themselves, so that even if they give little by way of clue,

[*] Corresponding author

within five minutes readers will already gain some understanding about what contents this news offers [6].

In summarizing Indonesian text, LDA and MMR approaches have never been comprehensively challenged. Although past studies have demonstrated that both methods are effective enough in themselves, the uniqueness of news topics and characteristics requires their combined use to completely handle [7]. Our aim is to bridge this gap, utilizing the strengths of both LDA and MMR with the same final goal of getting better quality and relevance on resulting summaries.

The present research aims at creating an automatic summarization system that employs LDA for topic modeling and then MMR for extractive summarization, to generate accurate summaries of Indonesian news articles. By offering concise language summaries focused on the topic, this method is intended to improve the efficiency of retrieving information and also promote a better reading experience for people who need it. The novelty and contribution of this paper lies in combining LDA and MMR. This is expected to push forward the development of text summarization models as an efficient approach for managing large amounts of data.

The remainder of this paper is organized as follows: Section II provides an in-depth literature review, analyzing various current methods and their limitations. In Section III, we describe the method we propose by combining Latent Dirichlet Allocation (LDA) with Maximum Marginality Relevant (MMR) for text summarization. Section IV outlines the experimental settings, while in Section V we report results along with a discussion of them. Comparison is given in Section VI. Finally, Section VII and VIII concludes the paper and points out future research directions.

II. Literature Review

Most of the contemporary means developed for automatic summarization are made to supply summaries at least on par with these extracted by people. Most of this research has focused on high-resource languages, although there are some studies for low-resource language such as Indonesian. This summarization has shown state of the art results in LDA, MMR: two example techniques that have successfully been applied to automatic text summarization across various languages.

Saikumar and Subathra (2020) introduced a set of summarization method using LDA, MMR and Text Rank (TR), proving that the generated summaries are much precise comparing to standalone use of MMR or TR techniques. Finally, the performance of this two-level document summarization (DS) method with LDA was compared to that based on MMR and TR [3]. TextRank and MMR were integrated to be used for summarization of Indonesian news articles by Gunawan, Harahap & Rahmat (2019) [8].

Tuhpatussania, Utami and Hartanto in 2022 [9]: In their work on summarization of online Indonesian news text they have compared the performance between LexRank dan MMR Algorithm: proof that mmmr is better than lexrank for precision, recall and f-measure. Musyaffanto, Herwanto and Riasetiawan (2019), on the other hand integrated MMR with Nonnegative Matrix Factorization (NMF) to ensure precision of online news articles [10].

LDA has been applied to text analysis problems in other research areas as well. To summarize Malayalam news documents closer to what human makes, Kondath, Suseelan and Idicula (2022) used LDA [5]. Rahman et al. (2021) used LDA to create summaries from Malay news documents that showed how system-generated news can save readers' time [6].

The other one is hybrid models for text summarization. Hybrid Approach Gurusamy, Rengarajan and Srinivasan [7] proposed a hybrid approach to this problem that combines semantic LDA and sentence concept mapping with transformer models for generation of coherent text summaries. LISJANA et al. [11], 2020 Classifiers used: They have also applied LDA classifiers with Latent Semantic Indexing (LSI), Similarity-Based Histogram Clustering (SHC) for multi-document text summarization.

Studies for enhancing MMR have started gaining momentum as well. Zheng, Liu, and Qin proposed an improved MMR algorithm that uses Word2Vec, TextRank with semantic information to make text summarization more efficient. [12]. Ramezani et al. (2023) [13] compared LSA vs. MMR in summarizing Persian broadcast news transcriptions, demonstrating LSA's superiority in generic summarization [13].

The above studies point to several research gaps which this study intends to address. Several studies have shown that LDA or MMR, both individually and combined with the other methods, are effective at detecting significant information in a document, but there is no unified approach combining these two aspects: prior to this study not yet proposed any research for Indonesian text summarization task. Besides, current studies usually do not consider the practical difficulty to generate concise summaries over articles of different subjects. This research tries to solve this problem by creating a new summarization method that combines the use of LDA (TextSum) and MMR, for Indonesian article summary generation can be done better with more relevant topics.

III. Proposed Methodology

Fig. 1 shows our proposed hybrid approach using Latent Dirichlet Allocation (LDA) and Maximum Marginal Relevance (MMR) in an Indonesian article summarisation system. It starts with the pre-processing of text documents, which consists of performing all the important steps needed to clean the data and make it ready for analysis. First, the text is broken down into small units such as sentences or words. Next, lowercase (): applies case folding to normalise the text by converting all characters to lowercase. Cleanup: removes all characters irrelevant to the analysis, such as unimportant symbols, punctuation or special characters. This is followed by the removal of stop words, which are common words that do not add any useful information to the summarisation process. The last step in the pre-processing is the stemming, where words are reduced to their root forms, so that different variants of a word can be treated equally.

In this step, the processed text is analysed using a technique called Latent Dirichlet Allocation and then Maximum Marginal

Relevance. We start by using LDA to find the natural topics in the text, which gives us an experienced explorer's view of what kind of information we want. A generative probability model helps to identify the content of the document, and it uses only key topics or words used to describe a single topic from the rest that was scattered. On the other hand, Maximum Marginal Relevance will select the sentences that are most relevant to the given query, i.e. the summary will be both comprehensive and contextually satisfying for the user. The combination of LDA and MMR helps to summarise the given text by capturing what is important in the text data. This dual approach not only increases summarisation accuracy, but also ensures that the output is contextually relevant and insightful optimizing real-world performance.

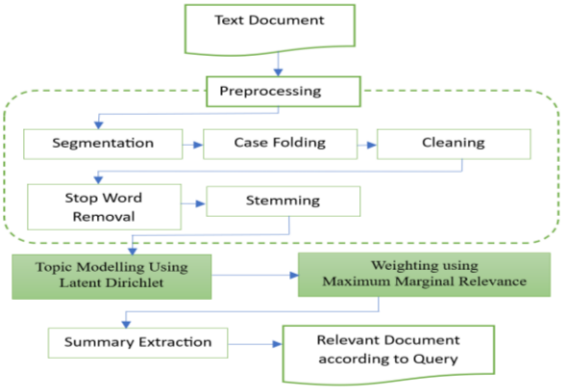
Text Document
Preprocessing
Segmentation
Case Folding
Cleaning
Stop Word Removal
Stemming
Topic Modelling Using Latent Dirichlet
Weighting using Maximum Marginal Relevance
Summary Extraction
Relevant Document according to Query

Fig. 1. Proposed system.

A. Dataset

Experimentally, we conducted the research on IndoSum dataset [14]. It was used in the current research. This corpus consists of around 14,290 news articles along with titles in which all are categorized into six classes (see Fig. 2) from ten different sources of the Indonesian press released to public [14]. The article URL and summary abstracts separately written by two native Indonesian speakers [14]. Rini Wijanti et al. used this dataset in their research [15]. Fig. 3 shows source of news.

The study carried out two testing experiments. This shown in experiment 2, reaching the best ROUGE measure by applying stemming without stopwords removal on test data.

B. Pre-processing

The preprocessing step is responsible for improving the structure of the input data. Preprocessing in NLP often involves tokenization. However, in this paper, we used data from the IndoSum dataset where tokenization has already been performed. Therefore, we can avoid repeating the process of tokenization in this study. Each paragraph consists of a list of clauses, with each clause containing a list of words (token). Segmentation, tokenization, case folding, stop word removal, stemming.

C. Segmentation

In the segmentation process, any paragraph separator is removed so that articles are divided directly on per sentence basis for further processing. During this stage, all paragraphs within each article are combined and then divided into individual sentences by the segmentation process. As a result, each article has sentences, and in turn the sentence will contain words or Tokens.

D. Case Folding

Stage 2 - Case folding at this stage of the preprocessing, we have to convert all uppercase words to lowercase. This process will help to avoid any confusion in the meaning of a word based on whether it's capitalized or not. Step 6 - Lower case processing of the segmentation output (sentence list with tokenized sentence) at the end of this process, we have a list of lowercase words.

E. Cleaning

After tokenization, the third step is data cleaning: This is because only characters are needed for input, not punctuation or numerals, so characters other than letters are removed from the record. And then they have a data cleansing purpose to remove other academic input needed for the program by keeping only clean text as input.

F. Stop Removal

The fourth stage is the elimination of stop words, which is the identification and elimination of words that are common and occur frequently in the text, but often without providing important information. Removing stop words is primarily aimed at cleaning up text and improving text analysis/modelling quality. It is useful to remove keywords to focus the analysis on more relevant or significant words. An example of stopword is a conjunction. Each token is checked whether it is on the stopword list or not, if it is, the token is deleted and not included in the next process. If it is not on the stopword list, it is passed on to the next process.

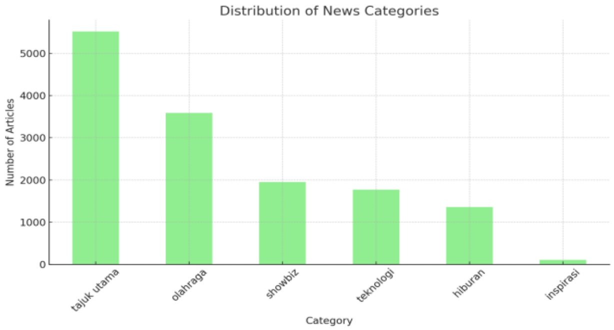

Fig. 2. Category of news.

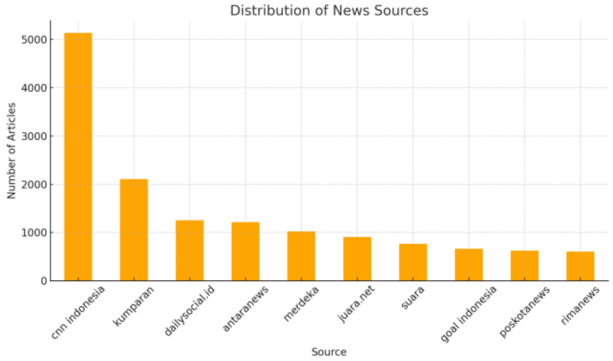

Fig. 3. Source of news.

G. Stemming

The next stage is stemming, which is a process in which words are transformed into their most basic forms. Removing inflections or affixes to ensure a consistent representation of words with a common root is the main goal of stemming. Each token is cross-referenced with the base word lexicon. If a token is missing, it is identified as an affixed word and stemming is initiated by deleting suffixes (-lah, -kah, -ku, -mu, -nya, -tah or -pun). Next, derivative affixes (-i, -kan, -an) are removed, followed by the removal of prefix affixes (be-, di-, ke-, me, pe-, se- and te-).

H. Final Preprocessing

The stemming process is carried out in the final preprocessing stage. The list of tokens will be transformed in this way. All empty strings and lists left over from the previous steps are removed in this final preprocessing step.

I. Latent Dirichlet Allocation

LDA is commonly used when the topics of the papers tend to cluster around a single focus. They are also used to produce topic model results in information technology papers, including domain-specific documents like research papers, news stories, and patents [16] [17] [18]. By using LDA on a set of documents, we can determine the distribution of hidden topics both across the set and within each individual paper. Each topic has its own probability distribution of words attached to it. It is a type of statistical modelling designed to represent the probability distribution of a set of data. This model is useful for generating new data [19], as it can generate data similar to the training data.

LDA models are shown in Fig. 4.

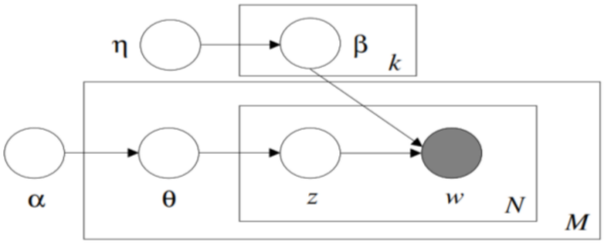

Fig. 4. Graphical representation of the smoothed LDA model [20].

The image shows the Latent Dirichlet Allocation (LDA) model, a generative statistical approach to uncover hidden thematic structure in text documents. The main elements of the model are α, θ, η, βk, z and w. The hyperparameters α and η affect the distributions of topics and words respectively. θ denotes the topic proportions for a document derived from a Dirichlet distribution with parameter α. βk represents the word distribution for a topic derived from a Dirichlet distribution with parameter η. The generative process involves selecting a topic z from θ and then a word w from βk, thereby linking words to topics and topics to documents [21].

Recent advances in LDA have addressed several challenges and expanded its applications. The empirical prior LDA (epLDA) model, which uses latent semantic indexing to obtain priors from the data, has shown notable improvements over the traditional LDA model [22].

The equation illustrates the joint probability distribution in Latent Dirichlet Allocation (LDA), a generative probabilistic approach to topic modelling. This model reveals thematic structures within a collection of documents by decomposing the joint probability into several elements: β_K (word distributions for each topic), θ_D (topic distributions for each document), Z_D (topic assignments for each word), and W_D (observed words). The hyperparameters α and η determine the Dirichlet priors for the topic and word distributions. This relationship is expressed in Eq. (1).

$$p(\beta_K, \theta_D, Z_D, W_D \mid \alpha, \theta) = \prod_{k=1}^{K} p(\beta_k | \eta) \prod_{d=1}^{D} p(\theta_d | \alpha) \prod_{d=1}^{N} p(Z_{d,n} | \theta_d) p(W_{d,n} | Z_{d,n,} B_{d,k}) \quad (1)$$

This product-based formula incorporates several essential elements: the distributions of words within topics (β_k), the distributions of topics within documents (θ_d the topic assignments for each word in each document ($z_{d,n}$), and the observed words in the documents ($w_{d,n}$).

In this model, $\beta_{1:K}$ represent the topic-word distributions, which are influenced by the parameter η and which determine the likelihood of words given topics. The term $\theta_{1:D}$ denotes the document-topic distributions, which are influenced by the parameter α and which determine the likelihood of topics given documents. The term $p(\beta_{k/\eta})$ denotes the probability of the word distribution for topic k given the Dirichlet prior η, while $p(\theta_d/\alpha)$ is the probability of the topic distribution for document d given the Dirichlet prior α. The term $p(z_{d,n}/\theta_d)$ represents the probability of assigning the n-th word in document d to a topic, based on the topic distribution for that document.

Finally, $p(w_{d,n} \ / \ z_{d,n}, \beta_{d:K})$ represents the probability of the n-th word in document d being assigned to topic $Z_{d,n}$, given the aforementioned topic assignment and the word distribution for that topic, β_k. Recent advances in the latent Dirichlet allocation (LDA) approach have led to the introduction of enhanced models and methodologies, including the empirical prior latent Dirichlet allocation (epLDA) and StreamFed-LDA. The epLDA model improves the computation of topics by employing latent semantic indexing to derive priors from data, thereby enhancing prediction accuracy [22].

- Hyperparameter Selection and Impact

The selection of hyperparameters in Latent Dirichlet Allocation (LDA) has a significant impact on the quality of the results, particularly in terms of topic coherence and relevance. In this study, the value of α (the Dirichlet prior for the distribution of topics per document) was set to 1/K, where K is the number of topics, and η (the Dirichlet prior for the distribution of words per topic) was set to 1/V, where V is the vocabulary size. These values were selected to ensure a balanced distribution of topics across documents and words across topics. The parameter α regulates the diversity of topics within a document; higher values of α result in more uniform topic distributions, thereby enabling documents to encompass a broader range of topics. Conversely, the influence of the parameter η on the distribution of words within each topic is inverse. Lower values of η result in sparser distributions, which in turn produce more focused and distinctive topics. The preliminary experiments demonstrated that varying α and η has a significant impact on the coherence of the topics and the relevance of the summaries produced. Further research could involve a more comprehensive investigation of these hyperparameters to enhance LDA performance in diverse contexts.

J. Maximum Marginal Relevance

The Maximum Marginal Relevance (MMR) algorithm is a well-established method in the field of information retrieval. The algorithm calculates a linear combination that includes both the relevance of the documents to the query and their similarity to previously chosen documents for summarization. This measure, known as 'edge correlation', is optimized during the retrieval and summarization processes to refine the final summary iteratively [23][24]. MMR summarizes text by evaluating the similarity between different parts of the text, showing efficiency in retrieving relevant information and uncluttering content. It identifies documents that match a specific query by combining two criteria: relevance and heterogeneity.

MMR summarizes text by evaluating the similarity between different parts of the text, thereby demonstrating efficiency in the retrieval of relevant information and the uncluttering of content. The method identifies documents that match a specific query by combining two criteria: relevance and heterogeneity.

In this framework, a linear combination is calculated in order to integrate a document's relevance to the query and its similarity to pre-selected documents for summarization. This metric, designated as 'edge correlation', is calibrated during the retrieval and summarization phases to incrementally refine the final summary. MMR employs a methodology whereby content is described by assessing the degree of similarity between text segments. This demonstrates the ability of the method to retrieve related data and avoid redundancy.

MMR employs a ranking system based on a combination of cosine similarity matrices in response to a given query. The calculation entails a comparison of the results pertaining to the relevance of the query with those concerning the similarity of sentences. A document is deemed to possess high marginal relevance if it exhibits a strong alignment with the document content and a high degree of similarity with the query. The MMR score can be calculated using the following Eq. (2) [25].

$$MMR = argmax\,[\lambda * Sim_1\,(S_i, Q) - (1 - \lambda) * maxSim_2\,(S_i, Summ)] \quad (2)$$

In this context, Si represents the candidate sentence, and Q is the query or main topic. The parameter λ (which ranges from 0 to 1) serves to regulate the equilibrium between relevance and diversity. The term Sim1(Si,Q) is employed to ascertain the degree of similarity between the candidate sentence Si and the query Q, thereby ensuring that the selected sentences are highly relevant. The term (1 − λ) · maxSim2(Si, Summ) is employed to ascertain the maximum similarity between the candidate sentence Si and the sentences that have already been included in the summary. This serves to minimize redundancy.

In this context, the term "Si" represents a sentence within the document, whereas "Summ" refers to the selected or extracted sentences. The relevance of a sentence is determined, and redundancy is minimized through the utilization of the coefficient λ.

The parameter λ is defined in the range of 0 to 1. When λ equals one, the MMR value is more pertinent to the original document. Conversely, when λ equals zero, the MMR value is more aligned with the previously extracted sentences. It is therefore recommended that λ be adjusted within this range to achieve optimal summarization. In the case of shorter texts, such as articles, the optimal value for λ is generally considered to be 0.7, which yields effective summary results [25].

IV. Experimental Results

The experiment was conducted on the initial 50 article data entries within the train.03 JSON file, which forms part of the IndoSum dataset. The test scenario was conducted on articles 1 to 50 to ascertain the ROUGE-1 value for each system-generated summary. Prior to the generation of summaries, each article was subjected to topic modelling using the Latent Dirichlet Allocation (LDA) technique, which yielded one topic and ten keywords. In the LDA topic modelling process, the alpha value was set to 1/K and the eta value was set to 1/V. The ten keywords were then employed as queries to generate summaries utilizing the Maximum Marginal Relevance (MMR) method. The MMR summarization process was conducted with three different lambda values, as follows: The values of 0.5, 0.7, and 0.9 were employed. The resulting summaries comprised either 50% or 30% of the total sentences in the original text.

A series of tests were conducted to compare the quality of human-generated summaries with those produced by the system. To obtain recall, precision, and F1-score values, this research employs the ROUGE-1 evaluation method. The greater the degree of alignment between the system summary and the human summary, the higher the recall value. Should the recall value attain its maximum value or a value of 1, it signifies that the entirety of the human summaries will be incorporated into the system summary. Conversely, if the precision value reaches the maximum value or 1, then the entire system summary will be included in the human summary. The combination of recall and precision, namely the F1-score, provides an overall picture of the system's ability to capture and present appropriate and relevant information in its summary.

The experiment was conducted using three values of λ: 0.5, 0.7, and 0.9. The results of Experiment 1 are presented in Table I.

TABLE I. ROUGE-1 EVALUATION RESULT EXPERIMENT 1LDRMMR

λ	**Compression rate**		
	50%		
	Average Recall	**Average Precision**	**Average F1-Score**
$\lambda = 0.5$	0.863	0.402	0.534
$\lambda = 0.7$	0.865	0.404	0.536
$\lambda = 0.9$	0.864	0.402	0.535

Table I illustrates the Rouge-1 assessment outcomes for Experiment 1LDRMMR, which evaluates a document summarisation system at a 50% compression rate utilising diverse values of the smoothing parameter, lambda (λ). In this comparison, the following metrics are considered: recall average, precision average and F1 score average for λ values of 0.5, 0.7 and 0.9. As λ increases, there is a modest enhancement in recall average, from 0.863 to 0.865, before a slight decline to 0.864. The value of the Precision Average is observed to increase from 0.402 to 0.404 at λ = 0.7 and then to remain at this level of 0.402 when λ = 0.9. The F1-score average, which weighs precision and recall equally, demonstrates the most optimal performance at λ=0.7. It increases from 0.534 to 0.536 and then experiences a slight decrease to 0.535.

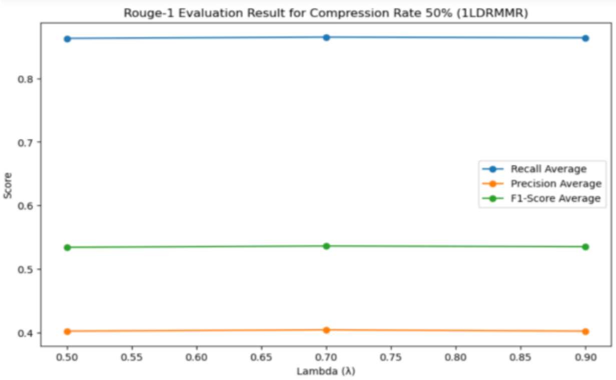

Fig. 5. Rouge-1 evaluation result experiment 1LDRMMR.

The graph in Fig. 5 demonstrates these trends, indicating that an increase in λ generally enhances the system's performance, with the optimal λ value being 0.7. At this value, the system attains the optimum balance between recall and precision, resulting in the highest F1 score. These findings suggest that while higher λ values produce marginal improvements in performance, the gains plateau beyond λ=0.7, indicating an optimal range for λ to optimise summarisation quality.

Additionally, in the second experiment, an evaluation was conducted with the results presented in Table II.

TABLE II. ROUGE-1 EVALUATION RESULT EXPERIMENT 2LDAMMR

λ	Compression rate		
	30%		
	Recall Average	Precision Average	Average F1-Score
λ = 0.5	0.778	0.485	0.580
λ = 0.7	0.775	0.484	0.578
λ = 0.9	0.781	0.486	0.581

The results of the Rouge-1 evaluation for the 2LDAMMR experiment, with a compression rate of 30%, are presented in Table II. The table presents a comparison of three distinct λ values: These values were 0.5, 0.7, and 0.9. For λ = 0.5, the recall average is 0.778, the precision average is 0.485, and the average F1-measure is 0.580. When λ=0.7, a slight decrease is observed in the Recall Average (to 0.775), while the Precision Average remains almost unchanged (at 0.484). The Average F1-Measure also decreases, albeit to a lesser extent (to 0.578). At λ=0.9, the recall average increases to 0.781, the precision average rises slightly to 0.486, and the average F1-measure also increases slightly to 0.581.

It can be concluded from these results that the λ value influences the evaluation metrics. While the changes observed are minor, higher λ values tend to result in slight improvements in both the Recall Average and Precision Average. However, these changes are not significant and there is consistency in the Average F1-Measure, indicating that the model demonstrates stable performance across different λ values. For a more comprehensive visual representation, please see the graph below, which illustrates the comparison of the metrics for each λ value.

Here is the graph that illustrates the comparison of metrics for each λ value in greater detail, as shown in Fig. 6.

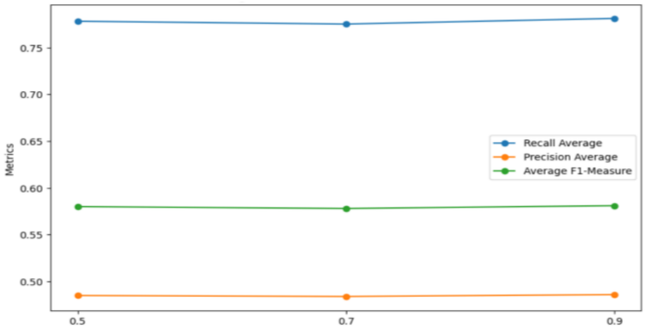

Fig. 6. Graph rouge-1 evaluation result experiment 2 LDAMMR.

Fig. 6 depicts the performance of Recall Average, Precision Average, and Average F1-Measure across λ values of 0.5, 0.7, and 0.9 for the 2LDAMMR experiment with a 30% compression rate. The Recall Average demonstrates a slight increase with elevated λ values, indicating enhanced recall capability, whereas the Precision Average remains stable, suggesting consistent precision. The Average F1-Measure, representing the harmonic mean of precision and recall, also shows minimal variation, indicating balanced performance. Overall, increasing λ results in minor improvements in recall and F1-Measure, demonstrating the model's robustness and consistent performance across the tested λ range.

Experiments were also conducted using MMR with title queries with the same dataset and λ value of 1, the results of which can be seen in Table III and Table IV.

TABLE III. ROUGE-1 EVALUATION RESULT EXPERIMENT 1MMR

Experiment 1MMR	Compression rate		
	50%		
	Average Recall	Average Precision	Everage F1-score
1 (λ = 0.5)	0.843	0.407	0.536
2 (λ = 0.7)	0.843	0.407	0.536
3 (λ = 0.9)	0.843	0.407	0.536

Table III presents the findings of the evaluation conducted on the 1MMR experiment, utilizing a compression rate of 50%. Three distinct λ values (0.5, 0.7, and 0.9) were subjected to evaluation. The results for Recall Average, Precision Average, and Average F1-Measure are consistent across all λ values. The Recall Average remains at 0.843, the Precision Average at 0.407, and the Average F1-Measure at 0.536 for each λ value. This consistency suggests that the λ parameter does not significantly affect the model's performance under these conditions.

TABLE IV. ROUGE-1 EVALUATION RESULT EXPERIMENT 2MMR

Experiment 2MMR	Compression rate		
	30%		
	Average Recall	**Average Precision**	**Average F1-score**
1 (λ = 0.5)	0.680	0.460	0.536
2 (λ = 0.7)	0.680	0.460	0.536
3 (λ = 0.9)	0.680	0.460	0.536

Table IV presents the results of the evaluation of the 2MMR experiment with a 30% compression rate. Once more, three distinct λ values (0.5, 0.7, and 0.9) are subjected to examination. As with the 1MMR experiment, the results for Recall Average, Precision Average, and Average F1-Measure are consistent across all λ values. The recall average is 0.680, the precision average is 0.460, and the average F1-measure is 0.536 for each λ value. The consistency across different λ values indicates that the λ parameter does not significantly impact the model's performance in the 2MMR experiment.

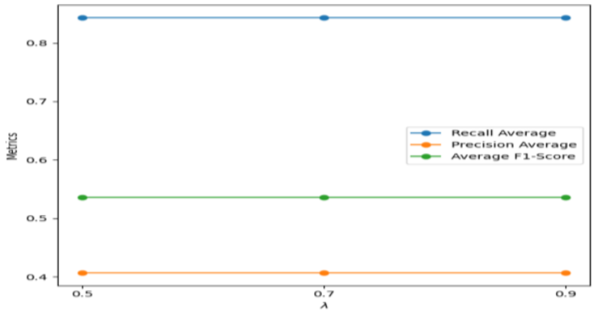

Fig. 7. Rouge-1 evaluation result experiment 1MMR.

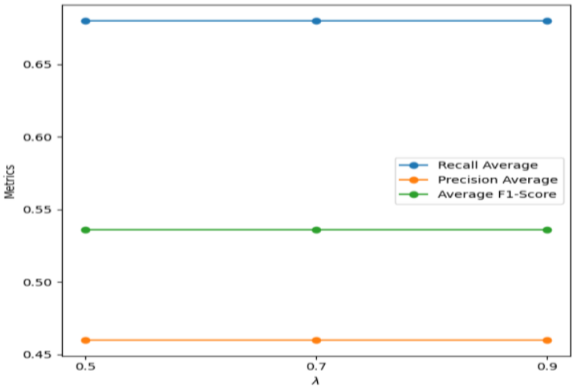

Fig. 8. Rouge-1 evaluation result experiment 2MMR.

Fig. 7 and Fig. 8 present a graphical representation of the performance of Recall Average, Precision Average, and Average F1-score across λ values of 0.5, 0.7, and 0.9 for both the 1MMR experiment with a 50% compression rate and the 2MMR experiment with a 30% compression rate. In the case of the 1MMR experiment, all metrics remain constant regardless of λ. The values for these metrics are as follows: Recall Average at 0.843, Precision Average at 0.407, and Average F1-Measure at 0.536. Similarly, in the 2MMR experiment, the metrics demonstrate no variation with different λ values, maintaining a Recall Average of 0.680, a Precision Average of 0.460, and an Average F1-Measure of 0.536. This consistency across both experiments indicates that λ has a negligible impact on model performance in these scenarios. At a 50% compression rate, 1MMR achieves a higher recall than 2MMR at a 30% compression rate.

TABLE V. COMPARISON BETWEEN MMR AND LDAMMR

Methods	F1-score Average	
	Compression Rate 50%	**Compression Rate 30%**
MMR	0.536	0.536
LDA & MMR	0.536	0.581

Table V presents a comparison of the average F1-score between two methods. The study compares the effectiveness of two approaches to text compression: Maximal Marginal Relevance (MMR) and a combined approach of MMR and Latent Dirichlet Allocation (LDA), applied at two different compression rates, 50% and 30%.

The average F1-score for the MMR method remains constant at 0.536 for both compression rates of 50% and 30%. This consistency indicates that the performance of MMR alone is not affected by different levels of compression, suggesting that MMR is robust in maintaining its effectiveness regardless of the compression rate applied.

In contrast, the combination of MMR and LDA yielded a notable enhancement in the average F1-score at the 30% compression rate, which increased to 0.581. However, at the 50% compression rate, the combination yields the same average F1-score of 0.536 as MMR alone. This suggests that the incorporation of LDA with MMR improves performance, particularly at the lower compression rate of 30%. This implies that LDA provides supplementary contextual information, enhancing the model's efficacy when the data is more condensed.

The experiments conducted utilising MMR with LDA queries yielded superior ROUGE-1 evaluation scores in comparison to those employing MMR with title queries. However, both systems exhibit a commendable ROUGE-1 score. According to Deutsch, the discrepancy in ROUGE-1 scores below 0.5 between systems is less indicative of the human perception of the same two systems [26].

The results presented in Table V illustrates that while the MMR method demonstrates consistent performance across varying compression rates, the integration of MMR with LDA markedly enhances the Average F1-Score at the 30% compression rate. This suggests that LDA improves MMR's capacity to capture pertinent information in a more condensed dataset. The consistency of results at the 50% compression rate indicates that the advantages of LDA are more evident when dealing with higher levels of data compression.

Fig. 9 illustrates the comparison of the Average F1-Score for each method under both compression rates.

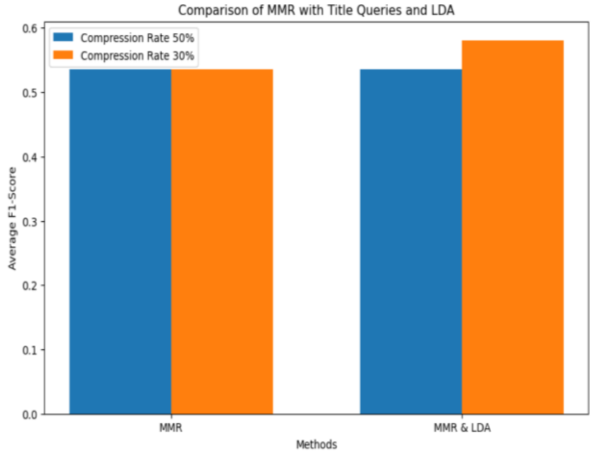

Fig. 9. Comparison of MMR with title query and LDA.

V. Discussion

The combination of Latent Dirichlet Allocation (LDA) and Maximal Marginal Relevance (MMR) offers notable advantages that make it a more effective approach than traditional text summarisation techniques. Although MMR has been extensively employed for the selection of pertinent sentences based on title queries, its principal limitation resides in its reliance on manually crafted or title-based queries, which frequently impede the precision of the resulting summaries, particularly when the article title does not fully encapsulate the content.

The incorporation of LDA enables the generation of contextually rich topic queries, which results in summaries that more accurately represent the contents and themes of the articles. This is especially advantageous in cases where article titles may be deceptive or inadequate in representing the actual content, such as in the context of clickbait headlines.

The combination of LDA and MMR has been demonstrated to achieve a higher F1-score at a 30% compression rate, indicating an improvement in both the quality and relevance of the summaries. These findings indicate that the proposed method is not only effective for summarising Indonesian news articles but could also be adapted to other languages and document types, thereby enhancing its overall applicability and versatility.

This approach represents a significant step forward in the creation of more accurate and contextually relevant summaries, particularly in cases where traditional methods may be inadequate.

VI. Comparison

The objective of this research is to develop a document summarization system that can extract the essential information from documents. The study presents a distinctive profile when compared to the various methods and tests identified in other studies. Table VI provides a summary of these differences. In their study [3] employs the Two-Level LDA method with customer opinion data on products and hotels, comparing LDA with MMR and TR. In contrast, our study employs the LDAMMR method and compares it solely with MMR. In contrast to the study [9] employs solely the MMR method, without comparison to other techniques. In contrast to the studies [6] does not undertake a comparative analysis; instead, it focuses on summarizing Malay language news articles using LDA. In their study [25] employed MMR and VSM to summarize students' final project abstracts. While existing research has demonstrated the effectiveness of LDA and MMR individually or in combination with other methods, there remains a lack of comprehensive approaches that specifically integrate LDA with MMR for summarizing Indonesian news articles. Additionally, existing studies often fail to thoroughly address the challenges of summarizing articles that cover a wide range of topics.

TABLE VI. Text Summarization Comparative Study

References	Object	Methods	Comparison methods	ROUGE-1 Results (F1-Score)
[3]	Two categories: products, hotels	Two Level LDA	Compare with MMR and TR summarization techniques	Not provided
[9]	Indonesian News Article	MMR	Not compare	Not provided
[6]	Malay News Article	LDA	Not compare	Not provided
[25]	One category: Students Final Project Abstracts	MMR & VSM	Not compare	Not provided
Ours	Four Category Indonesian News Article	LDAMMR	Compare LDA vs LDAMMR	0.536 (50% compression) 0.581 (30% compression)

VII. Conclusion

The integration of Latent Dirichlet Allocation (LDA) with Maximum Marginal Relevance (MMR) has been demonstrated to enhance text summarization. This is achieved by generating more accurate and relevant queries, reducing redundancy, and providing a contextual understanding of the document's themes. This combination of techniques improves efficiency through the automatic generation of queries, while maintaining a balance between precision and recall. The results of the research demonstrate that while MMR exhibited a constant average F1-score of 0.536, the integration of LDA resulted in an increase to an average F1-score of 0.581 at a 30% compression rate. This illustrates that LDA augments MMR's capacity to capture pertinent information in a more efficacious manner, thereby rendering summaries more succinct and contextually pertinent, particularly in the case of diverse and evolving subject matter such as Indonesian news articles.

The potential for application in multiple languages is a further advantage of this approach.

This passage presents the findings of a study on the summarization of Indonesian news articles, employing a methodology that is not language specific. The key techniques employed, namely Latent Dirichlet Allocation (LDA) and Maximal Marginal Relevance (MMR), are language-agnostic, meaning that they can be applied to different languages with some adjustments. These modifications include the adaptation of preprocessing procedures, such as tokenization, stopword removal and stemming, to align with the linguistic characteristics of the target language. The study posits that this approach could prove beneficial for languages with limited resources, where sophisticated text summarization tools are not as readily accessible. Furthermore, it urges future research to apply this methodology to multilingual datasets, which could facilitate the advancement of more versatile and globally applicable summarization techniques.

VIII. Future Work

The encouraging outcomes of this study suggest several avenues for future research. One avenue for further research would be to explore alternative topic modelling techniques, such as non-negative matrix factorisation (NMF) or latent semantic analysis (LSA), to ascertain whether they can enhance the quality of summaries even further. Furthermore, applying this method to a broader range of document types, including legal texts, scientific articles, or social media content, could serve to test its versatility and robustness across different contexts. Another promising avenue for future research is the integration of this method with transformer-based models, such as BERT or GPT, to develop a hybrid approach that combines the strengths of both extractive and abstractive summarisation. This could result in the generation of more coherent and contextually rich summaries, thereby advancing the state of the art in automatic text summarisation. Furthermore, adapting this model for real-time or streaming data could make it a valuable tool for dynamic content summarisation, providing immediate insights in fast-paced environments such as newsrooms or social media monitoring.

References

[1] V. Agate, S. Mirajkar, and G. Toradmal, “Book Summarization using NLP,” International Journal of Innovative Research in Engineering, pp. 476–480, Apr. 2023, doi: 10.59256/ijire.2023040218.

[2] U. Rani and K. Bidhan, “Comparative Assessment of Extractive Summarization: TextRank, TF-IDF and LDA,” Journal of scientific research, vol. 65, no. 01, pp. 304–311, 2021, doi: 10.37398/JSR.2021.650140.

[3] D. Saikumar and P. Subathra, “Two-Level Text Summarization Using Topic Modeling,” 2021, pp. 153–167. doi: 10.1007/978-981-15-5400-1_16.

[4] C. P. George and H. Doss, “Principled Selection of Hyperparameters in the Latent Dirichlet Allocation Model,” J. Mach. Learn. Res., vol. 18, pp. 162:1-162:38, 2017.

[5] M. Kondath, D. P. Suseelan, and S. M. Idicula, “Extractive summarization of Malayalam documents using latent Dirichlet allocation: An experience,” Journal of Intelligent Systems, vol. 31, no. 1, pp. 393–406, Mar. 2022, doi: 10.1515/jisys-2022-0027.

[6] N. A. Rahman, S. N. A. Ramlam, N. A. Azhar, H. M. Hanum, N. I. Ramli, and N. Lateh, “Automatic Text Summarization for Malay News Documents Using Latent Dirichlet Allocation and Sentence Selection Algorithm,” in 2021 Fifth International Conference on Information Retrieval and Knowledge Management (CAMP), IEEE, Jun. 2021, pp. 36–40. doi: 10.1109/CAMP51653.2021.9498029.

[7] B. M. Gurusamy, P. K. Rengarajan, and P. Srinivasan, “A hybrid approach for text summarization using semantic latent Dirichlet allocation and sentence concept mapping with transformer,” International Journal of Electrical and Computer Engineering (IJECE), vol. 13, no. 6, p. 6663, Dec. 2023, doi: 10.11591/ijece.v13i6.pp6663-6672.

[8] D. Gunawan, S. H. Harahap, and R. F. Rahmat, “Multi-document summarization by using textrank and maximal marginal relevance for text in Bahasa Indonesia,” in 2019 International conference on ICT for smart society (ICISS), 2019, pp. 1–5.

[9] S. Tuhpatussania, E. Utami, and A. D. Hartanto, “Comparison Of Lexrank Algorithm And Maximum Marginal Relevance In Summary Of Indonesian News Text In Online News Portals,” Jurnal Pilar Nusa Mandiri, vol. 18, no. 2, pp. 187–192, 2022.

[10] I. R. Musyaffanto, G. Budi Herwanto, and M. Riasetiawan, “Automatic Extractive Text Summarization for Indonesian News Articles Using Maximal Marginal Relevance and Non-Negative Matrix Factorization,” in 2019 5th International Conference on Science and Technology (ICST), IEEE, Jul. 2019, pp. 1–6. doi: 10.1109/ICST47872.2019.9166376.

[11] O. A. LISJANA, D. P. RINI, and N. YUSLIANI, “Multi-Document Text Summarization Based on Semantic Clustering and Selection of Representative Sentences Using Latent Dirichlet Allocation,” in Proceedings of the Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019), Paris, France: Atlantis Press, 2020. doi: 10.2991/aisr.k.200424.029.

[12] Y. Zheng, Y. Liu, and H. Qin, “Chinese News Text Abstract Extraction Using Improved MMR,” in 2021 International Conference on Electronic Information Engineering and Computer Science (EIECS), IEEE, Sep. 2021, pp. 601–607. doi: 10.1109/EIECS53707.2021.9587964.

[13] M. Ramezani, M.-S. Shahryari, A.-R. Feizi-Derakhshi, and M.-R. Feizi-Derakhshi, “Unsupervised Broadcast News Summarization; a Comparative Study on Maximal Marginal Relevance (MMR) and Latent Semantic Analysis (LSA),” in 2023 28th International Computer Conference, Computer Society of Iran (CSICC), IEEE, Jan. 2023, pp. 1–7. doi: 10.1109/CSICC58665.2023.10105403.

[14] K. Kurniawan and S. Louvan, “Indosum: A new benchmark dataset for Indonesian text summarization,” in 2018 International Conference on Asian Language Processing (IALP), 2018, pp. 215–220.

[15] R. Wijayanti, M. L. Khodra, and D. H. Widyantoro, “Single document summarization using bertsum and pointer generator network,” International Journal on Electrical Engineering and Informatics, vol. 13, no. 4, pp. 916–930, 2021.

[16] H. Jelodar et al., “Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey,” Multimed Tools Appl, vol. 78, pp. 15169–15211, 2019.

[17] C. Tian, J. Zhang, D. Liu, Q. Wang, and S. Lin, “Technological topic analysis of standard-essential patents based on the improved Latent Dirichlet Allocation (LDA) model,” Technol Anal Strateg Manag, pp. 1–16, 2022.

[18] J. Kim et al., “Trend Research on Maritime Autonomous Surface Ships (MASSs) Based on Shipboard Electronics: Focusing on Text Mining and Network Analysis,” Electronics (Basel), vol. 13, no. 10, 2024, doi: 10.3390/electronics13101902.

[19] H. Liu, T. Zhang, F. Li, M. Yu, and G. Yu, “A probabilistic generative model for tracking multi-knowledge concept mastery probability,” Front Comput Sci, vol. 18, no. 3, p. 183602, 2024.

[20] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.

[22] M. A. Adegoke, J. O. A. Ayeni, and P. A. Adewole, “Empirical prior latent Dirichlet allocation model,” Nigerian Journal of Technology, vol. 38, no. 1, p. 223, Jan. 2019, doi: 10.4314/njt.v38i1.27.

[23] J. Carbonell and J. Goldstein, “The use of MMR, diversity-based reranking for reordering documents and producing summaries,” in

Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998, pp. 335–336.

[24] J. Fan, X. Tian, C. Lv, S. Zhang, Y. Wang, and J. Zhang, “Extractive social media text summarization based on MFMMR-BertSum,” Array, vol. 20, p. 100322, 2023, doi: https://doi.org/10.1016/j.array.2023.100322.

[25] G. Gunawan, F. Fitria, E. Setiawan, and K. Fujisawa, “Maximum Marginal Relevance and Vector Space Model for Summarizing Students’ Final Project Abstracts,” Knowledge Engineering and Data Science, vol. 6, p. 57, 2023, doi: 10.17977/um018v6i12023p57-68.

[26] D. Deutsch, R. Dror, and D. Roth, “Re-examining system-level correlations of automatic summarization evaluation metrics,” arXiv preprint arXiv:2204.10216, 2022.