Text summarization: BART, RF, and hybrid BART-RF algorithm comparison

Muhammad Adib Zamzam, Agus Buono, Toto Haryanto
School of Data Science, Mathematics and Informatics, IPB University, Bogor, Indonesia

Article Info

Article history:

Received Oct 25, 2023
Revised Nov 16, 2025
Accepted Dec 15, 2025

Keywords:

Automatic text summarization
BART-RF
Hybrid
Random forest
ROUGE

ABSTRACT

Data and information accumulate quantitatively and qualitatively. Abundant text data are posted on the internet. The number correlates to the complexity of the summarization. Automatic text summarization (ATS) is one of the most challenging tasks in natural language processing (NLP). ATS approached in three ways: extractive, abstractive, and hybrid. Hybrid approach combines both extractive and abstractive. This research tests and compares performance of bidirectional auto-regressive transformer (BART) and random forest (RF) individually and the performance combination of hybrid BART and RF in ATS. The research shows that individually, BART and RF recall-oriented understudy for gisting evaluation (ROUGE) scores are having quite differences. Consecutively, ROUGE RF scores in R1, R2, and RL are 51.45, 45.52, and 54.58 respectively. Meanwhile, ROUGE BART scores are 32.78, 16.17, and 32.19. Consecutively, average ROUGE RF, BART, and RF×BART F-measure are 45.73, 21.38, and 31.31. RF has the highest average score. ATS hybrid RF×BART is shown to be performed better than the default BART. The average ROUGE F-measures for RF×BART obtain moderate score at 31.31. This score is better than the default BART's ROUGE score. RF×BART can be an alternative to the effective hybrid approach.

Corresponding Author:

Toto Haryanto
School of Data Science, Mathematics and Informatics, IPB University
Dramaga Bogor, West Java, 16680, Indonesia
Email: totoharyanto@apps.ipb.ac.id

1. INTRODUCTION

Data and information developed quantitatively and qualitatively. In the last decade, use of the internet has grown exponentially and has become an integral part of daily life [1]. Almost everyone from children, adolescent, adult, and old people use internet and generate massive amounts of data. The result of this increasing number of users is the significant increase in data which complicates information retrieval tasks. Therefore, utilizing modern information retrieval techniques becomes even more important.

Automatic text summarization (ATS) research has been conducted using various methods. A survey study conducted which stated that there are two kinds of abstractive approaches, namely structure-based and semantic-based [2]. The structural approach can have grammar problems because it does not consider the semantic representation of the document. The semantic approach has better linguistic quality because it involves semantic representations and semantic relations of text documents. The semantic method addresses the issue of a structural approach that reduces redundancies in the summary. The summary created produces better and denser cohesion. Research on ATS specifically on Indonesian is still relatively small [3].

A variation of the transformer neural network introduced, called bidirectional auto-regressive transformers (BART), which is a variation of bidirectional encoder transformers (BERT) [4]. BART used both pretraining and fine-tuning learning approach. Pretraining is done so that the model has a general understanding of the input documents. Fine tuning processed to serve specific purposes in a natural language processing (NLP) context. BART can accomplish specific goals such as generating abstract dialogue, question answering and summarization [4].

In the other hand, random forest (RF) was used in the ATS in a study and had good summary quality results in terms of relevance and novelty [5]. RF classifies whether a sentence should be included in the summary or not. The classifier is trained using the score features and summary information from each sentence from the document set previously created. Therefore, RF can be considered as the one approach in our next study. Automatically generating informative summaries remains a complex task. Previous paper stated that hybrid approach is one of good approach to get of both extractive and abstractive summarization advantages, but there is not enough proof of research conducted to support this argument. This is particularly true for ATS research in the Indonesian language, where the exploration of hybrid approaches is still limited. To address this gap, we propose a novel hybrid summarization approach for Indonesian text. Our model leverages the strengths of both extractive and abstractive techniques. The extractive component, powered by the RF, identifies the most relevant sentences within the document. The abstractive component, utilizing the BART, then generates a concise summary by reformulating and combining the key points extracted in the first stage.

2. RELATED WORKS

2.1. Automatic text summarization

ATS is the process of condensing information and ideas in a text into shorter texts [2]. Systematic literature review (SLR) studies on ATS have been conducted. General process of activities included are: representation, scoring, and text selection for summaries.

The first study on ATS was conducted in 1958. Experiments are conducted on extracting technical and magazine articles [6]. Statistical information such as the number of words in articles and their distribution are used. SLR conducted with a similar discussion, namely discussing approaches to topic representation, contextual influences on summaries, indicator-based representation approaches and the sentence selection process in summaries [7], [8]. A survey study the features of extractive ATS, supervised and unsupervised methods [8]. The extractive approach is classified as coherent in terms of the results of the summary. The supervised method requires mapped data from humans that marks sentences or parts of text that are included in the summary. The unsupervised approach is simpler than the supervised, as the label is not required in summary generation.

The ATS architecture in general includes a series of preprocessing, processing, and post processing [2]. Most ATS systems that perform well are focused on one goal, one method for a specific approach, one domain specific to the ATS. The preprocessing stage in ATS is to create a structured representation of the original text using various techniques such as segmentation, words tokenization, removal of stop-words, part-of-speech tagging, stemming and the like. The processing stage is the application of techniques for making a summary. The post-processing stage is solving the remaining problems is in the summary results such as sentence sequencing.

ATS is very often used in text mining and analytical applications such as information retrieval, information extraction, question answering. ATS has used infrared radiation (IR) techniques to strengthen search engine capabilities [9]. Further applications include media such as: news, opinion/sentiment, Microblog/Tweet, books, stories/novels, email, biomedical documents, legal documents, and scientific journals. Text summarization is also applicable for software summarization, as the artifacts or documentation and source code are basically text [10]. Software summarization is a field that is still under development. Software summarization discusses the process of generating representations of one or more software artifacts that contain information that stakeholders need to perform specific tasks in software engineering.

2.2. Random forest

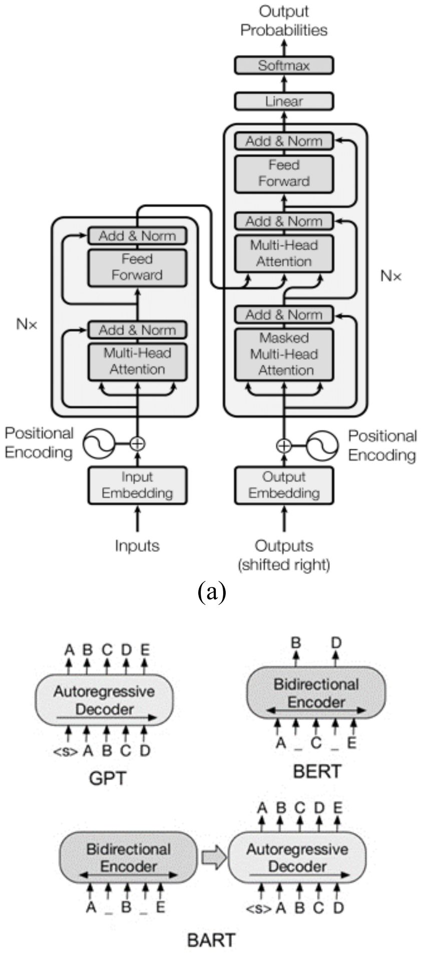
RF is an ensemble learning-based algorithm, which operates by building decision trees (DT). Each tree built from part of the dataset using the bootstrap aggregating (Bagging) method. The main advantages of RF are simplicity and effectiveness. The drawback of RF is low interpretability. The output class is the mode or average (mean) of all tree results that were built previously [11]. RF is often used in NLP cases, such as language modeling [12], sentiment analysis [13], detection of sarcasm [11], and summarization [3].

2.3. Bidirectional auto-regressive transformers

BART is a transformer model, a denoising autoencoder that can be used in a very wide range of tasks [4]. Transformers capable of handling sequential data using fewer resources than recurrent neural networks (RNN) or convolutional neural networks (CNN) [14]. Few of BART's predecessors are the BERT and GPT models [15], [16]. BART is particularly effective when fine-tuned for text generation. Finetuned

BART model can perform many NLP activities such as sequence classification, token classification, sequence generation and machine translation.

BART uses the standard seq2seq transformer architecture, utilizing GeLUs as activation function [4]. The basic model used is 6 layers of the encoder and decoder. The large model has 12 layers on the encoder and decoder. The architecture behaves like BERT with the differences of: i) each layer on the decoder performs cross-attention on the last hidden layer on the encoder, and ii) BERT uses an additional feed-forward network before word prediction, where BART is not. BART contains 10% more parameters than BERT. The brief comparison of BART and BERT is shown at Figure 1. Standard transformer architecture is shown in Figure 1(a). A general comparison of the BART architecture with GPT and BERT is shown in Figure 1(b).



(a)

(b)

Figure 1. Comparison of transformer-based models for NLP, highlighting their architectural differences: (a) standard transformer architecture [14] and (b) architecture comparison of GPT, BERT, and BART [4]

3. METHOD

The summary system process consists of several processes: i) input data in the form of text, ii) pre-process, iii) process each algorithm to produce an automatic summary, and iv) evaluation. The text data are news articles from the liputan6 dataset [17]. The preprocess consists of case folding, label mapping, tokenization and token removal. There are 3 scenarios of the algorithm that will be used, BART, RF, and a hybrid combination of RF and BART. Then, recall-oriented understudy for gisting evaluation (ROUGE) metric is then calculated on the summaries obtained. The summarization system proposed is shown in Figure 2. The data to be used is Liputan 6 Indonesian summarization dataset [17].

The proposed RF×BART hybrid model is defined by a two-stage process operating in an extractive-abstractive fashion. Crucially, the RF component functions as the first stage: an input selection layer. It ranks and selects the most salient sentences from the source document based on a set of linguistic and positional features. Only these high-quality, filtered sentences are passed to the BART model, which then generates the final abstractive summary. This architecture is designed to optimize the BART input, thereby mitigating redundancy and improving the overall content fidelity of the generated summary.

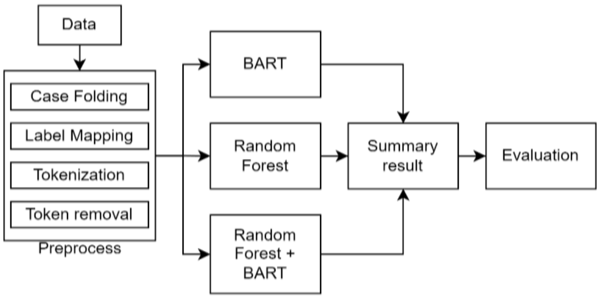


Figure 2. Proposed summarization system

3.1. Data preprocess

The preprocess consists of case folding, label mapping, tokenization and token removal. By converting all characters to lowercase (case folding), words differing only in capitalization are treated as identical. Label mapping is done on the original article data so that index of each reference sentence is preserved. Tokenization is the process of making the smallest elements of text into form of integer vectors. Token removal removes unnecessary words/elements. Token removal is done so that the data that is processed is substantial.

3.2. RF extractive summarization

For each document, we have to extract the information as the features. In this study, we used 8 attributes from sentences in one document refers to the [3], [18] as well:

i) TF/IDF: the value of this attribute is calculated based on the number of terms that appear in the sentence and the number in all sentences in the document. TF/IDF in this context can be referred to as (TF/ISF). The formula is written in (1).

$$TF(S_i) = \frac{(log(isf)\times(tf))}{len} \tag{1}$$

isf is the number of occurrences of the ith word in the entire sentence. *tf* is the occurrence of the word (term frequency) in the ith sentence. *len* is the number of words in the sentence.

ii) Uppercase: this attribute will have higher weighting value in sentences that have one or more capital letters such as the name of the person's subject, place or object name. Values are obtained using (2) and (3).

$$CW(s) = \frac{uppercase\ count\ at\ a\ sentence}{word\ counts\ at\ a\ sentence} \tag{2}$$

$$f2(s) = \frac{CW(s)}{max(CW(s))} \tag{3}$$

iii) Proper noun: the number of nouns in a sentence positively correlates with the value of this attribute. It is defined as ratio of noun word over the length of a sentence. The value is obtained using (4).

$$f3(s) = \frac{noun\ word\ count\ at\ a\ sentence}{word\ counts\ at\ a\ sentence} \tag{4}$$

iv) Cue phrases: in general, many sentences begin with phrases such as so, investigation and the like. Emphasis words such as "*terbaik*", "*terpenting*", "*paling penting*" which equivalent to "best" or "most important" in English and other phrases also show the good characteristic that a sentence is included in the summary. It is defined as ratio of phrase count over word count in a sentence. Attribute values are obtained using (5).

$$f4(s) = \frac{phrase\ count\ at\ a\ sentence}{word\ counts\ at\ a\ sentence} \tag{5}$$

v) Numerical data: numerical data/numbers indicate that a sentence may contain important information. The value of this attribute is obtained using (6).

$$f5(s) = \frac{numerical\ word/token\ in\ a\ a\ sentence}{word\ counts\ at\ a\ sentence} \tag{6}$$

vi) Sentence length: long sentences have a higher weight. It is defined as ratio of word count in a sentence over longest sentence in term of word count. Values obtained by (7).

$$f6(s) = \frac{word\ count\ in\ a\ sentence}{longest\ sentence\ in\ term\ of\ word\ count} \tag{7}$$

vii) Sentence position: the position of the sentence in a paragraph determines whether the sentence is considered important. It is assumed that the first sentence and the last sentence are important sentences. Values obtained by (8).

$$f7(s) = \{1,\ if\ first\ or\ last\ sentence\ \frac{N-P}{N}, if\ other \tag{8}$$

viii) Similarity to title: this attribute assesses the level of similarity in the title. It is defined as ratio of intersected word. Values obtained by (9).

$$f8(s) = \frac{keyword\ at\ a\ sentence \cap keyword\ at\ title}{keyword\ at\ a\ sentence \cup keyword\ at\ title} \tag{9}$$

RF will be trained to perform binary classification, which is to determine whether a sentence in an article is included in the summary. Labels 0 and 1 are used in the dataset to indicate whether a sentence is a summary reference. If the label is 0 then the sentence does not include a summary, if 1 then it is a summary.

3.3. BART abstractive summarization

Abstractive summarization is one of the text generation tasks. Summarizing is done by training BART with stages of pre-training and fine-tuning. Common pre-training approach refines its understanding by reconstructing corrupted text. Each method and pre-training example are shown in Figure 3. In the example in Figure 3, words or tokens are symbolized by the letters A through E. BART fine-tuning in the case of summarization is done in a specific way called sequence generation. Sequence generation is processed by manipulating the text of the input so that the results have similar meaning with shorter length.

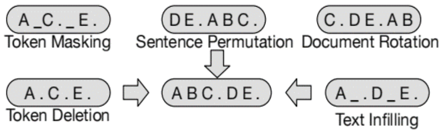


Figure 3. BART Pre-training example [4]

3.4. Hybrid summarization

Hybrid summarization is done by combining the RF and BART algorithms. RF is utilized at extractive summarization. The summary results of the RF are then processed on the BART for summary regeneration. Illustration hybrid summarization scenario is shown in Figure 4.

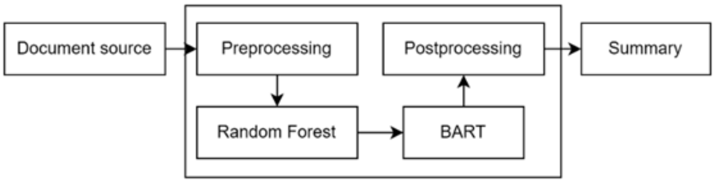


Figure 4. Hybrid summarization scenario

3.5. Recall-oriented understudy for gisting evaluation

ROUGE is one of the automatic measurement methods. In this method, the summary of the results of automation is compared to the summary made by humans [19]. The measurement is conducted by calculating the matching units such as n-grams (n-words), word order, and word pairs between the summary of the automated results compared to the reference summary. A higher ROUGE score indicates better summarization performance, reflecting a greater degree of overlap between the generated summary and the reference summary.

4. RESULTS AND DISCUSSION

4.1. Preprocessed data

Preprocessing consists of case folding, label mapping, tokenization and token removal. Pre-processing for RF model training data includes case folding, label mapping, and token removal. Label mapping is included in feature engineering. Tokenization is only prepared for the development of the BART model. Case folding is not used because it conflicts with the uppercase feature. The uppercase feature will always have a value of 0 if the case folding process is included.

On this occasion, comparison of the accuracy of the results of the RF model was carried out on two data treatments, namely with pre-processing and without pre-processing. Comparisons were made with random data of 1,000 and 5,000 points. Accuracy comparison of the two RF models is shown in Table 1.

Table 1. Accuracy comparison of RF model with and without preprocessing

No	Train data count	Test data count	Test scheme	Accuracy without preprocess (%)	Accuracy with reprocess (%)
1	1,000	10,972	train	76.55	73.32
2	1,000	10,972	validation	78.47	74.06
3	5,000	10,972	train	74.70	67.81
4	5,000	10,972	validation	75.06	74.61

Comparison shows the model that uses data without preprocessing has higher accuracy, therefore RF model without preprocessing will be used. Feature engineering process from the raw data includes the 8 features described in the previous section. Feature engineering is performed on all splits. Split train with a total of 193,883 data points produces 2,640,590 new data points which are a classification dataset of a sentence whether it is included in the summary or not. Each of these new data points is a numerical representation of each sentence. Data snippets are shown in Table 2.

Comparison of the amount of data based on the labels in Table 3 shows that the dataset has imbalance conditions. The condition can cause an overfit model that makes the classifier model lean towards only one class. Resampling the data with specific techniques allows us to train a model that is not biased by the initial imbalance.

Table 2. Snippet of new dataset

f1	f2	f3	f4	f5	f6	f7	f8	id_text	label
3.71	0.35	1	0	0	0.61	1	0.12	26408	1
2.56	0.14	0.29	0	0	0.25	0.33	0	26408	1
4.5	0.25	0.56	0	0	1	0.22	0	26408	0
4.47	0.19	1	0	0	0.49	1	0	26410	0
3.76	0.12	0.57	0	0.12	0.19	0.87	0.12	26410	1
3.83	0.09	0.7	0	0.09	0.26	0.67	0	26410	0

Table 3. Label count on new dataset

Label	Count
1	372,403
0	2,268,187

Tokenization is done using the tokenizer from the Indobart huggingface model. Tokenization creates the article representation to integer vector structure. Transformer is a model that processes integer vector structure input and produces vector output as well. The reverse process of tokenization is detokenization, which is the process of converting integer vectors to their original values. Detokenization is carried out at the output of the transformer model. Tokenization is done when BART is about to be trained. Tokenization is done on feature data and labels.

Initial BART training trials were conducted to determine the resources to be used. The training is conducted on infrastructure as a service (IaaS) which allows users to perform computing processes using the latest GPUs. The more training data that is processed, the higher the training costs. Data selection was carried out from 193,883 data to minimize model development costs. Data selection was carried out by utilizing the results of engineering features for previous RF models. The conditions used for the selection are an article must have feature values that are equal to or more than the average of each feature. 109,522 data points were obtained from this process.

4.2. Model comparison

RF has a weakness in the classification of imbalanced data. The class_weight=balanced parameter is used at the initialization of the RF model using sklearn python module, so that the classifier model performs training with balanced data automatically by using inversely proportional class weight. We used huggingface framework with BART model called Indobart [20]. The default configuration is: 16 batches, 12 hidden layers (6 encoders and 6 decoders). BART finetuned in 70 epochs and 115000 steps on 1×NVIDIA RTX A6000 48 GB. Beam search size of 5 is used in summary generation.

Area under receiver operator characteristic (AUROC/AUC ROC) is used to measure classifier quality [21]. The ROC graph is shown in Figure 5. A random classifier has a value of 0.5 and a perfect classifier has a value of 1. A classifier is considered effective if the AUROC has a value of more than 0.5 and preferably close to 1. The AUC RF value is shown to have a value > 0.5, so it is considered effective as a classifier, even though it is very close to 0.5.

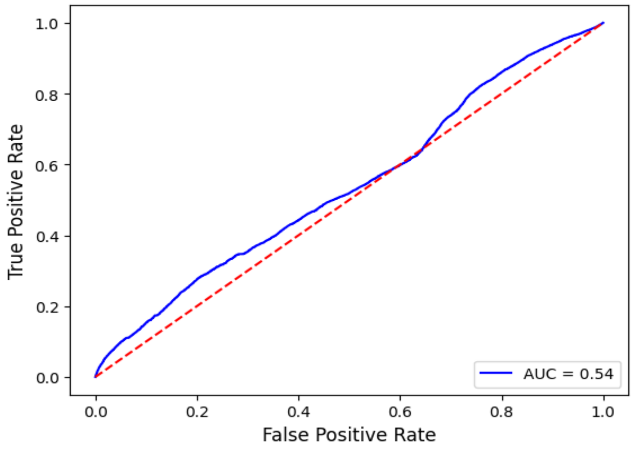


Figure 5. AUROC chart for RF model

The summary results of the comparison of the three scenarios are shown in Table 4. BART has results of abstractive summary. RF succeeded in summarizing extractive. RF×BART gives intermediate results, better than BART but lower in scores than RF. The results of the BART summary can capture the outline of the subject in sentences, such as "Kepolisian Daerah Riau" becomes "Kapolda Riau". BART generates fairly good abstractive summaries, but still generates characters or tokens and words that is not included in great dictionary of the Indonesian language or kamus besar bahasa Indonesia (KBBI) term such as *touwet*. At the end of a sentence BART still evokes meaningless symbols such as periods, exclamation points and question marks. The results of the summary of the hybrid RF×BART have similar characteristics to BART. In general, the statistical comparison of the count of words in the summary results is shown in Table 5.

4.3. Evaluation

ROUGE evaluation uses ROUGE n-gram, ROUGE longest common subsequence (RL) and ROUGE weighted LCS (RW). ROUGE recall n-grams are displayed with the metrics R1, R2 to Rn. RL considers the longest similar subsequence of two texts. RW also uses LCS with consideration of the weight in each LCS. Each metric has recall, precision, and f-measure measurements. Recall represents ratio between count overlapping word of reference and candidate summary over reference summary. Precision represents ratio between count overlapping word of reference and candidate summary over candidate summary. F-measure represents harmonic ratio between recall and precision.

Table 4. Sample summarization results

Part	Text (in Indonesian)	Text (in English)
Raw text	Liputan6.com, Jakarta*: Kepolisian Daerah Riau bertekad memberantas pelaku penyelundupan kayu yang kerap terjadi di Riau. Selain itu, Polda setempat juga akan memberangus menipulasi dana reboisasi dan iuran hasil hutan. Demikian ditegaskan Kepala Polda Riau Brigadir Jenderal Polisi Johny Yodjana, seusai dilantik menjadi Kapolda Riau oleh Kepala Polri Jenderal Polisi Suroyo Bimantoro, di Jakarta, baru-baru ini. Menurut Johny, pelaku tindak kriminal yang kerap menjarah kayu di Riau akan ditindak tegas. "Saya tak akan pandang bulu," janji Johny. Selain itu, ia bertekad menyelidiki dugaan manipulasi dana reboisasi dan iuran hasil hutan sebesar Rp 680 miliar yang dilakukan sebuah perusahaan kayu di Riau. Sementara itu, selain melantik Johny Yodyana, Kapolri juga melantik Inspektur Jenderal Polisi Firman Gani menjadi Kapolda Sulawesi Selatan dan Brigjen Pol. Eddy Darnadi menjadi Kapolda Maluku. Selain itu, Bimantoro juga melantik Komisaris Besar Pol. Totok Soenarjo menjadi Kapolda Jambi, Brigjen Pol. Sugiri menjadi Kapolda Lampung, dan Brigjen Pol. Dwi Purwanto menjadi Kapolda Bengkulu. (ICH/Edi Priyono dan Andi Azril).*	Liputan6.com, Jakarta: The Riau Regional Police are determined to crack down on timber smugglers who often operate in Riau. In addition, the local police will also crack down on the manipulation of reforestation funds and forest product levies. This was emphasized by Riau Regional Police Chief Brigadier General Johny Yodjana after being inaugurated as Riau Regional Police Chief by National Police Chief General Suroyo Bimantoro in Jakarta recently. According to Johny, criminals who frequently loot timber in Riau will be dealt with severely. "I will not discriminate," Johny promised. In addition, he is determined to investigate allegations of manipulation of reforestation funds and forest product levies amounting to Rp 680 billion by a timber company in Riau. Meanwhile, in addition to inaugurating Johny Yodyana, the National Police Chief also inaugurated Inspector General Firman Gani as South Sulawesi Police Chief and Brigadier General Eddy Darnadi as Maluku Police Chief. Furthermore, Bimantoro also appointed Senior Commissioner Totok Soenarjo as the Jambi Police Chief, Brigadier General Sugiri as the Lampung Police Chief, and Brigadier General Dwi Purwanto as the Bengkulu Police Chief. (ICH/Edi Priyono and Andi Azril).
Ext reference	Liputan6.com, Jakarta: *Kepolisian Daerah Riau bertekad memberantas pelaku penyelundupan kayu yang kerap terjadi di Riau. "Saya tak akan pandang bulu," janji Johny. Totok Soenarjo menjadi Kapolda Jambi, Brigjen Pol.*	Liputan6.com, Jakarta: The Riau Regional Police are determined to crack down on timber smugglers who often operate in Riau. "I will not discriminate," promised Johny. Totok Soenarjo became the Jambi Regional Police Chief, Brigadier General Pol.
Abs reference	*Kapolda Riau baru Brigjen Pol. Johny Yodjana bertekad memberantas pelaku penyelundupan kayu di Riau. Ia berjanji akan menindak tegas pelaku tanpa pandang bulu.*	The new Riau Police Chief, Brigadier General Johny Yodjana, is determined to crack down on timber smugglers in Riau. He has promised to take firm action against perpetrators without exception.
BART	*kapolda riau brigjen polisi johny yodjana berjanji menyelidiki dugaan manipulasi dana reboisasi dan iuran hasil hutan sebesar rp 680 miliar. polri memberangus menipulasi dan iuran hutan. kasus ini akan dilaporkan ke masyarakat. johny mengatakan tak akan pandang bulu... kejagung adalah pelaku tindak kriminal. polri......*	riau police chief brigadier general johny yodjana promised to investigate allegations of manipulation of reforestation funds and forest revenue contributions amounting to rp 680 billion. the national police will crack down on manipulation and forest revenue contributions. this case will be reported to the public. johny said he would not discriminate... the attorney general's office is the perpetrator of criminal acts. the national police...
RF	Liputan6. com, Jakarta*: Kepolisian Daerah Riau bertekad memberantas pelaku penyelundupan kayu yang kerap terjadi di Riau. Selain itu, Polda setempat juga akan memberangus menipulasi dana reboisasi dan iuran hasil hutan. Menurut Johny, pelaku tindak kriminal yang kerap menjarah kayu di Riau akan ditindak tegas "Saya tak akan pandang bulu," janji Johny*	Liputan6.com, Jakarta: The Riau Regional Police are determined to crack down on timber smugglers who frequently operate in Riau. In addition, the local police will also crack down on the manipulation of reforestation funds and forest product levies. According to Johny, criminals who frequently loot timber in Riau will be dealt with severely. "I will not discriminate," Johny promised.
RF×BART	*kepolisian daerah riau bertekad memberantas pelaku penyelundupan kayu yang kerap terjadi di riau. selain itu, polda setempat juga akan memberangus menipulasi dana reboisasi dan iuran hasil hutan. tni akan menindak tegas touwet-nya ?`' keterangan johny, pengedar akan*	the riau regional police are determined to crack down on timber smugglers who frequently operate in riau. in addition, the local police will also crack down on the manipulation of reforestation funds and forest product levies. will the tni take firm action against the smugglers? according to johny, the distributors will

Table 5. Statistical comparison of word count

Statistic	BART	RF	RF×BART
Test count	10972	10972	10972
Mean	47,99	64,76	50,73
Std dev	3,90	23,13	16,83
Min	32	7	6
25%	45	48	38
50%	48	62	49
75%	51	79	62
Max	61	183	132

The calculation of the ROUGE score in this hybrid method uses variables that match the context of the model. The ROUGE calculation for BART uses an abstract summary of the dataset as a reference. The ROUGE calculation for RF uses the extractive summary of the dataset as a reference. The ROUGE calculation for hybrid RF×BART uses the abstract and extractive summary of the dataset as a reference and the RF×BART results as a hypothesis summary. The hybrid calculation (RF×BART) is the average of the abstractive ROUGE score (RF×BART 1) and the extractive score (RF×BART 2). Comparison of the ROUGE score evaluation of the results of all models is shown in Table 6. The comparison of these scores shows that the RF model is the model with the highest ROUGE score of the majority of metrics. There is one RF×BART 1 score that can exceed the ROUGE RF score, namely the score on the R1 precision metric with a value of 43.99. Comparison of the average scores shown in Table 7.

Table 6. ROUGE scores of all models

ROUGE	Evaluation	BART	RF	RF×BART	RF×BART 1	RF×BART 2
R1	F-measure	32.78	51.45	40.81	46.61	35.02
	Precision	25.37	43.39	35.94	43.99	27.89
	Recall	46.97	70.06	52.54	54.05	51.02
R2	F-measure	16.17	45.52	28.54	38.39	18.70
	Precision	12.45	38.40	25.68	36.50	14.86
	Recall	23.43	62.16	35.81	44.20	27.43
R3	F-measure	8.93	43.33	22.93	34.94	10.92
	Precision	6.83	36.53	21.00	33.34	8.65
	Recall	13.09	59.51	28.17	40.19	16.16
RL	F-measure	32.19	54.48	41.50	48.64	34.36
	Precision	25.81	46.77	37.15	46.11	28.19
	Recall	43.17	70.04	50.60	54.68	46.51
RW	F-measure	16.81	33.88	22.75	27.17	18.34
	Precision	17.03	40.07	28.57	37.84	19.30
	Recall	16.85	31.97	20.66	22.69	18.64

Table 7. Average ROUGE scores of all models

Evaluation	BART	RF	RF×BART
F-measure	21.38	45.73	31.31
Precision	17.50	41.03	29.67
Recall	28.70	58.75	37.56

The highest RF score lies in the RL recall metric with a value of 70.06. The lowest RF score lies in the RW recall metric with a value of 28.50. The ROUGE RF score has the highest value because the model results have the most similarities with the extractive summary text. The ROUGE BART score has the lowest average, maximum and minimum values. The highest ROUGE BART score lies in the R1 recall metric with a value of 46.97. The lowest ROUGE BART score lies in the R3 precision metric with a value of 6.83. The ROUGE RF×BART score has sufficient average, maximum and minimum values. The highest ROUGE RF×BART score lies in the R1 recall metric with a value of 52.54. The lowest ROUGE RF×BART score lies in the RW recall metric with a value of 20.66.

Figure 6 presents a comparison graph of the average ROUGE F-measure scores for three summarization techniques: RF, BART, and RF×BART. Notably, RF consistently achieves the highest scores across all metrics. This can be attributed to RF focus on retrieving sentences most similar to the reference text through a classification process. While RF×BART performs better than BART, it still falls short of RF strong similarity to the reference.

A comparison of the RF×BART scores with other studies shown at Table 8. Multiple studies have shown that RF×BART outperforms BART on all ROUGE metrics. Some studies have a higher ROUGE score such as in [4], [22] on the R1 score. The BART model that surpasses all R1, R2, and RL RF×BART scores is the BART [23].

The RF×BART hybrid model is well-suited for deployment in practical environments requiring highly accurate and efficient summarization. Potential real-world applications include i) large-scale news aggregation, where the two-stage process allows the RF component to quickly filter key information for timely abstractive output; and ii) summarization of complex legal or medical documents, where the inherent precision of the extractive RF component ensures that critical clauses and facts are prioritized for inclusion in the final summary, making the resulting document both concise and legally traceable.

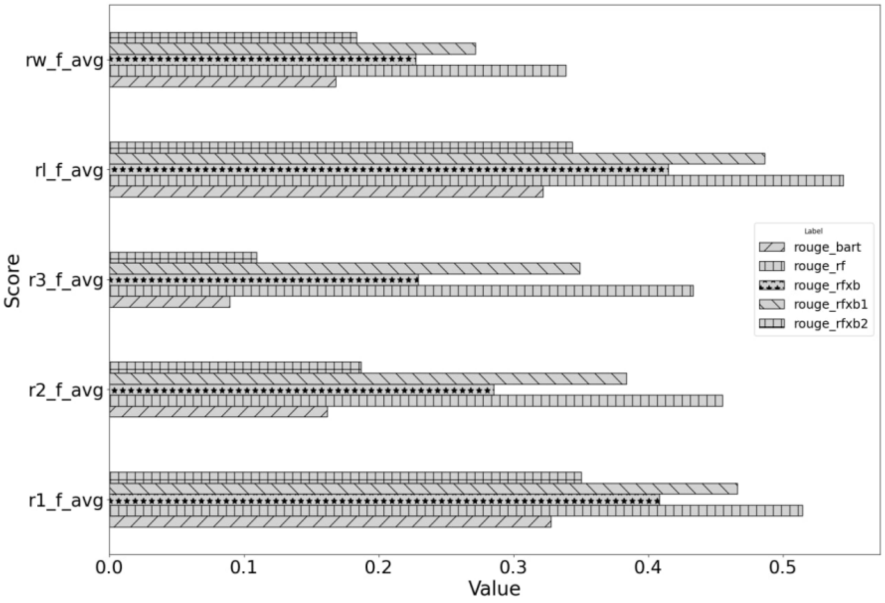


Figure 6. Comparison chart of average F-measure ROUGE RF, BART dan RF×BART scores

Table 8. Comparison of ROUGE scores of RF×BART with previous BART model

Researcher	Methods	R1	R2	RL
Upadhyay *et al.* [24]	BART	39.21	9.09	37.94
Lewis *et al.* [4]	BART	44.16	21.28	40.9
DeYoung *et al.* [25]	BART	27.56	9.4	20.8
Eddine *et al.* [22]	BART	42.4	28.8	40.3
Kondadadi *et al.* [23]	BART	60.51	48.14	57.65
This work	RF×BART	40.81	28.54	41.49

5. CONCLUSION

Research reveals a significant difference in ROUGE score between the performance of RF and BART on ATS tasks. ROUGE F1-scores R1, R2, and RL model of the RF are 51.45, 45.52, and 54.58. The BART scores for R1, R2, and RL for the RF model were 32.78, 16.17, and 32.19. The average ROUGE F-measure BART score is 45.73. The average ROUGE F-measure BART score is 21.38. RF has the highest average score. RF excels at extractive summarization, whereas BART can create summaries that express ideas in new ways (abstractive summarization). However, BART-generated text may include words or phrases not found in the KBBI. The combined RF×BART approach, while outscoring BART on the ROUGE metric, falls short of RF performance. The ROUGE F-measure RF×BART average score is 31.31. RF×BART has a moderate score. This score increases compared to the BART score independently. RF×BART can be an effective alternative to the hybrid method approach. The integration of several methods in the concept of hybrid summarization is still at the development stage. In future studies, it is expected that researchers to offer more diverse hybrid approach/architecture that is more robust than just relying on the output of extractive methods. A key benefit of the RF×BART architecture is its robust ability to balance interpretability and fluency. The RF component offers a high degree of interpretability and control; its feature-driven decision-making (e.g., explicitly weighting features like sentence position and key phrases) makes the extractive selection process transparent and traceable. Conversely, the BART model utilizes its powerful neural language generation capabilities to ensure the final summary is highly fluent and coherent, resulting in human-readable and grammatically correct prose. This successful integration merges the strengths of symbolic (RF) and neural (BART) techniques.

FUNDING INFORMATION

Funding source of this research has been enabled by the researchers themselves. Usage of resources that require fund are including virtual machine and remote GPU.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Muhammad Adib Zamzam	✓	✓	✓			✓	✓	✓	✓	✓	✓		✓	✓
Agus Buono	✓	✓		✓						✓		✓	✓	✓
Toto Haryanto		✓		✓		✓				✓		✓	✓	✓

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding acquisition
Fo	: **Fo**rmal analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, [MAZ].

REFERENCES

[1] E. S. E. L. Anderson and V. Stavropoulos, "Internet use and problematic internet use: a systematic review of longitudinal research trends in adolescence and emergent adulthood," *International Journal of Adolescence and Youth*, vol. 22, no. 4, pp. 430–454, 2017, doi: 10.1080/02673843.2016.1227716.

[2] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: a comprehensive survey," *Expert Systems with Applications*, vol. 165, 2021, doi: 10.1016/j.eswa.2020.113679.

[3] P. M. Sabuna and D. B. Setyohadi, "Summarizing Indonesian text automatically by using sentence scoring and decision tree," in *2017 2nd International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2017, pp. 1–6, doi: 10.1109/ICITISEE.2017.8285473.

[4] M. Lewis *et al.*, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880, doi: 10.18653/v1/2020.acl-main.703.

[5] A. John and M. Wilscy, "Random forest classifier based multi-document summarization system," in *2013 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, 2013, pp. 31–36, doi: 10.1109/RAICS.2013.6745442.

[6] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, 1958, doi: 10.1147/rd.22.0159.

[7] M. Allahyariet *et al.*, "Text summarization techniques: a brief survey," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, 2017, doi: 10.14569/ijacsa.2017.081052.

[8] K. Nenkova and K. McKeown, "A survey of text summarization techniques," in *Mining Text Data*, Boston, MA: Springer, 2012, pp. 43–76, doi: 10.1007/978-1-4614-3223-4_3.

[9] S. Tuarob, S. Bhatia, P. Mitra, and C. L. Giles, "AlgorithmSeer: a system for extracting and searching for algorithms in scholarly big data," *IEEE Transactions on Big Data*, vol. 2, no. 1, pp. 3–17, 2016, doi: 10.1109/TBDATA.2016.2546302.

[10] L. Moreno and A. Marcus, "Automatic software summarization: the state of the art," in *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)*, 2017, pp. 511–512, doi: 10.1109/ICSE-C.2017.169.

[11] T. Jain, N. Agrawal, G. Goyal, and N. Aggrawal, "Sarcasm detection of tweets: a comparative study," in *2017 Tenth International Conference on Contemporary Computing (IC3)*, 2017, pp. 1–6, doi: 10.1109/IC3.2017.8284317.

[12] I. Oparin, O. Glembek, L. Burget, and J. Cernocky, "Morphological random forests for language modeling of inflectional languages," in *2008 IEEE Spoken Language Technology Workshop*, 2008, pp. 189–192, doi: 10.1109/SLT.2008.4777872.

[13] S. E. Saad and J. Yang, "Twitter sentiment analysis based on ordinal regression," *IEEE Access*, vol. 7, pp. 163677–163685, 2019, doi: 10.1109/ACCESS.2019.2952127.

[14] A. Vaswani *et al.*, "Attention is all you need," *arXiv:1706.03762*, 2017.

[15] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, vol. 1, pp. 4171–4186.

[16] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *OpenAI,* pp. 1-12, 2018.

[17] F. Koto, J. H. Lau, and T. Baldwin, "Liputan6: a large-scale Indonesian dataset for text summarization," *1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 598–608, doi: 10.18653/v1/2020.aacl-main.60.

[18] N. S. Shirwandkar and S. S. Kulkarni, "Extractive text summarization using deep learning," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 2018, pp. 1–5 pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697465.

[19] C.-Y. Lin, “ROUGE: a package for automatic evaluation of summaries,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2004.

[20] S. Cahyawijaya *et al.*, “IndoNLG: benchmark and resources for evaluating Indonesian natural language generation,” in *2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8875–8898, 2021, doi: 10.18653/v1/2021.emnlp-main.699.

[21] S. Rastkar, G. C. Murphy, and G. Murray, “Automatic summarization of bug reports,” *IEEE Transactions on Software Engineering*, vol. 40, no. 4, pp. 366–380, 2014, doi: 10.1109/TSE.2013.2297712.

[22] M. K. Eddine, N. Tomeh, N. Habash, J. Le Roux, and M. Vazirgiannis, “AraBART: a pretrained Arabic sequence-to-sequence model for abstractive summarization,” in *Seventh Arabic Natural Language Processing Workshop (WANLP)*, 2022, pp. 31–42, doi: 10.18653/v1/2022.wanlp-1.4.

[23] R. Kondadadi, S. Manchanda, J. A. Ngo, and R. McCormack, “Optum at MEDIQA 2021: abstractive summarization of radiology reports using simple BART finetuning,” in *20th Workshop on Biomedical Language Processing*, 2021, pp. 280–284, doi: 10.18653/v1/2021.bionlp-1.32.

[24] A. Upadhyay, N. Bhavsar, A. Bhatnagar, M. Singh, and P. Motlicek, “Automatic summarization for creative writing: BART based pipeline method for generating summary of movie scripts,” in *Workshop on Automatic Summarization for Creative Writing*, 2022, pp. 44–50.

[25] J. DeYoung, I. Beltagy, M. V. Zuylen, B. Kuehl, and L. L. Wang, “MS^2: multi-document summarization of medical studies,” in *2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 7494–7513, doi: 10.18653/v1/2021.emnlp-main.594.

BIOGRAPHIES OF AUTHORS

Muhammad Adib Zamzam received Bachelor of Informatic Engineering in 2019 from Maulana Malik Ibrahim Public Islamic University, Indonesia. His research interest is in the machine learning, deep learning, and natural language processing. He can be contacted at email: adib35785@gmail.com.

Agus Buono received Bachelor of Computer Science from Bogor Agricultural University, Indonesia and a master degree from the same university. His Ph.D. obtained from Faculty of Computer Science, Universitas Indonesia. Currently, he is full professors in Computer science as School of Data Science, Mathematics and Informatics, IPB University. His research interest is computational intelligence, NLP, and voice recognition. He can be contacted at email: agusbuono@apps.ipb.ac.id.

Toto Haryanto received Bachelor of Computer Science in 2006 from Bogor Agricultural University, Indonesia and a master degree in 2011 from the same university. His Ph.D. obtained from Faculty of Computer Science, Universitas Indonesia in 2021. He is assistant professor at School of Data Science, Mathematics and Informatics, IPB University. He joins visiting researcher at University of Pardubice, Czech Republic for three months funded by Erasmus+ Program at 2018. His research interest is in high-performance computing for medical images, machine learning, and bioinformatics. He can be contacted at email: totoharyanto@apps.ipb.ac.id.