Journal of Information Systems and Informatics
Vol. 7, No. 2, June 2025 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v7i2.1082 Published By DRPM-UBD

Abstractive Text Summarization to Generate Indonesian News Highlight Using Transformers Model

I Gusti Agung Intan Utami Putri[1], I Nyoman Prayana Trisna[2], Ni Kadek Dwi Rusjayanthi[3]

[1,2,3]Information Technology Department, Udayana University, Bali, Indonesia
Email: [1]gungintan202@gmail.com, [2]prayana.trisna@unud.ac.id, [3]dwi.rusjayanthi@unud.ac.id

Abstract

The increasing volume of information has led to information overload, a condition where individuals struggle to filter and comprehend information efficiently within a limited time. To address this issue, automatic text summarization is essential, particularly for enhancing readability, supporting quick decision-making, and enabling real-time content processing in digital platforms. This research assesses the effectiveness of two transformer-based models, IndoT5 and mBART, for generating abstractive summaries (highlight) of Indonesian news articles. The abstractive approach allows models to generate new sentences with more natural language structures compared to extractive methods. Fine-tuning was conducted using a dataset comprising 10,410 news articles from Tempo.co, each containing full content and corresponding highlight as reference. ROUGE and BERT-Score metrics were employed to evaluate structural and semantic correspondence between the references and the generated summaries. Results show that IndoT5 outperformed in terms of ROUGE-1 (0.43087), ROUGE-2 (0.29143), ROUGE-L (0.39224), BERT-Score Recall (0.89130), and F1-Score (0.87708), indicating its capability to generate complete and relevant news highlight. Meanwhile, mBART achieved a higher BERT-Score Precision (0.86717) but tended to generate less informative outputs. The findings of this research are expected to aid in enhancing the coherence, readability, and real-time applicability of abstractive summarization systems.

Keywords: Abstractive Summarization, Transformers, mBART, IndoT5, News Highlight

1. INTRODUCTION

The rapid growth of information especially from online news media, has led to a condition known as information overload, where individuals are overwhelmed by the volume of data available and struggle to process it effectively [1]. A study by the Nielsen Norman Group shows that on average, readers only read 20–28% of the total text on a web page [2], indicating that most of the content remains unread. This phenomenon highlights the urgent need for systems that can automatically summarize long texts into concise, informative versions, so that enabling readers to quickly grasp essential information without reading the full content.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Automatic Text Summarization (ATS) focuses on condensing long documents into shorter versions that preserve the essential meaning [1]. ATS plays a critical role in enhancing reading efficiency, supporting decision-making, and improving accessibility to information. There are two main approaches in ATS: extractive and abstractive [3]. The extractive approach involves straightly choosing important part or segments from the source content, which often resulting in less coherent summaries due to the lack of restructuring [4]. On the other hand, abstractive summarization requires interpreting the original content and creating novel sentences to convey the main ideas using paraphrasing and generalization, which typically leads to more natural, fluent, and readable summaries [5].

Despite rapid progress in ATS for high-resource languages like English, abstractive summarization in low-resource languages such as Indonesian remains underexplored and faces several challenges [6]. The main challenges include the limited availability of high-quality annotated datasets, insufficient domain-specific pre-trained models, and linguistic characteristics unique to Indonesian such as flexible word order, affixation, and contextual pronoun usage, which complicate semantic representation and generation [7-8]. Most existing works rely on conventional approaches or pre-trained models not optimized for Indonesian. For example, earlier research utilized conventional models like BiGRU for Indonesian summarization, showed that the BiGRU model could generate individual words correctly, but the resulting summaries lacked coherence and were often difficult to read [9]. Ardyanti et al. (2024) compared LSTM with attention mechanisms and T5-Small on the Indosum dataset, finding that the LSTM-Attention model achieved a low ROUGE score, while the fine-tuned T5-Small outperformed with a significantly higher ROUGE, although the outputs were still less abstractive and tended to extract direct sentences without paraphrasing [10]. Purnama & Utami (2023) implemented T5 with the Indosum dataset but found the summaries repetitive and lacking informativeness [11]. Meanwhile, Itsnaini et al. (2023) explored IndoT5, a T5 variant pre-trained specifically for Indonesian, and reported promising results with high ROUGE scores, indicating strong surface-level performance, although there remains room for improvement in generating more abstract and paraphrased summaries [1].

Transformers have recently brought substantial improvements in abstractive summarization particularly for high-resource languages like English. Dharrao et al. (2024) compared BART, T5, and PEGASUS, finding T5 to be the most effective [12], while Lewis et al. (2019) reported that BART achieved the highest ROUGE scores on the XSum dataset [13]. Cao et al. (2024) introduced DMSeqNet-mBART, combining mBART with Adaptive-DropMessage techniques for abstractive summarization in Mandarin, demonstrating its effectiveness in handling syntactically complex languages [14], raising the possibility that mBART could also be effective for Indonesian. These findings suggest potential for cross-lingual

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

transfer, but their performance on Indonesian news summarization, especially in generating concise and informative highlights, remains underexplored and untested. Moreover, comparative studies involving multilingual models like mBART for Indonesian summarization are scarce, leaving a gap in understanding their applicability to low-resource languages.

This research seeks to address this limitation through the implementation and comparison of IndoT5 and mBART models in conducting abstractive summarization of Indonesian news articles. This work focuses explicitly on generating news highlights rather than full-length summaries. Highlights, as opposed to general summaries, are expected to capture the most critical elements such as main events, key facts, and issues within a few concise sentences. By leveraging a dataset sourced from Tempo.co, which provides structured news highlights, this research evaluates the model's effectiveness in producing concise yet informative summaries.

2. METHODS

To address the research objectives and evaluate the effectiveness of IndoT5 and mBART for abstractive highlight generation in Indonesian news articles, this study adopts an experimental research approach supported by a systematic and structured workflow. The research process is divided into several interconnected stages, as presented in Figure 1.

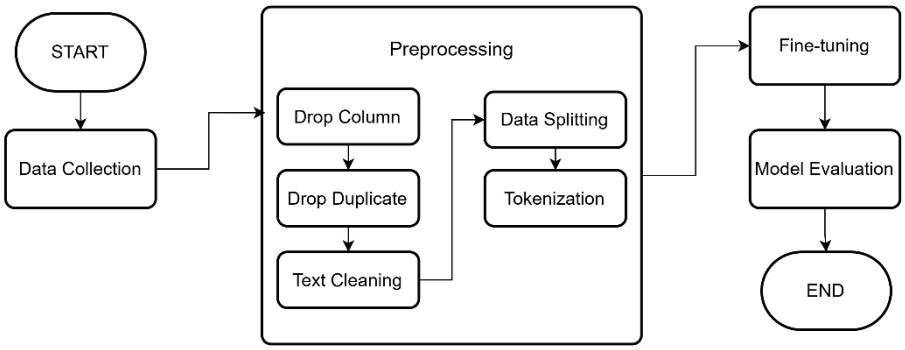

Figure 1. Research flow

The research workflow consists of four main stages: data collection, preprocessing, fine-tuning, and evaluation. The first stage, data collection, involves scraping news data from Tempo.co using BeautifulSoup. In the preprocessing stage, the data is cleaned, including steps such as drop column, drop duplicate, text cleaning, data splitting, and tokenization. Finally, the preprocessed data is used to fine-tune the IndoT5 and mBART50 models, followed by evaluation using ROUGE and

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

BERT-Score metrics to assess the quality of the generated summaries. The evaluation results are then analyzed to determine the effectiveness of each model in producing accurate and contextually relevant news highlights.

2.1. Data Collection

Data collection phase involves acquiring the data required for the modeling process. In this research, the dataset comprises 10,414 Indonesian news articles sourced from Tempo.co, collected from September to October 2024. The collected data consists of news articles covering various topics and varying text lengths. The data collection process was carried out using the web scraping method with the BeautifulSoup. Each news article contains news title, source URL, publication date, news content, and summary (highlight).

2.2. Preprocessing

After data collection, several preprocessing steps are applied to ensure data quality before training the model. The main steps include:

1) Drop Column: Irrelevant columns are removed to align the dataset with the model's requirements. The initial dataset consists of five columns: title, released (publication date), url, summary (news highlight), and content (full news text). However, only the summary and content columns are relevant for the model. The summary column serves as the target label, representing the expected output of the model, while the content column serves as the input or source text to be summarized.
2) Drop Duplicate: Duplicate entries are removed to enhance data quality and prevent the model from learning redundant patterns. Since scraped datasets often contain repeated records, a drop duplicate process is applied to eliminate these entries. From a total of 10,414 records, 4 duplicate entries are identified and removed, resulting in a final dataset of 10,410 unique entries.
3) Text Cleaning: Text cleaning is conducted to eliminate unnecessary elements that may introduce noise into the data. This process removes elements such as *"TEMPO.CO, "*, *"Pilihan Editor"*, "*Baca berita selengkapnya disini"*, non-ASCII characters like "Â", and excessive spaces. Removing these elements ensures that the dataset becomes more refined and appropriate for training, which in turn boosts the model's effectiveness in capturing meaningful information.
4) Data Splitting: An 80:10:10 ratio was used to split the dataset into training, validation, and test subsets, resulting in 8,328 samples for training, and 1,041 samples each for validation and testing.
5) Tokenization: Tokenization is applied to convert text into smaller units called tokens, that are easier for the model to process [15]. Each model

used in this study follows a different tokenization method. The IndoT5 model utilizes T5Tokenizer which is based on SentencePiece, while the mBART model employs mBARTTokenizer, which is based on Byte Pair Encoding (BPE). Tokenization ensures that each model processes the text efficiently and aligns with the pre-trained tokenizer architecture, allowing for optimal performance during training and inference.

The examples result of preprocessing steps including text cleaning and tokenization are shown in Table 1.

Table 1. Preprocessing Steps

Content	Text Cleaning	IndoT5 Tokenization	mBART Tokenization
TEMPO.CO, Jakarta -Â Menteri Pertahanan atau Menhan Prabowo Subianto mengakui bahwa cita-cita Indonesia memiliki pertahanan yang kuat masih belum tercapai. Hal itu disampaikan saat dia menghadiri...	*Menteri Pertahanan atau Menhan Prabowo Subianto mengakui bahwa cita-cita Indonesia memiliki pertahanan yang kuat masih belum tercapai. Hal itu disampaikan saat dia menghadiri...*	[82, 1687, 46, 1211, 1064, 4004, 15386, 2117, 883, 6836, 7, 4997, 100, 164, 1969, 17, 1175, 145, 538, 7913, 3, 663, 24, 18, 13, 14313, 633, 2003, 18843, 1717, 141, 8840, 166, 373, …]	[250004, 19143, 908, 47284, 66, 1166, 1111, 1121, 4061, 84201, 31, 8273, 14, 12806, 121241, 4238, 15869, 9, 30483, 3799, 5309, 178944, 119, 23565, 4315, 9558, 226306, 5, 4772, 752, 879, 174096, …]

2.3. Fine-tuning

The pre-trained IndoT5 and mBART50 models were fine-tuned using the preprocessed data to generate abstractive news highlights in Indonesian. Fine-tuning is a critical step that adapts a general pre-trained model to the specific target task by optimizing its parameters based on the given dataset [16]. To ensure consistency, the input token length was limited to 512, while the output was constrained to a maximum of 150 tokens. Training was performed over 20 epochs with an early stopping mechanism using a patience of 3. The patience parameter defines how many consecutive epochs without improvement in validation loss are tolerated before stopping training to prevent overfitting. Validation loss is evaluated at the end of each epoch, and training stops if no improvement is observed for 3 consecutive epochs. This strategy helps to halt training once the model stops improving on validation data, thereby improving generalization and saving computational resources.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

2.4. Evaluation

After the fine-tuning process, the performance of each model was evaluated using the test set to assess their capability in generating accurate and informative news highlights. The evaluation employed two metrics: ROUGE and BERT-Score. ROUGE measures the overlap of n-grams between generated summaries and reference summaries [17], providing an indication of lexical similarity. BERT-Score, on the other hand, computes token-level similarity using contextual embeddings from BERT, capturing semantic similarity.

2.5. T5 Based Indonesian Summarization (IndoT5)

IndoT5 is a transformer-based model tailored for summarizing Indonesian text. Developed by Cahya Wirawan, IndoT5 builds upon the foundational architecture of T5 (Text-to-Text Transfer Transformer) as presented in Figure 2, where all natural language tasks are formulated in a unified text-to-text format. The T5 model is built using an encoder-decoder transformer framework inspired by the architecture proposed by Vaswani et al [18]. The pre-trained IndoT5 model applied in this research is publicly available on Hugging Face at https://huggingface.co/cahya/t5-base-indonesian-summarization-cased.

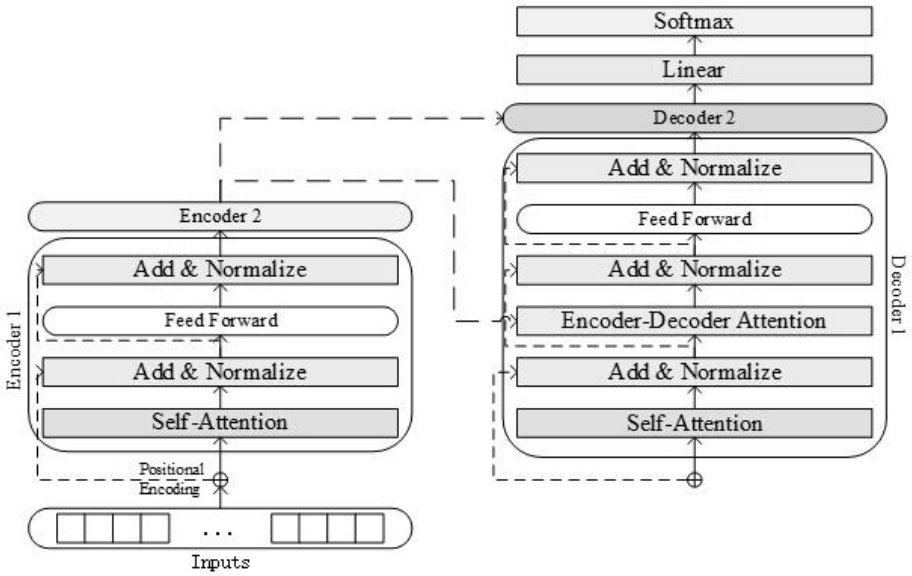

Figure 2. T5 Architecture

IndoT5 utilizes a unidirectional encoder, meaning that each token can only attend to preceding tokens during processing. The decoder is autoregressive, generating output one token at a time, with each generated token serving as input for the

subsequent prediction step until the entire summary is completed. The model is fine-tuned from the pre-trained checkpoint t5-base-bahasa-summarization-cased, originally developed by Husein Zolkepli. IndoT5 was further trained on a large-scale Indonesian news summarization dataset, particularly the id_liputan6 corpus, which comprises approximately 215,827 article-summary pairs [19].

2.6. mBART50 (Multilingual Bidirectional and Autoregressive Transformer)

mBART50 is a multilingual sequence-to-sequence transformer model introduced by Facebook AI in 2020, as an extension of the original mBART architecture. mBART50 adopts the core design of BART (Bidirectional and Auto-Regressive Transformer), comprising a bidirectional encoder paired with a decoder that operates autoregressively, as shown in Figure 3 [20]. The encoder is capable of capturing rich contextual representations by attending to both preceding and succeeding tokens, while the decoder generates output sequentially, one token at a time, ensuring syntactic and semantic coherence.

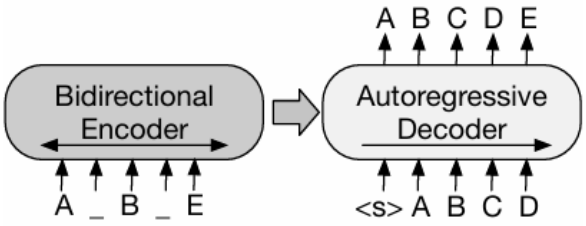

Figure 3. BART Mechanism

mBART50 is designed to support 50 languages, including low-resource languages like Indonesian. It requires an explicit language identifier token at the beginning of the input sequence to specify the target language for text generation. This multilingual capability allows mBART50 to perform various cross-lingual and generative tasks across diverse language pairs. The model was pre-trained using a denoising autoencoder objective on the CC25 dataset, which comprises monolingual corpora in 25 languages collected from Common Crawl. Pre-training involved two types of noise: random span masking and sentence permutation, enabling the model to learn robust linguistic representations. mBART50 expands upon the original mBART by incorporating an additional 25 languages using randomly initialized embedding layers. Further training was conducted using a combination of monolingual data from both CC25 and XLM-R corpora, allowing for broader linguistic coverage without necessitating training from scratch [21]. The pre-trained version of mBART50 used in this study is available at https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt.

2.7. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE serves as a standard metric in text summarization that quantifies the degree of overlap between system-generated summaries and human-written references [22]. ROUGE-N and ROUGE-L are the main variants of the ROUGE evaluation metric. ROUGE-N evaluates the similarity through n-grams, typically with ROUGE-1 and ROUGE-2 [23]. ROUGE-1 evaluates the overlap of unigrams between the generated summary and the reference, reflecting the level of informativeness and lexical accuracy. ROUGE-2 considers consecutive word pairs (bigrams), providing a better view of fluency and phrase-level coherence. Equation (1) presents the method used to compute ROUGE-N.

$$ROUGE\text{-}N = \frac{p}{q} \tag{1}$$

p refers to the overlapping n-grams, while q indicates the full count of n-grams within the reference text.

ROUGE-L measures summary by identifying the longest sequence of words in the same order shared by the two summaries [15]. This variant captures structural similarity and is effective for identifying logical consistency in the text. The ROUGE-L score is computed using Equation (2).

$$ROUGE\text{-}L = \frac{LCS}{m} \tag{2}$$

LCS refers to the length of the longest matching word sequence appearing in both summaries, and m represents the total word count of the reference summary.

2.8. BERT-Score

BERT-Score is a semantic evaluation metric that compares summaries based on contextual embeddings rather than surface-level token matches. Introduced by Zhang et al. (2019), BERT-Score utilizes pre-trained models like BERT to measure token-level similarity between generated summaries and their references. BERT-Score measures semantic similarity by comparing the meaning of words in context [24]. This allows it to capture nuanced relationships between texts, even when they use different surface forms.

In BERT-Score approach, each token in both the candidate and reference summaries is mapped into a high-dimensional vector representations through a language model like BERT. Then, for every token in the candidate summary, the cosine similarity with all tokens in the reference summary is calculated, and the

highest similarity score is selected. The cosine similarity score computed using Equation (3).

$$Cosine\ Similarity\ \left(x_i, \hat{x_j}\right) = \frac{x_i \cdot \hat{x_j}}{\|x_i\| \|\hat{x_j}\|} \quad (3)$$

Cosine similarity is computed by taking the dot product of the two embeddings and dividing it by the combined magnitudes of their vectors. In BERT-Score, a greedy matching mechanism is applied to find the best token pairs based on the highest cosine similarity scores. These maximum similarity values are then used to calculate recall, precision, and F1-score, which collectively indicate how semantically aligned the generated summary with its reference.

3. RESULTS AND DISCUSSION

3.1. Experimental Performance

The trained model was evaluated using test set to measure its performance in generate new summarization. The evaluation metrics were used ROUGE which measure lexical overlap, and BERT-Score which measures semantic similarity.

Table 2. Evaluation Result

Evaluation Metric	IndoT5	mBART
ROUGE-1	0.43087	0.42407
ROUGE-2	0.29143	0.28241
ROUGE-L	0.39224	0.38831
BERT-Score Precision	0.86362	0.86717
BERT-Score Recall	0.89130	0.88615
BERT-Score F1-Score	0.87708	0.87642

Table 2 presents a comparison of the evaluation results of IndoT5 and mBART models based on ROUGE and BERT-Score metrics. Overall, the results show that IndoT5 performs slightly better than mBART, although both models demonstrate relatively competitive performance. IndoT5 obtained higher ROUGE scores compared to mBART in all three ROUGE variants. This indicates that the summaries produced by IndoT5 tend to be more lexically similar to the reference summaries, across unigram, bigram, and longest common subsequence levels. In other words, IndoT5 tends to produce summaries that are closer to the structure and original words in the reference summary. Additionally, IndoT5 outperforms mBART in both recall and F1-score under the BERTScore metric. The higher recall indicates that IndoT5 is more effective producing summaries that are greater completeness and relevance to the source text, although they may be less concise compared to mBART.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

mBART obtained lower ROUGE scores compared to IndoT5, which may indicate that the model produces more abstractive summaries by paraphrasing or restructuring information in a form different from the reference text. mBART excels in precision scores on the BERT-Score metric. This suggests that mBART is more selective in choosing words in its summaries. mBART excels in ensuring that each generated word has a higher semantic similarity with the reference text, although the amount of information produced may be less compared to IndoT5. To further analyze model performance, qualitative comparisons were conducted using selected examples from the test set. The results were analyzed comparatively to evaluate each model's ability to generate news highlights in terms of completeness, clarity, and relevance.

Table 3. Comparison of Summary Results between IndoT5 and mBART Models

Example Result 1	
Input Text	*Garam adalah sumber natrium eksternal yang dibutuhkan tubuh dalam sehari-hari. Namun, garam berlebihan dalam tubuh dapat meningkatkan tekanan darah yang menjadi faktor risiko penyakit. Menurut Alison Brown, direktur program di Program Pencegahan dan Ilmu Kependudukan di Divisi Ilmu Kardiovaskular NHLBI, tidak mudah bagi sebagian besar orang menghindari makanan tinggi natrium. Sebab, garam terkandung dalam sebagian besar makanan sehingga seseorang memiliki potensi mengalami penyakit jantung dan stroke. Akibatnya, perlu beberapa cara ketat untuk mengurangi konsumsi garam berlebihan dalam tubuh. Lantas, bagaimana cara mengurangi garam berlebihan dalam tubuh? 1. Menerapkan Pola Makan DASH (Dietary Approaches to Stop Hypertension Brown) mengatakan, harus menentukan berapa banyak garam yang dimakan setiap hari dan menemukan sumber tinggi kandungan ini dalam makanan. Ahli diet terdaftar, Julia Zumpano, mengatakan, jika mencatat jumlah garam yang dimakan selama dua hari dalam seminggu, akan mendapatkan rata-rata...*
Reference from Tempo	*Jika mengalami tanda-tanda garam berlebihan dalam tubuh, terapkan cara berikut ini untuk menguranginya.*
IndoT5 Summary	*Perlu beberapa cara ketat untuk mengurangi garam berlebihan dalam tubuh. Berikut ini caranya. Apa saja cara ini? Berikut penjelasannya, mulai dari pola makan DASH, diet asin, hingga pengganti garam dalam makanan.*
mBART Summary	*Tidak mudah bagi sebagian besar orang menghindari makanan tinggi natrium. Berikut beberapa cara ketat untuk mengurangi konsumsi garam berlebihan dalam tubuh agar tidak menyebabkan penyakit jantung.*
Example Result 2	
Input Text	*Badan Pusat Statistik atau BPS melaporkan perekonomian Indonesia mengalami deflasi 0,12 persen secara bulanan pada September 2024. Dalam pemaparan Berita Resmi Statistik hari ini, disebutkan deflasi telah terjadi lima bulan beruntun sejak Mei. Deflasi 5 bulan berturut-turut ini menyerupai kondisi krisis, ujar Mohammad Faisal saat dihubungi, Selasa, 1 Oktober 2024 Pada kondisi normal, Indonesia dengan tingkat pertumbuhan ekonomi 5 persen seharusnya memang dapat menjinakkan inflasi. Inflasi rendah mestinya disebabkan kemampuan negara dalam mengendalikan harga-harga, bukan*

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

	Example Result 1
	pelemahan permintaan atau demand. Namun kini yang terjadi bukan hanya inflasi yang turun tapi malah deflasi bulanan beruntun. BPS melaporkan secara tahunan ekonomi Indonesia pada September telah mengalami inflasi 1,84 year on year (yoy)...
Reference from Tempo	*BPS mencatat Indonesia telah mengalami deflasi lima bulan berturut-turut yang menunjukkan terjadinya pelemahan daya beli konsumen.*
IndoT5 Summary	*BPS melaporkan perekonomian Indonesia mengalami deflasi 0,12 persen secara bulanan pada September 2024. Ini dampak dari lemahnya tingkat permintaan dan daya beli.*
mBART Summary	*Badan Pusat Statistik atau BPS melaporkan perekonomian Indonesia mengalami deflasi 0,12 persen secara bulanan pada September 2024, menurut pemaparan Berita Resmi Statistik hari ini.*
	Example Result 3
Input Text	*Duel Skotlandia vs Portugal akan tersaji pada pertandingan keempat UEFA Nations League 2024-2025. Laga ini akan berlangsung di Stadion Hampden Park, Selasa, 16 Oktober 2024, pada pukul 01.45 WIB. Nasib Portugal dan Skotlandia di UEFA Nations League A Grup 1 berbanding terbalik hingga tiga pertandingan yang telah mereka jalani. Selecao, julukan Portugal, masih sempurna dengan menyapu bersih kemenangan dalam tiga laga. Sedangkan Skotlandia selalu kalah dalam tiga pertandingan. Portugal memimpin klasemen Grup 1 dengan sembilan poin, diikuti Kroasia (6 poin) di peringkat kedua, Polandia (3) di posisi ketiga, sementara Skotlandia di dasar klasemen dengan 0 poin. Skotlandia gagal meraih kemenangan dalam lima laga terakhir. Setelah memperoleh skor 1-1 melawan Swiss di Euro 2024, The Tartan Army kemudian mengalami empat kekalahan beruntun. Pada pertemuan perdana melawan Portugal di UEFA Nations League musim ini, mereka kalah dengan skor 1-2. Saat laga terakhir, Skotlandia juga bertekuk lutut 1-2 di hadapan Kroasia. Sementara itu Portugal masih diliputi suasana positif...*
Reference from Tempo	*Duel Skotlandia vs Portugal akan tersaji pada pertandingan keempat UEFA Nations League 2024-2025. Mengapa Cristiano Ronaldo cs diunggulkan?*
IndoT5 Summary	*Duel Skotlandia vs Portugal akan tersaji pada pertandingan keempat UEFA Nations League 2024-2025. Posisi Diogo Costa sebagai penjaga gawang utama Selecao belum tergusur.*
mBART Summary	*Duel Skotlandia vs Portugal akan tersaji pada pertandingan keempat UEFA Nations League 2024-2025. Mengapa Tim Matador diunggulkan?*

Table 3 presents a comparison of news highlights generated by the IndoT5 and mBART models based on the test set. In the first example, the news discusses several recommended methods to reduce excessive salt intake in the body. IndoT5 successfully captures the main points by explicitly mentioning several key methods, such as the DASH diet, avoiding salty foods, and replacing salt-based seasonings, indicating its strength in covering a broader range of content. The summary also demonstrates abstraction capability by presenting a reorganized version of the original text, which reflect an attempt to synthesize multiple points cohesively. mBART focuses mainly on the difficulty of avoiding high-sodium foods and the

general need to reduce salt intake to prevent heart disease. While the output is more cohesive as a standalone paragraph, it lacks detail and omits most of the specific steps outlined in the original text. This indicates a narrower scope of information. Nonetheless, mBART demonstrates abstraction through its rephrased sentence structure and cause-effect linkage between salt intake and cardiovascular risk. This example suggests that IndoT5 is better at preserving the breadth of information, while mBART offers slightly better flow and narrative coherence.

The second example discusses Indonesia's current economic condition, particularly the occurrence of deflation. IndoT5 captures the core issue by mentioning the deflation rate of 0.12% and identifying weak consumer demand as a contributing factor. IndoT5's summary closely aligns with the reference from Tempo, which also highlights the deflation and its causes. In comparison, mBART focuses mainly on the BPS deflation report without explaining the causes, thus providing less insight into the overall issue. Additionally, the sentence structure in the mBART summary tends to resemble the original text without much rephrasing. Thus, IndoT5 is considered better at capturing the core of the problem and performing abstraction, although it still needs improvement.

The third example discusses the match between Scotland and Portugal in the UEFA Nations League 2024–2025. IndoT5 generates a highlight that accurately conveys the match context and includes additional engaging details, such as the condition of key player Diogo Costa. mBART produces a summary with a structure similar to the reference, but there is an information error by mentioning *"Tim Matador,"* which actually refers to Spain, not Portugal. This subject error indicates a failure in contextual understanding and abstraction, as the original article contains no reference to the Spanish team. Compared to mBART, IndoT5 has better performance to capture contextual meaning and key information.

3.2. Discussion

Evaluation results show that both the IndoT5 and mBART models have their respective strengths in performing the abstractive text summarization task. IndoT5 generally shows better performance, especially in understanding context, preserving information structure, and producing complete summaries. This is reflected in its higher ROUGE and BERT-Score Recall scores.

IndoT5's advantage can be attributed to its pretraining approach, which is tailored specifically for the Indonesian language. IndoT5 model was pre-trained on Liputan6 dataset, which containing 215.827 articles-summary pairs, allowing the model to understand sentence patterns, writing styles, and typical information structures in Indonesian-language news. These findings are consistent with prior

research by Itsnaini et al. (2023), which also demonstrated the superiority of IndoT5 in capture and retain relevant information in summarization tasks.

In addition, the evaluation results of IndoT5 in this study outperform those reported by Purnama & Utami (2023), who used the base T5 model for Indonesian summarization and achieved lower scores ROUGE-1 = 0.17866, ROUGE-2 = 0.16067, ROUGE-L = 0.16032. This contrast highlights the effectiveness of using IndoT5, a model specifically pre-trained on Indonesian corpora, compared to T5-base, which was not optimized for the language. Similarly, the mBART model in this study also achieved higher ROUGE scores compared to previous research by Astuti et al. (2024), which reported ROUGE-1 = 35.94, ROUGE-2 = 16.43, and ROUGE-L = 29.91 using mBART on a different dataset [25]. This reinforces mBART's potential as a competitive model for abstractive summarization in Indonesian, even though it still underperforms IndoT5 in terms of completeness and contextual relevance.

mBART, as a multilingual model, demonstrates advantages in information density and word-level precision, as reflected in the slightly higher BERT-Score Precision. mBART model tends to produce more concise and selective summaries. However, this tendency sometimes results in the loss of important details, making its summaries less complete. This is reflected in the lower ROUGE and BERT-Score Recall values compared to IndoT5. This limitation reflected in the lower ROUGE and BERT-Score Recall values compared to IndoT5.

Semantic errors were also found in the mBART summary, such as the use of the term "Matador Team" (which refers to the Spanish national football team) in the context of news about the Portuguese team. This indicates abstraction errors likely caused by insufficient understanding of the semantic and contextual nuances of the Indonesian language. The suboptimal performance of mBART in this task is likely caused by the limited amount of Indonesian training data during pretraining. Previous research by Cao et al. (2024) showed that mBART-50 is capable of generating abstractive and high-quality summaries for Mandarin news articles. The differences findings in this study can be explained by the amount of pretraining data used during mBART's pretraining phase. In the ML50 dataset used to train mBART, Mandarin is categorized as a high-resource language with over 42 million training samples, whereas Indonesian only has around 84,000 samples. This data imbalance significantly affects the model's capacity to learn and generalize language-specific structures. Languages with limited training data tend to yield less accurate summaries, as the model lacks sufficient examples to deeply understand the structure, vocabulary, and linguistic patterns to that language.

The models evaluated in this study demonstrated quite good performance however, still has several limitations. First, both IndoT5 and mBART tend to

generate summaries that are dominated by information from the initial paragraphs of the source text. This may be influenced by the inverted pyramid structure commonly used in news articles, where the most important information is placed at the beginning. In addition, the input length limitation also plays a role, as texts exceeding the maximum token limit are truncated and cannot be fully processed by the model. Second, the generated summaries are not fully abstractive. The model tends to retain many phrases from the source text, which is likely due to patterns in the training data that are not very different from the source text. Third, the evaluation used is solely based on ROUGE and BERTScore metrics, which cannot capture subjective aspects of summary quality. A more in-depth evaluation using human evaluation-based methods may be necessary for future research.

4. CONCLUSION

This study compares the performance of IndoT5 and mBART transformer-based models in generating abstractive Indonesian news highlights. Using a dataset of 10,410 articles from Tempo.co, both models were fine-tuned and evaluated using ROUGE and BERTScore metrics. IndoT5 achieved higher score in ROUGE, ROUGE-1 (0.43087), ROUGE-2 (0.29143), and ROUGE-L (0.39224), and outperformed mBART in BERTScore Recall (0.89130) and F1 (0.87708), showing stronger ability to produce complete and semantically accurate summaries. mBART, with a slightly higher BERTScore Precision (0.86717), was more concise but less comprehensive. These findings highlight IndoT5's effectiveness in summarizing Indonesian news, particularly in generating accurate and informative highlights, suggesting its strong potential for integration into journalistic tools, such as automatic news highlight generators. In contrast, mBART, while more abstractive and concise, occasionally generated semantically inconsistent outputs due to limited Indonesian training data. Nevertheless, mBART as a multilingual model, demonstrates potential for applications involving cross-lingual or multilingual news summarization, especially in producing short headline-style summaries across different languages. Future research is encouraged to incorporate human evaluation, and explore multi-domain datasets. Additionally, using training data with higher abstraction levels and more diverse expressions is recommended to help models generate more abstractive and generalized summaries.

REFERENCES

[1] Q. A. Itsnaini, M. Hayaty, A. D. Putra, and N. A. M. Jabari, "Abstractive Text Summarization using Pre-Trained Language Model 'Text-to-Text Transfer Transformer (T5)," *ILKOM Jurnal Ilmiah*, vol. 15, no. 1, pp. 124–131, Apr. 2023, doi: 10.33096/ilkom.v15i1.1532.124-131.

[2] Syinchan, "Teknik Skimming untuk Pembaca yang Sering 'Zoning Out,'" https://www.kompasiana.com/voyageonyx3449/67232577c925c4404c4eddf3/teknik-skimming-untuk-pembaca-yang-sering-zoning-out?page=1&page_images=1.

[3] A. R. Lubis *et al.*, "Enhancing Text Summarization with a T5 Model and Bayesian Optimization," *Revue d'Intelligence Artificielle*, vol. 37, no. 5, pp. 1213–1219, 2023, doi: 10.18280/ria.370513.

[4] N. Giarelis, C. Mastrokostas, and N. Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," Jul. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app13137620.

[5] A. N. Khasanah And M. Hayaty, "Abstractive-Based Automatic Text Summarization On Indonesian News Using GPT-2," *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, vol. 10, no. 1, pp. 9–18, Dec. 2023, doi: 10.33330/jurteksi.v10i1.2492.

[6] A. S. Wijaya and A. S. Girsang, "Augmented-Based Indonesian Abstractive Text Summarization using Pre-Trained Model mT5," *International Journal of Engineering Trends and Technology*, vol. 71, no. 11, pp. 190–200, 2023, doi: 10.14445/22315381/IJETT-V71I11P220.

[7] E. S. Lim *et al.*, "ICON: Building a Large-Scale Benchmark Constituency Treebank for the Indonesian Language," *Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories*, pp. 37–53, 2023.

[8] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 757–770, Dec. 2020.

[9] R. Adelia, S. Suyanto, and U. N. Wisesty, "Indonesian abstractive text summarization using bidirectional gated recurrent unit," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 581–588. doi: 10.1016/j.procs.2019.09.017.

[10] C. P. R. Ardyanti, Y. Wibisono, and R. Megasari, "Peringkasan Teks berita Berbahasa Indonesia Menggunakan LSTM dan Transformer," 2024.

[11] I. N. Purnama and W. Utami, "Implementasi Peringkas Dokumen Berbahasa Indonesia Menggunakan Metode Text To Text Transfer Transformer (T5)," 2023.

[12] D. Dharrao, M. Mishra, A. Kazi, M. Pangavhane, P. Pise, and A. M. Bongale, "Summarizing Business News: Evaluating BART, T5, and PEGASUS for Effective Information Extraction," *Revue d'Intelligence Artificielle*, vol. 38, no. 3, pp. 847–855, Jun. 2024, doi: 10.18280/ria.380311.

[13] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," Oct. 2019.

[14] K. Cao *et al.*, "DMSeqNet-mBART: A State-of-the-Art Adaptive-DropMessage Enhanced mBART Architecture for Superior Chinese Short News Text Summarization," May 03, 2024. doi: 10.36227/techrxiv.171470744.49740569/v1.
[15] G. Hartawan, D. Sa'adillah Maylawati, and W. Uriawan, "Bidirectional And Auto-Regressive Transformer (BART) For Indonesian Abstractive Text Summarization," 2024.
[16] B. Hanindhito, B. Patel, and L. K. John, "Large Language Model Fine-tuning with Low-Rank Adaptation: A Performance Exploration," in *Proceedings of the 16th ACM/SPEC International Conference on Performance Engineering*, New York, NY, USA: ACM, May 2025, pp. 92–104. doi: 10.1145/3676151.3719377.
[17] Y. M. Wazery, M. E. Saleh, A. Alharbi, and A. A. Ali, "Abstractive Arabic Text Summarization Based on Deep Learning," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/1566890.
[18] A. Vaswani *et al.*, "Attention Is All You Need," 2017.
[19] F. Koto, J. H. Lau, and T. Baldwin, "Liputan6: A Large-scale Indonesian Dataset for Text Summarization," Nov. 2020.
[20] Y. Liu *et al.*, "Multilingual Denoising Pre-training for Neural Machine Translation," Jan. 2020.
[21] Y. Tang *et al.*, "Multilingual Translation with Extensible Multilingual Pretraining and Finetuning," Aug. 2020.
[22] K. Ganesan, "ROUGE2.0: Updated and Improved Measures for Evaluation of Summarization Tasks," vol. 1, no. 1, Mar. 2018.
[23] B. Ay, F. Ertam, G. Fidan, and G. Aydin, "Turkish abstractive text document summarization using text to text transfer transformer," *Alexandria Engineering Journal*, vol. 68, pp. 1–13, Apr. 2023, doi: 10.1016/j.aej.2023.01.008.
[24] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BertScore: Evaluating Text Generation With BERT," Feb. 2020.
[25] R. H. Astuti, M. Muljono, and S. Sutriawan, "Indonesian News Text Summarization Using MBART Algorithm," *Scientific Journal of Informatics*, vol. 11, no. 1, pp. 155–164, Feb. 2024, doi: 10.15294/sji.v11i1.49224.