JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 10, No. 1, JANUARY, 2025, Pp. 110 – 122
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Enhancing Abstractive Multi-Document Summarization with Bert2Bert Model for Indonesian Language

Aldi Fahluzi Muharam [(1)], Yana Adita Gerhana [(2)], Dian Sa'adillah Maylawati [(3)*], Muhammad Ali Ramdhani [(4)], Titik Khawa Abdul Rahman [(5)]
[1,2,3,4] Department of Informatics, Faculty of Science and Technology, UIN Sunan Gunung Djati, Bandung, Indonesia
[5] Information and Communication Technology, Asia e University, Selangor, Malaysia
e-mail : 1207050008@student.uinsgd.ac.id,
{yanagerhana,diansm,m_ali_ramdhani}@uinsgd.ac.id, titik.khawa@aeu.edu.my.
* Corresponding author.
This article was submitted on 15 September 2024, revised on 19 Desember 2024, accepted on 28 Desember 2024, and published on 31 Januari 2025.

Abstract

This study investigates the effectiveness of the proposed Bert2Bert and Bert2Bert+Xtreme models in improving abstract multi-document summarization for Indonesians. This research uses the transformer model to develop the proposed Bert2Bert and Bert2Bert+Xtreme models. This research uses the Liputan6 data set which contains news data along with summary references for 10 years from October 2000 to October 2010 and is commonly used in many automatic text summarization research. The model evaluation results using ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore show that the proposed model has a slight improvement over previous research models, with Bert2Bert being better than Bert2Bert+Xtreme. Despite the challenges posed by limited reference summaries for Indonesian documents, content-based analysis using readability metrics, including FKGL, GFI, and Dwiyanto Djoko Pranowo, revealed that the summaries produced by Bert2Bert and Bert2Bert+Xtreme are at a moderate readability level, meaning they are suitable for mature readers and aligns with the news portal's target audience.

Keywords: Bert2Bert, Abstractive, Multi-document, Summarization, Transformer

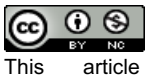
Abstrak

Studi ini menyelidiki efektivitas model Bert2Bert dan Bert2Bert+Xtreme yang diusulkan dalam meningkatkan peringkasan multi dokumen yang abstrak untuk bahasa Indonesia. Penelitian ini menggunakan model transformer sebagai dasar untuk mengembangkan model Bert2Bert dan Bert2Bert+Xtreme yang diusulkan. Kumpulan data Liputan6 yang berisikan data berita beserta referensi ringkasannya dengan periode 10 tahun mulai dari Oktober 2000 sampai Oktober 2010 digunakan dalam penelitian ini. Hasil evaluasi model menggunakan ROUGE-1, ROUGE-2, ROUGE-L, dan BERTScore menunjukkan bahwa model yang diusulkan memiliki sedikit peningkatan terhadap model penelitian terdahulu dengan Bert2Bert lebih baik daripada Bert2Bert+Xtreme. Meskipun tantangan yang ditimbulkan oleh ringkasan referensi terbatas untuk dokumen-dokumen Indonesia, analisis berbasis konten menggunakan metrik keterbacaan termasuk FKGL, GFI, dan Dwiyanto Djoko Pranowo mengungkapkan bahwa ringkasan yang dihasilkan oleh Bert2Bert dan Bert2Bert+Xtreme berada pada tingkat keterbacaan sedang, yang berarti cocok untuk pembaca dewasa dan selaras dengan target audiens portal berita.

Kata Kunci: Bert2Bert, Abstraktif, Multi-dokumen, Peringkasan, Transformer

1. INTRODUCTION

Natural language processing applications are being developed using artificial intelligence, which allows computers to have natural language processing capabilities like humans (Alquliti & Binti, 2019), one of which is automatic text summarization. With the development of the information age, the need for content processing and summarization is increasing. Therefore, there is a need for a document summarization tool that can automatically summarize information from various sources. Document summarization tools aim to generate relevant and appropriate text summaries for a given document or set (Jin & Wan, 2020a). There are two types of document summarization:



abstractive and extractive (Jin & Wan, 2020a). Extractive summarization methods focus on identifying and compiling key sentences from the source text to form a summary. This approach benefits from simpler implementation and direct utilization of text from the original document (Kuyate et al., 2023).

Meanwhile, abstractive summarization generates new sentences, often requiring complex models to paraphrase and condense the original text's meaning, which can lead to summaries that are more fluent and less redundant (Dangol et al., 2023). Extractive summarization directly retrieves meaningful sentences the model has selected from the original document (Goldstein et al., 2000; Shinde et al., 2022). Summaries written by humans are generally abstractive. Abstractive summarization allows computers to generate text summaries by creating a new set of sentences that represent the information contained in the source with a different form of presentation from the original text (Bing et al., 2015; Jin & Wan, 2020a; Li & Zhuge, 2021).

Every language has its uniqueness and characteristics, including Indonesian. Indonesian text summarization has seen significant advancements through the application of transformer models, addressing the challenges of processing a language spoken by nearly 200 million people but under-represented in NLP research (Devi & Suadaa, 2022). The existence of research on the development of Indonesian language data sets as an evaluation benchmark in the development of automatic text summarization tools opens opportunities for further research. IndoSum is a data set used as a new benchmark in Indonesian text summarization (Kurniawan & Louvan, 2018). Then, Liputan6 is used as a large-scale data collection tool for automatic text summarization (Koto, Rahimi, et al., 2020). Liputan6 was collected from data from an Indonesian language news portal, Liputan6.com, over ten years, so there were 215,827 text summary data documents (Koto, Rahimi, et al., 2020). In his research, a BERT-based single document summarization model was also developed using extractive and abstractive methods. MultilingualBERT and IndoBERT are used as pre-trained models in this model. Using IndoBERT produces good evaluation values for using the Liputan6 data set (Koto, Rahimi, et al., 2020). Then, there is also a comprehensive data set, Indonesia Language Evaluation Montage (IndoLEM), which includes seven NLP tasks and eight sub-data sets (Koto, Rahimi, et al., 2020). Other extractive text summarization models for the Indonesian language, such as the fine-tuning of Sentence Transformers (SBERT), have demonstrated improved performance in generating document summaries or snippets, particularly using Indonesian thesis documents to construct a new dataset for this task (Abka et al., 2022).

Based on the development of research on automatic text summarization for Indonesian, it was found that abstractive multi-document summarization in Indonesian research has not been carried out using a Transformer-based model, and it is a great opportunity to explore. This is due to the limited data sets that provide multi-document summary data (Jin & Wan, 2020a; J. Zhang et al., 2018b). In cases other than Indonesian, there is research on fine-tuning using the BERT Sentence Embedding Model on extractive summarization multi-document data sets (Lamsiyah et al., 2023). Then, there is also research by fine-tuning the Transformers model, which has been trained for single-document summarization of multi-document data sets by adjusting the Encoder and Decoder structure (Shen et al., 2023). Other research also adapts the Transformer model trained for single-document summarization to perform multi-document summarization tasks using the Decoding Controller (Jin & Wan, 2020a). All three studies used pre-trained models that were fine-tuned using multi-document data. From these researches, it emerged that the single-document summarization model has the potential to perform multi-document summarization tasks. Therefore, this study contributes to using a model trained with a single document summarization data set in an Indonesian multi-document summarization task with the Transformers model.

2. METHODS

2.1 Datasets

This research uses the Liputan6 data set in the training and evaluation process because the data set is 11 times larger and more abstract than the IndoSum data set (Koto, Rahimi, et al., 2020; Lucky & Suhartono, 2021). There are 193,883 training data and 10,972 data for development and

testing, respectively (Koto, Rahimi, et al., 2020). Each document has data for article text, extractive summary text, and abstractive summary text. The development and testing data have a more significant percentage of novel n-grams than the training data. Apart from the Canonical variant, there is an Xtreme variant in development and testing data, namely a data set that only contains more than 90% novel 4-grams. Hence, the data is more abstractive and produces a smaller data set. In the Liputan6 training data set, the number of words in article documents ranges from the smallest, 31 words, to the largest, 6,570 words, with an average of 195.74 words, a mode of 121 words, and a median of 163 words. Meanwhile, in reference documents, the abstractive summaries range from the smallest being 11 words to the largest being 80 words, with an average of 27.08 words, a mode of 27 words, and a median of 27 words.

The Canonical variant development data set on article text documents has the smallest 64 words and the largest, 1,567 words, with an average of 190.1 words, a mode of 141 words, and a median of 166 words. In the abstract summary document, the number of words ranges from 9 to 36, with an average of 21.88 words, a mode of 23, and a median of 22. For the Xtreme variant development data set, the smallest number of article text document words is 66 words, and the largest is 1567 words, with an average of 196.55 words, a mode of 141 words, and a median of 168 words. The number of words in abstractive summary text documents consists of the smallest 11 words, the largest 36 words, the average 21.16 words, the mode 22 words, and the median 21 words.

Meanwhile, in the Canonical variant test data set, the number of words in the article text document ranges from the smallest 62 words to the largest, 3,064 words, with an average of 181.53 words, a mode of 150 words, and a median of 158 words. The abstractive summary text ranges from 12 words to 35 words with an average of 23.18 words, a mode of 24 words, and a median of 23 words. In the Xtreme variant, the number of words in the article text document ranges from 63 to 3064 words with an average of 194.81 words, a mode of 132 words, and a median of 161 words. For abstractive summary text, the range is from 12 words to 33 words, with an average of 22.62 words, a mode of 22 words, and a median of 23 words.

2.2 Bert2Bert Model

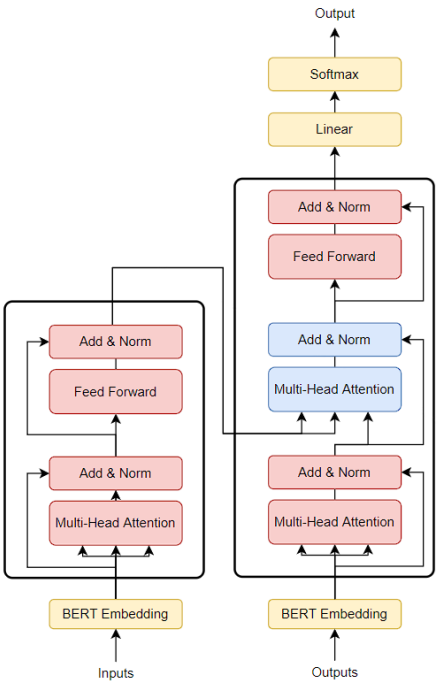


Figure 1 Transformer Model with IndoBERT Implementation

Transformer models thrive on automatic text summarization as NLP tasks. Transformer-based models are one type of model that can be used with the sequence-to-sequence learning (seq2seq) method. Using seq2seq-based models shows promising results on single document summarization tasks (Jin & Wan, 2020a; Koto, Rahimi, et al., 2020; J. Zhang et al., 2018b). The seq2seq model is a model that can be trained using varying input dimensions and with varying output dimensions as well (Sutskever et al., 2014). IndoBERT is applied as an encoder and decoder transformer with random or non-pretrained parameters to develop a single document summarization model. This research implements IndoBERT as an Encoder and Decoder using the Huggingface framework. BERT is a model designed to train in-depth two-way representations of unlabeled text by considering context from left and right at all layers (Devlin et al., 2018). The BERT model architecture represents the Encoder layer in a Transformer. The Encoder layer in the Transformer consists of a multi-head self-attention layer and a position-wise fully connected feed-forward network layer (Vaswani et al., 2017). Meanwhile, the Decoder layer in the Transformer is like the Encoder layer but modified by masking and adding a third layer, the multi-head cross-attention layer, which receives input from the Encoder output. Masking the Decoder layer, the multi-head self-attention layer that was originally bidirectional become unidirectional.

Figure 1 shows the architecture of the transformer model where the blue layer shows the parameter weights initiated randomly. In contrast, the red layer shows the parameter weights using the weights from the pre-trained BERT model. In the process of implementing BERT at the Encoder layer, there are no problems because the structure is the same. However, several adjustments are made to the model layer structure in the BERT implementation at the Decoder layer. In this process, a cross-attention layer is added between the self-attention and feed-forward layers (Rothe et al., 2020). The parameter weights in this additional layer are initiated randomly. Changes to the self-attention layer from bi-directional to unidirectional were also carried out without changing the parameter weights of that layer. Then, the LM (Language Model) Head layer was added as the final layer to define the conditional probability distribution of the output sequence. The parameter weights of this layer use the word embedding weights in the BERT Embedding layer.

The tokenizer from IndoBERT is used by setting the maximum Encoder token length to 512, the maximum Decoder token length to 256, and by adjusting the decoder_start_token_id with bos_token_id from the tokenizer. In the model settings also, adjustments are made to eos_token_id with eos_token_id from the tokenizer, pad_token_id with pad_token_id on the tokenizer, and vocab_size with encoder_vocab_size. Then, in the generative configuration, the maximum length is set to 80, the minimum length is 10, num_beams is 10, length_penalty is 2, no_repeat_n_gram_size is 3, and early_stopping is set to true. Adam optimization is used with a learning rate of 5×10-5 and lr_scheduler_type linear.

Two training methods are used in model development. In the first scenario, the model is trained with 8 epochs using 193,883 training data and a batch size of 18. Then, in the second scenario, the first model trained in the first scenario is trained again using the abstractive Xtreme variant development data set with 5 epochs and a batch size of 10. This data set has more than 90% novel 4-grams in each document, so it is assumed to produce a model that works more abstractive.

2.3 Evaluation Methods

The evaluation of model capabilities is divided into two categories: co-selection-based analysis and content-based analysis (Maylawati et al., 2024). Co-selection-based analysis compares the summary results against the reference summary. Meanwhile, the content-based analysis assesses the readability of summary results through sentence linkages without requiring a reference summary.

The co-selection-based evaluation process was carried out using ROUGE and BERTScore using single-document data from the Liputan6 data test. BERTScore is a text generation evaluation metric using context-based token representation to measure the similarity of text to its reference

(T. Zhang et al., 2019). The evaluation process is carried out using the F1 value from BERTScore. Then ROUGE or Recall-Oriented Understudy for Gisting Evaluation is a tool for measuring summarization results by comparing other ideal human summaries (Lin, 2004). This research uses three F1 values, namely ROUGE-1, ROUGE-2, and ROUGE-L, to measure the quality of the summarization results. ROUGE-1 and ROUGE-2 measure using N-grams that intersect each other between the summarization results and the reference with length N of 1 and 2 (Lin, 2004; Maylawati et al., 2024). The ROUGE-L measurement is based on the Longest Common Subsequence (LCS) or the longest substitution (Lin, 2004; Maylawati et al., 2024). This concept considers sentence structure in identifying similarities to consider words that are not sequential but still have the same or similar meaning in the context of the sentence.

Content-based evaluation in multi-document summarization results uses Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), and Dwiyanto Djoko Pranowo metric values. FKGL is a metric for measuring a text's difficulty level in understanding based on the length of words and sentences (Solnyshkina et al., 2017). Like FKGL, GFI is a metric used to measure a text's difficulty level in reading (Goh et al., 2007; Świeczkowski & Kułacz, 2021). Among the FKGL and FGIeveloped in English, Dwiyanto Djoko Pranowo is the creator of special measuring tools in Indonesian texts (Biddinika et al., 2016). There are thirteen indicators to assess readability in the Dwiyanto metric. The thirteen indicators are the average number of paragraphs, the average number of sentences in each paragraph, sentence length, percentage of continuation sentences, percentage of compound sentences, percentage of number of polysemic sentences, percentage of passive sentences, percentage of foreign words, percentage of abstract words, percentage of terms, percentage conjunctions, and percentage of loanwords (Pranowo, 2011). The readability measure is obtained by adding up all these indicators. The range 13.0 to 21.7 is defined as easy, the range 21.8 to 30.5 is interpreted as moderate, and the range 30.6 to 39.0 is interpreted as difficult.

3. RESULTS AND DISCUSSION

3.1 Result of Fine-Tuning Model

Previous research using IndoBERT for automatic text summarization was carried out for single documents (Koto, Rahimi, et al., 2020). Therefore, to determine the performance of the proposed model, namely Bert2Bert, the experiment was carried out using a single document with the same dataset as previous research, namely Liputan6 data. This model can accept input with a maximum length of 512 tokens. With that, documents whose length exceeds 512 after tokenization will have the excess removed. The model was then tested using Liputan6 test data, and the evaluation results are shown in Table 1. The evaluation was carried out using ROUGE and BERTScore.

Table 1 Evaluation result using ROUGE and BERTScore

Model	Canonical Test Set				Extreme Test Set			
	R1	R2	RL	BS	R1	R2	RL	BS
BERTABS (Koto, Lau, et al., 2020)	40.94	23.01	37.89	77.90	34.59	15.10	31.19	75.84
BERTEXTABS (Koto, Lau, et al., 2020)	41.08	22.85	38.01	77.93	34.84	15.03	31.40	75.99
Bert2Bert	41.13	22.85	34.65	73.30	34.86	15.29	27.89	72.54
Bert2Bert+Xtreme	39.21	20.04	32.51	72.66	34.45	14.56	27.37	72.10

In Table 1, the proposed model (Bert2Bert) has a higher R1 value in Canonical data testing, namely 41.13. It has the largest R1 and R2 values in the Xtreme data test, namely 34.86 and 15.29. Although the ROUGE and BERTScore results on Bert2Bert are like IndoBert, overall, Bert2Bert performs well like other models. This is indicated by the ROUGE and BERTScore values, which are not significantly different in the Canonical and Xtreme tests. Compared with the Bert2Bert+Xtreme model, another proposed model, the Bert2Bert model, performs better. This is shown in the ROUGE and BERTScore results, which are both better than those of

Bert2Bert+Xtreme. Table 2 shows examples of abstractive summary results produced by Bert2Bert and Bert2Bert+Xtreme.

Table 2 The Example of an Abstractive Summary with Bert2Bert

Reference Summary	Bert2Bert	Bert2Bert+Xtreme
Kendati Bank Sentral AS menurunkan suku bunganya, namun BI dinilai masih akan menemui masa sulit. Suku bunga Bank Sentral AS akan diturunkan menjadi empat persen. (Even though the US Central Bank has lowered its interest rate, BI is still considered to be facing difficult times. The US Central Bank interest rate will be reduced to four percent.)	menurut pengamat ekonomi didiek j. rachbini, bank indonesia akan menghadapi situasi sulit bila bi terus menurunkan tingkat suku bunga yang dimiliki.	bank indonesia dinilai masih akan menghadapi situasi sulit di tanah air. bahkan, tingkat suku bunga the fed akan diturunkan menjadi empat persen.
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 Regional Police Chief, Brigadier General Pol. Johny Yodjana, is determined to eradicate wood smuggling perpetrators in Riau. He promised to take firm action against the perpetrators without discrimination.)	polda riau bertekad memberantas para penyelundup kayu di riau. kapolda riau brigjen pol. johny yodjana berjanji tak akan pandang bulu.	kapolda riau brigjen pol. johny yodjana melantik kapolda baru di riau. selain itu, polri juga akan memberantas penipuan dana reboisasi dan iuran hasil hutan.

3.2 Result of Abstractive Multi-Document Summarization for Indonesian Language

The Bert2Bert and Bert2Bert+Extreme models have been proven to generate automatic abstractive summaries for single documents. Furthermore, this research applies the Bert2Bert and Bert2Bert+Extreme models for multi-document abstractive summarization in Indonesian using Transformers, which has never been done in previous research. Previous research conducts multi-document abstractive summarization without Transformers (Severina & Khodra, 2019). Then, most of the last research only focused on Indonesian abstractive summarization for single documents (Devianti & Khodra, 2019; Dewi & Widiastuti, 2022; Laksana et al., 2022; Sugiri et al., 2022; Wijayanti et al., 2021). Several previous multi-document studies also did not produce abstractive summarization but extractive summarization (D. Gunawan et al., 2019; Y. H. B. Gunawan & Khodra, 2021; Widjanarko et al., 2018). Table 3 and 4 shows the example of abstractive summarization results for Indonesian multi-documents with 2 Canonicals, 2 Xtremes, 3 Canonocals, and 3 Xtremes, where the bold means selected by Bert2Bert while underlined is selected by Bert2Bert+Xtreme.

3.3 Readability Evaluation

The limited number of summary references for multiple Indonesian documents makes conducting co-selection-based analysis evaluations such as ROUGE and BERTScore impossible. Therefore, the abstractive summary results of multiple Indonesian documents were evaluated using content-based analysis such as the readability metrics FKGL, GFI, and Dwiyanto Djoko Pranowo. Readability of summary results is a big challenge in automatic text summarization research, which is currently the focus of researchers (Maylawati, 2019; Verma et al., 2019; Verma & Om, 2019). Most automatic text summarization research involves humans evaluating the summary results, but more is needed on readability. As one of the contributions to this research, this section presents the results of multi-document abstractive summaries with Bert2Bert and Bert2Bert+Extreme along with the results of readability evaluation using FKGL, GFI, and Dwiyanto Djoko Pranowo metrics.

Figures 2 (a) and (b) show the FKGL and GFI evaluation results of Bert2Bert and Bert2Bert+Xtreme with Canonical and Xtreme data. Based on the FKGL evaluation results, Bert2Bert and Bert2Bert+Xtreme have a readability level of more than 18 for FKGL, which means

the resulting text is difficult to read or understand. The FKGL value indicates a text's readability based on the reader's age grade. For FKGL scores, 0-6 are categorized as easy, 6-12 as average, 12-18 as poor reading skills, and above 18 are considered difficult to read (Maylawati et al., 2024; Scott, 2024; Solnyshkina et al., 2017). The optimal value of GFI for the text that is categorized as readable is in the range of 7 to 8 (Maylawati et al., 2024; Scott, 2025; Świeczkowski & Kułacz, 2021). However, the average GFI value is 9-10, which means that readability is still acceptable because, according to GFI, texts that are very difficult to read have a value of more than 12. However, overall, the FKGL and GFI results cannot be categorized as hard to read to the adult age category because the data source is news portals whose readers are adults. So, the FKGL and GFI evaluation results can be accepted by news readers who are mostly adults.

Table 3 The Example of an Abstractive Summary with Bert2Bert

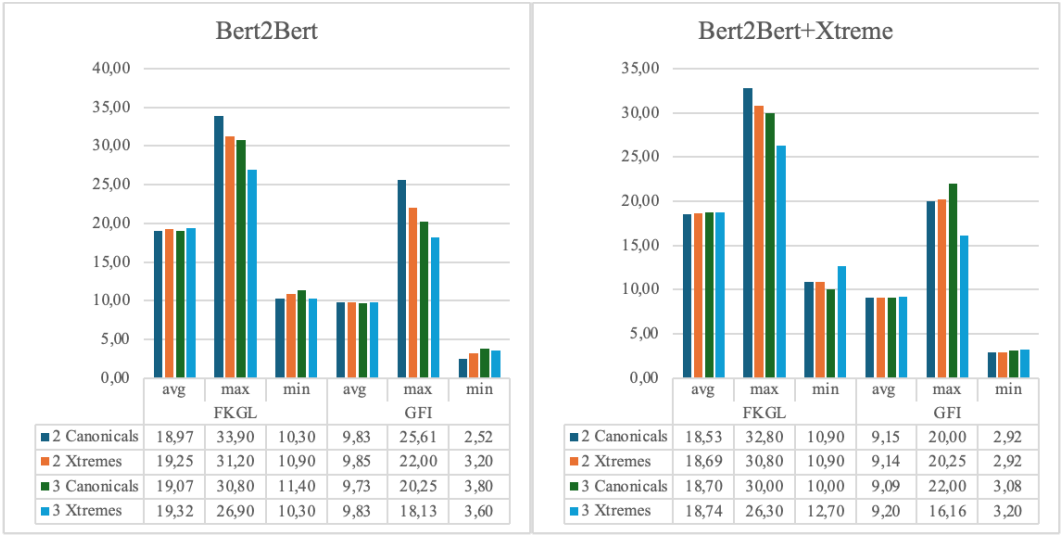
Type	Articles	Bert2Bert Summary	Bert2Bert+Extreme Summary
2 Canonicals	[Doc 1] Liputan6.com, Jakarta: **Kepolisian Daerah Riau bertekad memberantas pelaku penyelundupan kayu yang kerap terjadi di Riau.** Selain itu, Polda setempat juga akan memberangus menipulasi dana reboisasi dan iuran hasil hutan. 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. (ICH/Edi Priyono dan Andi Azril). [Doc 2] Liputan6.com, Jakarta: **Bank Indonesia dinilai masih akan menghadapi situasi sulit kendati Bank Sentral Amerika Serikat (The FED) terus menurunkan tingkat suku bunga yang dimiliki.** Penilaian itu dikemukakan pengamat ekonomi Didiek J. Rachbini, di Jakarta, baru-baru ini. Menurut perhitungan Didiek, dalam tahun ini, The FED telah lima kali menurunkan nilai suku bunga yang mereka miliki. Bahkan, Didiek memperkirakan, tingkat suku bunga The FED akan diturunkan hingga menjadi empat persen. Dengan keadaan itu, tambah Didiek, di atas kertas dapat dimanfaatkan BI untuk meningkatkan suku bunga BI sebagai upaya mempertahankan nilai tukar rupiah. Namun demikian, Didiek pesimistis, hal itu akan tercapai mengingat kondisi bangsa masih carut marut. "Jika keadaan terus seperti ini, tak tertutup kemungkinan, BI akan tetap memberlakukan nilai suku bunga tinggi," ujar Didiek. ICH/Fahmi Ihsan dan Donny Indradi).	polda riau bertekad memberantas para penyelundup kayu di riau. dalam tahun ini, the fed telah menurunkan tingkat suku bunga yang dimiliki the fed.	kapolda riau melantik kapolda baru menggantikan gubernur no. h. thobrak. didiek j. rachbini optimistis, bi akan tetap mempertahankan nilai suku bunga the fed.
2 Xtremes	[Doc 1] Liputan6.com, Jakarta: Romadhani alias Roban, **penjahat kelas kakap, tewas tertembak sesaat sebelum beraksi di kawasan Pasar Induk Kramatjati, Jakarta Timur,** Senin (31/12). Pria berusia 30 tahun tewas setelah peluru polisi bersarang di dadanya. Kepala Unit Reserse dan Intelijen Kepolisian Resor Jaktim Inspektur Satu Polisi Sudiono mengatakan, Roban terpaksa ditembak karena melawan ketika hendak ditangkap. (ZAQ/Nurul Amin dan Gatot Setiawan). [Doc 2] Liputan6.com, Jakarta: Menurut perhitungan Didiek, dalam tahun ini, **The FED telah lima kali menurunkan nilai suku bunga yang mereka miliki**. Bahkan, Didiek memperkirakan, tingkat suku bunga The FED akan diturunkan hingga menjadi empat persen. Dengan keadaan itu, tambah Didiek, di atas kertas dapat dimanfaatkan BI untuk meningkatkan suku bunga BI sebagai upaya mempertahankan nilai tukar rupiah. Namun demikian, Didiek pesimistis, hal itu akan tercapai mengingat kondisi bangsa masih carut marut. "Jika keadaan terus seperti ini, tak tertutup kemungkinan, BI akan tetap memberlakukan nilai suku bunga tinggi," ujar Didiek. (ICH/Fahmi Ihsan dan Donny Indradi).	seorang penjahat kelas kakap tewas ditembak di kawasan pasar induk kramatjati, jaktim. dalam tahun ini, the fed telah lima kali menurunkan suku bunga yang dimiliki the fed.	seorang penjahat kelas kakap tewas ditembak polisi saat beraksi di pasar kramatjati, jaktim. didiek j. rachbini pesimistis, bi akan mampu menurunkan suku bunga bank indonesia.

Table 4 The Example of an Abstractive Summary with Bert2Bert (Continued)

Type	Articles	Bert2Bert Summary	Bert2Bert+Extreme Summary
3 Canonicals	[Doc 1] Liputan6.com, Jakarta: Operasi Sadar Jaya yang dilancarkan Selasa (15/5) malam, sekitar pukul 23. 00 WIB, mengejutkan pengunjung Diskotik Millenium, yang berlokasi di Jalan Gajah Mada, Jakarta Pusat. Sebanyak 200 petugas gabungan dari Kepolisian Resor Metro Jakarta Pusat dan kesatuan Brigade Mobil Polda Metro Jaya menggeledah seluruh pengunjuk diskotik yang tengah asyik berdansa. Dari operasi tersebut, polisi menangkap 32 **pengunjung diskotik yang tertangkap basah membawa 66 butir pil ekstasi.** 14 orang di antara mereka adalah wanita muda. Para pengunjung yang tertangkap tampak pasrah saat dibawa ke kantor polisi. Sementara itu, kaca mobil patroli Polres Metro Jakpus yang digunakan untuk razia, terlihat pecah karena dilempar batu. Kaca mobil patroli pecah saat polisi menggeledah Diskotik Millenium. (COK/Christiyanto dan Johni Akbar). [Doc 2] Liputan6.com, Tangerang: **Empat warga negara asing terdakwa penyelundup heroin disidangkan di Pengadilan Negeri Tangerang**, Banten, Selasa (15/5). Keempat orang itu adalah Samuel Uwuchukwu Okoye dari Nigeria, Ozias Sibanda dari Zimbabwe, Hansen Antony Nwaolisa, dan Okwudili Ayotanze dari Liberia. Keempat tersangka diancam hukuman mati. (ICH/Roy Akhmad dan Agung Nugroho). [Doc 3] Liputan6.com, Jakarta: Tunggakan Kredit Usaha Tani (KUT) di Bank Rakyat Indonesia mencapai Rp 2, 4 triliun. Akibatnya, penyaluran Kredit Ketahanan Pangan (KKP) sebagai pengganti KUT untuk petani juga ikut terhambat. Demikian dikemukakan Direktur Utama BRI Rudjito, Rabu (16/5). Rudjito mengutip catatan pemerintah bahwa tunggakan KUT untuk musim tanam tahun 2000 mencapai Rp 3, 2 triliun. Sekitar 75 persen di antaranya KUT yang disalurkan BRI. (YYT/Olivia Rosalia dan Donni Indradi).	sebanyak 32 pengunjung diskotik millenium, jakpus, ditangkap karena kedapatan membawa 66 butir pil ekstasi. empat terdakwa penyelundup heroin disidangkan di pn tangerang.	diskotik millenium yang terletak di jalan gajah mada, jakarta pusat, dirazia ratusan polisi. empat tersangka penyelundup heroin disidangkan di pengadilan negeri tangerang.
3 Xtremes	[Doc 1] Liputan6.com, Jakarta: Romadhani alias Roban, **penjahat kelas kakap, tewas tertembak sesaat sebelum beraksi di kawasan Pasar Induk Kramatjati.** Jakarta Timur. Senin (31/12). Pria berusia 30 tahun tewas setelah peluru polisi bersarang di dadanya. Kepala Unit Reserse dan Intelijen Kepolisian Resor Jaktim Inspektur Satu Polisi Sudiono mengatakan, Roban terpaksa ditembak karena melawan ketika hendak ditangkap. (ZAQ/Nurul Amin dan Gatot Setiawan). [Doc 2] Liputan6.com, Jakarta: Operasi Sadar Jaya yang dilancarkan Selasa (15/5) malam, sekitar pukul 23. 00 WIB, mengejutkan pengunjung Diskotik Millenium, yang berlokasi di Jalan Gajah Mada, Jakarta Pusat. Sebanyak 200 petugas gabungan dari Kepolisian Resor Metro Jakarta Pusat dan kesatuan Brigade Mobil Polda Metro Jaya menggeledah seluruh pengunjuk diskotik yang tengah asyik berdansa. (COK/Christiyanto dan Johni Akbar). [Doc 3] Liputan6.com, Jakarta: **Tunggakan Kredit Usaha Tani (KUT) di Bank Rakyat Indonesia mencapai Rp 2, 4 triliun. Akibatnya, penyaluran Kredit Ketahanan Pangan (KKP) sebagai pengganti KUT untuk petani juga ikut terhambat.** Demikian dikemukakan Direktur Utama BRI Rudjito, Rabu (16/5). Rudjito mengutip catatan pemerintah bahwa tunggakan KUT untuk musim tanam tahun 2000 mencapai Rp 3, 2 triliun. (YYT/Olivia Rosalia dan Donni Indradi).	seorang penjahat yang kerap beraksi di pasar kramatjati, jakarta timur, tewas ditembak polisi. tunggakan kut di bri sebesar rp 2, 4 triliun membuat penyaluran kredit terhambat.	seorang penjahat kelas kakap tewas ditembak polisi di kawasan pasar induk kramatjati, jakarta timur. tunggakan kredit usaha tani untuk petani juga menghambat penyaluran kredit ketahanan pangan.

FKGL and GFI are readability measurement metrics used for English. However, several linguistic studies have adapted itnd is suitable for Indonesians (Fadziah et al., 2018; Mursyadah, 2021; Sari & Herri, 2020; Utami et al., 2021). A readability metric is used specifically for the Indonesian language, namely the Dwiyanto Djoko Pranowo metric. Therefore, this study's summary results of Bert2Bert and Bert2Bert+Xtreme also used Dwitanto Djoko Pranowo to measure readability. The evaluation results show that the summary of Bert2Bert and Bert2Bert+Xtreme, shown in

Table 5, are in the moderate readability range, namely 20.59 and 22.17. This indicates that the summary results of Indonesian multi-documents using Bert2Bert and Bert2Bert+Xtreme have good readability and are still understandable.



(a) (b)

Figure 2 FKGL and GFI Result: (a) Bert2Bert, (b) Bert2Bert+Xtreme

Table 5 Dwitanto's Evaluation of Bert2Bert and Bert2Bert+Extreme

Indicators	Bert2Bert				Bert2Bert+Xtreme			
	2 Canonicals	2 Xtremes	3 Canonicals	3 Xtremes	2 Canonicals	2 Xtremes	3 Canonicals	3 Xtremes
	Average of Dwiyanto's Score				Average of Dwiyanto's Score			
Paragraph	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Sentence Count	1.95	1.95	1.95	1.95	2.07	2.09	2.08	2.11
Sentence Length	12.80	12.76	12.40	12.19	11.54	11.41	11.37	10.97
Extension	0.74	0.75	0.75	0.76	0.74	0.74	0.75	0.75
Compound	0.76	0.76	0.76	0.77	0.75	0.75	0.75	0.77
Polysemy	0.76	0.76	0.76	0.77	0.76	0.76	0.76	0.77
Passive Sentence	0.63	0.63	0.63	0.64	0.62	0.63	0.63	0.64
Unfamiliar Word	0.01	0.01	0.02	0.02	0.01	0.02	0.02	0.02
Abstract Word	0.72	0.73	0.72	0.73	0.72	0.72	0.72	0.73
Terms	0.77	0.77	0.77	0.78	0.76	0.76	0.77	0.78
Conjunctions	0.74	0.74	0.75	0.75	0.73	0.73	0.73	0.74
Loan	0.70	0.70	0.70	0.70	0.69	0.69	0.69	0.70
Phrase	0.60	0.60	0.60	0.61	0.59	0.59	0.59	0.60
Dwiyanto's Total Score	22.17	22.17	21.82	21.66	20.99	20.87	20.85	20.59

4. CONCLUSIONS

The research findings highlight the performance of the proposed model, Bert2Bert, showcasing its efficacy in abstractive multi-document summarization tasks. While the ROUGE and BERTScore metrics show similarity between Bert2Bert and IndoBERT, Bert2Bert stands out with superior performance, especially compared to the Bert2Bert+Xtreme model. Despite the limitations in conducting co-selection-based analysis evaluations like ROUGE and BERTScore due to the lack of summary references for multiple Indonesian documents, this research adopts content-based analysis using readability metrics such as FKGL, GFI, and Dwiyanto Djoko

Pranowo. The results show that Bert2Bert and Bert2Bert+Xtreme produce summaries with an appropriate level of readability for adult readers, as expected from the target audience of the news portal. Overall, the moderate readability range of summary results suggests that the Bert2Bert and Bert2Bert+Xtreme models offer summaries that are understandable and accessible to their intended audience, aligning with the nature of news content targeted at adult readers. Future research could explore other transformer models for abstract summaries of several documents in Indonesian, such as GPT2GPT, BERT2GPT, or GPT2BERT. Further research could also contribute to preparing higher quality datasets for automatic text abstraction summarization of multiple documents to provide a more comprehensive evaluation using co-selection-based metrics and further improve model performance. By overcoming these challenges, future research can contribute to advancing automatic document summarization technology and be applied to real-world cases in Indonesia.

ACKNOWLEDGEMENT

We want to thank the various parties who supported this research, especially the Department of Informatics, UIN Sunan Gunung Djati Bandung, which funded the publication of this research.

REFERENCES

Abka, A. F., Azizah, K., & Jatmiko, W. (2022). Transformer-based Cross-Lingual Summarization using Multilingual Word Embeddings for English - Bahasa Indonesia. *International Journal of Advanced Computer Science and Applications*, *13*(12). https://doi.org/10.14569/IJACSA.2022.0131276

Alquliti, W. H., & Binti, N. (2019). Convolutional Neural Network based for Automatic Text Summarization. *International Journal of Advanced Computer Science and Applications*, *10*(4), 200–211. https://doi.org/10.14569/IJACSA.2019.0100424

Biddinika, M. K., Lestari, R. P., Indrawan, B., Yoshikawa, K., Tokimatsu, K., & Takahashi, F. (2016). Measuring the readability of Indonesian biomass websites: The ease of understanding biomass energy information on websites in the Indonesian language. *Renewable and Sustainable Energy Reviews*, *59*, 1349–1357. https://doi.org/10.1016/j.rser.2016.01.078

Bing, L., Li, P., Liao, Y., Lam, W., Guo, W., & Passonneau, R. (2015). Abstractive Multi-Document Summarization via Phrase Selection and Merging. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1587–1597. https://doi.org/10.3115/v1/P15-1153

Dangol, R., Adhikari, P., Dahal, P., & Sharma, H. (2023). Short Updates- Machine Learning Based News Summarizer. *Journal of Advanced College of Engineering and Management*, *8*(2), 15–25. https://doi.org/10.3126/jacem.v8i2.55939

Devi, K. U. S., & Suadaa, L. H. (2022). Extractive Text Summarization for Snippet Generation on Indonesian Search Engine using Sentence Transformers. *2022 International Conference on Data Science and Its Applications (ICoDSA)*, 181–186. https://doi.org/10.1109/ICoDSA55874.2022.9862886

Devianti, R. S., & Khodra, M. L. (2019). Abstractive Summarization using Genetic Semantic Graph for Indonesian News Articles. *Proceedings - 2019 International Conference on Advanced Informatics: Concepts, Theory, and Applications, ICAICTA 2019*. https://doi.org/10.1109/ICAICTA.2019.8904361

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://doi.org/10.48550/arXiv.1810.04805

Dewi, K. E., & Widiastuti, N. I. (2022). The Design of Automatic Summarization of Indonesian Texts Using a Hybrid Approach. *Jurnal Teknologi Informasi Dan Pendidikan*, *15*(1), 37–43. https://doi.org/10.24036/jtip.v15i1.451

Fadziah, Y. N., Rasim, R., & Fitrajaya, E. (2018). Penerapan Algoritma Enchanced Confix Stripping dalam Pengukuran Keterbacaan Teks Menggunakan Gunning Fog Index. *Jurnal Aplikasi Dan Teori Ilmu Komputer*, *1*(1), 14–22. https://doi.org/10.17509/jatikom.v1i1.25143

Goh, O. S., Fung, C. C., Depickere, A., & Wong, K. W. (2007). Using Gunnnig-Fog Index to Assess Instant Messages Readability from ECAs. *Third International Conference on Natural Computation (ICNC 2007) Vol V*, 480–486. https://doi.org/10.1109/ICNC.2007.800

Goldstein, J., Mittal, V., Carbonell, J., & Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. *NAACL-ANLP 2000 Workshop on Automatic Summarization -*, *4*, 40–48. https://doi.org/10.3115/1117575.1117580

Gunawan, D., Harahap, S. H., & Fadillah Rahmat, R. (2019). Multi-document Summarization by using TextRank and Maximal Marginal Relevance for Text in Bahasa Indonesia. *2019 International Conference on ICT for Smart Society (ICISS)*, *7*, 1–5. https://doi.org/10.1109/ICISS48059.2019.8969785

Gunawan, Y. H. B., & Khodra, M. L. (2021). *Multi-document Summarization using Semantic Role Labeling and Semantic Graph for Indonesian News Article*. https://doi.org/10.48550/arXiv.2103.03736

Jin, H., & Wan, X. (2020). Abstractive Multi-Document Summarization via Joint Learning with Single-Document Summarization. *Findings of the Association for Computational Linguistics: EMNLP 2020*, *2020*, 2545–2554. https://doi.org/10.18653/v1/2020.findings-emnlp.231

Koto, F., Lau, J. H., & Baldwin, T. (2020). Liputan6: A Large-scale Indonesian Dataset for Text Summarization. *Proceedings Ofthe 1st Conference Ofthe Asia-Pacific Chapter Ofthe Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 598–608. https://doi.org/10.48550/arXiv.2011.00679

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *Proceedings of the 28th International Conference on Computational Linguistics*, 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

Kurniawan, K., & Louvan, S. (2018). Indosum: A New Benchmark Dataset for Indonesian Text Summarization. *2018 International Conference on Asian Language Processing (IALP)*, 215–220. https://doi.org/10.1109/IALP.2018.8629109

Kuyate, S., Jadhav, O., & Jadhav, P. (2023). AI Text Summarization System. *International Journal for Research in Applied Science and Engineering Technology*, *11*(5), 916–919. https://doi.org/10.22214/ijraset.2023.51481

Laksana, M. D. B., Karyawati, A. E., Putri, L. A. A. R., Santiyasa, I. W., Sanjaya ER, N. A., & Kadnyanan, I. G. A. G. A. (2022). Text Summarization terhadap Berita Bahasa Indonesia menggunakan Dual Encoding. *JELIKU (Jurnal Elektronik Ilmu Komputer Udayana)*, *11*(2), 339. https://doi.org/10.24843/JLK.2022.v11.i02.p13

Lamsiyah, S., Mahdaouy, A. El, Ouatik, S. E. A., & Espinasse, B. (2023). Unsupervised extractive multi-document summarization method based on transfer learning from BERT multi-task fine-tuning. *Journal of Information Science*, *49*(1), 164–182. https://doi.org/10.1177/0165551521990616

Li, W., & Zhuge, H. (2021). Abstractive Multi-Document Summarization Based on Semantic Link Network. *IEEE Transactions on Knowledge and Data Engineering*, *33*(1), 43–54. https://doi.org/10.1109/TKDE.2019.2922957

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, 74–81. https://aclanthology.org/W04-1013/

Lucky, H., & Suhartono, D. (2021). Investigation of Pre-Trained Bidirectional Encoder Representations from Transformers Checkpoints for Indonesian Abstractive Text Summarization. *Journal of Information and Communication Technology*, *21*(1), 71–94. https://doi.org/10.32890/jict2022.21.1.4

Maylawati, D. S. (2019). Sequential Pattern Mining and Deep Learning to Enhance Readability of Indonesian Text Summarization. *International Journal of Advanced Trends in Computer Science and Engineering*, *8*(6), 3147–3159. https://doi.org/10.30534/ijatcse/2019/78862019

Maylawati, D. S., Kumar, Y. J., Kasmin, F., & Ramdhani, M. A. (2024). Deep sequential pattern mining for readability enhancement of Indonesian summarization. *International Journal of Electrical and Computer Engineering (IJECE)*, *14*(1), 782. https://doi.org/10.11591/ijece.v14i1.pp782-795

Mursyadah, U. (2021). Tingkat Keterbacaan Buku Sekolah Elektronik (BSE) Pelajaran Biologi Kelas X SMA/MA. *TEACHING : Jurnal Inovasi Keguruan Dan Ilmu Pendidikan*, *1*(4), 298–304. https://doi.org/10.51878/teaching.v1i4.774

Pranowo, D. D. (2011). *Alat ukur keterbacaan teks berbahasa Indonesia*.

Rothe, S., Narayan, S., & Severyn, A. (2020). Leveraging Pre-trained Checkpoints for Sequence Generation Tasks. *Transactions of the Association for Computational Linguistics*, *8*, 264–280. https://doi.org/10.1162/tacl_a_00313

Sari, M. P., & Herri, H. (2020). Analisa Konten Serta Tingkat Keterbacaan Pernyataan Misi dan Pengaruhnya Terhadap Kinerja Perbankan Indonesia. *Menara Ilmu: Jurnal Penelitian Dan Kajian Ilmiah*, *14*(1), 96–106. https://doi.org/10.31869/mi.v14i1.2003

Scott, B. (2024). *Learn How to Use the Flesch-Kincaid Grade Level Formula*. ReadabilityFormulas.Com. https://readabilityformulas.com/learn-how-to-use-the-flesch-kincaid-grade-level/

Scott, B. (2025). *The Gunning Fog Index (or FOG) Readability Formula*. ReadabilityFormulas.Com. https://readabilityformulas.com/the-gunnings-fog-index-or-fog-readability-formula/

Severina, V., & Khodra, M. L. (2019). Multidocument Abstractive Summarization using Abstract Meaning Representation for Indonesian Language. *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 1–6. https://doi.org/10.1109/ICAICTA.2019.8904449

Shen, C., Cheng, L., Nguyen, X.-P., You, Y., & Bing, L. (2023). *A Hierarchical Encoding-Decoding Scheme for Abstractive Multi-document Summarization*. https://doi.org/10.48550/arXiv.2305.08503

Shinde, K., Roy, T., & Ghosal, T. (2022). An Extractive-Abstractive Approach for Multi-document Summarization of Scientific Articles for Literature Review. *Proceedings of the Third Workshop on Scholarly Document Processing*, 204–209. https://aclanthology.org/2022.sdp-1.25/

Solnyshkina, M. I., Zamaletdinov, R. R., Gorodetskaya, L. A., & Gabitov, A. I. (2017). Evaluating Text Complexity and Flesch-Kincaid Grade Level. *Journal of Social Studies Education Research*, *8*(3), 238–248. http://www.jsser.org/index.php/jsser/article/view/225

Sugiri, Eko Prasojo, R., & Alfa Krisnadhi, A. (2022). Controllable Abstractive Summarization Using Multilingual Pretrained Language Model. *2022 10th International Conference on Information and Communication Technology (ICoICT)*, 228–233. https://doi.org/10.1109/ICoICT55009.2022.9914846

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to Sequence Learning with Neural Networks*. http://arxiv.org/abs/1409.3215

Świeczkowski, D., & Kułacz, S. (2021). The use of the Gunning Fog Index to evaluate the readability of Polish and English drug leaflets in the context of Health Literacy challenges in Medical Linguistics: An exploratory study. *Cardiology Journal*, *28*(4), 627–631. https://doi.org/10.5603/CJ.a2020.0142

Utami, S. D., Dewi, I. N., & Efendi, I. (2021). Tingkat Keterbacaan Bahan Ajar Flexible Learning Berbasis Kolaboratif Saintifik. *Bioscientist : Jurnal Ilmiah Biologi*, *9*(2), 577. https://doi.org/10.33394/bioscientist.v9i2.4246

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. https://doi.org/10.48550/arXiv.1706.03762

Verma, P., & Om, H. (2019). A novel approach for text summarization using optimal combination of sentence scoring methods. *Sādhanā*, *44*(5), 110. https://doi.org/10.1007/s12046-019-1082-4

Verma, P., Pal, S., & Om, H. (2019). A Comparative Analysis on Hindi and English Extractive Text Summarization. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *18*(3), 1–39. https://doi.org/10.1145/3308754

Widjanarko, A., Kusumaningrum, R., & Surarso, B. (2018). Multi document summarization for the Indonesian language based on latent dirichlet allocation and significance sentence. *2018 International Conference on Information and Communications Technology (ICOIACT)*, 520–524. https://doi.org/10.1109/ICOIACT.2018.8350668

Wijayanti, R., Khodra, M. L., & Widyantoro, D. H. (2021). Indonesian Abstractive Summarization using Pre-trained Model. *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 79–84. https://doi.org/10.1109/EIConCIT50028.2021.9431880

Zhang, J., Tan, J., & Wan, X. (2018). *Towards a Neural Network Approach to Abstractive Multi-Document Summarization*. https://doi.org/10.48550/arXiv.1804.09010

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. *8th International Conference on Learning Representations, ICLR 2020*. https://doi.org/10.48550/arXiv.1904.09675