

Hybrid model for extractive single document summarization: utilizing BERTopic and BERT model

Maryanto, Philips, Abba Suganda Girsang

Department of Computer Science, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Feb 20, 2023

Revised Oct 17, 2023

Accepted Nov 16, 2023

Keywords:

BERTopic

BERT

Cable news network

Daily mail

Extractive summarization

ABSTRACT

Extractive text summarization has been a popular research area for many years. The goal of this task is to generate a compact and coherent summary of a given document, preserving the most important information. However, current extractive summarization methods still face several challenges such as semantic drift, repetition, redundancy, and lack of coherence. A novel approach is presented in this paper to improve the performance of an extractive summarization model based on bidirectional encoder representations from transformers (BERT) by incorporating topic modeling using the BERTopic model. Our method first utilizes BERTopic to identify the dominant topics in a document and then employs a BERT-based deep neural network to extract the most salient sentences related to those topics. Our experiments on the cable news network (CNN)/daily mail dataset demonstrate that our proposed method outperforms state-of-the-art BERT-based extractive summarization models in terms of recall-oriented understudy for gisting evaluation (ROUGE) scores, which resulted in an increase of 32.53% of ROUGE-1, 47.55% of ROUGE-2, and 16.63% of ROUGE-L when compared to baseline BERT-based extractive summarization models. This paper contributes to the field of extractive text summarization, highlights the potential of topic modeling in improving summarization results, and provides a new direction for future research.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Abba Suganda Girsang

Department of Computer Science, School of Computer Science, Bina Nusantara University

Jakarta-11480, Indonesia

Email: agirsang@binus.edu

1. INTRODUCTION

In recent years, the advancement in technology has made access to information much easier. Textual material has been thriving on the internet and has only grown ever since. The total number of scholarly articles on the internet has reached at least 27 million by 2014 [1]. With the rapid growth of digital information, the demand for effective automatic text summarizer has increased, leading to research of more efficient and effective text summarization approaches [2].

Automatic text summarizer is one of the subsets of natural language processing (NLP), which is used to create a more concise form of long textual content for easier digestion by humans [3]. Text summarization can be done in two main approaches: extractive and abstractive. Extractive text summarization is an approach of text summarization that extracts important features, i.e. words or sentences, from a source document without alteration to be combined into a more concise summary [4]. On the other hand, abstractive text summarization extracts important information from the source document and generates its own summary, which may generate novel words in the process [5]. Based on the amount of documents involved, automatic text summarizer can also be divided into two categories: single document and multi-document

summarization. Single document summarization is associated with generating short summaries from long text documents, while multi-document summarization works with multiple documents that are correlated with unspecified themes. Recently, deep learning approaches have shown great potential in text summarization, particularly with the advent of pre-trained models such as bidirectional encoder representations from transformers (BERT) [6]. BERT is a transformer-based architecture that has achieved state-of-the-art results in various NLP tasks, including text summarization tasks [7]–[9].

This paper proposed an extractive-based text summarization method that leverages both BERT and BERTopic, a variant of BERT specifically designed for topic modeling. Our method first applies BERTopic to identify the underlying topics and form sentence clusters. Then BERT is applied to summarize each sentence cluster. Finally, the summary from each cluster is combined to form the final summary. To the best of our knowledge, this is the first implementation of such an approach.

Experiments were conducted on the cable news network (CNN)/daily mail dataset, and the results show that our proposed method outperforms BERT when it is used on its own in terms of recall-oriented understudy for gisting evaluation (ROUGE) scores, a widely used evaluation metric for summarization. By implementing these improvements, we were able to improve the performance of BERT extractive summarization to 32.53% of ROUGE-1, 47.55% of ROUGE-2, and 16.63% of ROUGE-L. In conclusion, our proposed method demonstrates the effectiveness of combining BERT and BERTOPIC for extractive text summarization. The results show that BERT can effectively capture the semantic information of the text and BERTOPIC can effectively identify the underlying topics, leading to a more informative and coherent summary. This work provides a new direction for future research in the field of extractive text summarization and highlights the potential of leveraging pre-trained models for this task.

2. RELATED WORKS

2.1. Topic modeling

Topic modeling is a text-mining technique used in NLP and computer science to discover and extract the latent topics (i.e. word clusters) present in a large collection of text documents [10]. It is a type of unsupervised learning, which entails that it is performed without labeled data [10]. In recent years, it has gained significant popularity due to its efficiency in handling large volumes of unstructured text data and its ability to automatically identify patterns and themes within that data. It has a wide range of applications, including information retrieval, document classification, sentiment analysis, and social media analysis [11]–[13].

Several studies have also demonstrated the viability of a topic modeling approach on extractive summarization [14]–[16]. Those studies incorporated bag-of-words approach-based algorithms, such as latent dirichlet allocation (LDA) and latent semantic analysis (LSA). These algorithms differ in their mathematical formulation and the assumptions they make about the underlying structure of the text data, but they all share the goal of identifying a set of latent topics that can be used to describe the collection of text documents. However, these techniques disregard the grammar of the words, hence losing the contextual information of the text. This often results in a false interpretation of the text. In conclusion, topic modeling is a powerful and widely used technique for uncovering the underlying themes and patterns in large collections of text data. Its applications are numerous, and it continues to be an active area of research in the field of NLP and computer science.

2.2. Bidirectional encoder representations from transformers

BERT is a pre-trained bidirectional encoder published in 2018 that is based on the transformer architecture. BERT was intended for NLP tasks, including sentiment analysis and question-answering. Since its initial release, BERT has become one of the most widely used language models in NLP and has been incorporated in numerous studies. In this literature review, we will discuss some of the key contributions and findings related to BERT.

BERT's bidirectional training mechanism is a critical factor in its popularity. By processing input sequences from both left and right sides, BERT can capture the context in a way that unidirectional models cannot. Another important contribution of BERT is its use of attention mechanisms. The attention mechanism is highly parallelizable, unlike sequential models like recurrent neural network (RNN), which may be difficult to scale with large corpus [17]–[18]. This allows BERT to attend to different parts of the input sequence, regardless of the document size. As a result, BERT can better capture the relationships between words and improve its representations. This has been shown to result in improved performance on many benchmark NLP tasks, such as general language understanding evaluation (GLUE), multi-genre natural language inference (MultiNLI), and stanford question answering dataset (SQUAD) v1.1 [19].

Finally, BERT has also been used as a pre-trained model in many NLP applications, including question-answering, text classification, and text summarization. In these applications, BERT has been shown

to outperform other pre-trained models and even task-specific models that have been trained from scratch [6]. In conclusion, BERT has had a significant impact on the NLP community and is effective in a wide range of NLP tasks. Its bidirectional training mechanism and attention mechanisms have been identified as key factors in its success, and its ability to be fine-tuned for specific tasks has made it a popular choice for NLP practitioners.

2.3. BERTopic

Topic modeling is a technique originated from text mining that allows latent topics to be extracted from a document or a collection of documents. The aim is to uncover the underlying structure of the text by discovering the words and phrases that are most representative of a particular topic. Topic modeling has numerous applications in various domains, such as text classification, information retrieval, and text summarization. Recent advancement in topic modeling is the discovery of BERTopic, which leverages the power of BERT, a transformer-based deep learning model. BERTopic is a pre-trained language model that has been fine-tuned on a large corpus of text data and is specifically designed for topic modeling. Leveraging BERT's attention mechanism, BERTopic is able to capture semantic relationships between words and phrases in the text, hence having a better understanding of the text, which leads to improved performance on topic modeling tasks [20]. In a recent study, BERTopic was compared to LDA and other topic modeling methods on several datasets for text classification and text summarization. The results showed that BERTopic outperformed the other methods in terms of accuracy and coherence, indicating that it is well-suited for text classification and summarization tasks [21].

BERTopic works in a few steps: first, the document is embedded using a sentence transformer model. Next, the embeddings will be reduced in dimensions using uniform manifold approximation and projection (UMAP). This process will simplify the embedding, hence reducing computation complexities for the next process. A clustering technique, in this case hierarchical density-based spatial clustering of applications with noise (HDBSCAN) is then applied to the reduced embeddings to form clusters of documents. Finally, term frequency-inverse document frequency (TF-IDF) will be applied and the most relevant topic representations will be extracted.

In addition to its performance, BERTopic has several other advantages over traditional topic modeling methods. Firstly, BERTopic is highly scalable and can be applied to large datasets, making it ideal for big data applications [22]. Secondly, BERTopic is easy to use and does not require extensive hyperparameter tuning, making it a user-friendly tool for topic modeling. Finally, BERTopic is highly interpretable, as it provides not only the topics but also the words and phrases that are most representative of each topic, making it a valuable tool for understanding the underlying structure of the text [23]. To conclude, BERTopic is a powerful tool for topic modeling that leverages the strengths of BERT to provide improved performance over traditional topic modeling methods. Its scalability, ease of use, and interpretability make it a valuable tool for a variety of NLP applications, including text classification, information retrieval, and text summarization.

2.4. Recall-oriented understudy for gisting evaluation

ROUGE is a metric that is widely used to evaluate text summarization tasks. ROUGE works by measuring the overlapping phrases between the generated summary and reference summary. It is based on the concept of precision and recall, in which the former measures how much of the produced summary is relevant to the reference summary, meanwhile the latter measures how much of the reference summary the produced summary is capturing [24].

Due to its simplicity, ROUGE has been widely used in text summarization. It is determined by counting the overlapped n-grams, word sequences, and word pairs between the generated summary and the reference summary. ROUGE offers a variety of metrics, such as ROUGE-L, which evaluates the length of the longest common subsequence between the two summaries, and ROUGE-N, which measures the overlap in n-grams. However, despite its popularity, ROUGE possesses some major limitations. Due to how it works, ROUGE only measures recall, disregarding other important aspects of text summarization such as coherence, fluency, and relevance. Additionally, ROUGE also doesn't understand the concept of synonyms, hence not suitable for evaluating abstractive summarization where new words are generated.

In conclusion, ROUGE is used widely as an evaluation metric for extractive text summarization. Although ROUGE has some limitations such as it only measures recall and does not take account of coherence, relevance, and fluency which are important to summarization. it is still an important tool for evaluating extractive summarization performance.

2.5. Cable news network/daily mail dataset

The CNN/daily mail dataset is a widely used benchmark dataset for the task of text summarization. It was introduced in the paper "abstractive text summarization using sequence-to-sequence RNNs and

beyond” and it consists of news articles and their corresponding highlights or summaries [25]. Since its introduction, the CNN/daily mail dataset has been used extensively in the field of text summarization to evaluate the performance of various models.

3. METHOD

In this paper, we attempted to implement BERTopic to model the topic of each sub-document, from which summary sentences will be decided based on their similarity with the topic. However, this method is not feasible because, for shorter documents, the topic tends to not be generated. Currently, we are working on alleviating this issue by exploring other methods. In this case, we are replacing the BERTopic clustering technique from HDBSCAN to k-nearest neighbor (KNN) algorithm as HDBSCAN does not force documents into a cluster where they might not belong [26]. This is irrelevant to our case where every sentence in the news should have correlation with one another. As a result, this will improve our resulting topic representation. Figure 1 shows how our algorithm works.

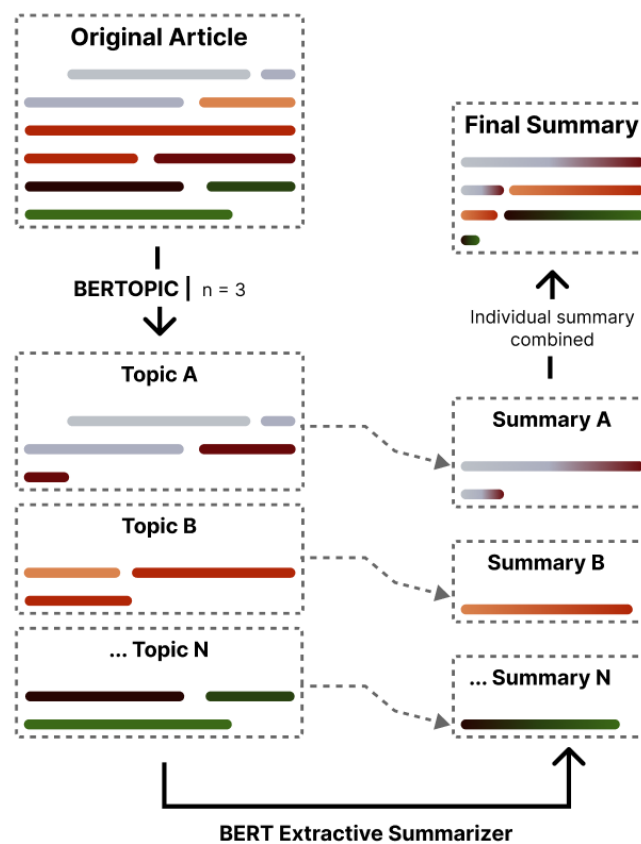


Figure 1. Illustrates the proposed method

First, the original article are split into an array of sentences. These sentences are fed to our custom BERTopic model which will result in three clusters of topics that contain sentences. These topic clusters are then summarized by the BERT extractive summarization model, and one sentence is produced from each. These sentences are then joined together to form a final summarization. Following is an example of how a text will be processed by our method:

Step 1: break the original article dataset into sentences, as shown in Figure 2.

Step 2: feed the sentences into the BERTopic model, as shown in Figure 3.

Step 3: compile sentences in clusters of topics, as shown in Figure 4.

Step 4: summarize each topic and join into the final summarization, as shown in Figure 5.

Step 5: compare with base BERT extractive summarization model result and compute ROUGE-1, ROUGE-2, ROUGE-L. In this paper, we compared the summary generated from our proposed model with the original BERT model. The results of this comparison are shown in Figure 6.

After testing on a sample dataset, we ran the model on 10.000 CNN/daily mail test dataset to perform a summarization of the given article and evaluate each ROUGE-1, ROUGE-2, and ROUGE-L F1-scores before averaging them to get the final model performance.

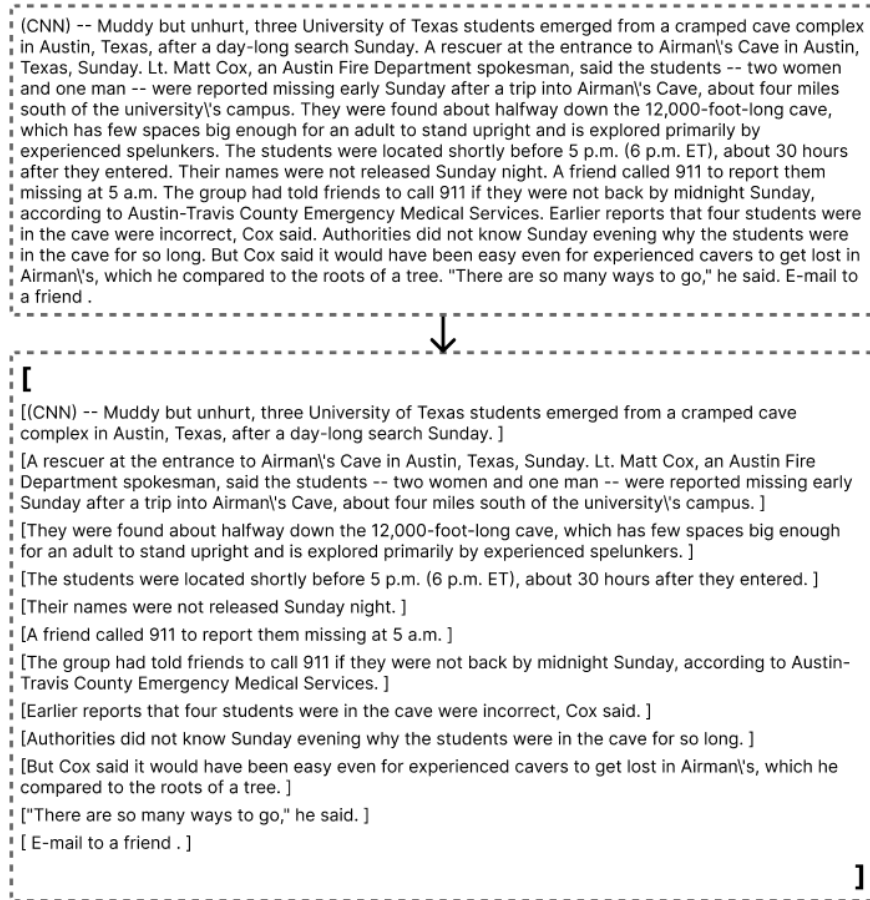


Figure 2. Illustrates text splitting process

	0	1
0	(CNN) -- Muddy but unhurt, three University of...	0
1	A rescuer at the entrance to Airman's Cave in ...	1
2	Lt.	1
3	Matt Cox, an Austin Fire Department spokesman,...	1
4	They were found about halfway down the 12,000-...	0
5	The students were located shortly before 5 p.m...	2
6	Their names were not released Sunday night.	2
7	A friend called 911 to report them missing at ...	1
8	Earlier reports that four students were in the...	0
9	Authorities did not know Sunday evening why th...	2
10	But Cox said it would have been easy even for ...	0
11	"There are so many ways to go," he said.	0
12	E-mail to a friend .	1

Figure 3. Topic extracted from text

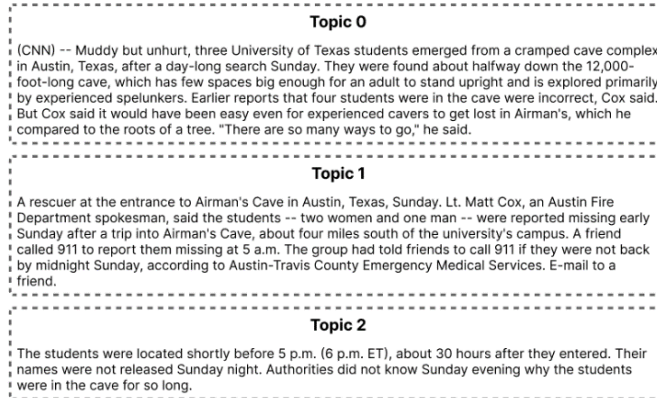


Figure 4. Sentence topic clusters

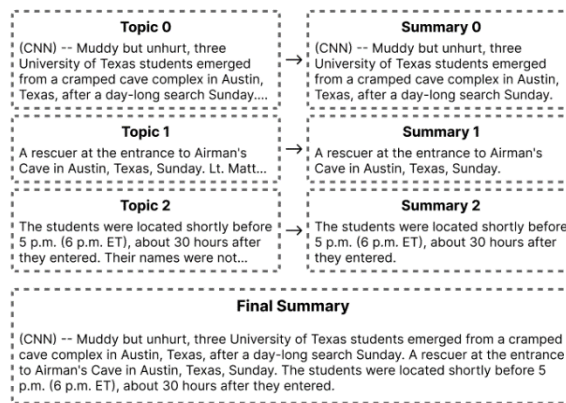


Figure 5. Summary being generated from each sentence cluster

BERTOPIC + BERT Summary
 (CNN) -- Muddy but unhurt, three University of Texas students emerged from a cramped cave complex in Austin, Texas, after a day-long search Sunday. A rescuer at the entrance to Airman's Cave in Austin, Texas, Sunday. The students were located shortly before 5 p.m. (6 p.m. ET), about 30 hours after they entered.

BERT Summary
 (CNN) -- Muddy but unhurt, three University of Texas students emerged from a cramped cave complex in Austin, Texas, after a day-long search Sunday. Their names were not released Sunday night. A friend called 911 to report them missing at 5 a.m. The group had told friends to call 911 if they were not back by midnight Sunday, according to Austin-Travis County Emergency Medical Services.

Golden Summary
 "Three Austin cave explorers are safe and are out of the cave, officials said. The University of Texas students went into Airman's Cave Saturday. The 12,000 foot long cave complex has tight twists and turns. Unclear exactly how the students lost their way, but cave is complex, official said."

Base BERT Model With CNN Daily Mail Dataset Metrics	With Topic Modeling		Without Topic Modeling
ROUGE-1 precision	0.285714		0.179104
ROUGE-1 recall	0.313725		0.235294
ROUGE-1 fmeasure	0.299065		0.20339
ROUGE-2 precision	0.127273		0.0606061
ROUGE-2 recall	0.14		0.08
ROUGE-2 fmeasure	0.133333		0.0689655
ROUGE-L precision	0.178571		0.119403
ROUGE-L recall	0.196078		0.156863
ROUGE-L fmeasure	0.186916		0.135593
ROUGE-Lsum precision	0.25		0.134328
ROUGE-Lsum recall	0.27451		0.176471
ROUGE-Lsum fmeasure	0.261682		0.152542

Figure 6. Summary generated from various model

4. RESULTS AND DISCUSSION

Evaluation results of the proposed method are presented in Table 1. In this study, we attempted three different runs using two, three, and four numbers of topics respectively. One summary sentence is selected from each topic. Hence the summary length is directly proportional to the amount of topics. As can be seen in the table, the best result is obtained when three topics are selected. According to Table 1, the F1-score for ROUGE-2 increases proportionately with the number of topics, or in other words. It means that the greater the number of topic clusters, which means an increase in summary length in our approach, the overlapping of bigrams between the system and reference summaries also increases. In this experiment, the highest value of overlapping bigrams is 14.48%, contained in the summary results of four topics.

Table 1. Evaluation of the model for different number of topics count

Number of topics	ROUGE-1 (%)	ROUGE-2 (%)	ROUGE-L (%)
2	34.02	13.20	21.74
3	35.59	14.27	21.79
4	35.13	14.48	21.02

We evaluate our model and baseline BERT summarization model which results are shown in Table 2. Compared with baseline models, our proposed solution using BERTopic topic modeling achieves better scores in overall metrics. This result is expected because our model focuses on clustering sentences into their relevant latent topics before proceeding to summarization. This will generally produce a summarization where each sentence represents each main event in the article, and thus it achieves good performance on this comparison.

- Baseline model: as seen from Table 3, the summary generated by the baseline model manages to include one sentence that contains the relevant information to the golden summary, pertaining to how the suspect is found guilty, while the rest were deviated. The longest pair of consecutive words that overlap with the golden summary is two (bigrams), with 3 occurrences, while the pair of consecutive words with the highest occurrence belongs to unigram, with 15. It is worth mentioning that a large portion of unigrams that occurred in the baseline summary are stop words, not keywords.
- Proposed model: the summary generated by the proposed model manages to include most sentences that are relevant to the golden summary. Our proposed model manages to reach 11-grams with one occurrence. Additionally, the model also manages to reach a high number of unigrams, with a minimum amount of stop words. The summary generated was of high proximity to the golden summary, though still lacking fluency.

Table 2. The performance of the baseline BERT summarization model compared to our proposed method

Model	ROUGE-1 (%)	ROUGE-2 (%)	ROUGE-L (%)
BERT	26.86	9.67	18.68
BERT+BERTopic	35.59	14.27	21.79

Table 3. Comparison of summaries generated by models with golden summary

Golden summary	Baseline BERT summarization model	Our proposed model
“Dzhokhar Tsarnaev is found guilty on all 30 charges he faced. Seventeen counts were capital charges, meaning he is eligible for the death penalty.”	“After deliberating for 11½ hours, jurors found Dzhokhar Tsarnaev guilty on Wednesday of all 30 counts he faced in the Boston Marathon bombing trial. He was there, Tsarnaev’s defense attorney Judy Clarke conceded as the trial opened, but many say the defense strategy always had been to focus on persuading the jury to spare Tsarnaev’s life. The younger man, Clarke said, was only following his older brother.”	“Seventeen of the 30 counts were capital charges, meaning he is eligible for the death penalty. After the verdict, CNN legal analyst Paul Callan said Clarke now faces an uphill battle. After deliberating for 11½ hours, jurors found Dzhokhar Tsarnaev guilty on Wednesday of all 30 counts he faced in the Boston Marathon bombing trial.”

5. CONCLUSION




Based on the result of the experiment, it can be concluded that we can achieve higher performance in extractive summarization by leveraging BERTopic to cluster the original text before performing summarization. We were able to improve the performance of BERT in the extractive summarization tasks by 32.53% for ROUGE-1, 47.55% for ROUGE-2, and 16.63% for ROUGE-L. Although this method has yielded a high performance, there is still room for improvement and further research. In this research paper, we only experimented using BERTopic as the topic modeling model and BERT as the summarization model, while

there are other newer models such as ALBERT, TinyBERT, or DistilBERT which can perform summarization at similar performance without using too many resources. Future work also can experiment with using our method for abstractive text summarization to further improve the result of the abstractive text summarization process.




REFERENCES

- [1] M. Khabsa and C. L. Giles, "The number of scholarly documents on the public web," *PLoS ONE*, vol. 9, no. 5, pp. 1-6, May 2014, doi: 10.1371/journal.pone.0093949.
- [2] G. Salton, A. Singhal, M. Mitra, and C. Buckley, "Automatic text structuring and summarization," *Information Processing and Management*, vol. 33, no. 2, pp. 193–207, Mar. 1997, doi: 10.1016/S0306-4573(96)00062-3.
- [3] P. Patil, S. Dalmia, S. A. A. Ansari, T. Aul, and V. Bhatnagar, "Automatic text summarizer," in *Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2014*, Sep. 2014, pp. 1530–1534, doi: 10.1109/ICACCI.2014.6968629.
- [4] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," in *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, Jan. 2017, pp. 1–6, doi: 10.1109/ICCCSP.2017.7944061.
- [5] N. Moratanch and S. Chitrakala, "A survey on abstractive text summarization," in *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, Mar. 2016, pp. 1–7, doi: 10.1109/ICCPCT.2016.7530193.
- [6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, 2019.
- [7] D. William, S. Achmad, D. Suhartono, and A. P. Gema, "Leveraging BERT with extractive summarization for depression detection on social media," in *2022 International Seminar on Intelligent Technology and Its Applications: Advanced Innovations of Electrical Systems for Humanity, ISITIA 2022-Proceeding*, Jul. 2022, pp. 63–68, doi: 10.1109/ISITIA56226.2022.9855370.
- [8] T. Niu, C. Xiong, and R. Socher, "Deleter: leveraging BERT to perform unsupervised successive text compression," *arXiv-Computer Science*, pp. 1-5, 2019.
- [9] D. Miller, "Leveraging BERT for extractive text summarization on lectures," *arXiv-Computer Science*, pp. 1-7, 2019.
- [10] G. Miner, D. Delen, J. Elder, A. Fast, T. Hill, and R. A. Nisbet, *Practical text mining and statistical analysis for non-structured text data applications*. Massachusetts, USA: Academic Press, 2012.
- [11] F. Jian, J. X. Huang, J. Zhao, T. He, and P. Hu, "A simple enhancement for ad-hoc information retrieval via topic modelling," in *SIGIR 2016-Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2016, pp. 733–736, doi: 10.1145/2911451.2914748.
- [12] Thielmann, C. Weisser, A. Krenz, and B. Säfken, "Unsupervised document classification integrating web scraping, one-class SVM and LDA topic modelling," *Journal of Applied Statistics*, vol. 50, no. 3, pp. 574–591, Apr. 2023, doi: 10.1080/02664763.2021.1919063.
- [13] M. Choirul Rahmadan, A. Nizar Hidayanto, D. Swadani Ekasari, B. Purwandari, and Theresiawati, "Sentiment analysis and topic modelling using the LDA method related to the flood disaster in Jakarta on Twitter," in *Proceedings-2nd International Conference on Informatics, Multimedia, Cyber, and Information System, ICIMCIS 2020*, Nov. 2020, pp. 126–130, doi: 10.1109/ICIMCIS51567.2020.9354320.
- [14] D. F. O. Onah, E. L. L. Pang, and M. El-Haj, "A data-driven latent semantic analysis for automatic text summarization using LDA topic modelling," in *Proceedings-2022 IEEE International Conference on Big Data, Big Data 2022*, Dec. 2022, pp. 2771–2780, doi: 10.1109/BigData55660.2022.10020259.
- [15] K. A. R. Issam, S. Patel*, and S. C. N., "Topic modeling based extractive text summarization," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 6, pp. 1710–1719, Apr. 2020, doi: 10.35940/ijtee.f4611.049620.
- [16] R. Rani and D. K. Lobiyal, "An extractive text summarization approach using tagged-LDA based topic modeling," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 3275–3305, Sep. 2021, doi: 10.1007/s11042-020-09549-3.
- [17] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1-11, 2017.
- [18] T. Mikolov, M. Karafiát, L. Burget, C. Jan, and S. Khudanpur, "Recurrent neural network based language model," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, Sep. 2010, pp. 1045–1048, doi: 10.21437/interspeech.2010-343.
- [19] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2019, pp. 3730–3740, doi: 10.18653/v1/d19-1387.
- [20] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv-Computer Science*, pp. 1-10, 2022.
- [21] A. Abuzayed and H. Al-Khalifa, "BERT for Arabic topic modeling: an experimental study on BERTopic technique," *Procedia CIRP*, vol. 189, pp. 191–194, 2021, doi: 10.1016/j.procs.2021.05.096.
- [22] G. Hristova and N. Netov, "Media coverage and public perception of distance learning during the COVID-19 pandemic: a topic modeling approach based on BERTopic," in *Proceedings-2022 IEEE International Conference on Big Data, Big Data 2022*, Dec. 2022, pp. 2259–2264, doi: 10.1109/BigData55660.2022.10020466.
- [23] A. Thielmann, C. Weisser, T. Kneib, and B. Säfken, "Coherence based document clustering," in *Proceedings-17th IEEE International Conference on Semantic Computing, ICSC 2023*, Feb. 2023, pp. 9–16, doi: 10.1109/ICSC56153.2023.00009.
- [24] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proceedings of the Workshop on Text Summarization Branches Out*, 2004, pp. 74–81.
- [25] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gülçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *CoNLL 2016-20th SIGNLL Conference on Computational Natural Language Learning, Proceedings*, 2016, pp. 280–290, doi: 10.18653/v1/k16-1028.
- [26] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, no. 11, pp. 1-2, Mar. 2017, doi: 10.21105/joss.00205.




BIOGRAPHIES OF AUTHORS

Maryanto    is a student enrolled in the Bina Nusantara University Graduate Program (BGP) and is currently in his seventh semester at Bina Nusantara University, located in Jakarta, Indonesia. In 2021, he worked as a software engineer at Coding Studio, and starting in 2022, he has been employed as a software engineer at tiket.com. His professional interests lie in machine learning, software engineering, web programming, and artificial intelligence. He can be contacted at email: maryanto001@binus.ac.id.



Philips    is a student enrolled in the Bina Nusantara University Graduate Program (BGP) and is currently in his seventh semester at Bina Nusantara University, located in Jakarta, Indonesia. From 2022 until 2023, he worked as a frontend developer at Fotoyu, and starting from 2023, he has been employed as a UI/UX designer at Kotakode. His professional interests lie in machine learning, UI/UX design, web programming, and artificial intelligence. He can be contacted at email: philips@binus.ac.id.



Abba Suganda Girsang    has been serving as a lecturer at the Master in Computer Science program at Bina Nusantara University in Jakarta, Indonesia, since 2015. He earned his Ph.D. in 2015 from the Institute of Computer and Communication Engineering within the Department of Electrical Engineering at National Cheng Kung University in Tainan, Taiwan. He completed his undergraduate studies in Electrical Engineering at Gadjah Mada University (UGM) in Yogyakarta, Indonesia, in 2000. Following that, he pursued his master's degree in Computer Science at the same university from 2006 to 2008. His professional journey includes roles as a staff consultant programmer at Bethesda Hospital in Yogyakarta in 2001 and as a web developer from 2002 to 2003. Later, he joined the faculty of the Department of Informatics Engineering at Janabadra University, where he worked as a lecturer from 2003 to 2015. His research interests encompass swarm intelligence, combinatorial optimization, and decision support systems. He can be contacted via email at agirsang@binus.edu.