Journal of Engineering Science and Technology Review 17 (5) (2024) 199-210

Review Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

Readability Evaluation Metrics for Indonesian Automatic Text Summarization: A Systematic Review

Dian Sa'adillah Maylawati[1,2,*], Yogan Jaya Kumar[1], Fauziah Binti Kasmin[1] and Muhammad Ali Ramdhani[2]

[1]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Jalan Hang Tuah Jaya, Melaka, 76100, Malaysia
[2]Department of Informatics, Faculty of Science and Technology, UIN Sunan Gunung Djati Bandung, Jalan A. H. Nasution 105, Bandung, 40614, Indonesia

Received 15 October 2022; Accepted 24 September 2024

Abstract

Producing a readable summary from an automatic text summarization system is a big challenge, especially for the Indonesian language. The readability of the generated summary in automatic text summarization is important to reach a quality text summary that is easy to understand. Therefore, this research aims to prepare and investigate the evaluation metrics of the readability aspect of automatic Indonesian text summary results. This research used PRISMA 2020 to conduct a systematic review. We searched Elsevier (SCOPUS), Web of Science, Google Scholar, Science and Technology Index (SINTA), IEEE Xplore, arXiv, and forward and backward references for studies published about readability evaluation for automatic text summarization in the last five years until July 2022. We found that completed readability evaluation in automatic text summarization studies, especially for Indonesian text, is rarely measured. Most studies (94,23% of 52) only use co-selection-based analysis. However, co-selection-based analysis is not adequate to evaluate the readability, it needs content-based analysis and human evaluation. Therefore, this study contributes to the design of the concept of readability evaluation metrics based on a systematic review of Indonesian automatic text summarization and readability evaluation for Indonesian text. This research gives benefits to provides a foundation for future studies to build upon, offering a clear direction for developing and evaluating readability metrics in automatic text summarization, not just for Indonesian, but for other languages facing similar challenges.

Keywords: automatic text summarization; Indonesian language; readability; readability evaluation; text summary.

1. Introduction

The automatic text summarization technique has developed rapidly. Automatic text summarization is a part of Natural Language Processing (NLP) [1]. NLP is a branch of linguistics, computer science, and artificial intelligence concerned with computer-human interaction, specifically how to design computers to handle and evaluate massive amounts of natural language data [2-5]. There are two types of automated text summarization: extractive and abstractive [6]. Extractive summarization generates a sequential summary based on the document's source and with no changes to the sentence's word structure [7-8]. The final summary consists of the original document sentences. Extractive summaries or extracts are created by recognizing relevant sentences that are directly selected from the original content. While abstractive summarization produces the summary with modification [9], for example, paraphrasing. So, the result of abstractive summarization does not have similar sentences or structure to the original document but still has the same meaning.

Many methods are used in automatic text summarization today. At least there are three approaches to producing a summary automatically: the feature-based approach, the graph-based approach, and the popular one is machine learning approach. The feature-based approach uses document components as features (such as words, sentences, phrases, and so on) to be considered and calculated in the process of automatic text summarization [10]. Sentence Scoring [11-12], SumBasic [13], and Latent Semantic Analysis (LSA) [14-16] are feature-based approach algorithms that can be used for text summarization. In the process of text summarization with a graph-based approach, the text data in the document is converted to numeric data. Each sentence in the document will be converted into a node, and the edge value will represent the degree of similarity between two sentences in the document. The algorithms in the graph-based approach that usually used for automatic text summarization such as TextRank [17-18], LexRank [19], and Bellman-Ford [20-22]. Then, the popular automatic text summarization is a machine learning approach that uses deep learning. Deep learning is the development of an artificial neural network that can produce a better summary than the other methods based on several kinds of research. Researches on automatic text summarization that use deep learning methods, such as neuro-fuzzy [23], Recurrent Neural Networks (RNN) [24-25], Convolutional Neural Networks (CNN) [26], deep reinforcement learning [27], graph convolutional neural networks [28], Deep Belief Networks (DBN) [29], Bidirectional Encoder Representations from Transformers (BERT) [30], NeuralSum [31], and so on.

The great challenge in automatic text summarization research is how to produce a readable generated summary [24], [32-33]. The readability of the summary result is a fundamental factor for evaluating the effectiveness of automatic text summarization because it assures that it is readable and understandable [34-36]. As a result, the

[*] E-mail address: diansm@uinsgd.ac.id

ISSN: 1791-2377 © 2024 School of Science, DUTH. All rights reserved.
doi:10.25103/jestr.175.25

summary should be simple to comprehend (readable). A digestible summary is composed of sentences that build on the previous statements and themes to make them easier to understand. In the summary, complex words and grammatical errors should be avoided. The readability of a document can be determined by the content and relationship between sentence aspects. The difficulty of the language's syntax and vocabulary, as well as if the relatedness between sentences (between the previous and next sentence) exhibits reading fluency, determine the readability of the reference summary and the system's summary result.

There are not many automatic text summarization studies that focus on evaluating summary results on the readability aspect, but this is an important thing. Especially for automatic text summarization in Indonesian, no research focuses on assessing the readability aspect of the summary results using special measurements for text readability. Compared to more widely spoken languages like English, Indonesian has fewer linguistic resources (limited NLP resources), such as large annotated corpora and pre-trained models, which makes it harder to train effective summarization algorithms that can account for the unique characteristics of the language. Therefore, this study investigates and prepares a readability evaluation concept that can be used for automatic Indonesian text summarization using systematic review. As far as this research is concerned, the Indonesian automatic text summarization research does not focus on measuring the readability of the generated summary results. Therefore, the novelty of this study is that the results of a systematic review can be a concept for measuring the readability of a comprehensive summary of Indonesian language results. The next section will explain about methods, results, and discussion that were used and found in this systematic research.

2. Materials and Methods

This section provides the research activity, materials, and method. Figure 1 presents the research activities that started from collecting the related works about the current research about automatic text summarization, the summary evaluation method of automatic text summarization, and the readability evaluation method for text, especially for Indonesian text. Then, the next process is to investigate the related works and prepare the readability evaluation concept to evaluate Indonesian automatic text summarization.

This research uses the checklist and flow of PRISMA 2020 to arrange a systematic review. Figure 2 presents the PRISMA flow diagram of this research. In the collecting related works process, the publication articles referenced are obtained from Google Scholar, Scopus, Web of Science, IEEE, arXiv, and Science and Technology Index (SINTA) the Ministry of Education, Culture, Research, and Technology of Indonesia Republic. The sources of these journal articles are used to maintain the quality of publications made by previous researchers. The year of publication of the journal articles that are the maximum reference for the last 5 years. Where the reference search focuses on research topics related to automatic text summarization comprehensively, to find out the current technological developments, Indonesian automatic text summarization to find out the development of the research in Indonesia, and measurements for text readability, especially for the Indonesian language. There are 52 studies from 253 studies included in this systematic review. The next activity after various related works have been collected is reviewing and investigating methods, datasets, and evaluations carried out for automatic summary results. Then the results of these investigations are mapped so that they can become a concept for measuring text readability that can be implemented on the results of automatic summarization.

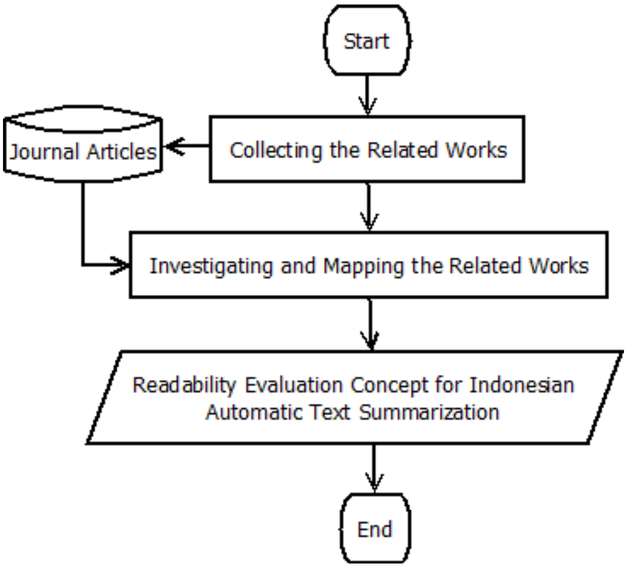
Start
Journal Articles
Collecting the Related Works
Investigating and Mapping the Related Works
Readability Evaluation Concept for Indonesian Automatic Text Summarization
End

Fig. 1. Research Activities

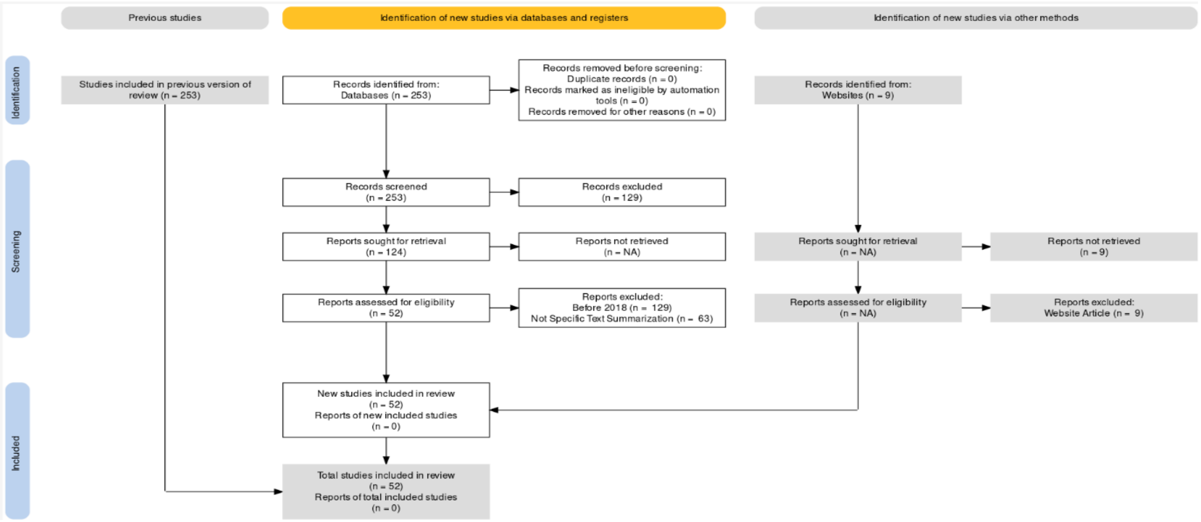
Previous studies
Identification of new studies via databases and registers
Identification of new studies via other methods
Identification
Screening
Included
Studies included in previous version of review (n = 253)
Records identified from: Databases (n = 253)
Records removed before screening: Duplicate records (n = 0) Records marked as ineligible by automation tools (n = 0) Records removed for other reasons (n = 0)
Records identified from: Websites (n = 9)
Records screened (n = 253)
Records excluded (n = 129)
Reports sought for retrieval (n = 124)
Reports not retrieved (n = NA)
Reports sought for retrieval (n = NA)
Reports not retrieved (n = 9)
Reports assessed for eligibility (n = 52)
Reports excluded: Before 2018 (n = 129) Not Specific Text Summarization (n = 63)
Reports assessed for eligibility (n = NA)
Reports excluded: Website Article (n = 9)
New studies included in review (n = 52) Reports of new included studies (n = 0)
Total studies included in review (n = 52) Reports of total included studies (n = 0)

Fig. 2. PRISMA flow diagram

3. Result and Discussion

This section presents the result of the review and investigation of several sources of literature or related works which will then be discussed in this research. The presented results and discussion are organized according to the overall development of automatic text summarization, the development of Indonesian automatic text summarization, readability evaluation for text used for automatic text summarization, and discussion of the concept of readability evaluation for Indonesian automatic text summarization.

A. Readability Evaluation for Automatic Text Summarization

Based on Table 1, the most automatic text summarization evaluation that is used is ROUGE metrics with its variations. There is limited research that uses other evaluation metrics besides ROUGE, especially readability metrics. Moreover, there is a survey for automatic text summarization in many languages, including Hindi, Punjabi, Kannada, Assamese, Konkani, Nepali, Tamil, Marathi, Odia, Sanskrit, Sindhi, Telugu and Gujarati, Bengali, Malayalam, Arabic, Chinese, Greek, Persian, Turkish, Spanish, Czech, Rome, Urdu, Indonesian, and many more do not use readability evaluation [37]. Those researchers use ROUGE, similarity evaluation, precision, recall, f-measure, accuracy, and only a view that involves a reader or expert to evaluate the summary result. There are three types of summary evaluation [6], [38-41]: co-selection-based analysis, content-based analysis, and human readability evaluation or human evaluation.

1) Co-selection-based analysis

Co-selection-based analysis is a summary evaluation with a reference summary. The co-selection-based evaluation is based on the co-occurrence of terms in the system summary and requires a document comparison summary. The evaluation is carried out by picking the system summary and reference summary's common phrases, respectively. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is widely used for text summarization evaluation. It can be said that ROUGE is always used as a metric for evaluating summary results. Similar to Precision, Recall, and F-measure evaluation, ROUGE is also used for informativeness evaluation. There are 5 ROUGE main types, including ROUGE-N, ROUGE-S, ROUGE-SU, ROUGE-L, and ROUGE-W [6], [42].

ROUGE-N is an N-gram measurement between summary result and summary reference collections, where N is determined by N-gram's length, such as ROUGE-1 for unigram and ROUGE-2 for bigram. ROUGE-S is a skip-bigram co-occurrence statistics that evaluate the proportion of skip bigram, where any word pair in the sentence is the skip bigram for random gaps. The ROUGE-SU is developed from ROUGE-S that evaluates the average between ROUGE-1 (unigram) and ROUGE-S, and it extends ROUGE-S with counting unit as a unigram. ROUGE-L computes the LCS metric, where LCS is the maximum length of a common subsequence for two given sequences X and Y. ROUGE-W is a development of ROUGE-L that can be computed as weighted longest common subsequence metrics using dynamic programming. ROUGE-WE then extends ROUGE by employing soft lexical matching based on Word2Vec cosine similarity [43].

Besides ROUGE, the summary result evaluation that always conducted as a co-selection-based analysis of Precision, Recall, F-measure, and Accuracy, which is also accomplished with reference summary [6], [36], [44]. Precision is equal to the sum of retrieved correct sentences (RC) divided by the sum of retrieved correct sentences and incorrect sentences in the document (RI), where Precision determines whether the sentences chosen by the human and produced by the system are correct. The Recall is equal to the sum of retrieved and non-retrieved correct sentences (NC) in a document, where Recall evaluates the proportion of sentences chosen by humans that are produced by the system. Precision and recall are combined in the F-measure (F1 score), while Accuracy is the ratio of total correct and incorrect retrieved text divided by the total text in a document. Precision, Recall, F-measure, and Accuracy have mathematical formula (1)-(4).

$$precision = \frac{|RC|}{|RC|+|RI|} \tag{1}$$

$$recall = \frac{|RC|}{|RC|+|NC|} \tag{2}$$

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{3}$$

$$accuracy = \frac{|RC|+|NI|}{|D|} \tag{4}$$

Another co-selection-based analysis such as: (1) SUMMAC evaluation [45]; (2) S^3 [46] is a model-based measure that predicts the evaluation score using previously proposed evaluation metrics as input features, such as ROUGE and ROUGE-WE. The model was trained using human judgment datasets from TAC conferences; (3) BERTScore calculates similarity scores by token-level alignment of generated and reference summaries. The cosine similarity between contextualized token embeddings from BERT is maximized by computing token alignments greedily [47]; (4) BLEU (Bilingual Evaluation Understudy) is a method of evaluating automatic machine translation that is fast, inexpensive, and language-independent, correlates favorably with human evaluation, and has a low marginal cost per run [48]; (5) BARTScore is a conceptually simple and empirically effective system with a variety of versions that may be used unsupervised to evaluate text from many angles (e.g. informativeness, fluency, or factuality) [49]; (6) MoverScore [50], the word mover's distance is used to calculate the semantic distance between a summary and a reference text using n-gram embeddings aggregated from BERT representations; (7) BLEURT is an evaluation with a few thousand potentially biased training examples, a learned assessment metric based on BERT can simulate human judgments. An innovative pre-training technique that leverages millions of synthetic instances to help the model generalize is a major component of this approach [51]; (8) PRISM considers the task of evaluating machine translation output as one of grading it with a sequence-to-sequence paraphraser based on a human reference; (9) If there is no reference summary collection to compare, PYRAMID can be used as a metrics evaluation [6], [52]; and many more evaluation metrics that can use for automatic text summarization.

2) Content-based analysis

Content-based evaluation is conducted without reference summary to evaluate the readability of the summary result. Readability is an important parameter to measure the performance of automatic text summarization that makes sure the summary result can be read and understandable. The readability of text can be evaluated based on content and relatedness between sentence aspects [35]. Readability between reference summary and summary result from the system depends on the complexity of syntax and vocabulary of the language and whether the relatedness between sentences (between previous and next sentence) shows the fluency of reading.

Readability can be evaluated using content-based analysis metrics evaluations, such as FKGL, GFI, SMOGI, CLI, ARI, and RPS [35], [36], [44]. All of those metrics are developed for English but can be used for another language also, even Hindi [35]. Flesch-Kincaid Grade Level (FKGL) uses word complexity and sentence length to evaluate the readability of text, where complex words and longer sentences influence the reader's concentration to understand the meaning of the text [53-54]. FKGL has several processes [55]: calculate the average number of words per sentence, calculate the average number of syllables per word, and then calculate FKGL using the formula (5). Gunning Fog Index (GFI) is used to count three or more syllables of words and determine the grade levels with the total number of a sentence [54], where the calculation formula is available in formula (6). SMOG Index (SMOGI) counts polysyllabic words in a fixed number of sentences and gives an index of the relative difficulty of the text [54], [56]. The processes of SMOG begin by defining a sentence as a string of words punctuated with an until count for all the polysyllabic words and the number of the sentence in the text using formula (7), where SMOGI has a conversion table to compare the number of polysyllabic words with approximate grade level [57].

$$FKGL = 0.39\left(\frac{W}{|Sum|}\right) + 11.8\left(\frac{Syl}{W}\right) - 15.59 \tag{5}$$

$$GFI = 0.4\left[\left(\frac{W}{|Sum|}\right) + 100\left(\frac{cw}{W}\right)\right] \tag{6}$$

$$SMOGI = 1.0430 \times \sqrt{\left(Syl \times \frac{30}{|Sum|}\right)} + 3.1291 \tag{7}$$

Instead of counting syllables and sentence length, the Coleman-Liau Index (CLI) only uses length in characters in text because, in this formula estimate, the word length in letters is better than the word length in syllables to evaluate readability [58]. Therefore, CLI has a simple formula such as formula (8) that only uses the average number of letters per 100 words and the average number of sentences per 100 words [59]. Automated Readability Index (ARI) is a readability measurement to evaluate reading difficulty levels of text [60]. Similar as CLI, ARI produces a number that represents the age needed to understand the text with Formula (9) [61]. Last, Relatedness with Previous Sentence (RPS) is the sentence readability measurement that developed based on cosine similarity with a formula (10) [36].

$$CLI = 0.0588 \times AC - 0.296 \times AS - 15.8 \tag{8}$$

$$ARI = 4.71\left(\frac{C}{W}\right) + 0.5\left(\frac{W}{|Sum|}\right) - 21.43 \tag{9}$$

$$Readability(Sum) = \frac{\sum_{j \epsilon Sum} Sim_{cos}(S_j, S_{j+1})}{\max Sim_{cos}(S_j)} \tag{10}$$

Other readability metrics that can be used such as Flesch Reading Ease [62-63], FORCAST [64], Fry Graph [65], New Dale-Chall [66-67], New Fog Count [68], Raygor Estimate [69], and Linsear Write Calculation [70]. Those readability metrics are used for English, but several are suitable for another language, including Indonesian. The top five best measurements that are widely used include FKGL, Flesh Reading Ease, GFI, SMOGI, and FORCAST [71].

3) Human readability testing

Human evaluation is also important to ensure the readability of the summary result. Human readability evaluation is conducted with expert/ native/ reader. Evaluation involving language experts is also important but rarely done. This is related to the limited time and expert resources to measure the readability of the text one by one, especially for a large number of documents. Elements of subjectivity can also affect the results of this manual measurement. Therefore, manual measurements are better carried out by experts in odd numbers and have no personal interest.

The evaluation criteria for measuring text readability are quite varied. It differs from one researcher to another. However, automated text summarization research that performs human evaluation provides several scoring criteria. The experts will give a rating for each summary result produced by the system. The ratings are readable, partially readable, and non-readable with several conditions as consideration [34], [36], [72]: (1) the summary should be understandable, non-redundant, and focused on the main topic; (2) summary sentences should be complete and related to one another; and (3) the summary should not include complex sentences. If the summary meets all of those criteria, it is considered readable. If the summary meets half of those requirements, it is classified as partially readable. Otherwise, it is classified as unreadable.

Because human judgement is not always consistent, measuring inter-judge agreement between experts can be conducted. It can use Kappa statistics to measure how well the expert judges agree on readability with the formula (11).

$$\text{kappa (K)} = \frac{\mathrm{P(A)} - P(E)}{1 - \mathrm{P(E)}} \tag{11}$$

Where P(A) is a proportion of the time, the judges agreed, and P(E) is the proportion of the time the judges agree by chance. If the value of the Kappa formula is between 0.67-1, then the judgment is acceptable.

B. Development of Automatic Text Summarization Research

Automatic text summarization has been studied since the late 1950s. Since then, numerous scholars have been working on the topic of automatic text summarizing to find new ways to automate text summarization. This chapter provides an overview of the text summarizing research. It encompasses text summarization research (other than the Indonesian language) using several types of summaries, including extractive and abstractive summarization, at least in the last five years. The methodologies and assessments that were employed, as well as some of the places where text summarization was used, are also presented in this chapter. Table 1 shows the summary of the development of automatic summarization research.

Table 1. Summary of Development of Automatic Text Summarization Research

Methods	Evaluations			Dataset	Language	Ref.
	Co-selection-based Analysis	Content-based Analysis	Human Evaluation			
Graph-based using Itemset Mining and Sentence Clustering	✓	N/A	N/A	BioMed Central's open access corpus	English	[73]
Statistical Model TF-IDF, and Deep Learning using Seq2Seq Model	✓	N/A	N/A	Titles and Abstracts of the Web of Science	English	[41]
Encoder-Decoder Long Short-Term Memory (LSTM)	N/A	N/A	N/A	DUC 2005, NewsIR '16	English	[74]
Betweeness Centrality	✓	N/A	N/A	DUC 2002	English	[75]
Pattern-Growth Sentence Compression	✓	N/A	N/A	Malay News Dataset	Malay	[76]
Adaptive Neuro-Fuzzy Inference System (ANFIS), K-Means and Hierarchical Clustering (HC)	✓	N/A	N/A	DUC 2002	English	[77]
Sentence Scoring for Internal and external information	✓	N/A	N/A	DUC 2001, and DUC 2002	English	[78]
Convolutional Neural Network (CNN)	✓	N/A	N/A	DUC 2002	English	[26]
Fuzzy Analysis	✓	N/A	N/A	Virtual Learning Environment (VLE) dataset for Portuguese	Portuguese	[79]
LSTM	✓	N/A	N/A	Amazon Fine Food Reviews	English	[80]
Semantic Role Labeling (SLR) and Explicit Semantic Analysis (ESA)	✓	N/A	N/A	DUC 2002	English	[81]
Convolutional Neural Network	✓	N/A	N/A	DUC 2002	English	[26]
Unsupervised Neural networks using Auto-Encoder	✓	N/A	N/A	EASC Dataset, Summarization and Keyword Extraction (SKE) from Emails Dataset	Arabic, and English	[82]
Adaptive, knowledge-based event-index cognitive mode	✓	N/A	N/A	DUC 2001	English	[8]
TextRank, LexRank, ChunkRank, Luhn, LSA, Edmundson, TGraph, NN-ED, NN-SE, UniRank, FE-SE, SummaRuNNer, and MMR-SE	✓	✓	✓	Forum of Information Retrieval Evaluation (FIRE) conference data (2011)	Hindi and English	[35]
Optimal Combination of Sentence Scoring	✓	✓	✓	DUC 2006 and DUC 2007	English	[36]
F-RBM (Fuzzy Restricted Boltzmann Machine)	✓	N/A	N/A	UC Irvine Machine Learning Repository, BBC news and DUC 2004	English	[83]
EdgeSumm (combination of graph-based, statistical-based, semantic-based, and centrality-based)	✓	N/A	N/A	DUC 2001 and DUC 2002	English	[84]
Feature-based approach with sentence clustering	N/A	✓	N/A	TripAdvisor.com	English	[85]
TextRank and Recurrent Neural Network (RNN)	✓	N/A	N/A	ScienceDirect article 2012-2018	English	[86]
Modified of PageRank	✓	N/A	N/A	Essex Arabic Summaries Corpus (EASC)	Arabic	[87]
Textual Graph and Maximum Independent Sets.	✓	N/A	N/A	DUC 2002 and DUC 2004	English	[88]
Vector Space Model (VSM)	✓	N/A	N/A	BBC-Urdu (collected by own)	Urdu	[89]
CNN	✓	N/A	N/A	Multilingual Single-document Summarization (MSS)	English, Malayalam, and Hindi	[90]
Seq2Seq Model	✓	N/A	N/A	Chinese dataset LCSTS (Large-scale Chinese Short Text Summarization)	Chinese	[91]
Weighted word embedding based method	✓	N/A	N/A	DUC 2007	English	[92]
Tagged-Latent Dirichlet Allocation (Tagged-LDA)	✓	N/A	N/A	Hindi Novels and Stories	Hindi	[93]
T-BERTSum based on BERT	✓	N/A	N/A	CNN/Daily mail and XSum dataset	English	[94]

Methods	Evaluations			Dataset	Language	Ref.
	Co-selection-based Analysis	Content-based Analysis	Human Evaluation			
Deep Learning Modified Neural Network Classifier (DLMNN)	✓	N/A	N/A	DUC (not specific)	English	[95]
Combine the LDA with classification technique	✓	N/A	N/A	DUC 2001, DUC 2002, DUC 2006, and DUC 2007	English	[96]
TextRank and Bayesian Additive Regression Trees (BART)	✓	N/A	N/A	CNN/Daily mail	English	[97]

From Table I, automatic text summarization research are developed rapidly, with Recall-Oriented Understudy for Gisting Evaluation (ROUGE) as the most used evaluation metric. Based on this research, ROUGE is used for multiple languages. ROUGE evaluates the performance of summary results that are produced automatically from the system and compared to the reference summary. Readability evaluation focuses on the content of the summary result whether the meaning of the summary result is not lost and whether the structure of the text is easy to understand and read. Therefore, it is not enough to evaluate the summary results only with ROUGE to determine the readability of the summary results. And of most automated text summarization studies, the readability evaluation of summaries is very limited.

C. Development of Indonesian Automatic Text Summarization Research

Indonesian automatic text summarization has also developed rapidly. Many techniques used begin with basic or common text summarization, such as sentence scoring, then graph-based, until using a machine learning technique. Today, deep learning is also already used for Indonesian automatic text summarization. Also, the dataset is collected from various sources, such as the abstract of a scientific article, social media, and most of the research uses news article documents. Some research related to Indonesian automatic text summarization is available in Table 2.

Table 2. Summary of Development of Indonesian Automatic Text Summarization Research

Methods	Evaluations			Dataset	Ref.
	Co-selection-based Analysis	Content-based Analysis	Human Evaluation		
Vector Space Model (VSM)	✓	N/A	N/A	Indonesian publication articles (collected by own)	[98]
Graph method and Ant Colony algorithm	✓	N/A	✓	Not clear	[99]
Sentence Scoring and Decision Tree	✓	N/A	N/A	Indonesian news articles (collected by own)	[12]
NeuralSum	✓	N/A	N/A	Indonesian news articles called IndoSum (collected by own)	[31]
Cross Latent Semantic Analysis	✓	N/A	N/A	Indonesian news articles (collected by own)	[100]
Semantic Network using Maximum Marginal Relevance (MMR)	✓	N/A	N/A	Indonesian news articles (collected by own) and WordNet Bahasa	[101]
Bidirectional Gated Recurrent Unit (BiGRU)	✓	N/A	N/A	Indonesian Journal document (collected by own)	[102]
Graph Convolutional Network	✓	N/A	N/A	Indonesian news article (collected by own)	[28]
IndoBERT Indonesian version of Bidirectional Encoder Representations from Transformers (BERT)	✓	N/A	N/A	IndoSum and Liputan6 (Indonesian news articles that collected by own)	[30]
Bellman-Ford Algorithm	✓	N/A	N/A	Sahih Bukhari Muslim Hadith (collected by own)	[22]
TextRank	✓	N/A	N/A	al-Misbah interpretation book (collected by own)	[17]
Bellman-Ford Algorithm	✓	N/A	N/A	Indonesian publication articles (collected by own)	
Decoder-Encoder model called IndoBART	✓	N/A	N/A	Liputan6	[21]
BERT	✓	N/A	N/A	IndoSum	[103]
TextRank	✓	N/A	N/A	Undiksha Academic Information System	[104]
Multi-featured based on the regression model	✓	N/A	N/A	IndoSum	[105]
Cosine Similarity and Maximum Marginal Relevance (MMR)	✓	N/A	N/A	Health Ethics Protocol Document (collected by own)	[106]
LSTM and Gated Recurrent Units (GRUs)	✓	N/A	N/A	Indonesian publication articles (collected by own)	[107]
Feature-based POS tagging and sentence relevance	✓	N/A	N/A	11 groups of Indonesian news documents (collected by own)	[108]
BERT	✓	N/A	N/A	IndoSum	[109]

Methods	Evaluations			Dataset	Ref.
	Co-selection-based Analysis	Content-based Analysis	Human Evaluation		
Support Vector Machine (SVM) and K-Nearest Neighbor (KNN)	✓	N/A	N/A	Game Stam Review Document (collected by own)	[110]

Indonesian automatic text summarization research is currently quite developed. Starting from the provision of datasets that are currently a benchmark for research on automatic summarization of Indonesian texts such as IndoSum [31] and Liputan6 [30]. Likewise with the various methods used, currently the use of deep learning is becoming popular for use in Indonesian automatic text summarization. However, in the aspect of evaluating summary results, especially readability, of the many Indonesian automatic text summarization studies, none has focused on evaluating the readability. Most of them use ROUGE, similarity evaluation, recall, precision, and F-measure value [12], [28], [31], [111-114]. While the readability is simply evaluated by the expert reader or native without specific metrics for readability [112], [115-116]. Because, to evaluate readability is not only to evaluate the performance of summary results, such as the use of ROUGE.

D. Concept of Readability Evaluation for Indonesian Automatic Text Summarization

Readability evaluation of text summary results is a big challenge in this research. During this research, there was very limited research on automatic text summarization that focused on the readability of summary results. In fact, there is no automatic text summarization research for the Indonesian language that focuses on evaluating the readability of the summary results (also based on Table 2). Therefore, this research contributes to starting the importance of readable automatic text summary results for the Indonesian language. The difference between this study and other systematic reviews of automatic text summarization is this study reveals the importance of measuring the readability quality of automatic text summarization results. However, the fact shows that there are not many automatic text summary studies that are concerned with measuring the readability of the generated summary. Therefore, from the results of a systematic review and further exploration of the measurement of the readability of Indonesian texts, this study resulted in a novel concept for evaluating the readability of Indonesian automatic text summarization results.

As previously stated, in the Indonesian automatic text summarization research, no one has focused on evaluating the readability of the summary results. However, there is a study that assessed the readability of an Indonesian language website conducted by Biddinika et al [71]. Because measuring the readability of summary results is not enough only with co-selection-based analysis using ROUGE, BLEU, and other metrics, but also content-based analysis is necessary to evaluate the readability of Indonesian text. Actually, SMOGI, GFI, and FKGL are used for English. In many research those readability metrics are used for English, but in Indonesian language research metrics that suitable are FKGL, SMOGI, and GFI [71]. Since evaluation with SMOGI can be effectively carried out with a minimum of 30 sentences, the summary result should be more than 30 sentences. Then came GFI and FKGL, which were chosen for their advantage of simple calculations and interpretations rather than their popularity among the other participants in the readability study [117].

Because this research focuses on the Indonesian language, there should be a text readability measurement that is specifically used for the Indonesian language. The available metrics are very limited. Several Indonesian language studies use English readability metrics to evaluate the readability of Indonesian text and most of them evaluate the readability with survey. Those studies : (1) evaluate the readability of Indonesian text using GFI and Cloze technique [118]; (2) conduct readability evaluation of Indonesian text using gap test and fry chart or Cloze test, but this research mentions that SMOGI and FKGL can be used to evaluate Indonesian text [119]; (3) FKGL, GFI, and SMOGI can use to evaluate readability of Indonesian text, although in this research use fry graphic and survey [120]; and (4) GFI also used to evaluate readability of Indonesian text [121] [122].

Of the many Indonesian studies that use surveys, questionnaires, and several measuring tools for English, there is a specific measuring instrument for the Indonesian language made by Dwiyanto Djoko Pranowo [71], [123]. There are thirteen indicators to evaluate the readability of Indonesian text with Dwiyanto's metrics. The thirteen indicators are categorized into three, namely easy, medium, and difficult to read. By adding up all indicators, based on the range of criteria values, the readability of the Indonesian text can be determined. Table 3 and formula (12) provide the calculation of Dwiyanto's indicators.

Table 3. Readability Score of Dwiyanto's Evaluation

Indicators	Criteria	Category	Score
The typical number of paragraphs	≤ 5 6 ≥ 7	Easy Medium Hard	1 2 3
The average number of sentences in each paragraph	6.0-7.1 4.7-5.9 3.6-4.6	Easy Medium Hard	1 2 3
A sentence's length	7.2-8.5 8.6-9.8 9.9-11	Easy Medium Hard	1 2 3
Percentage of sentences that were extended	79.0-85.6% 85.7-92.4% 92.3-99%	Easy Medium Hard	1 2 3
Compound sentence percentage	38-42% 43-46% 47-50%	Easy Medium Hard	1 2 3
The percentage of sentences that contain polysemy	44.2-56.3% 32.1-44.1% 19.9-32%	Easy Medium Hard	1 2 3
Passive sentence percentage	11.3-17.7% 17.8-24.1% 24.2-30.5%	Easy Medium Hard	1 2 3
Percentage of words that are unfamiliar	7.5-11.6% 11.7-15.7% 15.8-19.7%	Easy Medium Hard	1 2 3
The proportion of abstract words	15-20.7% 20.8-26.4% 26.5-32.2%	Easy Medium Hard	1 2 3
Terms as a percentage	1.4-4.7% 4.8-8.1% 8.2-11.4%	Easy Medium Hard	1 2 3
The proportion of conjunctions	3-4.4% 4.5-5.9% 6-7.3%	Easy Medium Hard	1 2 3
Loan word percentage	1.7-2.7% 2.8-4.8% 4.9-7.8%	Easy Medium Hard	1 2 3
Phrase percentage	2.2-3.1% 3.2-4% 4.1-4.9%	Easy Medium Hard	1 2 3

$$Dwiyanto's = \sum_{i=1}^{13} indicator_i \qquad (12)$$

Where, the conversion value of Dwiyanto's measurement is 13.0-21.7 means Easy, 21.8-30.5 means Medium, and 30.6-30 means Hard to read.

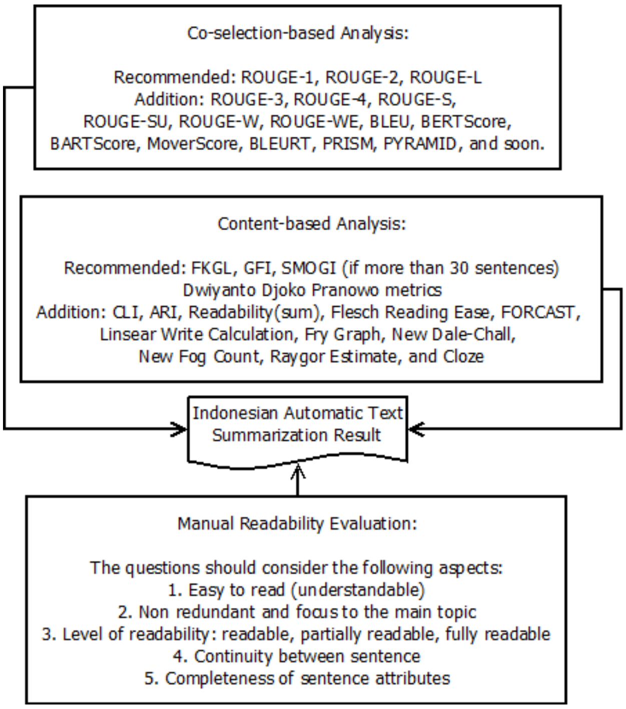
Co-selection-based Analysis:
Recommended: ROUGE-1, ROUGE-2, ROUGE-L
Addition: ROUGE-3, ROUGE-4, ROUGE-S,
ROUGE-SU, ROUGE-W, ROUGE-WE, BLEU, BERTScore,
BARTScore, MoverScore, BLEURT, PRISM, PYRAMID, and soon.
Content-based Analysis:
Recommended: FKGL, GFI, SMOGI (if more than 30 sentences)
Dwiyanto Djoko Pranowo metrics
Addition: CLI, ARI, Readability(sum), Flesch Reading Ease, FORCAST,
Linsear Write Calculation, Fry Graph, New Dale-Chall,
New Fog Count, Raygor Estimate, and Cloze
Indonesian Automatic Text
Summarization Result
Manual Readability Evaluation:
The questions should consider the following aspects:
1. Easy to read (understandable)
2. Non redundant and focus to the main topic
3. Level of readability: readable, partially readable, fully readable
4. Continuity between sentence
5. Completeness of sentence attributes

Fig. 3. Concept of Readability Evaluation for Indonesian Automatic Text Summarization

Then, involving language experts in assessing the readability of Indonesian automatic text summarization results is also important. Figure 3 shows the concept or scenario of automatic text summary readability evacuation for the Indonesian language based on a review of previous research. Starting from minimal co-selection-based analysis and content-based analysis that can be used, to human evaluations that involve Indonesian language experts. ROUGE-1, ROUGE-2 and ROUGE-L are minimal metrics that can be used to measure the performance of summary results, while FKGL, GFI, SMOGI (for summaries that more than 30 sentences), and Dwiyanto's can be used to measure the readability of summary results. Then, human readability evaluation can be done by involving Indonesian language experts. The human assessment carried out by experts in measuring the readability of the automatic summary results does not actually have a standard format, but the assessment can consider several aspects of readability such as: (1) the question that aims to get the opinion about whether summary result is easier to read or directly easy to understand the content; (2) the question that aims to know the opinion of readability level of summary result that have are three options: Readable, Partially Readable, or Unreadable [35]; (3) the question that aims to know whether the evaluators thinks that the summary result contains content that is in accordance with the main topic [35]; (4) the question that aims to see the continuity between sentences in the summary results because continuity between sentences is one indicator of the readability of a text [35]; and (5) the question that aims to know whether summary result has complete sentence attributes so that it becomes a text that is both structurally and the content conveyed can be easily read and understood.

The limitation of research in measuring the readability of automatic summary results in Indonesian is the limited readability measurement metrics for Indonesian text. Only Dwiyanto's metric currently exists to measure the readability of Indonesian text, other measurements are adaptations of readability metrics for English. Indonesian lacks well-established readability models compared to languages like English, where metrics such as the Flesch-Kincaid score or Gunning Fog Index are widely used. The absence of such models for the Indonesian language likely discourages researchers from prioritizing readability evaluation in their studies. Readability is a key indicator of a summary's usefulness, and neglecting it can lead to an incomplete or inaccurate evaluation of a summarization system's performance. While the systematic review covers a broad range of sources, it is possible that relevant studies published in non-indexed databases or gray literature (e.g., conference papers, reports) were missed. This could limit the comprehensiveness of the findings.

Therefore, future research should focus on developing readability metrics specifically tailored to the Indonesian language, accounting for its unique linguistic features. This would provide a foundation for more comprehensive evaluations in future summarization studies. Future studies could investigate how to incorporate readability directly into the training and optimization of text summarization models. This might involve integrating human-like readability scores into the training objectives to create models that prioritize clarity and ease of understanding. By addressing these gaps, future research could help ensure that automatic summarization systems generate summaries that are not only accurate but also easy to read and understand, enhancing their usability in real-world applications.

4. Conclusion

The readable summary is an important aspect that should be reached in the automatic text summarization technique. In the context of Indonesian text summarization research, readability evaluation has received limited attention, with most studies relying solely on co-selection-based analysis approaches, such as ROUGE metrics, and minimal human evaluation. Therefore, it is an opportunity and challenge to evaluate Indonesian summary results not only using a co-selection-based analysis and human readability approach but also using a content-based analysis approach. So, the evaluation of the readability of Indonesian automatic text summarization is complete and comprehensive. This research provides the concept of Indonesian automatic text summarization in the aspect of readability based on previous works. There are ROUGE-1, ROUGE-2, and ROUGE-L that are recommended for co-selection-based analysis for Indonesian. FKGL, GFI, SMOGI, and Dwiyanto's for content-based analysis. Then five aspects are considered for human readability evaluation with language experts: ease of reading, level of readability, continuity between sentences, relation with the main topic, and the completeness of sentence attributes. For further works, this concept can be implemented to evaluate the Indonesian automatic summary result, so that the readability aspect of the summary can be reached. The practical implications of these findings are significant for both developers and researchers in the field of automatic summarization. For developers, implementing this comprehensive framework can lead to the creation of summarization systems that not only generate concise summaries but also ensure they are easily understandable to readers. For researchers, this study highlights the need to integrate more nuanced and holistic readability metrics into their evaluation processes, paving the way for more user-centered, effective summarization tools.

Acknowledgment

The author would like to thank for the support of the Center for Advanced Computing Technology, Universiti Teknikal Malaysia Melaka so that this article can be published. Also thank you for support from the academic community of UIN Sunan Gunung Djati Bandung.

References

[1] A. Kumar, M. Xu, and Z. Luo, "Text Summarization using Natural Language Processing," Worcester Polytechnic Institute, 2018.

[2] G. G. Chowdhury, "Natural language processing," *Annual Review Info Sci Amp. Tec*, vol. 37, no. 1, pp. 51–89, Jan. 2003, doi: 10.1002/aris.1440370103.

[3] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, Jul. 2015, doi: 10.1126/science.aaa8685.

[4] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: An introduction," *J. Am. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 544–551, Oct. 2011, doi: 10.1136/amiajnl-2011-000464.

[5] E. Roche and Y. Schabes, "Finite-state language processing," *Comp. Mathem. Applic.*, vol. 35, no. 9, Art. no. 142, May 1998, doi: 10.1016/S0898-1221(98)90701-5.

[6] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," *Artif Intell Rev*, vol. 47, pp. 1–66, Jan. 2017, doi: 10.1007/s10462-016-9475-9.

[7] V. Gupta and G. S. Lehal, "A Survey of Text Summarization Extractive techniques," *JETWI*, Aug. 2010, pp. 258–268. doi: 10.4304/jetwi.2.3.258-268.

[8] M. Rajangam and C. Annamalai, "Extractive document summarization using an adaptive, knowledge based cognitive model," *Cogn. Syst. Res.*, vol. 56, pp. 56–71, Mar. 2019, doi: 10.1016/j.cogsys.2018.11.005.

[9] N. R. Kasture, N. Yargal, N. N. Singh, N. Kulkarni, and V. Mathur, "A Survey on Methods of Abstractive Text Summarization," *Int. J. Res. Emerg. Sci. Techn.*, vol. 1, no. 6, pp. 54–57, Nov. 2014.

[10] J. Jayabharathy, S. Kanmani, and Buvana, "Multi-document Summarization Based on Sentence Features and Frequent Itemsets," in *Adv. Comp. Sci., Eng. Applic.*, vol. 166, D. C. Wyld, J. Zizka, and D. Nagamalai, Eds., *Adv. Intellig. Soft Comput.,* vol. 166, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 657–671. doi: 10.1007/978-3-642-30157-5_66.

[11] T. Sri, R. Raju, and B. Allarpu, "Text Summarization using Sentence Scoring Method," *Int. J. Res. Emerg. Sci. Techn.*, vol. 4, no. 4, pp. 1777–1779, Apr. 2017.

[12] P. M. Sabuna and D. B. Setyohadi, "Summarizing Indonesian text automatically by using sentence scoring and decision tree," in *2017 2nd Int. Conf. Inform. Technol., Inform. Sys. Electr. Engin. (ICITISEE)*, Yogyakarta: IEEE, Nov. 2017, pp. 1–6. doi: 10.1109/ICITISEE.2017.8285473.

[13] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, "Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion," *Inf. Process Manag.*, vol. 43, no. 6, pp. 1606–1618, Nov. 2007, doi: https://doi.org/10.1016/j.ipm.2007.01.023.

[14] M. G. Ozsoy, F. N. Alpaslan, and I. Cicekli, "Text summarization using Latent Semantic Analysis," *J. Inform. Sci.*, vol. 37, no. 4, pp. 405–417, Aug. 2011, doi: 10.1177/0165551511408848.

[15] K. Merchant, "NLP Based Latent Semantic Analysis for Legal Text Summarization," in *Int. Conf. of ICACCI*, pp. 1803–1807, Sep. 2018, doi: https://doi.org/10.1109/ICACCI.2018.8554831.

[16] K. Merchant and Y. Pande, "NLP Based Latent Semantic Analysis for Legal Text Summarization," in *2018 Int. Conf. Adv. Comp., Communic. Inform. (ICACCI)*, Bangalore: IEEE, Sep. 2018, pp. 1803–1807. doi: 10.1109/ICACCI.2018.8554831.

[17] M. F. Fakhrezi, M. A. Bijaksana, and A. F. Huda, "Implementation of Automatic Text Summarization with TextRank Method in the Development of Al-Qur'an Vocabulary Encyclopedia," *Procedia Comput Sci*, vol. 179, pp. 391–398, Jan. 2021.

[18] P. Wongchaisuwat, "Automatic Keyword Extraction Using TextRank," in *2019 IEEE 6th Int. Conf. Industr. Engin. Applicat. (ICIEA)*, Tokyo, Japan: IEEE, Apr. 2019, pp. 377–381. doi: 10.1109/IEA.2019.8714976.

[19] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *J. Artf. Intell. Res.*, vol. 22, pp. 457–479, Dec. 2004.

[20] D. S. Maylawati, C. N. Alam, M. F. Muharram, M. A. Ramdhani, A. S. Amin, and H. Aulawi, "The Purpose of Bellman-Ford Algorithm to Summarize the Multiple Scientific Indonesian Journal Articles," in *Int. Conf. of ICWT*, IEEE, Nov. 2020, pp. 1–5.

[21] M. F. Muharram *et al.*, "Automatic Text Summarization for Multiple Scientific Indonesian Journal Article using Bellman-Ford Algorithm," in *Int. Conf. of ICIIC*, Melaka: Universiti Teknikal Malaysia Melaka, Sep. 2021, pp. 1-6.

[22] W. W. Adytoma *et al.*, "Automatic Text Summarization for Hadith with Indonesian Text using Bellman-Ford Algorithm," in *Int. Conf. of ICCED*, IEEE, Oct. 2020, pp. 1–6.

[23] M. Azhari and Y. Jaya Kumar, "Improving text summarization using neuro-fuzzy approach," *J. Inf. Technol. Commun.*, vol. 1, no. 4, pp. 367–379, Aug. 2017, doi: 10.1080/24751839.2017.1364040.

[24] J. ge Yao, X. Wan, and J. Xiao, "Recent advances in document summarization," *Knowl. Inf. Sys.*, vol. 53, pp. 297–336, Mar. 2017, doi: 10.1007/s10115-017-1042-4.

[25] G. Rossiello, "Neural abstractive text summarization," in *CEUR Workshop Proceedings*, Nov. 2016.

[26] W. H. Alquliti and N. Binti, "Convolutional Neural Network based for Automatic Text Summarization," *Int. J. Adv. Comp. Sci. Appl.*, vol. 10, no. 4, pp. 200 – 211, Apr. 2019, doi: 10.14569/IJACSA.2019.0100424.

[27] K. Yao, L. Zhang, T. Luo, and Y. Wu, "Deep reinforcement learning for extractive document summarization," *Neurocomputing*, vol. 284, April, pp. 52–62, Apr. 2018, doi: 10.1016/j.neucom.2018.01.020.

[28] G. Garmastewira and M. L. Khodra, "Summarizing Indonesian news articles using Graph Convolutional Network," *J. Inform. Commun. Techn.* , vol. 18, no. 3, pp. 345–365, Jun. 2019, doi: 10.32890/jict2019.18.3.6.

[29] E. Karari Kinyanjui, M. Ndenga, and H. O. Nyongesa, "Hybridization of DBN with SVM and its Impact on Performance in Multi-Document Summarization," *Mach. Learn. Applic. Int. J.*, vol. 8, no. 2/3, pp. 37–51, Sep. 2021, doi: 10.5121/mlaij.2021.8304.

[30] F. Koto, J. H. Lau, and T. Baldwin, "Liputan6: A Large-scale Indonesian Dataset for Text Summarization," 2020, *arXiv*. doi: 10.48550/ARXIV.2011.00679.

[31] K. Kurniawan and S. Louvan, "Indosum: A New Benchmark Dataset for Indonesian Text Summarization," in *2018 Int. Conf. Asian Lang. Process. (IALP)*, Bandung, Indonesia: IEEE, Nov. 2018, pp. 215–220. doi: 10.1109/IALP.2018.8629109.

[32] Y. J. Kumar, O. S. Goh, H. Basiron, N. H. Choon, and P. C. Suppiah, "A Review on Automatic Text Summarization Approaches," *J. Comp. Sci.*, vol. 12, no. 4, pp. 178–190, Apr. 2016, doi: 10.3844/jcssp.2016.178.190.

[33] K. Nandhini and S. R. Balasundaram, "Improving readability through extractive summarization for learners with reading difficulties," *Egypt. Inform. J.*, vol. 14, no. 3, pp. 195–204, Nov. 2013, doi: 10.1016/j.eij.2013.09.001.

[34] P. Verma and H. Om, "MCRMR: Maximum coverage and relevancy with minimal redundancy based multi-document summarization," *Expert Syst Appl*, vol. 120, pp. 43–56, Apr. 2019, doi: 10.1016/j.eswa.2018.11.022.

[35] P. Verma, S. Pal, and H. Om, "A Comparative Analysis on Hindi and English Extractive Text Summarization," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 18, no. 3, pp. 1–39, Sep. 2019, doi: 10.1145/3308754.

[36] P. Verma and H. Om, "A novel approach for text summarization using optimal combination of sentence scoring methods," *Sadhana*, vol. 44, no. 5, Art. no. 110, Apr. 2019, doi: 10.1007/s12046-019-1082-4Sad.

[37] Y. Kumar, K. Kaur, and S. Kaur, "Study of automatic text summarization approaches in different languages," *Artif Intell Rev*, vol. 54, no. 8, pp. 5897–5929, Dec. 2021, doi: 10.1007/s10462-021-09964-4.

[38] V. Gupta, N. Bansal, and A. Sharma, "Text Summarization for Big Data: A Comprehensive Survey," in *Int. Conf. of ICICC,* vol. 56, pp. 503–516, Jan. 2019, doi: 10.1007/978-981-13-2354-6.

[39] M. Zopf, E. Loza Mencía, and J. Fürnkranz, "Which Scores to Predict in Sentence Regression for Text Summarization?," in *Proc. 2018 Conf. North American Chapter Assoc. Comput. Linguis.: Human Lang. Techn., Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent, Eds., New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1782–1791. doi: 10.18653/v1/N18-1161.

[40] M. E. Johnson, "Automatic Summarization of Natural Language," 2018, *arXiv*. doi: 10.48550/ARXIV.1812.10549.

[41] M.-Y. Day and C.-Y. Chen, "Artificial Intelligence for Automatic Text Summarization," in *2018 IEEE Int. Conf. Inform. Reuse Integr. (IRI)*, Salt Lake City, UT: IEEE, Jul. 2018, pp. 478–484. doi: 10.1109/IRI.2018.00076.

[42] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[43] J.-P. Ng and V. Abrecht, "Better Summarization Evaluation with Word Embeddings for ROUGE," Aug. 25, 2015, *arXiv*: arXiv:1508.06034. Accessed: Oct. 26, 2024. [Online]. Available: http://arxiv.org/abs/1508.06034

[44] R. Rautray and R. C. Balabantaray, "An evolutionary framework for multi document summarization using Cuckoo search approach: MDSCSA," *Appl. Comput. Inform.*, vol. 14, no. 2, pp. 134–144, Jul. 2018.

[45] I. Mani, G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim, "SUMMAC: a text summarization evaluation," *Nat. Lang. Eng.*, vol. 8, no. 1, pp. 43–68, Jun. 2002.

[46] M. Peyrard, T. Botschen, and I. Gurevych, "Learning to Score System Summaries for Better Content Selection Evaluation.," in *Proc. Works. New Frontiers Summariz.*, L. Wang, J. C. K. Cheung, G. Carenini, and F. Liu, Eds., Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 74–84. doi: 10.18653/v1/W17-4510.

[47] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," Feb. 24, 2020, *arXiv*: arXiv:1904.09675. Accessed: Oct. 26, 2024. [Online]. Available: http://arxiv.org/abs/1904.09675

[48] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. 40th Annual Meet. Assoc. Computat. Ling. - ACL '02*, Philadelphia, Pennsylvania: Association for Computational Linguistics, 2001, p. 311. doi: 10.3115/1073083.1073135.

[49] W. Yuan, G. Neubig, and P. Liu, "BARTScore: Evaluating Generated Text as Text Generation," 2021, *arXiv*. doi: 10.48550/ARXIV.2106.11520.

[50] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, "MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance," Sep. 26, 2019, *arXiv*: arXiv:1909.02622. Accessed: Oct. 26, 2024. [Online]. Available: http://arxiv.org/abs/1909.02622

[51] T. Sellam, D. Das, and A. P. Parikh, "BLEURT: Learning Robust Metrics for Text Generation," May 21, 2020, *arXiv*: arXiv:2004.04696. Accessed: Oct. 26, 2024. [Online]. Available: http://arxiv.org/abs/2004.04696

[52] H. Saggion and T. Poibeau, "Automatic Text Summarization: Past, Present and Future," in *Multi-source, Multil. Inform. Extract. Summariz.* T. Poibeau, H. Saggion, J. Piskorski, and R. Yangarber, Eds., in *Th. Applic. Nat. Lang. Proces.*., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 3–21. doi: 10.1007/978-3-642-28569-1_1.

[53] A. J. Choudhry *et al.*, "Readability of discharge summaries: with what level of information are we dismissing our patients?," *Am. J. Surg.*, vol. 211, no. 3, pp. 631–636, Mar. 2016.

[54] S. Ayoub, E. Tsui, T. Mohammed, and J. Tseng, "Readability Assessment of Online Uveitis Patient Education Materials," *Ocul. Immunol. Inflamm.*, vol. 27, no. 3, pp. 399–403, Dec. 2019.

[55] Readability Formulas, "The Flesch Grade Level Readability Formula," readabilityformulas.com. [Online]. Available: https://readabilityformulas.com/flesch-grade-level-readability-formula.php. [Accessed: May 04, 2020].

[56] G. H. Mc Laughlin, "SMOG grading-a new readability formula," *J. Read*, vol. 12, no. 8, pp. 639–646, May 1969.

[57] Readability Formulas, "The SMOG Readability Formula, a Simple Measure of Gobbledygook," readabilityformulas.com. [Online]. Available: https://readabilityformulas.com/smog-readability-formula.php. [Accessed: May 04, 2020].

[58] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring.," *J. Appl. Psychol.* vol. 60, no. 2, Art. no. 283, Apr. 1975.

[59] Readability Formulas, "The Coleman-Liau Readability Formula (also known as The Coleman-Liau Index)," readabilityformulas.com. [Online]. Available: https://readabilityformulas.com/coleman-liau-readability-formula.php. [Accessed: May 04, 2020].

[60] J. P. Kincaid and L. J. Delionbach, "Validation of the Automated Readability Index: A follow-up," *Hum. Factors*, vol. 15, no. 1, pp. 17–20, Feb. 1973.

[61] Readability Formulas, "The Automated Readability Index (ARI)," readabilityformulas.com. [Online]. Available: https://readabilityformulas.com/automated-readability-index.php. [Accessed: May 04, 2020].

[62] J. N. Farr, J. J. Jenkins, and D. G. Paterson, "Simplification of Flesch reading ease formula.," *J. Appl. Psychol.,* vol. 35, no. 5, Art. no. 333, Oct. 1951.

[63] "Flesch Reading Ease and the Flesch Kincaid Grade Level," Readable. [Online]. Available: https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/. [Accessed: Aug. 20, 2021].

[64] J. Caylor, T. Stict, and J. P. Ford, "The FORCAST readability formula," *Lit. Discuss.*, vol. 67, Art. no. 68, Jun. 1973.

[65] J. A. Longo, "The Fry graph: Validation of the college levels," *J. Read*, vol. 26, no. 3, pp. 229–234, Dec. 1982.

[66] M. Martin and W. Lee, "Sample Frequency in Application of Dale-Chall Readability Formula," *Educ. Res. Bull.*, vol. 40, no. 6, pp. 146–149, Sep. 1961.

[67] W. D. Lee and B. R. Belden, "A cross-validation readability study of general psychology textbook material and the Dale-Chall Readability Formula," *J. Educ. Res.* vol. 59, no. 8, pp. 369–373, Dec. 2014, doi: 10.1080/00220671.1966.10883383.

[68] D. R. Hansberry *et al.*, "Analysis of the readability of patient education materials from surgical subspecialties," *The Laryngoscope*, vol. 124, no. 2, pp. 405–412, Jun. 2014.

[69] R. S. Baldwin and R. K. Kaufman, "A concurrent validity study of the Raygor readability estimate," *J. Read*, vol. 23, no. 2, pp. 148–153, Nov. 1979.

[70] M. Khosrow-Pour, D.B.A., Ed., *Advanced Methodologies and Technologies in Modern Education Delivery:* in Advances in Educational Technologies and Instructional Design. IGI Global, 2019. doi: 10.4018/978-1-5225-7365-4.

[71] M. K. Biddinika, R. P. Lestari, B. Indrawan, K. Yoshikawa, K. Tokimatsu, and F. Takahashi, "Measuring the readability of Indonesian biomass websites: The ease of understanding biomass energy information on websites in the Indonesian language," *Renew. Sustain. Energy Rev.*, vol. 59, pp. 1349–1357, Jun. 2016.

[72] W. H. DuBay, *The Principles of Readability*. ERIC Clearinghouse, 2004. [Online]. Available: https://books.google.gr/books?id=Aj0VvwEACAAJ

[73] M. Nasr Azadani, N. Ghadiri, and E. Davoodijam, "Graph-based biomedical text summarization: An itemset mining and sentence clustering approach," *J. Biomed. Informat.*, vol. 84, pp. 42–58, Aug. 2018, doi: 10.1016/j.jbi.2018.06.005.

[74] M. Patel, A. Chokshi, S. Vyas, and K. Maurya, "Machine Learning Approach for Automatic Text Summarization Using Neural Networks," *Int. J. Adv. Res. Comp. Commun. Eng.,* vol. 7, no. 1, pp. 194–202, Jan. 2018, doi: 10.17148/IJARCCE.

[75] G. L. De La Peña Sarracén and P. Rosso, "Automatic Text Summarization based on Betweenness Centrality," in *Proc. 5th Spanish Conf. Inform. Retr.*, Zaragoza Spain: ACM, Jun. 2018, pp. 1–4. doi: 10.1145/3230599.3230611.

[76] S. Alias, S. K. Mohammad, K. H. Gan, and T. T. Ping, "MYTextSum : A Malay Text Summarizer Model Using a Constrained Pattern- Growth Sentence Compression Technique,"

in *Int. Conf. on Computational Science and Technology*, Singapore: Springer, Feb. 2017, pp. 141–150. doi: 10.1007/978-981-10-8276-4.

[77] M. Azhari, Y. Kumar, O. Goh, and B. Raza, "Text Summarization Evaluation Based on Sentence Scoring and Clustering," *J. Eng. Sci. Techn.,* Vol. 13, pp. 144–151, Jul. 2018.

[78] K. Sarkar, "Automatic Text Summarization Using Internal and External Information," in *2018 Fifth Int. Con.f on EAIT*, IEEE, Sep. 2018.

[79] F. B. Goularte, S. M. Nassar, R. Fileto, and H. Saggion, "A text summarization method based on fuzzy rules and applicable to automated assessment," *Expert Syst. Appl.*, vol. 115, pp. 264–275, Jan. 2019, doi: 10.1016/j.eswa.2018.07.047.

[80] P. M. Hanunggul and S. Suyanto, "The impact of local attention in lstm for abstractive text summarization," in *Int. Conf. of ISRITI*, IEEE, Mar. 2019, pp. 54–57.

[81] M. Mohamed and M. Oussalah, "SRL-ESA-TextSum: A text summarization approach based on semantic role labeling and explicit semantic analysis," *Inf. Process. Manage.*, vol. 56, no. 4, pp. 1356–1372, Jul. 2019, doi: 10.1016/j.ipm.2019.04.003.

[82] N. Alami, M. Meknassi, and N. En-nahnahi, "Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning," *Expert Syst. Appl.*, vol. 123, pp. 195–211, Jun. 2019.

[83] B. Sharma, M. Tomer, and K. Kriti, "Extractive text summarization using F-RBM," *J. Stat. Manag. Syst.*, vol. 23, no. 6, pp. 1093–1104, Sep. 2020.

[84] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "EdgeSumm: Graph-based framework for automatic text summarization," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, doi: 10.1016/j.ipm.2020.102264.

[85] C. F. Tsai, K. Chen, Y. H. Hu, and W. K. Chen, "Improving text summarization of online hotel reviews with review helpfulness and sentiment," *Tourism Manag.,* vol. 80, Oct. 2020, doi: 10.1016/j.tourman.2020.104122.

[86] H. C. Wang, W. C. Hsiao, and S. H. Chang, "Automatic paper writing based on a RNN and the TextRank algorithm," *Appl. Soft Comput.*, vol. 97, Dec. 2020, doi: 10.1016/j.asoc.2020.106767.

[87] R. Elbarougy, G. Behery, and A. El Khatib, "Extractive Arabic Text Summarization Using Modified PageRank Algorithm," *Egypt. Inform. J.*, vol. 21, no. 2, pp. 73–81, Jul. 2020, doi: 10.1016/j.eij.2019.11.001.

[88] T. Uçkan and A. Karcı, "Extractive multi-document text summarization based on graph independent sets," *Egypt. Inform. J.*, vol. 21, no. 3, pp. 145–157, Jan. 2020, doi: 10.1016/j.eij.2019.12.002.

[89] A. Nawaz, M. Bakhtyar, J. Baber, I. Ullah, W. Noor, and A. Basit, "Extractive Text Summarization Models for Urdu Language," *Inf. Process. Manage.*, vol. 57, no. 6, Art. no. 102383, Nov. 2020, doi: 10.1016/j.ipm.2020.102383.

[90] R. Bhargava and Y. Sharma, "Deep Extractive Text Summarization," *Procedia Comp. Sci.*, vol. 167, pp. 138–146, Mar. 2020, doi: 10.1016/j.procs.2020.03.191.

[91] Z. Deng, F. Ma, R. Lan, W. Huang, and X. Luo, "A Two-stage Chinese text summarization algorithm using keyword information and adversarial learning," *Neurocomputing*, vol. 425, pp. 117–126, Feb. 2021, doi: 10.1016/j.neucom.2020.02.102.

[92] R. Rani and D. K. Lobiyal, "A weighted word embedding based approach for extractive text summarization," *Expert Sys. Appl.*, vol. 186, p. 115867, Dec. 2021.

[93] R. Rani and D. K. Lobiyal, "An extractive text summarization approach using tagged-LDA based topic modeling," *Multimed. Tools Appl.*, vol. 80, no. 3, pp. 3275–3305, Feb. 2021.

[94] T. Ma, Q. Pan, H. Rong, Y. Qian, Y. Tian, and N. Al-Nabhan, "T-bertsum: Topic-aware text summarization based on bert," *IEEE Trans Comput Soc Syst*, vol. 9, no. 3, pp. 879–890, Jun. 2021.

[95] B. Muthu *et al.*, "A framework for extractive text summarization based on deep learning modified neural network classifier," *TALLIP*, vol. 20, no. 3, pp. 1–20, May. 2021.

[96] R. K. Roul, "Topic modeling combined with classification technique for extractive multi-document text summarization," *Soft Computing*, vol. 25, no. 2, pp. 1113–1127, Oct. 2021.

[97] Y. Chen and Q. Song, "News Text Summarization Method based on BART-TextRank Model," in *Int. Conf. of IAEAC*, IEEE, Apr. 2021, pp. 2005–2010.

[98] C. Slamet, A. R. Atmadja, D. S. Maylawati, R. S. Lestari, W. Darmalaksana, and M. A. Ramdhani, "Automated Text Summarization for Indonesian Article Using Vector Space Model," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 288, Art. no. 012037, Jan. 2018, doi: 10.1088/1757-899X/288/1/012037.

[99] I. W. A. Setyadi, D. C. Khrisne, and I. M. A. Suyadnya, "Automatic Text Summarization Menggunakan Metode Graph dan Metode Ant Colony Optimization," *J. Techn. Eng.*, vol. 17, no. 1, pp. 124–130, May. 2018.

[100] G. Mandar and G. Gunawan, "Peringkasan dokumen berita Bahasa Indonesia menggunakan metode Cross Latent Semantic Analysis," *Register*, vol. 3, no. 2, pp. 94–104, Jul. 2018.

[101] W. Yulita, S. Priyanta, and S. N. Azhari, "Automatic Text Summarization Based on Semantic Networks and Corpus Statistics," *Indonesian J. Comp. Cybern.*, vol. 13, no. 2, pp. 137–148, Apr. 2019.

[102] R. Adelia, S. Suyanto, and U. N. Wisesty, "Indonesian Abstractive Text Summarization Using Bidirectional Gated Recurrent Unit," *Procedia Comput. Sci.*, vol. 157, pp. 581–588, Oct. 2019.

[103] S. Cahyawijaya *et al.*, "IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation," 2021, *arXiv*. doi: 10.48550/ARXIV.2104.08200.

[104] R. Wijayanti, M. L. Khodra, and D. H. Widyantoro, "Indonesian Abstractive Summarization using Pre-trained Model," in *2021 3rd East Indonesia Conf. Comp. Inform. Techn. (EIConCIT)*, Surabaya, Indonesia: IEEE, Apr. 2021, pp. 79–84. doi: 10.1109/EIConCIT50028.2021.9431880.

[105] I. G. M. Darmawiguna, G. A. Pradnyana, and I. B. Jyotisananda, "Indonesian sentiment summarization for lecturer learning evaluation by using textrank algorithm," *J. Phys.: Conf. Ser.*, vol. 1810, no. 1, Art. no. 012024, Mar. 2021, doi: 10.1088/1742-6596/1810/1/012024.

[106] N. Lin, J. Li, and S. Jiang, "A simple but effective method for Indonesian automatic text summarisation," *Connection Science*, vol. 34, no. 1, pp. 29–43, Dec. 2022, doi: 10.1080/09540091.2021.1937942.

[107] D. P. Purbawa, R. N. E. Anggraini, and R. Sarno, "Automatic Text Summarization using Maximum Marginal Relevance for Health Ethics Protocol Document in Bahasa," in *Int. Conf. of ICTS*, IEEE, Nov. 2021, pp. 324–329.

[108] D. Fitrianah and R. N. Jauhari, "Extractive text summarization for scientific journal articles using long short-term memory and gated recurrent units," *Bull. Electr. Eng. Inform.*, vol. 11, no. 1, pp. 150–157, Feb. 2022, doi: 10.11591/eei.v11i1.3278.

[109] M. Z. Abdullah and C. Fatichah, "Feature-based POS tagging and sentence relevance for news multi-document summarization in Bahasa Indonesia," *Bulletin EEI*, vol. 11, no. 1, pp. 541–549, Feb. 2022, doi: 10.11591/eei.v11i1.3275.

[110] F. Halim, L. Liliana, and K. Gunadi, "Ringkasan Ekstraktif Otomatis pada Berita Berbahasa Indonesia Menggunakan Metode BERT," *Jurnal Infra*, vol. 10, no. 1, pp. 162–168, Jan. 2022.

[111] A. Najibullah, "Indonesian Text Summarization based on Naïve Bayes Method," in *Proc. Int. Sem. Interr. Relig. Sci. Cult. Econ.*, pp. 67–78, Sep. 2015.

[112] M. Fachrurrozi, N. Yusliani, and R. U. Yoanita, "Frequent Term based Text Summarization for Bahasa Indonesia," in *ICE/ITMC*, pp. 30–32, Dec. 2013, doi: 10.15242/IIE.E1213550.

[113] D. Gunawan, A. Pasaribu, R. F. Rahmat, and R. Budiarto, "Automatic Text Summarization for Indonesian Language Using TextTeaser," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 190, Art. no. 012048, Apr. 2017, doi: 10.1088/1757-899X/190/1/012048.

[114] D. T. Massandy and M. L. Khodra, "Guided summarization for Indonesian news articles," in *2014 Int. Conf. Adv. Informat.: Conc., Theory Applic. (ICAICTA)*, Bandung, Indonesia: IEEE, Aug. 2014, pp. 140–145. doi: 10.1109/ICAICTA.2014.7005930.

[115] F. Christie and M. L. Khodra, "Multi-document summarization using sentence fusion for Indonesian news articles," in *2016 Int. Conf. Adv. Inform.: Conc., Theory Applic. (ICAICTA)*, Penang, Malaysia: IEEE, Aug. 2016, pp. 1–6. doi: 10.1109/ICAICTA.2016.7803134.

[116] C. Slamet, A. R. Atmadja, D. S. Maylawati, R. S. Lestari, W. Darmalaksana, and M. A. Ramdhani, "Automated Text Summarization for Indonesian Article Using Vector Space Model," *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 288, p. 012037, Jan. 2018, doi: 10.1088/1757-899X/288/1/012037.

[117] D. SJ, H. DM, L. KF, and S. MJ, "Financial opacity and firm performance: the readability of REIT annual reports," *J. Real Estate Financ.*, vol. 45, no. 2, pp. 450–470, Aug. 2012.

[118] Y. N. Fadziah and E. F. Rahman, "Penerapan Algoritma Enchanced Confix Stripping dalam Pengukuran Keterbacaan Teks Menggunakan Gunning Fog Index," *J. Aplik. Teori Ilm. Komp.*, vol. 1, no. 1, pp. 15–24, Mar. 2018.

[119] U. Mursyadah, "Tingkat Keterbacaan Buku Sekolah Elektronik (BSE) Pelajaran Biologi Kelas X SMA/MA," *Teaching*, vol. 1, no. 4, pp. 298–304, Dec. 2021.

[120] S. D. Utami, I. N. Dewi, and I. Efendi, "Tingkat Keterbacaan Bahan Ajar Flexible Learning Berbasis Kolaboratif Saintifik," *Bioscientist*, vol. 9, no. 2, pp. 577–587, Dec. 2021.

[121] M. P. Sari and H. Herri, "Analisa Konten Serta Tingkat Keterbacaan Pernyataan Misi dan Pengaruhnya terhadap Kinerja Perbankan Indonesia," *Menara Ilmu*, vol. 14, no. 1, pp. 96 – 106, Jul. 2020.

[122] R. M. Sareb Putra, "Fog Index dan Keterbacaan Berita Utama (Headline) Suara Merdeka 03 Mei 2013," *J. Ilmu Komm.*, vol. 10, no. 1, pp. 41 – 48, Jun. 2013, doi: 10.24002/jik.v10i1.152.

[123] D. D. Pranowo, "Instrument of Indonesian Texts Readability," UNY, 2011. [Online]. Available: http://staff.uny.ac.id/sites/default/files/Readability instrument-thesis.pdf