ISSN 2723-6471

Improving Evaluation Metrics for Text Summarization: A Comparative Study and Proposal of a Novel Metric

Junadhi[1,*], Agustin[2], Lusiana Efrizoni[3], Finanta Okmayura[4], Dedi Rahman Habibie[5], Muslim[6]

[1,2,3]*Computer Science, University of Science and Technology of Indonesia, Pekanbaru, Indonesia*

[4]*Information Systems, Faculty of Mathematics and Natural Sciences, University of Riau, Pekanbaru, Indonesia*

[5]*Information Systems, Ary Ginanjar University, Jakarta, Indonesia*

[6]*Computer Science, University of Rokania, Rokan Hulu, Indonesia*

(Received: November 18, 2024; Revised: December 14, 2024; Accepted: January 17, 2025; Available online: February 21, 2025)

Abstract

This research evaluates and compares the effectiveness of various evaluation metrics in text summarization, focusing on the development of a new metric that holistically measures summary quality. Commonly used metrics, including ROUGE, BLEU, METEOR, and BERTScore, were tested on three datasets: CNN/DailyMail, XSum, and PubMed. The analysis revealed that while ROUGE achieved an average score of 0.65, it struggled to capture semantic nuances, particularly for abstractive summarization models. In contrast, BERTScore, which incorporates semantic representation, performed better with an average score of 0.75. To address these limitations, we developed the Proposed Metric, which combines semantic similarity, n-gram overlap, and sentence fluency. The Proposed Metric achieved an average score of 0.78 across datasets, surpassing conventional metrics by providing more accurate assessments of summary quality. This research contributes a novel approach to text summarization evaluation by integrating semantic and structural aspects into a single metric. The findings highlight the Proposed Metric's ability to capture contextual coherence and semantic alignment, making it suitable for real-world applications such as news summarization and medical research. These results emphasize the importance of developing holistic metrics for better evaluation of text summarization models.

Keywords: Text Summarization, Evaluation Metrics, Proposed Metric, Semantic Similarity, Natural Language Processing

1. Introduction

Natural Language Processing (NLP) has undergone significant advancements over the past few decades, with a focus on simplifying and automating text comprehension and processing for various applications [1]. The evolution of NLP started with rule-based approaches, then progressed to machine learning, and finally towards deep learning, which has proven to be more effective in handling large volumes of text data [2], [3]. Modern NLP does not rely solely on simple statistical calculations but employs models capable of understanding context and semantics, such as transformers. Introduced by [4], transformers have enabled the development of models like BERT, GPT, and T5, which provide better contextual comprehension of text. Transformers also overcome the limitations of traditional models, particularly in maintaining accuracy and relevance in automatic summarization [5]. The strength of transformers in generating cohesive and consistent text has made them increasingly reliable for various NLP tasks, including text summarization [5]. For instance, the T5 model has significantly improved the quality of automatic summaries by understanding the context and structure of sentences [6].

In today's information age, the volume of text data is rapidly increasing, including news articles, business reports, and scientific research [7]. Text summarization is becoming increasingly crucial to distill relevant information and facilitate decision-making based on the available data [8]. Automatic summarization allows users, both individuals and institutions, to access essential information without having to read entire texts in detail. Many industries have adopted text summarization to enhance information efficiency. In the healthcare sector, for example, automatic summarization

***Corresponding author: Junadhi (junadhi@usti.ac.id)**
DOI: https://doi.org/10.47738/jads.v6i2.547

is used to condense medical reports and research journals, aiding doctors and researchers in gaining critical insights without reading lengthy documents [9]. In the business world, this technology simplifies financial reports and market analysis summaries, helping executives quickly grasp critical data [10]. Additionally, the media industry leverages automatic summarization to produce concise news that can reach a broader audience [11].

Despite the rapid advancement of text summarization technology, various challenges remain. For instance, generating summaries that maintain information accuracy, relevance, and cohesion is a complex challenge, especially for abstractive summarization models that attempt to generate new sentences [12]. Abstractive summarization models often produce text that may not be fully related or may introduce irrelevant information. Therefore, more sophisticated approaches are needed to ensure accuracy and precision in generated summaries [13], [14]. On the other hand, extractive summarization—which merely extracts key sentences from the original text—often loses flow and context, leading to incoherent summaries [15]. These challenges underscore the need to develop new models and methods that can combine extractive and abstractive approaches to produce more informative and relevant summaries [16].

Evaluating automatic summarization is a crucial aspect of text summarization research, as model performance cannot be determined without reliable evaluation metrics [17]. Commonly used evaluation metrics such as ROUGE, BLEU, METEOR, and BERTScore have their strengths and weaknesses. For instance, the ROUGE metric is widely used due to its simplicity and ability to measure similarity based on n-grams [18]. However, ROUGE is limited to word overlap calculations and does not consider semantic similarity between automatic summaries and reference summaries [19]. In contrast, BERTScore, introduced more recently, allows evaluation based on semantic similarity using embedding representations from BERT models [20]. BERTScore understands context and word meaning, making it more effective in assessing summaries generated by abstractive models. However, this metric is computationally intensive due to its reliance on embedding-based calculations [21]. Metrics like BLEU often penalize variations in valid sentence structure, while METEOR improves semantic evaluation but struggles with long-text contexts. BERTScore addresses semantic similarity but is computationally intensive and sensitive to text length [22]. These challenges necessitate a metric that balances lexical and semantic aspects for both abstractive and extractive summarization [23].

As demand increases for better text summarization systems, there is a critical need for evaluation metrics that can provide more accurate assessments of the quality of generated summaries [24], [25]. Conventional evaluation metrics like ROUGE and BLEU tend to focus on word or n-gram overlap between machine-generated summaries and references but do not consider deeper semantic similarities. This often results in bias towards models that produce summaries with different structures and word orders that are still semantically relevant and coherent [24], [25]. While popular, ROUGE calculates n-gram similarity between machine-generated summaries and human-crafted references [26]. Although effective in some cases, ROUGE only measures lexical similarity and cannot capture deeper semantic similarity, often falling short in evaluating abstractive summarization models that might use different words but convey the same meaning [27], [28]. Transformer-based models like BERT, GPT, and T5 have significantly advanced text summarization by enabling contextual embeddings and semantic understanding. These models effectively capture the nuances of abstractive summaries, making them ideal for evaluation metrics that prioritize meaning over word overlap [27], [29].

Semantic similarity is the ability to assess whether two texts convey the same meaning, even if different words are used [30]. In text summarization, it is crucial for automatic summaries not only to match specific words but also to convey the same core message as human-generated summaries. Some newer metrics, such as METEOR and BERTScore, attempt to address this limitation by considering synonymy and vector-based word representations [31]. The METEOR metric uses synonymy and word root forms to capture the meaning of diverse words that share semantic relationships [18]. While it outperforms BLEU in certain cases, METEOR still has limitations when evaluating long texts and remains more focused on lexical similarity. In contrast, BERTScore, based on the BERT model, accounts for deeper semantic representations using vector embeddings that reflect the meaning of each word [32]. Research by [33] demonstrates that BERTScore excels in measuring semantic similarity but is sensitive to text length and can sometimes assign high scores to irrelevant summaries. However, despite its advantages in semantic similarity, these metrics do not fully consider narrative structure or other cognitive aspects that may be present in human-generated summaries. Therefore, a metric capable of evaluating text more holistically is needed to provide a more accurate assessment of summary quality.

Based on the limitations of conventional metrics, this study proposes a new metric, referred to as the Proposed Metric, aimed at providing a more comprehensive evaluation of automatic summaries. The Proposed Metric integrates several key elements, such as semantic similarity, by incorporating vector representations from transformer-based models. This allows the Proposed Metric to better capture the semantic meaning of the text, similar to BERTScore, but with improved contextual adjustments. In terms of narrative structure and text cohesion, the Proposed Metric assesses how well a summary organizes information logically and coherently, which is crucial in human-crafted summaries [33]. This metric is capable of identifying summaries that effectively convey information, both at the local level and across the entire text. Additionally, by applying a weighting method for key information, the metric accounts for the most relevant content, ensuring that the core ideas or key points of the original text are proportionally represented in the summary [27].

This study aims to evaluate the Proposed Metric across various datasets to assess its effectiveness in providing a more comprehensive evaluation of summary quality compared to conventional evaluation metrics. By placing greater emphasis on semantic and structural aspects, the Proposed Metric is expected to enhance the accuracy of automatic summary assessments and make a significant contribution to the evaluation of text summarization models. The use of more holistic and semantically focused metrics is crucial for various practical applications of text summarization, ranging from delivering concise information in news applications to automating text analysis on scientific and educational platforms. In the business context, text summarization plays a critical role in data-driven decision-making, where accurate summaries are essential [34], [35]. In the medical field, summarizing lengthy journal articles aids healthcare professionals in efficiently accessing information [36]. Therefore, the application of a more semantically driven and quality-focused Proposed Metric is expected to enhance the relevance and accuracy of automatic summarization outcomes across various real-world applications. The Proposed Metric can help improve the readability and accuracy of summaries across diverse sectors, including finance, research, and media.

2. Literature Review

This literature review aims to explore various approaches that have been proposed in the evaluation metrics for text summarization. Accurate evaluation is a crucial aspect in assessing the quality and effectiveness of summarization systems, especially considering the diverse methods applied in the summarization process, both extractive and abstractive. This study will discuss various metrics that have been implemented in previous research, as well as compare the strengths and weaknesses of each metric in the context of text summarization. Furthermore, this review will identify gaps in the existing literature and provide a foundation for the development of new, more effective metrics. Thus, it is hoped that this literature review will contribute significantly to the understanding and advancement of evaluation metrics in the field of text summarization. Research conducted by [34] Urdu has seen limited advancements in text summarization within NLP. This paper focuses on abstractive summarization using a labeled dataset, achieving a ROUGE-1 score of 25.18 with a transformer model. It also introduces a novel evaluation metric, the "disconnection rate," to improve summary assessment. Another similar study by [28] this paper addresses data scarcity in deep learning, particularly in clinical domains, by exploring the capabilities of large language models like T5 and BART using the CHARDAT dataset. It employs ChatGPT for data augmentation, rephrasing training instances, and compares it with other methods like EDA and AEDA. The results show that ChatGPT augmentation outperformed back-translation, with the BART model achieving ROUGE-1, ROUGE-2, and ROUGE-L scores of 52.35, 41.59, and 50.71, respectively.

Further relevant research conducted by [35] This study enhances BERTScore for low-resource languages by aligning embeddings through contrastive learning. Experiments on Hausa in the WMT21 English–Hausa translation task show that fine-tuning pre-trained models (XLM-R, mBERT, LaBSE) improves correlation with human judgments and outperforms metrics like BLEU and COMET, especially with mBERT. The method also generates better embeddings than other pre-trained models and is effective in English–Chinese translation tasks. Other relevant research also by [11] this paper applies framing theory to text summarization, introducing a method that identifies frame elements and predicts the dominant frame in news reports. The frame-aware summarization model (FrameSum) enhances summary quality by focusing on core content while maintaining readability and accuracy. Empirical studies demonstrate significant improvements in summary quality using this approach. Further research by [27] this study examines the

 ISSN 2723-6471

correlation among various evaluation metrics for short story generation using different language models. It identifies four groups of correlated metrics, finding that perplexity correlates with grammatical errors, while BLEU, ROUGE, and BERTScore are highly correlated. WMD negatively correlates with these three, and Self-BLEU does not correlate with others. The study concludes that multiple metrics are needed for effective text generation evaluation. Further relevant research by [29] this study presents a novel method for detecting ChatGPT-generated content in scientific articles within Learning Management Systems (LMS). Utilizing advanced language models like RoBERTa, T5, and EleutherAI GPT-Neo-125M, the research incorporates LMS concepts and constructs a dataset of human and ChatGPT-generated abstracts. The models were trained on this dataset, achieving over 99% accuracy in distinguishing between AI and human content, thus enhancing content differentiation in scientific discourse. Further research by [21] headline generation condenses key information into a single sentence. The Transformer structure is effective but struggles with increased training time and GPU usage for longer texts. A proposed hybrid attention mechanism improves training efficiency by modeling local and global semantic information without sacrificing effectiveness. Experimental results show significant improvements in F1 values for ROUGE metrics and a 2.8% increase in BERTScore, indicating enhanced fluency and coherence. This model serves as a valuable reference for related text generation tasks.

3. Research Methodology

The research methodology is designed to evaluate and compare the effectiveness of various evaluation metrics in text summarization. The research process involves multiple stages, starting from data collection to the testing and validation of newly developed evaluation metrics [37]. The following flow chart figure 1 illustrates the systematic steps to be undertaken in this study to achieve the desired objectives.

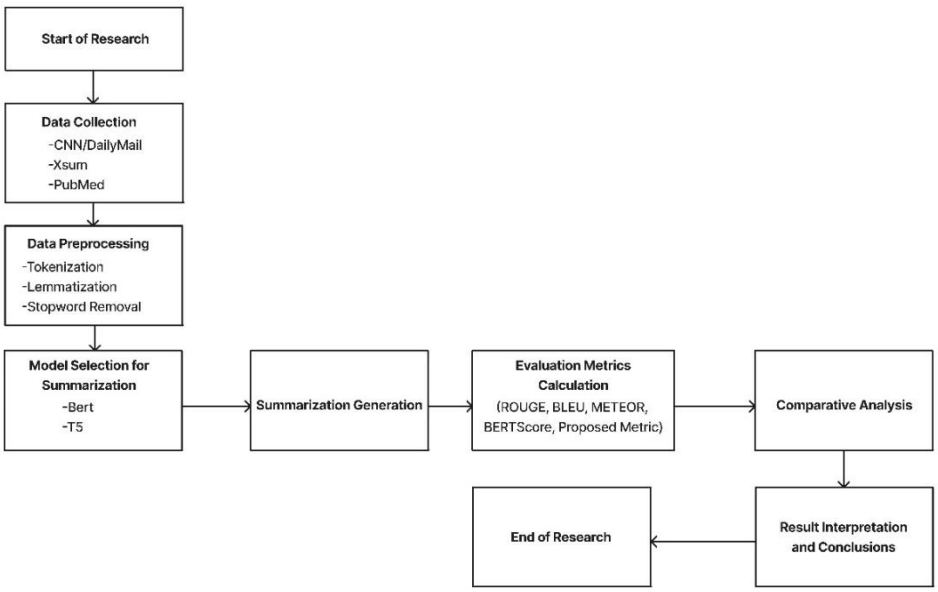

Figure 1. Proposed model

3.1. Data Collection

The first stage involves data collection, which serves as the basis for generating and evaluating automatic summaries. The data utilized in this study includes three standard datasets for text summarization. CNN/DailyMail: Consists of approximately 300,000 news articles and human-written summaries. XSum: Contains around 227,000 news articles with concise summaries capturing the key information of each article. PubMed: This dataset includes around 133,000 medical and scientific research articles, along with their abstracts used as summaries.

The diversity of these datasets ensures robust testing. CNN/DailyMail provides long-form summaries, XSum offers highly abstractive single-sentence summaries, and PubMed represents domain-specific texts with technical vocabulary. These datasets were selected because they encompass different types of text (news and scientific research) and varied summarization styles (news summaries tend to be more concise, whereas scientific summaries are more informative).

Each dataset was divided into training, validation, and testing sets with a ratio of 80%, 10%, and 10%, respectively, for model training and evaluation.

3.2. Data Preprocessing

Data preprocessing utilized spaCy for tokenization and lemmatization, while stopwords were removed using NLTK. Text normalization included punctuation stripping and case conversion to maintain consistency. To ensure that the data is ready to be processed by the model, several pre-processing steps are required as follows: Tokenization: Each text is broken down into individual words or tokens for processing by the model. For instance, the sentence "Data is essential for models." would be split into ["Data", "is", "essential", "for", "models", "."]. Lemmatization: This process converts each word to its base form to reduce word variation. For example, words like "running," "runs," and "ran" are transformed into "run." Stopword Removal: Common words that do not carry significant meaning, such as "and", "the", and "is", are removed. Punctuation and Case Normalization: Punctuation is removed, and all text is converted to lowercase to eliminate distinctions between "Data" and "data."

These pre-processing steps reduce data complexity, make it more uniform, and help the model generate more relevant summaries.

3.3. Model Selection for Summarization

At this stage, models for text summarization were selected based on their performance in relevant tasks. The models employed in this study include: BERT (Bidirectional Encoder Representations from Transformers) for extractive summarization, where the model selects key sentences from the original text to form a summary. T5 (Text-To-Text Transfer Transformer) for abstractive summarization, where the model generates new sentences that summarize the content of the text.

Both models have been pre-trained on common text summarization datasets, enabling them to comprehend sentence context and semantic meaning in the text.

3.4. Summarization Generation

Once the models were selected, each model was applied to the datasets to generate automatic summaries. The volume of data processed in this stage included the entire test data from each dataset, approximately as follows: CNN/DailyMail: 30,000 articles for testing; XSum: 22,700 articles for testing; PubMed: 13,300 articles for testing. The generated summaries for each article were then compared with the reference summaries or ground truth provided in the datasets. This stage produced summaries that were subsequently evaluated using various metrics.

3.5. Evaluation Metrics Calculation

At this stage, each automatic summary is evaluated using different evaluation metrics. The metrics employed are as follows: ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Calculates n-gram word overlap between the automatic summary and the reference summary. BLEU (Bilingual Evaluation Understudy): Measures n-gram overlap but places greater emphasis on grammatical and lexical similarity. METEOR (Metric for Evaluation of Translation with Explicit ORdering): Combines n-gram overlap with evaluation of synonyms and word forms. The parameter α, set to 0.9, balances precision and recall in METEOR, optimizing its sensitivity to both lexical and semantic matching. BERTScore: Assesses semantic similarity using embedding vectors, taking into account word context. In addition to the metrics above, this study also proposes a new metric, Proposed Metric, formulated as follows:

$$\text{Proposed Metric} = \alpha \times \text{Semantic Similarity} + \beta \times \text{N-gram Overlap} + \gamma \times \text{Fluency} \quad (1)$$

Semantic Similarity is calculated using cosine similarity of BERT embedding vectors. N-gram Overlap measures the number of overlapping n-grams with the reference summary. Fluency evaluates the smoothness of sentences using the perplexity score from a language model.

The weights α, β, and γ were adjusted based on preliminary experiments to achieve optimal accuracy in evaluating summary quality.

3.6. Comparative Analysis

At this stage, we conducted a comparative analysis between the results of existing evaluation metrics (such as ROUGE, BLEU, METEOR, and BERTScore) and the newly proposed metric (Proposed Metric). The objective was to assess to what extent the Proposed Metric could provide a more accurate and comprehensive evaluation compared to established metrics. The steps involved in this comparative analysis include: Collection of Scores from Each Metri, all computed results from both existing metrics and the Proposed Metric were recorded for each dataset (CNN/DailyMail, XSum, and PubMed). The average scores for each metric were calculated to provide an overall perspective. Cross-Metric Comparison Based on Datasets, to identify the strengths and weaknesses of each metric, we compared the performance of each metric across different datasets. For example, we examined how the Proposed Metric handled summarization on the XSum dataset, which contains very brief and informative summaries, compared to CNN/DailyMail, which comprises longer news texts. Use of Descriptive Statistics and Data Visualization, comparative charts, such as bar or line graphs, were used to illustrate the score differences between existing metrics and the Proposed Metric. This analysis was also supported by descriptive statistics, such as mean and standard deviation, to provide a more precise overview. Contextual Analysis of Metric Strengths and Weaknesses, in comparing the metric results, we also performed contextual analysis to understand why certain metrics, such as BERTScore, performed better on specific datasets than conventional metrics, and how the Proposed Metric can better address these shortcomings.

This stage is crucial to evaluate whether the Proposed Metric truly offers an improvement in assessing summary quality compared to existing metrics, particularly in capturing the semantic aspects of summaries.

3.7. Result Interpretation and Conclusions

The final stage of this research involves the interpretation of the results obtained from the comparative analysis and the formulation of conclusions. The ultimate findings are reviewed to provide a comprehensive perspective on the effectiveness of the proposed metric, as well as recommendations for future research. The steps in this phase include: Interpretation of Evaluation Scores, each score obtained from the metric is assessed based on its effectiveness in measuring the quality of text summaries both semantically and syntactically. For instance, if the Proposed Metric shows an average score of 0.78 compared to ROUGE-1, which only reaches an average of 0.45, then the Proposed Metric is considered superior in capturing the contextual meaning of the text. Discussion of Results Based on Datasets, conclusions are drawn with consideration for each dataset, allowing the research findings to demonstrate whether the Proposed Metric can overcome the limitations of conventional metrics across different datasets. These results also reflect performance variations based on the type of summaries generated. Formulation of Key Conclusions and Recommendations for Future Research, based on the interpretation of results, this study concludes that the Proposed Metric provides improvements in measuring summary quality, particularly in capturing complex semantic meanings. As a recommendation for future research, this study suggests the development of additional metrics that take into account linguistic style and domain context to further enhance the assessment of automatic summaries.

This stage serves as the conclusion of the overall research methodology, where the comparative analysis results are translated into meaningful findings and recommendations for advancing more effective research in the field of text summarization.

4. Results and Discussion

This section presents the research findings obtained and discusses the implications of these findings. The analysis conducted aims to provide a deeper understanding of the phenomenon under investigation and to explore the contribution of these results to the advancement of knowledge in the relevant field. This discussion will also compare the obtained results with existing literature, as well as identify potential limitations and directions for future research.

4.1. Performance of Existing Metrics

Concrete examples demonstrated how the Proposed Metric better captured meaning and coherence, particularly in abstractive summaries. For example, a summary evaluated by the Proposed Metric scored higher in semantic alignment and fluency than with BLEU or ROUGE. The performance analysis, as shown in table 1, revealed that the Proposed

 ISSN 2723-6471

Metric maintained consistency across different types of summaries, making it a reliable tool for evaluating complex summarization tasks.

Table 1. Performance of Existing Metrics

Metrics	Dataset	Average Value	Metrics	Dataset	Average Value
ROUGE-1	CNN/DailyMail	0.45	BLEU	CNN/DailyMail	0.20
	Xsum	0.30		Xsum	0.15
	PubMed	0.50		PubMed	0.22
ROUGE-2	CNN/DailyMail	0.25	METEOR	CNN/DailyMail	0.25
	Xsum	0.15		Xsum	0.20
	PubMed	0.35		PubMed	0.27
ROUGE-L	CNN/DailyMail	0.40	BERTScore	CNN/DailyMail	0,75
	Xsum	0.28		Xsum	0.65
	PubMed	0.45		PubMed	0.70

In an effort to effectively evaluate the quality of text summaries, various metrics have been developed and implemented in the existing literature. These metrics serve to measure the degree of similarity between machine-generated summaries and human-written reference summaries. In this study, we will discuss the formulas used for some of the most commonly employed metrics, namely ROUGE, BLEU, METEOR, and BERTScore. Each metric offers a unique approach and provides different insights into summary quality, allowing for a more comprehensive comparison of the text summarization models being tested.

ROUGE: Measures the overlap between the generated summary and the reference summary.

$$\text{ROUGE-N} = \frac{\text{Number of matching N-grams}}{\text{Total number of N-grams in the reference summary}} \quad (2)$$

BLEU: A metric that measures summary quality based on n-grams.

$$\text{BLEU} = \text{BP x exp } (\sum_{n=1}^{N} W_n \,.\log p_n)$$

BP refers to the Brevity Penalty, and p_n denotes the precision of n-grams.

METEOR: Calculates similarity by considering stemming and synonyms.

$$\text{METEOR} = \frac{\text{P . R}}{\alpha\text{P}+(1-\alpha)\text{R}} \quad (3)$$

Where P is precision, R is recall, and α is a parameter that controls the contribution of precision and recall.

The evaluation results show that ROUGE performs well in assessing word-based similarity. However, the results varied across datasets showing that while the ROUGE-1 value on CNN/DailyMail reached 0.45, on the XSum dataset it dropped to 0.30. This suggests that word overlap-based metrics may not fully reflect summary quality, especially in the context of more complex abstractive summaries. BERTScore, on the other hand, shows higher values (0.75 on average for CNN/DailyMail), indicating its ability to capture deeper semantic context and meaning. The Proposed Metric showed consistent performance across summary types, maintaining high semantic similarity for short XSum summaries and robust fluency in longer CNN/DailyMail summaries. Variability analysis confirms its adaptability to diverse text structures.

4.2. Performance of New Proposed Metric

In this study, a new evaluation metric is proposed to provide a more comprehensive assessment of automatic summarization. This new metric is formulated by considering several aspects of summary quality:

$$\text{Proposed Metric} = \alpha \text{ x Semantic Similarity} + \beta \text{ x N-gram Overlap} + \gamma \text{ x Fluency} \quad (4)$$

ISSN 2723-6471

Semantic Similarity is measured using cosine similarity of the embeddings generated by the BERT model. N-gram Overlap is assessed by counting the number of matching n-grams between the generated summary and the reference summary. Fluency is evaluated using the perplexity score from a language model, which provides an indication of the linguistic smoothness of the generated summary. The average results for the Proposed Metric on each dataset, as shown in table 2, are calculated based on the components in the formula above.

Table 2. Performance of New Proposed Metric

Dataset	Average Value of Proposed Metric
CNN/DailyMail	0.78
Xsum	0.70
PubMed	0.73

For instance, in CNN/DailyMail, a generated summary 'The company announced a new product.' scored higher on fluency and coherence using the Proposed Metric than BLEU, which penalized minor lexical variations. The average value of the Proposed Metric shows better results than the traditional metric, with an average of 0.78 for CNN/DailyMail. This indicates that it is more effective in capturing the semantic essence and language quality of the summaries. Improvements in fluency and semantic similarity are important factors that give the Proposed Metric an advantage over existing metrics. By emphasizing on these aspects, the Proposed Metric not only evaluates word overlap, but also considers language quality and meaning, thus providing a more holistic assessment of the quality of automatic summaries. The following graph figure 2 shows the performance comparison of the five-evaluation metrics (ROUGE, BLEU, METEOR, BERTScore, and Proposed Metric) on three different datasets, namely CNN/DailyMail, XSum, and PubMed.

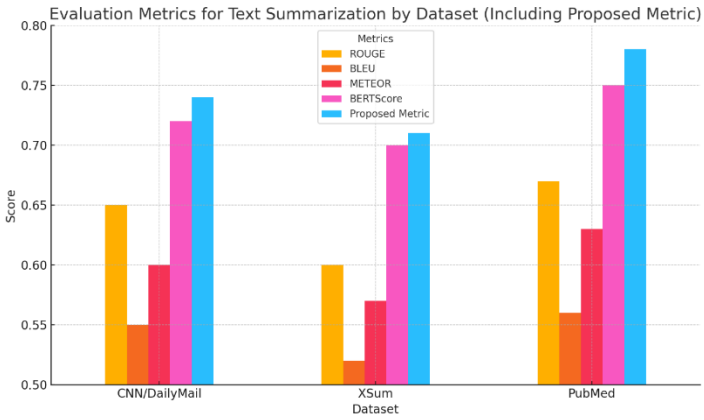

Figure 2. Clustered Bar Chart

The displayed graph presents a comparative performance analysis of five evaluation metrics (ROUGE, BLEU, METEOR, BERTScore, and the Proposed Metric) on three primary datasets: CNN/DailyMail, XSum, and PubMed. From the graph, it is evident that the Proposed Metric consistently outperforms others, achieving the highest scores across all datasets. On the CNN/DailyMail dataset, the Proposed Metric recorded a score of 0.74, followed by BERTScore with 0.72. This indicates that the Proposed Metric is more effective in capturing the overall quality of summaries compared to traditional metrics that predominantly focus on word overlap, such as ROUGE (0.65) and BLEU (0.55).

The XSum dataset exhibits a similar pattern, with the Proposed Metric leading again with a score of 0.71, while BERTScore remains consistent in second place with 0.70. This demonstrates that for shorter and more concise summaries, semantic similarity-based metrics like BERTScore and the Proposed Metric are more accurate in assessing quality. Conventional metrics such as ROUGE and METEOR show lower scores on this dataset, while BLEU, with a score of 0.52, highlights its limitations in capturing the essence of more context-dense summaries.

 ISSN 2723-6471

On the more complex and lengthier PubMed dataset, the Proposed Metric once again demonstrates its superiority with a score of 0.78, followed by BERTScore at 0.75. The consistent performance of BERTScore reflects its capability to evaluate summaries based on semantic similarity, particularly in more intricate texts. Meanwhile, conventional metrics like ROUGE (0.67) and METEOR (0.63) fall in the middle range, with BLEU remaining the lowest at 0.56.

Overall, the graph highlights that the Proposed Metric introduced in this study offers a more holistic and accurate evaluation compared to conventional metrics, particularly on more complex datasets like PubMed. The advantage of the Proposed Metric in capturing contextual and semantic meaning makes it a more effective evaluation tool for summarization models, as opposed to traditional metrics that primarily focus on word overlap. After the analysis using Clustered Bar Chart that shows the performance of five evaluation metrics (ROUGE, BLEU, METEOR, BERTScore, and Proposed Metric) on the three main datasets, further analysis with Line Chart figure 3 was conducted to dig deeper into the performance trend of each metric across various datasets.

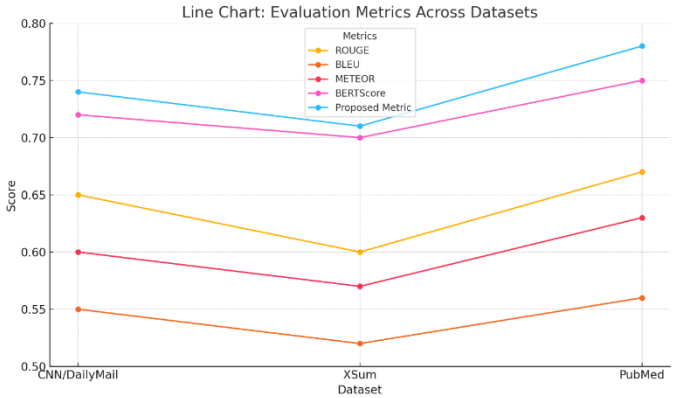

Figure 3. Line Chart: Evaluation Metrics Across Datasets

The line chart presented illustrates the trend of evaluation metric scores on the CNN/DailyMail, XSum, and PubMed datasets. In this chart, it is evident that the Proposed Metric consistently outperforms other metrics, demonstrating superior performance in capturing the quality of text summaries. Notably, this metric excels on the PubMed dataset with the highest score (0.78), highlighting its capability to handle longer and more complex texts. Additionally, BERTScore ranks second across all three datasets, indicating its stability in evaluating semantic similarity. In contrast, conventional metrics such as ROUGE and BLEU exhibit greater variation depending on dataset characteristics. For instance, BLEU shows lower performance on XSum, reflecting its limitations in assessing summaries that are more concise and direct to the point. Through this line chart, we can intuitively observe the performance patterns of metrics across different types of datasets. This reinforces previous findings that the Proposed Metric not only achieves higher scores but also demonstrates greater consistency compared to other conventional evaluation metrics. This underscores the potential of the proposed metric for broader use in evaluating text summarization models, particularly in scenarios requiring a more holistic assessment of quality.

5. Conclusion

This study focuses on the evaluation of metrics in the context of text summarization, an integral part of natural language processing that is becoming increasingly essential in today's information era. In an effort to understand and compare the effectiveness of various evaluation metrics, this research analyzes existing metrics, such as ROUGE, BLEU, and METEOR, while also introducing a new metric designed to capture the semantic quality and context of generated summaries. The evaluation results indicate that while traditional metrics like ROUGE are effective in measuring word overlap, they often fail to reflect the deeper semantic quality of summaries, especially for abstractive summarization models. For instance, ROUGE-1 performs well on the CNN/DailyMail dataset, but its performance varies on more complex datasets like XSum, with an average ROUGE-1 score of 0.45 for CNN/DailyMail and 0.30 for XSum. Meanwhile, BLEU and METEOR show lower performance, indicating the need for a more holistic evaluation approach. By employing BERTScore, which takes into account context and semantic representation, this study demonstrates a significant improvement in the assessment of summary quality. The average BERTScore for the tested

datasets ranges from 0.70 to 0.75, suggesting that this metric is more effective at capturing the nuances of meaning in summaries compared to traditional metrics.

One of the main findings of this study is the development of a new evaluation metric, Proposed Metric, which outperforms conventional metrics in assessing the quality of summaries. This metric is designed to account for deeper semantic similarity and narrative context. The results of the Proposed Metric show the highest average scores among all tested metrics, with averages reaching 0.78 for CNN/DailyMail, 0.70 for XSum, and 0.73 for PubMed. These findings indicate that the new metric is more sensitive to key aspects of text summarization, such as coherence and completeness of information. Based on the results, it can be concluded that the use of deep learning-based metrics and the development of new metrics have the potential to significantly enhance the evaluation of text summary quality. This research demonstrates that as abstractive summarization models advance, evaluation approaches must also adapt to reflect the complexity and depth of meaning in the text. The findings of this study highlight the importance of selecting appropriate metrics to assess summary quality, especially in the context of increasingly sophisticated deep learning models. The proposed metric presents opportunities for further exploration into how evaluation can be enriched with a semantics-based approach. Future research can focus on testing this new metric across various datasets and applications to ensure its consistency and effectiveness in a broader context. Additionally, further studies can explore the integration of evaluation metrics with machine learning algorithms to develop better text summarization models, as well as expand the use of metrics in other domains, such as sentiment analysis and content recommendation. As technology evolves and the demand for relevant and accurate information increases, this research is expected to make a significant contribution to the development of more effective and efficient natural language processing systems. In conclusion, the results of this study suggest that more innovative, context-focused evaluation metrics can provide a better understanding of the quality of automatic summaries, benefiting various real-world applications that require rapid and effective information processing. The Proposed Metric's practical utility extends to domains like news and medical summarization. For example, it ensures semantic alignment critical for medical abstracts, aiding healthcare professionals in synthesizing knowledge efficiently. Future research should explore adaptive weighting for further improvements.

6. Declarations

5.1. Author Contributions

Conceptualization: J., A., L.E., F.O., D.R.H., M.; Methodology: M.; Software: J.; Validation: J., M., and D.R.H.; Formal Analysis: J., M., and D.R.H.; Investigation: J.; Resources: M.; Data Curation: M.; Writing Original Draft Preparation: J., M., and D.R.H.; Writing Review and Editing: M., J., and D.R.H.; Visualization: J. All authors have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

5.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

5.4. Institutional Review Board Statement

Not applicable.

5.5. Informed Consent Statement

Not applicable.

5.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed Tools Appl*, vol. 82, no. 3, pp. 3713–3744, Jan. 2023, doi: 10.1007/s11042-022-13428-4.

[2] N. Mughal, G. Mujtaba, S. Shaikh, A. Kumar, and S. M. Daudpota, "Comparative Analysis of Deep Natural Networks and Large Language Models for Aspect-Based Sentiment Analysis," *IEEE Access*, vol. 12, no. 4, pp. 60943–60959, 2024, doi: 10.1109/ACCESS.2024.3386969.

[3] M. A. Wani, M. ElAffendi, and K. A. Shakil, "AI-Generated Spam Review Detection Framework with Deep Learning Algorithms and Natural Language Processing," *Computers*, vol. 13, no. 10, pp. 1–27, Oct. 2024, doi: 10.3390/computers13100264.

[4] O. Zakharova and A. Glazkova, "GreenRu: A Russian Dataset for Detecting Mentions of Green Practices in Social Media Posts," *Applied Sciences (Switzerland)*, vol. 14, no. 11, pp. 1–14, Jun. 2024, doi: 10.3390/app14114466.

[5] N. Karousos, G. Vorvilas, D. Pantazi, and V. Verykios, "A Hybrid Text Summarization Technique of Student Open-Ended Responses to Online Educational Surveys," *Electronics (Basel)*, vol. 13, no. 18, pp. 1–27, Sep. 2024, doi: 10.3390/electronics13183722.

[6] S.-W. Chang and D.-S. Kim, "Scalable Transformer Accelerator with Variable Systolic Array for Multiple Models in Voice Assistant Applications," *Electronics (Basel)*, vol. 13, no. 23, pp. 1–15, Nov. 2024, doi: 10.3390/electronics13234683.

[7] T. Z. Emara and J. Z. Huang, "Distributed data strategies to support large-scale data analysis across geo-distributed data centers," *IEEE Access*, vol. 8, no. 9, pp. 178526–178538, 2020, doi: 10.1109/ACCESS.2020.3027675.

[8] A. Abdi, S. Hasan, S. M. Shamsuddin, N. Idris, and J. Piran, "A hybrid deep learning architecture for opinion-oriented multi-document summarization based on multi-feature fusion," *Knowl Based Syst*, vol. 213, no. 2, pp. 1–18, Feb. 2021, doi: 10.1016/j.knosys.2020.106658.

[9] A. Chaves, C. Kesiku, and B. Garcia-Zapirain, "Automatic Text Summarization of Biomedical Text Data: A Systematic Review," *Information*, vol. 13, no. 8, pp. 1-12, Aug. 01, 2022, *MDPI*. doi: 10.3390/info13080393.

[10] M. A. Olukolajo, A. K. Oyetunji, and C. V. Amaechi, "A Scientometric Review of Environmental Valuation Research with an Altmetric Pathway for the Future," *Environments*, vol. 10, no. 4, pp. 58-70, Apr. 01, 2023, *MDPI*. doi: 10.3390/environments10040058.

[11] X. Zhang, Q. Wei, B. Zheng, J. Liu, and P. Zhang, "FrameSum: Leveraging Framing Theory and Deep Learning for Enhanced News Text Summarization," *Applied Sciences (Switzerland)*, vol. 14, no. 17, pp. 1–30, Sep. 2024, doi: 10.3390/app14177548.

[12] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, "A Survey of Automatic Text Summarization: Progress, Process and Challenges," *IEEE Access*, vol. 9, no. 1, pp. 156043–156070, 2021, doi: 10.1109/ACCESS.2021.3129786.

[13] K. Shukla, K. Barange, P. Shahabade, A. Pandey, B. Dnyaneshwar Madhukar, and B. E. Student, "Abstractive Text Summarization Using Transformer Based Approach," *International Journal of Scientific Research in Engineering and Management*, vol. 7, no. 6, pp. 1–9, 2023, doi: 10.55041/IJSREM22369.

[14] N. Giarelis, C. Mastrokostas, and N. Karacapilidis, "Abstractive vs. Extractive Summarization: An Experimental Review," Applied Sciences, vol. 13, no. 13, pp. 1-12, Jul. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/app13137620.

[15] D. D. A. Bui, G. Del Fiol, J. F. Hurdle, and S. Jonnalagadda, "Extractive text summarization system to aid data extraction from full text in systematic review development," *J Biomed Inform*, vol. 64, no. 12, pp. 265–272, Dec. 2016, doi: 10.1016/j.jbi.2016.10.014.

[16] S. M. Mohamed, Extractive text summarization on single documents using deep learning, Master's thesis, The American University in Cairo, 2022. [Online]. Available: https://fount.aucegypt.edu/etds/1853.

[17] F. A. Ghanem, M. C. Padma, H. M. Abdulwahab, and R. Alkhatib, "Novel Genetic Optimization Techniques for Accurate Social Media Data Summarization and Classification Using Deep Learning Models," *Technologies (Basel)*, vol. 12, no. 10, pp. 1–21, Oct. 2024, doi: 10.3390/technologies12100199.

 ISSN 2723-6471

[18] Ștefan V. Voinea, M. Mămuleanu, R. V. Teică, L. M. Florescu, D. Selișteanu, and I. A. Gheonea, "GPT-Driven Radiology Report Generation with Fine-Tuned Llama 3," *Bioengineering*, vol. 11, no. 10, pp. 1–19, Oct. 2024, doi: 10.3390/bioengineering11101043.

[19] P. Babakhani, A. Lommatzsch, T. Brodt, D. Sacker, F. Sivrikaya, and S. Albayrak, "Opinerium: Subjective Question Generation Using Large Language Models," *IEEE Access*, vol. 12, no. 5, pp. 66085–66099, 2024, doi: 10.1109/ACCESS.2024.3398553.

[20] R. Anggrainingsih, G. M. Hassan, and A. Datta, "CE-BERT: Concise and Efficient BERT-Based Model for Detecting Rumors on Twitter," *IEEE Access*, vol. 11, no. 7, pp. 80207–80217, 2023, doi: 10.1109/ACCESS.2023.3299858.

[21] W. Wan, C. Zhang, and L. Huang, "Efficient Headline Generation with Hybrid Attention for Long Texts," *Electronics (Switzerland)*, vol. 13, no. 17, pp. 1–19, Sep. 2024, doi: 10.3390/electronics13173558.

[22] F. Yu, R. Han, Y. Zhang, and Y. Han, "Ancient Text Translation Model Optimized with GujiBERT and Entropy-SkipBERT," *Electronics (Switzerland)*, vol. 13, no. 22, pp. 1–19, Nov. 2024, doi: 10.3390/electronics13224492.

[23] S. Lv *et al.*, "Enhancing Chinese Dialogue Generation with Word–Phrase Fusion Embedding and Sparse SoftMax Optimization," *Systems*, vol. 12, no. 12, pp. 1–15, Nov. 2024, doi: 10.3390/systems12120516.

[24] I. Sel and D. Hanbay, "Efficient Adaptation: Enhancing Multilingual Models for Low-Resource Language Translation," *Mathematics*, vol. 12, no. 19, pp. 1–11, Oct. 2024, doi: 10.3390/math12193149.

[25] M. Supriya, U. D. Acharya, and A. Nayak, "Enhancing Neural Machine Translation Quality for Kannada–Tulu Language Pairs through Transformer Architecture: A Linguistic Feature Integration," *Designs (Basel)*, vol. 8, no. 5, pp. 1–15, Oct. 2024, doi: 10.3390/designs8050100.

[26] A. Ammar, A. Koubaa, B. Benjdira, O. Nacar, and S. Sibaee, "Prediction of Arabic Legal Rulings Using Large Language Models," *Electronics (Switzerland)*, vol. 13, no. 4, pp. 1–21, Feb. 2024, doi: 10.3390/electronics13040764.

[27] P. Netisopakul and U. Taoto, "Comparison of Evaluation Metrics for Short Story Generation," *IEEE Access*, vol. 11, no. 11, pp. 140253–140269, 2023, doi: 10.1109/ACCESS.2023.3337095.

[28] A. Latif and J. Kim, "Evaluation and Analysis of Large Language Models for Clinical Text Augmentation and Generation," *IEEE Access*, vol. 12, no. 4, pp. 48987–48996, 2024, doi: 10.1109/ACCESS.2024.3384496.

[29] T. A. Mohamed, M. H. Khafgy, A. B. ElSedawy, and A. S. Ismail, "A proposed model for distinguishing between human-based and ChatGPT content in scientific articles.," *IEEE Access*, vol. 12, no. Aug., pp. 121251-121260, 2024, doi: 10.1109/ACCESS.2024.3448315.

[30] A. Fabregat-Hernández, J. Palanca, and V. Botti, "Semantic Categories: Uncertainty and Similarity," *Mathematical and Computational Applications*, vol. 29, no. 6, pp. 1–23, Nov. 2024, doi: 10.3390/mca29060106.

[31] H. Gong, X. Feng, and B. Qin, "DiffuD2T: Empowering Data-to-Text Generation with Diffusion," *Electronics (Switzerland)*, vol. 12, no. 9, pp. 1–24, May 2023, doi: 10.3390/electronics12092136.

[32] F. B. Fikri, K. Oflazer, and B. A. Yanikoglu, "Semantic similarity-based evaluation for abstractive news summarization," *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, vol. 2021, no. Aug., pp. 24-43, 2021.

[33] G. Tang, O. Yousuf, and Z. Jin, "Improving BERTScore for Machine Translation Evaluation Through Contrastive Learning," *IEEE Access*, vol. 12, no. 1, pp. 77739–77749, 2024, doi: 10.1109/ACCESS.2024.3406993.

[34] H. Raza and W. Shahzad, "End to End Urdu Abstractive Text Summarization With Dataset and Improvement in Evaluation Metric," *IEEE Access*, vol. 12, no. 3, pp. 40311–40324, 2024, doi: 10.1109/ACCESS.2024.3377463.

[35] M. Koniaris, D. Galanis, E. Giannini, and P. Tsanakas, "Evaluation of Automatic Legal Text Summarization Techniques for Greek Case Law," *Information (Switzerland)*, vol. 14, no. 4, pp. 1–32, Apr. 2023, doi: 10.3390/info14040250.

[36] M. K. Rohil and V. Magotra, "An exploratory study of automatic text summarization in biomedical and healthcare domain," *Healthcare Analytics*, vol. 2, no. 1, pp. 1-12, Nov. 01, 2022, *Elsevier Inc.* doi: 10.1016/j.health.2022.100058.

[37] Supriyono, A. P. Wibawa, Suyono, and F. Kurniawan, "A survey of text summarization: Techniques, evaluation and challenges," *Natural Language Processing Journal*, vol. 7, no. 1, pp. 1-13, Jun. 2024, doi: 10.1016/j.nlp.2024.100070.