# sBERT score: Evaluating text generation using Sentence BERT

**Group 3**
Gaurav Sharma, Shikher Srivastava, Gautam Srinidhi Iruvanti, Surya Pratap Singh Parmar

## Abstract

Text generation is a challenging task that requires evaluating the quality of generated text. Existing automatic metrics, such as ROUGE and BLEU scores, lack semantic understanding and struggle with paraphrased text. To address these limitations, semantic evaluation metrics like BERTScore have been proposed, but their computational cost remains a challenge. In this paper, we introduce sBERTScore, a novel evaluation metric inspired by BERTScore, aimed at reducing computational time while maintaining performance. We evaluate sBERTScore on machine translation and summarization tasks using standard datasets, comparing it with exact string matching metrics like ROGUE Score and semantic metrics like BERTScore. Our results demonstrate the effectiveness of sBERTScore in capturing semantic similarities between generated and reference text, making it a promising evaluation metric for text generation.

## 1 Introduction

Text generation is a complex and evolving field that plays a crucial role in various natural language processing tasks, including machine translation, summarization, and dialogue systems. Evaluating the quality of generated text is of paramount importance to ensure the effectiveness and reliability of text generation models. Traditional evaluation metrics, such as ROUGE(Lin, 2004) and BLEU(Papineni et al., 2002) scores, have been widely used but are limited in their ability to capture the semantic understanding of generated text and often struggle with paraphrased content. As a result, there is a pressing need for advanced evaluation metrics that can overcome these limitations and provide a more accurate assessment of the quality of generated text.

To address the shortcomings of traditional evaluation metrics, researchers have proposed semantic evaluation metrics that leverage contextual token representations instead of relying solely on exact string matching. One such metric is BERTScore(Zhang et al., 2019), which operates on contextualized embeddings to capture semantic similarities between candidate and reference text. BERTScore has shown promising results in correlating with human annotations and providing a more robust evaluation metric for text generation.

However, a significant challenge with BERTScore and other similar metrics is their computational cost. The pairwise comparison of token representations in BERTScore requires considerable computational resources, making it impractical for large-scale evaluation scenarios. Therefore, there is a need for a more computationally efficient evaluation metric that can achieve comparable performance to BERTScore.

In this paper, we propose sBERTScore, a novel evaluation metric inspired by BERTScore that aims to reduce computational time while maintaining performance. The key objective of sBERTScore is to capture semantic similarities between generated and reference text while significantly decreasing the time required for evaluation. By employing sentence embeddings and pairwise cosine similarity, sBERTScore calculates the semantic similarity between candidate and generated text.

To evaluate the effectiveness of sBERTScore, we conduct experiments on two widely studied text generation tasks: machine translation and summarization. For machine translation evaluation, we utilize the WMT18(Bojar et al., 2018) dataset, which provides a standard benchmark for assessing translation performance. Additionally, we employ the SummEval(Fabbri et al., 2020) dataset for text summarization, which includes human judgment scores, allowing for correlation analysis with the proposed metric. By comparing sBERTScore with traditional exact string matching metrics such as ROUGE score and semantic metrics like BERTScore, we aim to demonstrate the superiority of sBERTScore in capturing semantic understanding and providing

a more efficient evaluation metric.

Our evaluation focuses on two key criteria: correlation with human annotations, measured by Pearson correlation between the sBERTScore and human annotations, and the average time taken per sample, providing insights into the speed of our proposed metric. Furthermore, we benchmark sBERTScore against existing evaluation metrics to highlight its advantages and improvements over traditional approaches.

In summary, this paper introduces sBERTScore, a novel evaluation metric for text generation that addresses the limitations of existing metrics by reducing computational time while maintaining performance. Through extensive experiments and comparisons on machine translation and summarization tasks, we demonstrate the effectiveness and efficiency of sBERTScore in capturing semantic similarities between generated and reference text. The introduction of sBERTScore contributes to advancing the field of text generation evaluation by providing a more accurate and scalable metric, facilitating the development and assessment of high-quality text generation models. The code for reproducibility is shared at https://github.com/multitude00999/MSAI_337_project

## 2 Related work

Text generation systems are generally evaluated using annotated reference text. Various metrics have been proposed for comparing the quality of generated text with respect to the reference text. One class of such metric is exact string matching based metrics such s BLEU (Papineni et al., 2002) for machine translation, ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005). Nothjer type of metrcis are the edit distance based metrics such as TER (Schwartz et al., 2006). These metrics evaluate similarity by the amount of edit operations needed to go from candidate text to reference text. Recently, Embedding based metrics such as BERTScore (Zhang et al., 2019) have shown higher correlation with human evaluations. BERTScore uses contextualized token representation for matching instead of string matching.

## 3 Model

Given a reference text $R$ and a candidate text $C$, First we are tokenizing both of these piece of texts into sentences using sentence tokenizer (Loper and Bird, 2002). After tokenizing, the reference text and candidate text can be represented as $R = \langle R_1, R_2, ..., R_p \rangle$ and $C = \langle C_1, C_2, ..., C_q \rangle$ respectively. Here, $R_i$ and $C_i$ are the $i^{th}$ sentence in reference and candidate text respectively. $p$ and $q$ are the total number of sentences in reference and candidate text respectively.

**Sentence representation** Next we're feeding tokenized $R = \langle R_1, R_2, ..., R_p \rangle$ and $C = \langle C_1, C_2, ..., C_q \rangle$ into a pre-trained sentence BERT model (Reimers and Gurevych, 2019). The sentence BERT returns a 384 dimensional real valued vector for each sentence in $R$ and $C$ as depicted in Figure 1. We've experimented with different models for encoding the sentences and found out that the distill roberta (Sanh et al., 2019) gave best results, see section 6.

**Similarity measure** We're calculating pairwise cosine similarity between each of the sentence representations in the candidate and generated text. This results in a $pxq$ dimensional 2D matrix representing pairwise cosine similarity depicted in Figure 1 as similarity matrix.

**sBERTScore** The final score matches each of the sentences in $C$ to $R$ for calculating recall denoted by $R_{sBERTScore}$ and each of sentence in $R$ to $C$ for calculating precision $P_{sBERTScore}$ . Finally the precision and recall are combined to calculate $F_{sBERTScore}$. In figure 1 calculation of $R_{sBERTScore}$ is shown.

Our proposed metric sBERTScore is computationally more efficient than BERTScore (Zhang et al., 2019) because in BERTScore metric pairwise cosine similarity metric is calculated between each of the tokens in candidate and reference text. On the other hand, in sBERTScore pairwise cosine similarity is calculated between each of the sentences of candidate and generated text. This hugely decreases the number of cosine similarity calculations as well as size of the similarity metric see Figure 1. All of these factors significantly reduces the computational time of sBERTScore as compared to BERTScore.

## 4 Dataset

We conducted experiments across two Text Generation tasks: Machine Translation and text summarization .

**Text Summarization**: We employed the Summeval(Fabbri et al., 2020) dataset and the CNN/Daily Mail(See et al., 2017) dataset for our evaluations. The Summeval dataset served as our

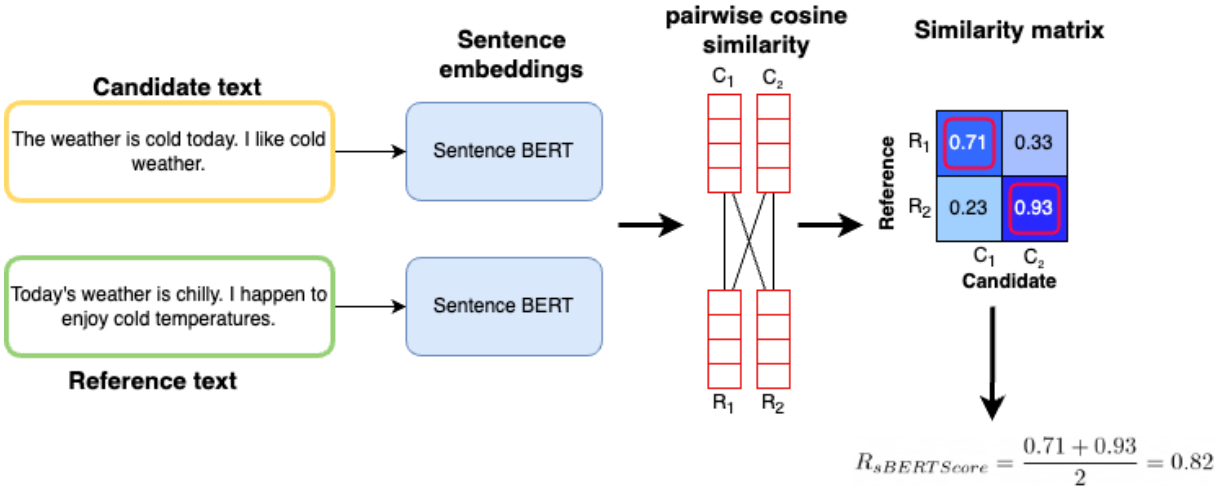


Figure 1: Illustration of calculation of recall metric $R_{sBERTScore}$. For recall maximum score is taken along the reference axis in similarity matrix. Precision can be calculated by taking maximum score along the candidate text axis.

evaluation benchmark for calculating correlation with human annotations. It consists of summaries generated by 16 different models, based on 100 source news articles, resulting in a total of 1600 examples. Each summary underwent rigorous evaluation by 5 independent crowdsource workers and 3 expert evaluators, resulting in 8 annotations per summary. The evaluations covered four dimensions: coherence, consistency, fluency, and relevance. Additionally, each source news article in the Summeval dataset was accompanied by an original reference summary from the CNN/DailyMail dataset and 10 additional crowdsourced reference summaries.

The time efficiency evaluation was performed using the CNN/Daily Mail dataset.We chose the CNN/Daily Mail dataset as it provides a diverse collection of news articles and corresponding reference summaries. This dataset allowed us to perform comprehensive evaluations of the time efficiency of different scoring methods.

**Machine Translation**: We have used the WMT18 dataset(Bojar et al., 2018) for evaluating the time taken by sBERTScore. The dataset consists of parallel corpora, which are sets of source texts and their corresponding translations. These parallel corpora are typically extracted from sources like news articles, websites, books, and other text resources.

## 5 Experiments

We conducted a series of experiments to evaluate the performance of our metric on machine translation and text summarization. Our primary objectives were to measure the correlation between our metric and human judgment and assess the time efficiency of score generation.

**Text Summarization**: We have conducted a series of experiments to evaluate the performance of our metric on text summarization. Our primary objectives were to measure the correlation between our metric and human judgment and assess the time efficiency of score generation. We employed the Summeval dataset(Fabbri et al., 2020) and the CNN/Daily Mail(See et al., 2017) dataset for our evaluations.

For this evaluation, we have selected the summaries generated by the first model in the Summeval(Fabbri et al., 2020) dataset. The quality of these summaries was assessed by computing the average human judgment scores across the four dimensions.The original reference summary from the CNN/DailyMail(See et al., 2017) dataset served as a reference point for comparison.We measured the correlation between our proposed metric, sBERTScore, and the human judgment scores across the four dimensions.To provide a comprehensive evaluation, we compared our metric with two other commonly used metrics: ROUGE score and BERTScore.

In addition to evaluating correlation, we investigated the influence of different sBERT models on the correlation scores. Six distinct variants of sBERT were selected for comparison. We computed the correlation between each sBERT model's scores and the human judgment scores across the

four dimensions.

To assess the time efficiency of score generation, we compared our proposed sBERT metric with the ROUGE score and BERT score.The evaluation was performed using the CNN/Daily Mail dataset. We recorded the time taken to generate scores using each method for all the summaries in the dataset.

**Machine Translation**: To assess the time efficiency of score generation for machine translation, we conducted experiments comparing our proposed sBERT metric with the ROUGE score and BERT score. Our objective was to evaluate the computational efficiency of these scoring methods.

## 6 Results and analysis

In this section, we present the results of our experiments comparing the time efficiency and correlation scores of different evaluation metrics for machine translation and text summarization tasks.

### 6.1 Computation time

For the machine translation task, we evaluated the time taken by three metrics: ROUGE Score, BERT Score, and sBERTScore. The results indicate that ROUGE Score had the shortest execution time, taking approximately 0.488 seconds for 200 examples. In contrast, BERT Score took around 6.872 seconds, while sBERTScore took approximately 2.944 seconds. These findings demonstrate that sBERTScore provides a notable reduction in computational time compared to BERT Score, without sacrificing performance.

For the text summarization task, a similar pattern emerged in terms of time efficiency. The execution time of ROUGE Score was approximately 0.635 seconds, while BERT Score took around 7.747 seconds. The proposed metric, sBERTScore , achieved a faster execution time of approximately 3.409 seconds for 200 examples. These results confirm the computational advantage of sBERTScore over BERT Score in the context of text summarization. A visual comparison can be found in Figure 2

### 6.2 Correlation with human annotations

To assess the correlation between the evaluation metrics and human judgment, we calculated Pearson correlation coefficients for four dimensions: relevance, coherence, consistency, and fluency. The correlation scores for different models are presented in Table 1. From the correlation results, we observe that BERT f1 achieved the highest correlation scores for all dimensions, indicating a stronger alignment with human judgment. The ROUGE metrics (rouge1, rouge2, rougeL, rougeLsum) showed lower correlation scores compared to the BERT-based metrics. Among the BERT-based models, BERT prec and BERT recall exhibited similar correlation scores, while all-distilroberta-v1 prec (Sanh et al., 2019) and all-distilroberta-v1 (Sanh et al., 2019) f1 achieved relatively higher correlation scores across the dimensions. More detailed comparison of other BERT based models is illustrated in Appendix A.2

Overall, the correlation analysis demonstrates the effectiveness of the proposed sBERTScore metric in capturing semantic similarities between generated and reference text. It showcases its potential as a promising evaluation metric for text generation tasks, providing a balance between computational efficiency and correlation with human judgment.

These results collectively highlight the advantages of sBERTScore over existing metrics, such as ROUGE and BERTScore, by addressing the limitations of computational cost and semantic understanding. The proposed sBERTScore metric offers a more efficient and comparable evaluation approach for text generation models, facilitating improved assessment of the quality and effectiveness of generated text.

## 7 Conclusion

In this paper, we introduced sBERTScore, a novel evaluation metric inspired by BERTScore, aimed at addressing the limitations of existing metrics in evaluating text generation. Our objective was to develop a metric that combines computational efficiency with semantic understanding to provide a more accurate assessment of the quality of generated text.

Through extensive experiments on machine translation and text summarization tasks, we demonstrated the effectiveness of sBERTScore in capturing semantic similarities between generated and reference text. The correlation analysis revealed that sBERTScore achieved competitive correlation scores with human judgment across multiple dimensions, highlighting its ability to align with human assessment of text quality.

One significant advantage of sBERTScore is its computational efficiency. The time comparison results clearly showed that sBERTScore outper-

| Model | relevance | coherence | consistency | fluency |
|---|---|---|---|---|
| BERT-prec | 0.3407 | 0.3206 | 0.1729 | 0.2256 |
| BERT-recall | 0.3408 | 0.2755 | 0.1548 | 0.2384 |
| BERT-F1 | 0.3764 | 0.3302 | 0.1804 | 0.2563 |
| ROUGE1 | 0.2816 | 0.2253 | 0.1677 | 0.1342 |
| ROUGE2 | 0.2146 | 0.1809 | 0.0502 | 0.1576 |
| ROUGEL | 0.2893 | 0.2972 | 0.1467 | 0.0901 |
| sBERTScore Prec | 0.3008 | 0.3462 | 0.1912 | 0.2110 |
| sBERTScore Recall | 0.2785 | 0.2380 | 0.1759 | 0.1256 |
| sBERTScore F1 | 0.3087 | 0.3161 | 0.1903 | 0.1796 |

Table 1: Pearson correlation results with human annotations on SummEval(Fabbri et al., 2020) dataset. For BERTScore and sBERTScore correlation is calculated between precion, recall, F1 and human annotator score on different criterias i.e relevance, coherence, consisitence and fluency.

formed BERTScore, providing a considerable reduction in computational time without compromising performance by much. This efficiency makes sBERTScore a more practical and scalable metric for evaluating text generation models, especially in scenarios where large-scale evaluation is required.

By leveraging sentence embeddings and pairwise cosine similarity, sBERTScore leverages contextual token representations to capture semantic understanding, overcoming the limitations of exact string matching metrics such as ROUGE. It offers a more nuanced evaluation of text generation by considering the semantic similarity between candidate and reference text.

Future research can focus on further improving the efficiency and effectiveness of evaluation metrics for text generation. Exploring alternative approaches to semantic evaluation and considering additional linguistic aspects, such as style and coherence, could enhance the evaluation process and provide a more comprehensive understanding of text quality.

In conclusion, sBERTScore presents a promising solution for evaluating text generation models. Its combination of computational efficiency and semantic understanding offers an improved evaluation metric that can benefit researchers, practitioners, and developers in the field of natural language processing.

## References


Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ond rej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Richard Schwartz, B Dorr, R Schwartz, L Micciulla, and J Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

# A Appendix

## A.1 Computation time results

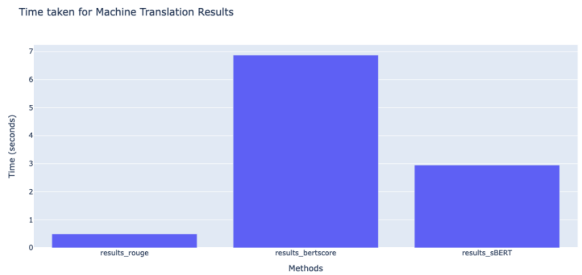


Figure 2: Computational time comparison of ROUGE, BERTScore and sBERTScore metric on 200 examples of WMT 18 dataset

## A.2 Pearson correlation results of other BERT models

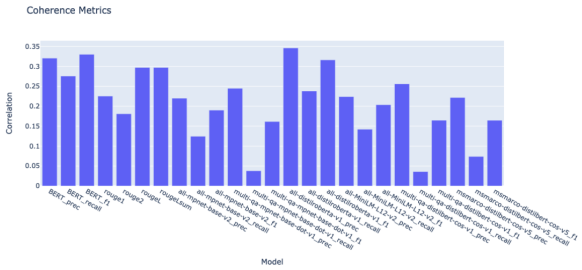


Figure 3: Comparison of pearson correlation of various models along coherence on SummEval dataset

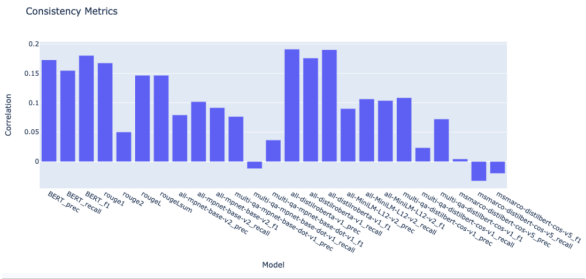


Figure 4: Comparison of pearson correlation of various models along consistency on SummEval dataset

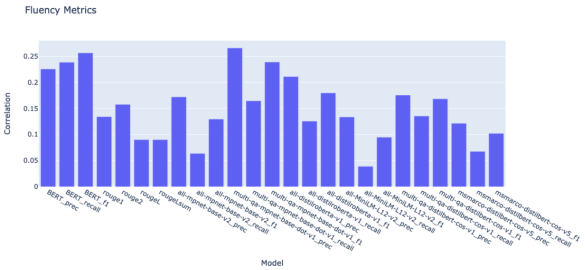


Figure 5: Comparison of pearson correlation of various models along fluency on SummEval dataset

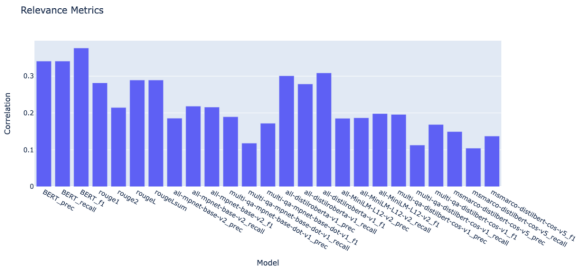


Figure 6: Comparison of pearson correlation of various models along relevance on SummEval dataset