

Default Project: Explorations in Text Summarization

Nick Lam
nyl2120

Matthew Labasan
mjl2278

Abstract

This project evaluates the performance of BART and Pegasus large language models (LLMs) for summarizing text from Reddit microblogs and books, two domains with distinct challenges in text length, structure, and summarization needs. We compare five models—BART-large, BART-large-CNN, BART-Extractive, Pegasus-large, and Pegasus-CNN-DailyMail—on the Reddit_TIFU/long corpus and the BookSum corpus, using ROUGE and BERTScore metrics. Results show that BART models seem to fair better in both abstract and extractive summarization for both the Reddit and BookSum datasets when compared to Pegasus models. Our qualitative analysis of the generated summaries also shows that BART faired better in developing abstract summaries for longer texts compared to Pegasus, which was a surprise.

1 Introduction

The goal of this project is to evaluate the capabilities of large language models (LLMs) in generating effective and accurate summaries across two distinct domains: Reddit microblogs and books. These domains were chosen because they represent extreme variations in text length, structure, and summarization requirements. While Reddit posts are short, informal, and topic-specific, books are long, narratively complex, and often require a higher degree of abstraction to distill essential themes and ideas. By systematically comparing the performance of different models on these datasets, we aim to better understand how model architectures and fine-tuning strategies influence summarization quality.

From a technical perspective, this research is important because it addresses the challenges associated with summarizing texts of varying lengths and formats using LLMs. Existing summarization techniques often excel in one domain but may fail

to generalize across others, especially when dealing with long-form texts that exceed standard token limitations. Understanding these limitations and their effects on extractive and abstractive summarization will inform future advancements in model architectures and training paradigms.

On a broader scale, effective summarization has practical applications in content aggregation, information retrieval, and accessibility. Automating high-quality summarization for diverse content types has the potential to save time, improve access to knowledge, and support decision-making across domains.

The main research questions this project seeks to address are:

1. How do LLMs perform on short-text summarization (e.g., Reddit) compared to long-text summarization (e.g., books)?
2. What are the strengths and limitations of each model architecture (e.g., BART-large, Pegasus-large, Pegasus-X) in handling the respective challenges of each domain?

By addressing these questions, this research will provide insights into how well different architectures summarize varying text lengths, contributing to the development of more robust and versatile summarization frameworks.

2 Methods

2.1 Datasets

The Reddit_TIFU/long corpus (Kim et al., 2019) consists of 42,139 Reddit posts, each paired with a TLDR style-reference summary written by users. This dataset offers a unique challenge due to its informal and topic-specific nature, and was originally collected to test a novel abstractive summarization model.

The BookSum Chapter Corpus (Kryscinski et al., 2021) consists of 12,629 samples, split into 9,712

Dataset	Size (samples)	Avg Input Length (words)	Avg Summary Length (words)
Reddit_TIFU/long	42,139	150–300	20–30
BookSum (chapters)	12,629	1,000	500

Table 1: Dataset statistics for Reddit and BookSum corpora.

for training, 1,485 for validation, and 1,432 for testing. The chapter-specific dataset includes lengthy chapters from narrative stories, alongside detailed summaries. This dataset was created to benchmark abstractive summarization methods for long, narrative texts and presents challenges like token length limits for many models.

More statistics on both of our datasets are included in Table 1. Truncated sample text and summaries for each model are included below:

Reddit_TIFU/long

- Post: "unlike most tifu this actually just happened. it is around 3:30am where i live. just as i'm about to enjoy a nice youtube binge watch before falling asleep i notice a black spot on my ceiling. being someone who has white walls and white ceilings this could only mean one of two things. either i knocked something into my ceiling or... well it was the second thing. a f***ing spider [...]"
- Reference Summary: "noticed a spider above my bed,decided to overcome my fear and try to kill it, ended up having it fall towards me get,startled and fall of my bed hurting my foot pretty bad destroying any dignity i had left and just increasing my fear."

BookSum

- Chapter:"Mine ear is open, and my heart prepared: The worst is worldly loss thou canst unfold: Say, is my kingdom lost?" SHAKE-SPEARE. It was a feature peculiar to the colonial wars of North America, that the toils and dangers of the wilderness were to be encountered before the adverse hosts could meet [...]"
- Reference Summary: "Before any characters appear, the time and geography are made clear. Though it is the last war that England and France waged for a country that neither would retain, the wilderness between the forces still has to be overcome first. Thus it is in 1757, in the New York area between the head waters of the Hudson River and Lake George to the north [...]"

2.2 Models/Approach

We evaluate five pre-trained language models for their summarization performance across the Reddit and BookSum corpora. Each model is selected based on its architecture, training objectives, and suitability for summarization tasks. The models include BART-large, Pegasus-large, BART-Extractive, and fine-tuned versions of BART and Pegasus on the CNN Daily Mail corpus. Details of the models are provided below:

- **BART-large (406M parameters)** (Lewis et al., 2020): BART is an encoder-decoder model trained using denoising autoencoding. It is particularly effective in abstractive summarization tasks for short- to medium-length texts. For this project, we use the facebook/bart-large checkpoint from HuggingFace. BART's pre-training objective of reconstructing corrupted text gives it an advantage in understanding context and generating coherent summaries.
- **BART-Extractive (406M parameters)**: To have better control on developing extractive summarization, we also use a fine-tuned version of BART designed for extractive summarization tasks. We used this specific model¹ for our experiments. We expect this model to provide a comparative baseline for extractive vs. abstractive focused models on both short-form and long-form text.
- **BART-large-CNN (406M parameters)** (Lewis et al., 2020): This model is built on BART-large, and fine-tuned on the CNN Daily Mail corpus for summarization tasks. This model will provide us with a secondary benchmark for BART-large, as it is fine tuned to summarize a dataset consisting of short to mid-length texts.
- **Pegasus-large (~570M parameters)** (Zhang et al., 2020): Pegasus employs an encoder-

¹The BART Extractive model was used from Hugging Face, available at https://huggingface.co/jordanfan/bart_extractive_1024_750.

decoder architecture with a pre-training objective called Gap Sentence Generation (GSG), which focuses on predicting key sentences in a document. This design makes Pegasus particularly effective for abstractive summarization on medium- to long-form texts. We use the google/pegasus-large checkpoint for this evaluation.

All models are evaluated using two metrics to assess summarization quality:

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** (Lin, 2004): Measures n-gram overlap between the generated and reference summaries. It is particularly effective for extractive summarization evaluation.
- **BERTScore** (Zhang et al., 2019): Evaluates semantic similarity between generated and reference summaries, making it better suited for abstractive summarization tasks.

Our model selection ensures a comprehensive analysis of summarization performance across domains and text lengths. The architecture, training objectives, and input size limitations of each model will be compared to identify their strengths and weaknesses in different summarization tasks.

2.3 Experiments

In our experiment, we primarily generated summaries of varying lengths for each model on each dataset and assessed their quality using established metrics, such as BERTScore and ROUGE. First, we set up a Python environment and installed all necessary dependencies. We downloaded the Reddit_tifu dataset locally based on directions found on TensorFlow, and the Booksum dataset through Hugging Face, selecting the ‘train’ split for both of them. Any split would have worked, as we are purely generating summaries.

We began by generating summaries with the Reddit_tifu dataset. Using the HuggingFace ‘transformers’ and ‘pipeline’ modules, we set up each model for evaluation. We selected the first 100 samples from the dataset and split each sample into the source text and the reference summary. A loop was used to iterate through each sample, where for each iteration, we generated a summary for the text portion. To encourage more abstractive and extractive summaries, we varied

the token range across three configurations: 10-50 tokens, 10-150 tokens, and 40-150 tokens. We then evaluated the generated summary by comparing it against the reference summary using BERTScore and ROUGE-1. For the Reddit dataset experiments, the ROUGE scores were most important, as the Reddit summaries were more extractive or a hybrid of abstractive and extractive.

Then, we repeated the same process with the Booksum dataset. Since this dataset contains longer text inputs, we had to self-tokenize, truncate, and detokenize the source text to a maximum of 1024 tokens for each sample to ensure model compatibility. This was also done with the HuggingFace ‘transformers’ module. We followed the same steps as above for summary generation and evaluation, though we used token ranges of 100-300, 200-500, and 500-1000 tokens to accommodate the longer texts while still encouraging both abstractive and extractive summaries. For the Booksum dataset, the BERTScores were most important, as the BookSum summaries were more abstractive.

3 Results/Analyses

3.1 Quantitative Results

We present the ROUGE-1 and BERTScore scores for different models across various summary token ranges and length penalties. The models evaluated include **BART-large**, **BART-large-CNN**, **BART-Extractive**, **PEGASUS-large**, and **PEGASUS-CNN-DailyMail**. Table 2 shows the ROUGE-1 scores for each model under different summary token ranges, while Table 3 shows the BERTScores.

3.2 ROUGE-1 Insights

Table 2 shows the performance of BART-large, BART-large-CNN, BART-Extractive, PEGASUS-large, and PEGASUS-CNN-DailyMail on the ROUGE-1 metric across different summary token ranges.

ROUGE-1 measures word-level overlap, which is important for extractive summaries since they often copy exact words or phrases from the source text.

Reddit summaries are more extractive, so when compared to the Reddit reference summaries, we see that BART-large-cnn is best at developing similar extractive summaries as it has a higher correlation with the reference summaries. We also see that the Pegasus-CNN model (trained on the same dataset), performs comparatively low in this

Summary Token Range	BART-large	BART-large-CNN	BART-Extractive	PEGASUS-large	PEGASUS-CNN-DailyMail
Reddit					
10-50 tokens	0.173	0.219	0.158	0.206	0.166
10-150 tokens	0.14	0.227	0.166	0.191	0.176
40-150 tokens	0.14	0.221	0.172	0.218	0.185
BookSum					
100-300 tokens	0.264	0.248	0.291	0.247	0.215
200-500 tokens	0.282	0.302	0.337	0.306	0.24
500-1000 tokens	0.262	0.257	0.356	0.253	0.177

Table 2: ROUGE-1 Scores for Different Models and Summary Token Ranges.

Summary Token Range	BART-large	BART-large-CNN	BART-Extractive	PEGASUS-large	PEGASUS-CNN-DailyMail
Reddit					
10-50 tokens	0.828	0.858	0.86	0.848	0.842
10-150 tokens	0.808	0.858	0.858	0.846	0.843
40-150 tokens	0.808	0.858	0.842	0.851	0.843
BookSum					
100-300 tokens	0.777	0.822	0.837	0.821	0.817
200-500 tokens	0.775	0.819	0.833	0.822	0.806
500-1000 tokens	0.775	0.793	0.822	0.795	0.771

Table 3: BERTScore for Different Models and Summary Token Ranges.

regard, showing that the model is unable to create extractive summaries as well as BART-large-CNN, which we may be able to attribute to differences in their architecture. An interesting thing to note, though, is that BART-large performs the worst among them all in each token range, indicating that having some sort of pretraining will allow one to exploit its extractive summarization better.

BookSum summaries are more abstractive, so we could expect ROUGE-1 scores to be higher in Pegasus than BART, indicating that Pegasus models generated summaries that were more closely related to the reference summaries provided. However, that is not the case. The data shows that the BART-extractive model performed the best, which could indicate that the BookSum summaries are more extractive-leaning, or that BART models have a higher degree of flexibility in abstract vs extractive summarization when it comes to longer content. However, it is important to note that ROUGE-1 is not an accurate way to see which models faired better in abstract summaries, as ROUGE-1 compares similar words, while abstract summaries are based on meaning. To this end, we must look at the BERTScores.

3.3 BERTScore Insights

BERTScore measures the semantic similarity between two texts, something important for abstractive summaries. We see that while the BART models faired better in the ROUGE-1 scores (meaning they were better at extracting text), the Pegasus scores can capture a relatively similar amount of semantic meaning (as seen in Table 3), indicating that

BART and Pegasus have similar performance in generating abstract summaries. Interestingly, Pegasus fails to do better than the two fine-tuned BART models (which are not specifically fine-tuned for abstract summarization) in both datasets. This indicates that Pegasus was unable to capture as much semantic meaning as BART did when compared to the reference summaries in either dataset.

3.4 Qualitative Results and Trends

Upon inspecting our generated summaries for the Reddit-TIFU dataset, the summaries built off of BART models all seem to repeat details from the text in a more explicit fashion than the Pegasus models. An example is included below from the 40-150 token range (indicating a preference towards extractive summaries):

- **BART-large-CNN:** This is my first post on reddit, so i'm sorry if i do anything wrong. Due to a heart condition i'm required to take medication to slow my heart down and lower my blood pressure. When i miss my meds i am highly active and excitable. My friends call me the anti-womanizer because i have the social skills of an autistic turtle but somehow i attract girls and not realize it [...]
- **Pegasus-CNN:** i just started in a new school last week, and on fridays we are allowed to eat upstairs in desks. so last week on friday i was eating lunch with the people who's desks were in my vicinity. now one of my new friends decided to bring her friends along [...]
- **Reference Summary:** a girl tried to give me

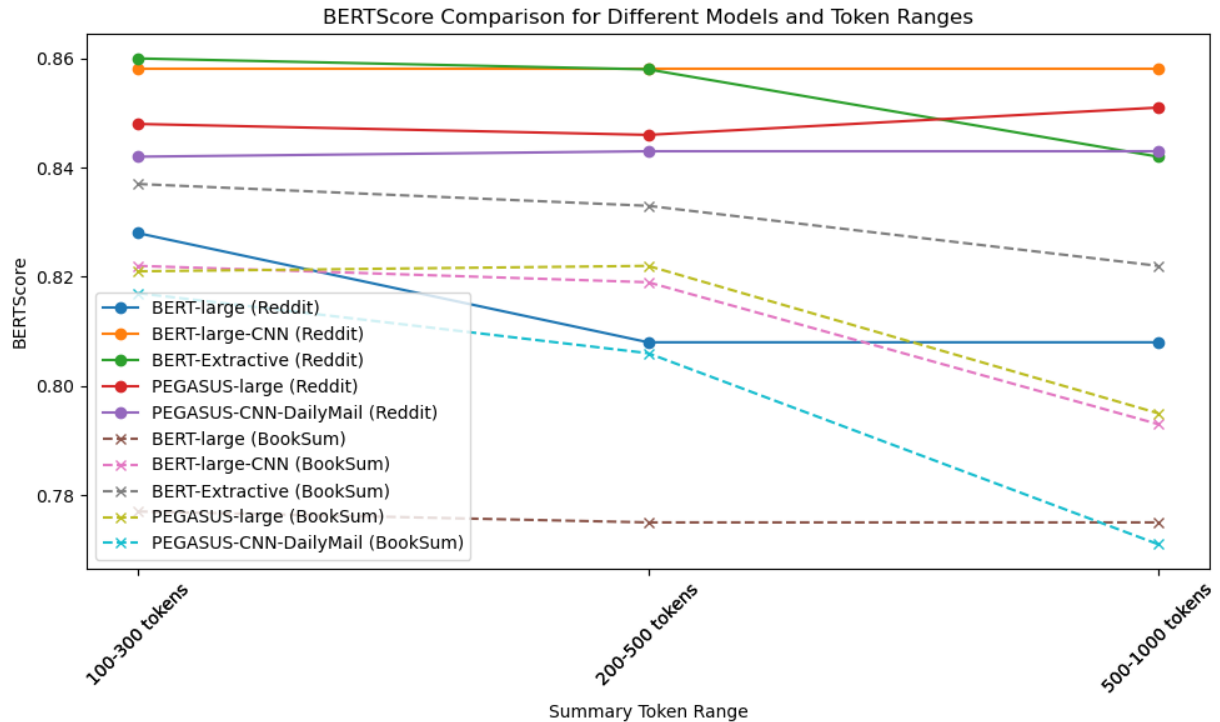


Figure 1: BERTScore Comparison for Different Models and Token Ranges.

a banana, smashed a different banana onto my desk.

However, when we look at the summaries for the BookSum dataset, the summaries built off of Pegasus models seem to be much more extractive. An example is included below from the 100-300 token range (indicating a preference towards abstract summaries):

- **Reference:** Machiavelli prefaces *The Prince* with a letter to "the Magnificent Lorenzo de Medici." In fact, the first edition of *The Prince* was dedicated to Guiliano de Medici, son of Lorenzo the Magnificent. Guiliano died in 1516, and so Machiavelli rededicated the book to one of Lorenzo's grandsons, the Duke of Urbino, who was also named Lorenzo. The opening letter is abstract enough to allow for these changes: [...]
- **BART-extractive:** This article addresses the governance of principalities (i.e., hereditary or new principalities) in Italy. Specifically, the author defines principalities as hereditary, in which the family has been long established, or new. The new are either entirely new, as was Milan to Francesco Sforza, or they are, as it were, members annexed to the hereditary state

of the prince who has acquired them, or as was Naples to that of Louis the Twelfth, King of France, in '84.

- **Pegasus-large:** I say at once there are fewer difficulties in holding hereditary states, and those long accustomed to the family of their prince, than new ones; for it is sufficient only not to transgress the customs of his ancestors, and to deal prudently with circumstances as they arise, for a prince of average powers to maintain himself in his state, unless he be deprived of it by some extraordinary and excessive force; and if he should be so deprived of it, whenever anything sinister happens to the usurper, he will regain it.

This is interesting to see, suggesting that BART works better at abstracting longer texts, while Pegasus works better at abstracting shorter texts. However, as these results are inconsistent with what is found in other research papers, our results may be due to a faulty experiment or analysis.

3.5 Overall Analysis

According to our ROUGE-1 and BERTScore results, fine-tuned for summarization BART models seem to fair better in both abstract and extractive summarization for both the Reddit and BookSum

datasets when compared to Pegasus models. Our qualitative analysis of the generated summaries also shows that BART fared better in developing abstract summaries for longer texts compared to Pegasus, which was a surprise.

4 Related Works

4.1 Challenges in Short and Long Text Summarization

Recent research has explored various challenges in summarizing short and long texts, focusing on how different models handle text length, structure, and content coherence. (Devlin et al., 2019) found that BART-based models are effective at short-text summarization, but struggle with longer documents, often failing to capture long-range dependencies across paragraphs. In contrast, long-form text summarization requires models that can retain global context and ensure coherence throughout the summary. This was different from our findings, where BART performed well across the board, while Pegasus showed weaker performance for longer texts.

Another study by (Cao et al., 2020) emphasized the limitations of many existing models when dealing with extremely long documents, highlighting the need for techniques that can handle longer input sequences without truncation. This challenge was evident in our project, where PEGASUS-CNN-DailyMail faced truncation issues on long inputs, which negatively impacted its performance. BART-Extractive, on the other hand, demonstrated more robustness in these situations, suggesting that extractive methods might be better suited for long texts.

4.2 Effectiveness of Hybrid Models

Hybrid models that combine extractive and abstractive summarization techniques have garnered attention due to their ability to leverage the strengths of both approaches. For example, (Liu et al., 2021) introduced an extractive-abstractive hybrid model that outperforms traditional models in multi-domain summarization tasks. This approach is particularly useful when summarizing texts with varied structures and lengths, as it allows the model to first extract key sentences before generating a more coherent abstract summary. In our experiments, BART-Extractive showed solid performance on long texts from the BookSum dataset, and combining extractive methods with abstractive strategies could improve performance further, especially

for tasks requiring a balance between informativeness and brevity.

Additionally, (Wang et al., 2020) explored how hybrid models could be used for summarizing short-form content such as news articles and social media posts. They found that combining extractive and abstractive approaches can enhance both the precision and fluency of summaries. This finding aligns with our own work, where BART’s extractive variant demonstrated robustness on Reddit posts, while PEGASUS’s abstractive approach worked better for BookSum’s longer, more complex narratives when compared to its own results on shorter texts. This suggests that hybrid models might be a promising direction for future research, particularly when handling both short and long texts simultaneously.

4.3 Evaluating Summarization Models on Real-World Datasets

Several studies have examined the performance of summarization models on real-world datasets that are not confined to the traditional benchmarks. For example, (Ding et al., 2021) evaluated various summarization models on user-generated content, such as social media posts and forum discussions. They noted that models trained on formal, edited text often underperform when applied to noisy, informal datasets, such as Reddit posts. Our study contributes to this line of work by applying BART and PEGASUS models to the Reddit_TIFU/long dataset, which contains informal, noisy user-generated content. We found that BART’s ability to maintain coherence in short, noisy summaries made it well-suited for this type of dataset, while also surprisingly outdoing PEGASUS in long-form content.

In a similar vein, (Belz and Gardent, 2021) focused on evaluating summarization models on domain-specific datasets, showing that models need to be fine-tuned for specific types of content to achieve optimal performance. This supports our approach of using domain-specific datasets, like Reddit and BookSum, and fine-tuning the models accordingly to account for unique text characteristics. Their work highlights the importance of customizing models to fit the dataset’s domain, and we find that BART and PEGASUS perform optimally when fine-tuned for each dataset’s specific characteristics, whether it’s short-form or long-form content.

4.4 Future Directions in Text Summarization

The challenges identified in these studies point to several future directions for text summarization research. As noted by (Cao et al., 2020), improving models’ ability to handle longer documents without truncation remains a key challenge. Our findings indicate that extractive methods, like BART-Extractive, offer a solution to this issue, though combining these with abstractive methods may further improve performance.

Moreover, (Ding et al., 2021) emphasized the need for models to be more adaptable to domain-specific content, particularly when dealing with noisy or unstructured data. In future work, we plan to further explore fine-tuning strategies tailored to different domains to improve the performance of both BART and PEGASUS across diverse text types and summarization tasks.

5 Conclusion

5.1 Project Goals and Research Questions

This project evaluates BART and PEGASUS large language models (LLMs) for summarizing Reddit microblogs and books, focusing on model performance across varying text lengths and structures. The key research questions include:

1. How do LLMs perform on short-text summarization (e.g., Reddit) compared to long-text summarization (e.g., books)?
2. What are the strengths and limitations of each model architecture (e.g., BART-large, Pegasus-large, Pegasus-X) in handling the respective challenges of each domain?

5.2 Main Findings

We evaluated five models—BART-large, BART-large-CNN, BART-Extractive, Pegasus-large, and Pegasus-CNN-DailyMail—and concluded:

Quantitative Results

- **Reddit Dataset:** BART-large-CNN outperforms Pegasus-CNN in extractive summarization, showing a higher correlation with reference summaries. Pegasus models struggle to create extractive summaries as effectively as BART models, possibly due to architectural differences. BART-large performs the worst across all token ranges, highlighting how fine-tuning impacts the summarization task.

- **BookSum Dataset:** Despite expectations that Pegasus would perform better in abstractive summarization, BART-extractive models scored higher on ROUGE-1 (meaning the generated summaries were similar to the abstractive reference). This suggests that BookSum summaries are more extractive than abstractive, or that BART is more flexible across abstract and extractive tasks for longer content. BERTScore shows that BART and Pegasus perform similarly in capturing semantic meaning for abstractive summaries. However, both fine-tuned BART models outperform Pegasus in semantic similarity despite not being specifically tuned for abstractive summarization.

Quantitative Results

- **Reddit Dataset:** BART models generate summaries that explicitly repeat details from the text, favoring an extractive approach. Pegasus summaries are less detailed and more paraphrased, reflecting weaker extractive summarization performance.
- **BookSum Dataset:** Pegasus summaries are unexpectedly more extractive than BART summaries, despite expectations that Pegasus would generate better abstract summaries. BART demonstrates better abstraction capabilities for longer texts.

5.3 Limitations and Future Work

Our limitations included the following:

- **Truncation:** The selected models were unable to use more than 1024 tokens, preventing us from using the full BookSum text.
- **Resource Limitations:** There was difficulty running summary generations locally and on the Google Colab, so we resorted to 100 summary generations for each dataset.

Future work could explore a much larger set of summaries (more than 100 samples), go into deeper qualitative analysis of the summaries, utilize faster and stronger machines to test larger models that can support more tokens such as Pegasus-X, and **evaluate the generated summaries against the actual text instead of the reference summaries**. We recognize that one mistake in our experimental design was that we computed evaluation scores on generated and reference summaries, which would

not provide the most accurate information. Evaluating alongside the original text would allow for a better analysis of whether a model did better at abstraction or extraction.

6 Contribution Statement

Both team members ideated and settled on this project idea. Nick took the lead on the Project Proposal One Pager, and Matthew took the lead on the Project Update. For the presentation in class, slides were split evenly. For this Final Project, Matthew analyzed the data and wrote the results, while Nick worked on formatting and writing the report.

References

- Anja Belz and Claire Gardent. 2021. [The domain-dependency of abstractive summarization: A survey and a case study](#). *Computational Linguistics*, 47(4):935–977.
- Yao Cao, Xiang Ren, and William Yang Wang. 2020. [Long document summarization with memory-augmented transformers](#). In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of NAACL-HLT 2019*. Association for Computational Linguistics.
- Jie Ding, Yuqing Kong, Wenjie Li, and Xuanjing Huang. 2021. [A survey of social media summarization: Tasks, methods, and applications](#). In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM 2021)*. ACM.
- Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. [Abstractive summarization of reddit posts with multi-level memory networks](#). *Preprint*, arXiv:1811.00783.
- Wojciech Kryscinski, Nazneen Fatema Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir R. Radev. 2021. [Booksum: A collection of datasets for long-form narrative summarization](#). *CoRR*, abs/2105.08209.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Ankur Mohamed, Matthew E Peters, Luke Zettlemoyer, and Kyunghyun Cho. 2020. [Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension](#). *arXiv preprint arXiv:1910.13461*.
- Chin-Yew Lin. 2004. [Rouge: A package for automatic evaluation of summaries](#). In *Proceedings of the Workshop on Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.
- Jianfei Liu, Maosong Sun, and Xipeng Qiu. 2021. [Extractive summarization with pre-trained models and generative training objectives](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Lewei Wang, Shujie Zhang, Jun Liu, Ming Zhou, Jiang Xu, and Qun Liu. 2020. [Exploring the potential of bart for extractive and abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Jiangning Zhang, Yung-Song Chuang, Hung-Yi Lee, and Hsin-Hsi Chen. 2019. [Bertscore: Evaluating text generation with bert](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1697–1708. Association for Computational Linguistics.
- Jingqing Zhang, Yao Liu, Ruslan Salakhutdinov, and Mirella Lapata. 2020. [Pegasus: Pre-training with extracted gap-sentences for abstractive summarization](#). *arXiv preprint arXiv:1912.08777*.