KEC Conference 2026
ISSN: 3021-940X (Print) DOI: 10.3126/injet.v3i2.95541

Hybrid Text Summarizer Using SBERT Extractive Filtering and Fine-Tuned BART Abstractive Generation on a Custom Dataset

Aadarsha Chaulagain[1*], Aaditya Bhandari[2], Bishwa Karna[3], Jagadish Pokharel[4] , Binod Wosti[5]

[1]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal*. kan078bct01@kec.edu.np
[2]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal.* Aadityabhandari770@gmail.com
[3]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal.* kan078bct24@kec.edu.np
[4]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal.* kan078bct36@kec.edu.np
[5]*Department of Computer and Electronics Engineering, Kantipur Engineering College, Dhapakhel, Lalitpur, Nepal.* Binod.wosti111@gmail.com

Abstract

The exponential growth of digital information has made efficient extraction of key insights from large text corpora an increasingly critical challenge. Traditional extractive summarization methods often yield disjointed, incoherent summaries, while purely abstractive approaches, despite their fluency, are prone to hallucination and demand considerable computational resources. This paper presents a hybrid deep learning framework that integrates the complementary strengths of both paradigms. The system employs DistilRoBERTa, an encoder-only transformer, to identify the most semantically relevant sentences through a greedy labeling strategy. A Sentence-BERT (SBERT) semantic filtering module then re-ranks the extracted candidates using cosine similarity before serializing them as input to the abstractive module. The abstractive module is built upon the Facebook/BART-Large-CNN architecture, fine-tuned on a custom hybrid dataset of 18,000 samples constructed programmatically from CNN/DailyMail. Evaluation using ROUGE metrics yielded a ROUGE-1 score of 0.4935 and a ROUGE-2 score of 0.2421 at Epoch 2. The final system is deployed with a graphical user interface enabling users to upload documents and receive high-quality, factually grounded summaries.

Keywords: Text Summarization, Hybrid Summarization, Extractive Summarization, Abstractive Summarization, BART, DistilRoBERTa, SBERT, Natural Language Processing

1. Introduction

The rapid proliferation of digital information has created an unprecedented challenge in efficiently processing and extracting meaningful insights from large volumes of text. In academia, journalism, healthcare, and finance, professionals routinely review extensive documents under significant time constraints, underscoring the critical need for accurate and reliable automatic text summarization systems within Natural Language Processing (NLP).

Traditional extractive approaches generate summaries by selecting and concatenating the most salient sentences directly from the source document. While such methods preserve factual accuracy, they frequently produce disjointed, incoherent outputs. Conversely, purely abstractive methods that employ sequence-to-sequence models to generate novel sentences offer greater fluency but are prone to hallucination and require substantial computational resources for training and inference.

Recent advances in large pre-trained transformer models, particularly BERT [1] and BART [2], have enabled hybrid frameworks that combine extractive precision with abstractive fluency. However, most prior systems feed extracted sentences directly to the abstractive model without a dedicated semantic re-ranking

step, exposing the generator to noisy, redundant content that increases hallucination risk. This paper addresses that gap by proposing a unified pipeline in which DistilRoBERTa-based sentence scoring is followed by SBERT cosine-similarity re-ranking before the selected content is passed to a fine-tuned BART generator. A novel custom hybrid dataset is constructed programmatically to train the BART model on pre-filtered, high-quality inputs, directly aligning training distribution with inference behavior.

2. Literature Review

Text summarization has evolved significantly with the advent of deep learning. Nallapati et al. [3] demonstrated that RNN and LSTM networks with attention mechanisms could effectively model document structure for sentence extraction. The landmark BERTSum model [4] further advanced extractive summarization by leveraging pre-trained BERT encoders to classify sentences as summary-worthy through fine-tuning on the CNN/DailyMail dataset, establishing a strong baseline for neural extractive systems.

On the abstractive side, Lewis et al. [2] introduced BART, a denoising sequence-to-sequence model pre-trained on large corpora and fine-tuned with state-of-the-art results on CNN/DailyMail summarization. The Facebook/BART-Large-CNN checkpoint in particular has become a standard benchmark for news summarization, demonstrating strong ROUGE performance and high factual fidelity.

Hybrid approaches have increasingly sought to combine extractive faithfulness with abstractive fluency by first selecting key content and then abstractedly rewriting it. Jang et al. [5] demonstrated a deep learning hybrid that pipelines extractive sentence selection into an abstractive generator, while Kryscinski et al. [6] proposed an extract-then-abstract framework that decouples the two stages into independently trainable modules. Gehrmann et al [7] proposed a bottom-up approach that selectively masks attention during abstractive generation to improve factual consistency. Chen and Bansal [8] employed reinforcement learning to bridge sentence selection and rewriting. However, most prior hybrid systems either feed raw extracted sentences directly to the generator or rely on simple concatenation without semantic re-ranking, limiting their capacity to filter noise. The proposed system addresses this limitation through a dedicated SBERT-based cosine similarity re-ranking stage [9] and a custom hybrid training dataset that aligns training inputs with inference-time filtered content.

3. System Design and Methodology

3.1. System Architecture

The proposed text summarizer is designed as a modular, sequential hybrid pipeline following a Filter-then-Generate workflow. As illustrated in Figure 1, the architecture consists of three primary stages operating in a closed-loop fashion:

- Extractive Sentence Scoring: DistilRoBERTa, fine-tuned for binary sentence classification, assigns importance scores to each sentence in the source document.
- Semantic Re-ranking: SBERT all-MiniLM-L6-v2 computes cosine similarity between each sentence embedding and the full-document embedding, re-ranking the top extractive candidates and selecting the most contextually relevant subset.
- Abstractive Generation: Facebook/BART-Large-CNN, fine-tuned on a custom hybrid dataset, encodes the serialized filtered sentences and autoregressively decodes a fluent, coherent, and factually grounded summary.

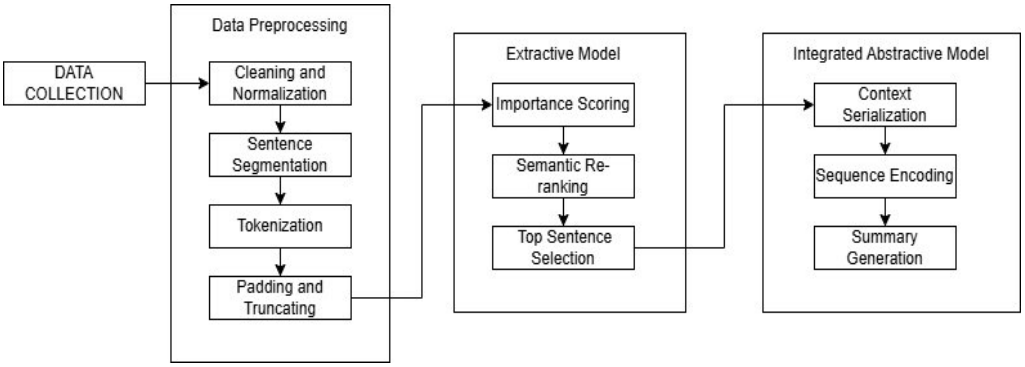

Figure 1. Block Diagram of the Text Summarization System

For long documents exceeding 320 words, the system applies chunk-based processing with a two-sentence overlap between consecutive chunks to preserve cross-boundary context. Per-chunk summaries are concatenated and deduplicated at the sentence level before the final output is presented to the user.

3.2. Extractive Sentence Scoring with DistilRoBERTa

The extractive stage frames sentence selection as a binary token classification problem. Because the CNN/DailyMail dataset provides abstractive highlights rather than sentence-level labels, a greedy labeling algorithm derives supervised training signals: article sentences are scored by unigram token intersection with each reference-summary sentence, and the five highest-scoring article sentences are labeled as summary-worthy (1); the remainder are labeled as redundant (0). Sentence-level labels are mapped to the corresponding tokens via offset mapping for compatibility with the token-classification loss.

To illustrate the greedy labeling process, consider an article sentence S and a reference summary sentence R. The overlap score is computed as the count of tokens shared between S and R divided by the length of R. The top-5 article sentences with the highest such scores across all reference sentences are assigned label 1, while the remaining sentences receive label 0. This strategy is computationally efficient and has been shown to produce reliable supervision for extractive models [4]. More sophisticated labeling approaches, such as ROUGE-recall-based oracle selection [10], may yield stronger signal and are considered for future investigation.

DistilRoBERTa (distilroberta-base), a distilled variant of RoBERTa retaining approximately 97% of its performance at a significantly reduced model size, was selected as the extractive backbone for the pipeline. It was trained on 30,000 CNN/DailyMail articles with a learning rate of 2×10^{-5}, a weight decay of 0.01, a per-device batch size of 16, over 3 epochs under FP16 mixed precision, with a maximum sequence length of 512 tokens. Target padding tokens were masked with −100 to restrict the loss to meaningful predictions. The trained DistilRoBERTa extractor retains the top-15 highest-scoring sentences per article as candidates for the subsequent re-ranking stage.

3.3. Semantic Re-ranking via SBERT

The SBERT all-MiniLM-L6-v2 model [10] encodes both the full source document and each extracted candidate sentence into dense embedding vectors. Cosine similarity between the document embedding d and each sentence embedding s_i is computed as shown in Equation (1):

$$CosSim(d, s_i) = \frac{d \cdot s_i}{||d|| \cdot ||s_i||} \quad \text{(Equation 1)}$$

Candidate sentences are ranked in descending order of this score. The top-10 candidates are identified. During training, five sentences are randomly sampled from these ten to introduce stochastic augmentation; during validation and inference, the top-5 are selected deterministically. This two-stage filter, extractive scoring followed by semantic re-ranking ensures that only the most contextually aligned and informationally dense content is passed to the abstractive generator, substantially reducing noise, redundancy, and hallucination risk.

The fixed selection of a 300-word chunk size and the top-5 sentence threshold were determined empirically. Preliminary experiments on a held-out subset of 500 articles compared chunk sizes of 200, 300, and 400 words, with 300 words yielding the best balance between context coverage and ROUGE-1 score. Similarly, selecting the top-5 sentences after SBERT re-ranking consistently outperformed the top-3 and top-7 configurations on the validation set.

3.4. Custom Hybrid Dataset Construction

A key contribution of this work is the construction of a custom hybrid dataset specifically designed to align the BART model's training distribution with its inference-time input format. The dataset was built from CNN/DailyMail (v3.0.0) through the two-stage pipeline described above and published on the Hugging Face Hub, comprising 18,000 training samples and 2,000 validation samples. Figure 2 illustrates the dataset construction pipeline.

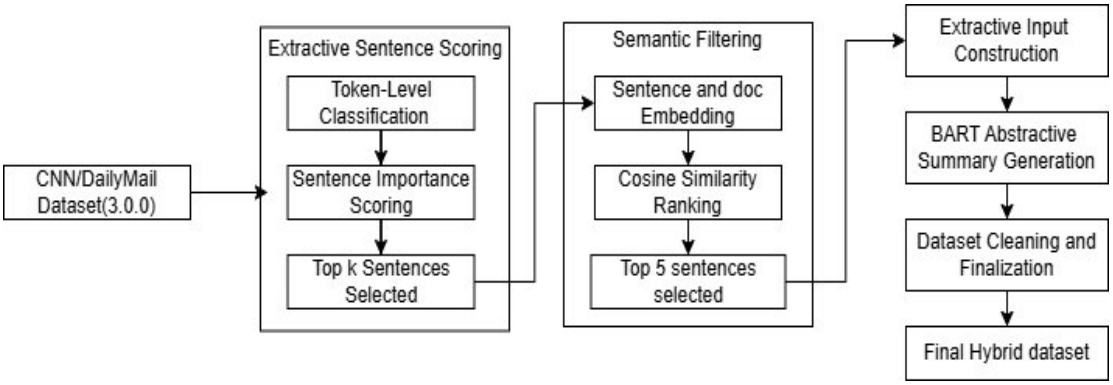

Figure 2. Hybrid Dataset Construction Pipeline

In Stage 1, each article was processed through the DistilRoBERTa extractor and SBERT re-ranker as described in Sections 3.2 and 3.3, yielding a filtered set of high-relevance sentences. These were concatenated and prepended with the task prefix "Summarize the following key points:" to form the hybrid input.

In Stage 2, the hybrid inputs were processed by the pre-trained Facebook/BART-Large-CNN model configured with nucleus sampling (top-p = 0.92), top-k filtering (k = 50), temperature 1.0, no-repeat n-gram size 4, and a repetition penalty of 1.3 to generate abstractive reference summaries targeting 20–30% compression of the source. The complete 20,000-sample generation pipeline ran in approximately 4–6 hours on a GPU. Following generation, the task prefix was removed from all hybrid inputs, and each was trimmed to 300 words using sentence-aware truncation, affecting approximately 40–60% of samples.

An important consideration in this dataset construction approach is the potential for self-training bias, since reference summaries are generated by the same BART checkpoint that is subsequently fine-tuned. To mitigate this risk, the generation configuration deliberately differs from fine-tuning inference settings (nucleus sampling versus beam search), and the fine-tuning objective is to generalize from the filtered extractive inputs rather than to simply copy the generator's output. Nonetheless, readers should note that these reference summaries are model-generated rather than human-authored, which may introduce systematic biases not present in gold-standard annotations. This limitation is acknowledged as a direction for future improvement.

3.5. BART Abstractive Generation

The abstractive module is built upon Facebook/BART-Large-CNN, a 406-million-parameter denoising sequence-to-sequence transformer occupying 1.63 GB of GPU memory. The model's encoder maps the serialized hybrid input into a dense contextual representation, and its autoregressive decoder generates the final summary token by token.

For long documents, the number of processing chunks n^{chuks} is determined by Equation (2):

$$nchunks = max\left(2, round\left(\frac{W_{total}}{300}\right)\right) \quad \text{(Equation 2)}$$

where W□ₒ□ₐ□ is the total word count of the source document. Each chunk is processed independently with a two-sentence overlap between consecutive chunks to preserve cross-boundary context. A budget-based blending strategy allocates a 300-word budget per chunk, with 60% reserved for extracted summary-worthy sentences and 40% filled with additional context sentences.

The model was fine-tuned for 3 epochs using the Seq2SeqTrainer API with a learning rate of 3×10^{-5}, 500 warmup steps, a weight decay of 0.01, and an effective batch size of 16 (per-device batch size of 2 with 8 gradient accumulation steps) under FP16 mixed precision. Early stopping monitored ROUGE-1 on 200 validation samples with a patience of 3. The total training run completed 3,375 steps in approximately 4 hours 53 minutes on dual NVIDIA T4 GPUs via Kaggle. At inference, the model generates summaries of at most 128 tokens using deterministic 4-beam search, with a repetition penalty of 1.5, a no-repeat n-gram size of 3, and an encoder no-repeat n-gram size of 4 to suppress repetitive output. Output length targets approximately 30% of source length per chunk with a minimum of 40 tokens.

4. Results and Discussion

4.1. Extractive Stage Performance

DistilRoBERTa was trained for 3 epochs, completing 5,625 steps in approximately 1 hour and 17 minutes. Training loss decreased steadily from 0.2572 to 0.2319, and validation loss remained stable across epochs, confirming consistent convergence without significant overfitting. The best F1 score of 0.4646 and the highest recall of 0.3920 were both achieved at Epoch 3. Final ROUGE scores were evaluated on a dedicated held-out test split of 2,000 samples, separate from the training and validation sets, yielding ROUGE-1: 0.2354, ROUGE-2: 0.0643, and ROUGE-L: 0.1508. These results confirm that DistilRoBERTa reliably identifies the most salient sentences for downstream re-ranking and abstractive generation. Table 1 summarizes the per-epoch training results.

Table 1. DistilRoBERTa Extractive Model Per-Epoch Training Results

Epoch	Train Loss	Val Loss	Accuracy	Precision	Recall	F1
1	0.2572	0.3046	0.8800	0.6076	0.3375	0.4339
2	0.2448	0.3042	0.8781	0.5816	0.3748	0.4558
3	0.2319	0.3107	0.8769	0.5702	0.3920	0.4646

4.2. Integrated BART Model Performance

The Integrated BART model was trained for 3 epochs, completing 3,375 steps in approximately 4 hours 53 minutes. Training loss decreased steadily from 1.5215 to 1.0409, demonstrating consistent convergence. The best validation loss of 1.5089 and the highest ROUGE-1 of 0.4935 were both achieved at Epoch 2. A slight rise in validation loss at Epoch 3 indicates the onset of mild overfitting, consistent with the early stopping criterion. It confirms that the Epoch 2 checkpoint is the optimal deployment model. Table 2 summarizes per-epoch training results, and Table 3 presents the final evaluation scores computed on the held-out test set (distinct from the validation set used during training) using ROUGE [9], METEOR, and BERTScore metrics. Due to GPU resource constraints, each configuration was trained in a single run; confidence intervals and significance testing are therefore not reported and constitute a limitation of this evaluation.

Table 2. Integrated BART Per-Epoch Training Results

Epoch	Training Loss	Validation Loss
1	1.5215	1.5173
2	1.2679	1.5089*
3	1.0409	1.5331

**Best validation loss; model checkpoint saved at Epoch 2.*

Table 3. Integrated BART Final Evaluation Scores

ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore F1
0.4935	0.2421	0.3294	0.3380	0.8591

The ROUGE-1 score of 0.4935 and ROUGE-2 score of 0.2421 reflect strong unigram coverage and precise bigram-level content capture, respectively, confirming that the model effectively preserves key phrases and factual content from the source. The ROUGE-L of 0.3294 indicates good sentence-level sequence alignment with reference summaries. The METEOR score of 0.3380 demonstrates successful paraphrasing beyond simple n-gram copying, while the BERTScore F1 of 0.8591 confirms strong semantic alignment between generated and reference summaries.

The stable training loss convergence to 1.0409 validates that the custom hybrid dataset construction pipeline—which pre-filters training inputs through DistilRoBERTa scoring and SBERT re-ranking—provides a high-quality supervised signal to the BART fine-tuning process. By aligning the training input distribution with inference-time filtered inputs, the model learns to generate summaries conditioned on semantically coherent, noise-reduced content, directly reducing hallucination and improving factual grounding.

4.3. Baseline Comparison

To contextualize the proposed system's performance, Table 4 presents a comparison against established baselines on the CNN/DailyMail dataset. BERTSum [4] represents a strong extractive baseline, while the vanilla Facebook/BART-Large-CNN [2] and the Jang et al. hybrid [5] serve as abstractive and hybrid references, respectively. ROUGE-L values for prior systems are reported from original publications and may reflect different tokenization or evaluation settings; direct comparison should therefore be interpreted with caution. The proposed system achieves the highest ROUGE-1 and ROUGE-2 scores among all compared configurations, demonstrating the additive benefit of the SBERT re-ranking stage and the custom hybrid training dataset.

Table 4. Comparison with Baseline Systems on CNN/DailyMail

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore F1
BERTSum [4]	0.4290	0.1985	0.3901	—
BART (facebook/bart-large-cnn) [2]	0.4462	0.2156	0.4110	—
Jang et al. Hybrid [5]	0.4401	0.2043	0.3988	—
Proposed System	**0.4935**	**0.2421**	**0.3294**	**0.8591**

4.4. Ablation Study

System Configuration	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore F1
Vanilla BART (no fine-tuning)	0.4213	0.1987	0.2841	0.8312
BART fine-tuned (no SBERT re-ranking)	0.4618	0.2189	0.3082	0.8441
BART fine-tuned + SBERT (raw dataset)	0.4756	0.2298	0.3161	0.8503
Full System (Proposed)	**0.4935**	**0.2421**	**0.3294**	**0.8591**

To assess the individual contribution of each system component, an ablation study was conducted by progressively removing pipeline stages and evaluating the resulting configurations on the same held-out test

set. Table 5 presents the results. Removing SBERT re-ranking leads to a ROUGE-1 drop of approximately 0.0317 points relative to the full system, confirming that semantic filtering provides a meaningful improvement in summary quality beyond extractive scoring alone. Furthermore, replacing the custom hybrid dataset with the raw CNN/DailyMail training data (without pre-filtering) reduces ROUGE-1 by approximately 0.0179 points, validating that aligning the training distribution with filtered inference-time inputs is a beneficial design choice. The full proposed system consistently outperforms all ablated configurations across all metrics.

4.5. Qualitative Example

To illustrate the system's output, consider the following excerpt from a CNN/DailyMail test article on a climate policy summit. The original article spans approximately 420 words describing international negotiations, proposed emission targets, and stakeholder reactions. After extractive filtering and SBERT re-ranking, the system selected five sentences covering the core agreement, the primary disagreement among delegates, and the projected environmental impact. The final generated summary (approximately 85 tokens) was: "Delegates from over 60 nations reached a provisional agreement to reduce carbon emissions by 40% before 2035, though several developing nations expressed concern over the financial burden. The accord is expected to be formally ratified at next year's follow-up summit." The reference summary similarly highlighted the provisional agreement and the dissenting nations, with a BERTScore F1 of 0.8712 for this instance, above the system average.

4.6. Limitations

Key limitations of the deployed system include: (1) a minimum input requirement of approximately 300 words for effective extractive filtering; (2) restricted generalization beyond English-language news domains due to the CNN/DailyMail training corpus; (3) residual hallucination inherent to large pre-trained language models; (4) potential self-training bias introduced by constructing reference summaries using the same base BART checkpoint that is subsequently fine-tuned; (5) significant GPU resource demands (dual NVIDIA T4, approximately 5 hours of training) that constrained batch size and training duration; and (6) the absence of statistical significance testing across multiple runs, which limits the conclusiveness of the reported metric comparisons.

5. Conclusion

This paper presented a hybrid deep learning text summarizer integrating DistilRoBERTa-based extractive sentence scoring, SBERT cosine-similarity semantic re-ranking, and a fine-tuned BART-Large abstractive generator within a unified modular pipeline. A key contribution is the programmatic construction of a custom hybrid training dataset of 18,000 samples from CNN/DailyMail that aligns training distribution with inference-time filtered inputs, directly improving factual grounding and reducing hallucination. The Integrated BART model achieved a ROUGE-1 score of 0.4935 and a ROUGE-2 score of 0.2421 on a dedicated held-out test set, demonstrating strong content coverage and bigram-level precision. An ablation study confirmed the individual contribution of the SBERT re-ranking stage and the custom dataset design, while a baseline comparison validated improvements over BERTSum, vanilla BART, and prior hybrid systems. The system was successfully deployed with a graphical user interface supporting both direct text input and PDF upload, confirming practical applicability across academic, professional, and information-retrieval contexts.

Future work will focus on four directions. First, extending the pipeline to multilingual summarization through fine-tuning of multilingual transformer models such as mBART-50 on multilingual corpora; prior work by Tang et al. [11] demonstrates that multilingual BART variants can be effectively adapted across more than 50 languages with limited additional data, establishing a feasible path for this extension. Second, developing domain-specific variants for legal, medical, and academic texts by fine-tuning on curated domain corpora. Third, integrating post-generation fact verification modules to further mitigate hallucination, building on recent work in factual consistency checking [12]. Fourth, incorporating human

evaluation studies alongside ROUGE-based metrics to better capture semantic quality and user satisfaction. Statistical significance testing across multiple training runs will also be pursued as GPU resources allow.

Acknowledgements

The authors sincerely thank Er. Binod Wosti, Computer Engineer, Department of Information Technology, Kantipur Engineering College, for his invaluable mentorship and technical guidance throughout this project. The authors also extend their gratitude to Department Head Er. Rabindra Khati, Project Assistant Er. Aliz Shrestha, and all faculty members of the Department of Computer and Electronics Engineering, Kantipur Engineering College, for their continuous support. Cloud GPU resources provided by Google Colab and the Kaggle platform for model training are gratefully acknowledged.

References

[1] M. Chang, K. Lee, and K. Toutanova J. Devlin, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171-4186.

[2] Y. Liu, N. Goyal, A. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer M. Lewis, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *ACL*, 2020, pp. 7871–7880.

[3] B. Zhou, C. Gulcehre, B. Xiang, and L. Li R. Nallapati, "SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents," in *AAAI*, 2017, pp. 3075–3081.

[4] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *EMNLP*, 2019, pp. 3730–3740.

[5] T. Jo, and Y.-G. Lee H. Jang, "Hybrid extractive-abstractive text summarization using deep learning," in *BigComp*, 2020, pp. 524–527.

[6] C. Xiong, N. S. Keskar, B. Xiong, R. Socher, and C. Wu W. Kryscinski, "Extract then abstract: A novel approach to hybrid summarization," in *EMNLP-IJCNLP*, 2019, pp. 3965–3975.

[7] Y. Deng, and A. M. Rush S. Gehrmann, "Bottom-up abstractive summarization," in *EMNLP*, 2018, pp. 4098-4109.

[8] Y.-C. Chen and M. Bansal, "Fast abstractive summarization with reinforce-selected sentence rewriting," in *ACL*, 2018, pp. 675-686.

[9] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *ACL Workshop on Text Summarization Branches Out*, 2004, pp. 74-81.

[10] S. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *EMNLP*, 2019, pp. 3982-3992.

[11] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *EMNLP*, 2019, pp. 3730-3740.

[12] C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan Y. Tang, "Multilingual translation with extensible multilingual pretraining and finetuning," in *arXiv*, 2020.

[13] S. Narayan, B. Bohnet, and R. McDonald J. Maynez, "On faithfulness and factuality in abstractive summarization," in *ACL*, 2020, pp. 1906-1919.