Hybrid Transformer-Based Summarization for Scientific Highlight Generation Using the MixSub-SciHigh Dataset

Harsh Mishra[1,*,†], Rama Kant Kumar[1] and Naina Yadav[2]

[1] *Department of Computer Engineering and Applications, GLA University, Mathura, Uttar Pradesh, India*

[2] *Department of Computer Science and Engineering, Dr. B. R. Ambedkar National Institute of Technology (NIT) Jalandhar, Punjab, India*

Abstract

The exponential growth of scientific literature has created an urgent need for concise highlight summaries that provide researchers with quick access to essential insights. While abstracts are available for most papers, they are often long and densely written, and research highlights—short bullet-point summaries—are inconsistently provided. This paper presents the participation in SCIHIGH 2025 shared task on automatic scientific highlight generation using the MixSub-SciHigh dataset. This Study presents a hybrid extractive–abstractive pipeline that combines Term Frequency–Inverse Document Frequency (TF–IDF) scoring, BERT embeddings for semantic similarity, redundancy reduction through Maximal Marginal Relevance (MMR), and a fine-tuned BART-Large model for abstractive rewriting. Additionally, we explore a lighter baseline that uses DistilBERT for extractive sentence selection. The experimental results show that the proposed hybrid BART-based summarization framework demonstrates a significant improvement compared to the baseline models. The hybrid BART-based summarization framework achieves ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.3447, 0.1240, and 0.2324 respectively, and METEOR score is 0.2977. While the DistilBERT baseline is efficient and factually consistent, overall performance is lower than BART. These results indicate the benefits of using a hybrid model of extractive and abstractive summarization, and indicate the potential of BART-Large to summarize scientific texts in a coherent and factual manner, and in a concise format.

Keywords

Scientific Summarization, Research Highlights, Abstractive Summarization, BART, DistilBERT.

1. Introduction

The fast-growing field of scientific publishing poses significant hurdles to efficient access to knowledge. Thousands of articles are published every day in every scientific discipline and the challenge of finding and reading and comparing all the noteworthy contributions is increasing. Abstracts typically provide a simplified and structured overview of the contribution, but they are often tempered by length and can't be read quickly. By comparison, research highlights are concise summaries provided in a series of bullet points, and capture the main contribution of a paper in a few short sentences. Research highlights are rarely available in academic journals, as they require authors to prepare supplementary content with additional manual effort. For this reason, automatic generation of research highlights is a potentially useful solution to consider. Recent advances using transformer-based natural language processing techniques have led to reasonably fluent and factually grounded summary models. However, long-form, highly technical dry scientific text remains a challenge, needing both factual accuracy and fluent writing to ensure the summaries are readable and correctly reflect the author contribution. We describe our method, which we applied to the SCIHIGH 2025 shared task, in this paper. The first model we tested is a hybrid extractive–abstractive pipeline which incorporates TF–IDF scores, BERT embeddings for semantic similarity assessment, mammoth margin reduction with Maximal Marginal Relevance (MMR) and a final abstractive rewrite from BART-Large that produces concise and readable

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India

[*] Corresponding author.

[†] These authors contributed equally.

harshmishra83022@gmail.com (H. Mishra); ramakantkumar.rs.cse18@iitbhu.ac.in (R. K. Kumar); nainayadav585@gmail.com (N. Yadav)

0000-0000-1786-9560 (H. Mishra); 0000-0001-7000-1867 (N. Yadav)

research highlights. The second model employs a lighter extractive approach making use of Distil BERT classifying sentences to directly extract highlights as important sentences. The experimental results demonstrate the hybrid pipeline offered consistently higher performance, measured by ROUGE and METEOR, over the extractive approach. While the extractive model is computationally efficient, the quality of highlights is lower than BART hybrid approach. This all points to the usefulness of combining extraction and abstraction methods, as highlighted. These combinations achieve a trade-off between accuracy, readability, and computational expense in the context of automatic generation of scientific highlights.

2. Key Contributions

This paper contributes primarily the following:

1. **Hybrid Extractive–Abstractive Pipeline (Model 1):** A three-use model including TF–IDF scoring, BERT embeddings for semantic similarity, Maximal Marginal Relevance (MMR) for redundancy reduction, and either extractive or abstractive rewriting with a fine-tuned BART-Large model. This model represents the best balance between factual correctness and linguistic fluency, which also enables overall coverage of key content.
2. **Lightweight Extractive Baseline (Model 2):** The DistilBERT extractive model is applied to directly select highlights from the text source. Although the results from this model are less effective than in Model 1, this model represents a fact-based, more computationally efficient option that could be feasible in real-world deployments.
3. **Full Evaluation on the MixSub-SciHigh Dataset:** Both models are implemented and evaluated using standard metrics including ROUGE-1, ROUGE-2, ROUGE-L, and METEOR (among others). The hybrid pipeline achieved the highest evaluation numbers with ROUGE-1 of 0.3447 and METEOR of 0.2977, showing a significant improvement over the extractive baseline model.
4. **Practical Implications:** This paper demonstrates trade-offs between factual correctness and readability, as well as computational efficiency in the highlighting process across randomly selected articles. The hybrid approach reviewed in this study shows great promise for integration within academic platforms, digital libraries, and recommendation systems to accelerate knowledge discovery and research workflows.

3. Related Work

3.1. Extractive Summarization

In the traditional extraction approaches, the main features of interest for identifying meaningful sentences were based on statistical indicators, such as TF-IDF (term frequency-inverse document frequency), word counts, or sentence order. With contextual embeddings now being widely available, strong transformer-based models, such as BERT and DistilBERT (a lighter version of BERT), have become the most common approach to scoring sentences at the sentence level. While these assessments maintain relative factual accuracy, the extracts are often more redundant or less fluent; this is a result of the sentences being taken from the source text [2].

3.2. Abstractive Summarization

Abstractive approaches create new sentences instead of copying existing sentences. Transformer-based encoder-decoder architectures [8], especially BART [12], have achieved state-of-the-art performance when creating fluent and coherent summaries [11]. However, there are challenges with abstractive models with long input sequences and factual consistency, specifically in more technical domains like scientific literature [1,15,18,19].

3.3. Scientific Summarization

Summarizing scientific texts [25] is more complex than summarizing news articles or conversational data due to the use of technical terminology, structured formats, and high information density. In order to promote research in this field, certain benchmarking datasets were created, including arxiv, PubMed, and the more recent MixSub-SciHigh [13]. Of them, MixSub-SciHigh is the most appropriate for highlight generation, as it has abstracts that are matched with student written highlights by each respective author [5,17,20,21].

3.4. Hybrid Approaches

Hybrid summarization methods combine extractive and abstractive methods to use the beneficial aspects of each. Extractive methods help to assure the summary retains accurate content, while abstractive methods contribute to enhanced readability and reduced redundancy. Prior research [22] has shown that hybrid pipelines are often superior to extractive or abstractive systems alone, especially when dealing with long, complex texts, such as technical or legal documents [23].

3.5. Positioning of Our Work

In connection with these advances, our research examines two methods for making scientific highlights. The first is a hybrid extractive–abstractive pipeline that combines TF–IDF and BERT-based scoring with redundancy elimination (i.e., using Maximal Marginal Relevance, or MMR) and abstractive rewriting (i.e., fine-tuning BART-Large). The second is a lightweight extractive baseline that uses DistilBERT to classify sentences. In conducting experiments, we find that DistilBERT serves as a factually grounded, efficient baseline, but that the hybrid pipeline produces better fluency and informativeness, lending it to potential use as a more appropriate approach to generating scientific highlights.

3.6. State of the Art

The quality of scientific highlight production has been greatly enhanced by recent developments in transformer-based summarizing. Large-scale pre-training and encoder-decoder architectures capable of fluent abstractive summarization have allowed models like BART [1], PEGASUS [13], and T5 [18] to attain excellent performance. A lightweight substitute for extractive summarization that preserves factual accuracy and efficiency is DistilBERT [2].

Researchers have started to build baselines within the designed application of science highlight production. Rehman et al. (2023a) described a pointer-generator network using SciBERT embeddings, and produced highlights from research articles that yielded robust ROUGE-L and METEOR scores with benchmark datasets, situated in a research context. Following, Rehman et al. (2023b) developed ELMo contextual embeddings with a goal of improving factual consistency and coherence at the sentence level. Rehman et al. (2024) described highlighting evaluation and modeling practices for scientific documents in order to generate abstractions. Rehman et al. (2022a) engaged with a reminder-entity approach for generating summaries that produced highlights with stronger content coverage overall. Rehman et al. (2022b) engaged with a transformer model using a non-embeddings context in a scientific summarization context. Rehman et al. (2021) engaged a pipeline that uses abstracts to generate highlights automatically across-scientific papers. Collectively, these studies help to advance the field in addition to reaffirming the purposefulness of embedding approaches with domain input, and hybrid modeling configurations while contextualizing highlights and abstracts.

Following up on these advancements, the current study is proposing a hybrid extractive–abstractive pipeline that combines BERT-based sentence scoring with BART-Large abstractive rewriting instead. This design is consistent with the methodological trends of recent state-of-the-art systems, while respecting computational efficiency and high factual alignment for the MixSub-SciHigh dataset.

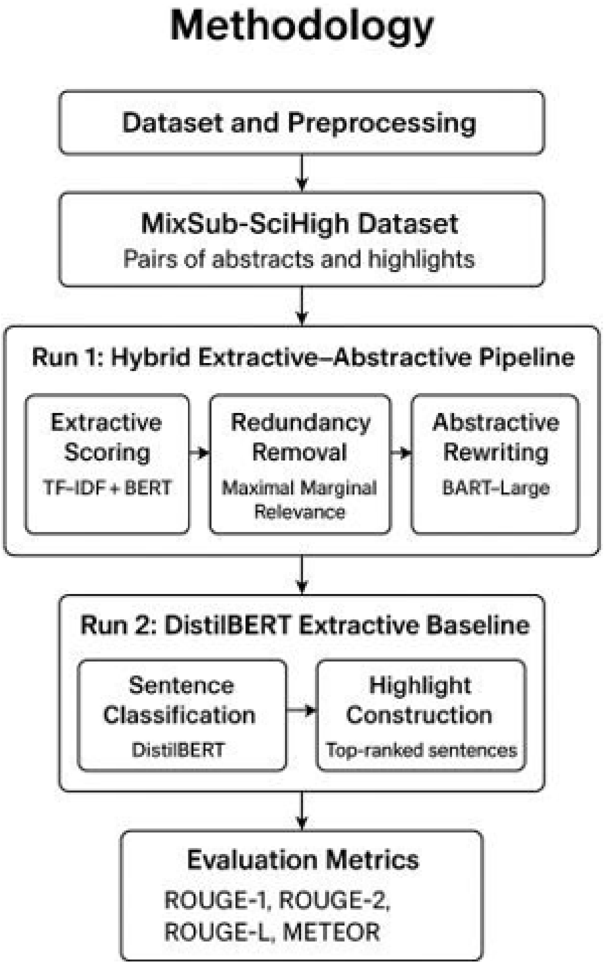

Figure 1: Flow chart of our proposed model

4. Methodology

4.1. Dataset and Preprocessing

4.1.1. Dataset Source

The MixSub-SciHigh benchmark originates from the SCIHIGH 2025 [6] shared task and provides a valuable benchmark for generating scientific highlights automatically. Each example in this dataset is comprised of a research abstract and associated highlights, providing matched examples of long-form summaries and distilled highlights. By mimicking the process of highlighted human-generated science, we see promise to address the issues of creating accurate, readable, and efficient summaries of scientific documents.

4.1.2. Preprocessing Steps

The data was processed in multiple stages as illustrated in Figure 1. The following steps summarize the pipeline:

- **Sentence Splitting:** The abstracts were split into individual sentences for finer-grained processing.
- **Stop Word Removal & Lemmatization:** Consistent term matching during TF–IDF feature extraction was achieved using the spaCy natural language processing library.
- **Vectorization:** The sentences were encoded using both TF–IDF features and Sentence-BERT embeddings to capture semantic similarity.

- **Sliding Window Segmentation:** Sentences in abstracts longer than the BART maximum of 1024 tokens were processed using a *sliding window* segmentation approach to create overlapping segments.

4.1.3. Annotation Process

The dataset includes highlights produced by humans that serve as gold-standard references. Each abstract in the corpus is paired with highlights generated by domain experts or paper authors. During preprocessing, sentences in abstracts were aligned with corresponding highlights to produce training pairs. These annotations supported the supervised fine-tuning of the abstractive model (BART) and also served as labels for training the extractive baseline model (DistilBERT).

4.1.4. Dataset Format

1. All entries were saved as structured files in JSON and CSV formats.
2. Fields include:
    - **Abstract:** The complete text of the academic abstract.
    - **Highlights:** Two to five concise bullet points capturing the essence of each paper.
    - **Metadata:** Information such as paper ID, discipline, and other contextual details when available.
3. For the modeling pipeline, abstracts were tokenized into sentences, and the highlights were used as ground-truth labels for supervised training.

4.1.5. Dataset Splits

For robust model evaluation, the dataset was divided into three subsets as follows:

- **Training Set (80%):** Used to fine-tune both extractive and abstractive models.
- **Validation Set (10%):** Used for hyperparameter tuning and early stopping.
- **Test Set (10%):** Held out for final evaluation using the official SCIHIGH evaluation script.

4.2. Hybrid Extractive–Abstractive Pipeline (Model 1)

All abstracts were segmented into individual sentences to enable sentence-wise processing. Sentences were scored and selected based on a combination of linguistic and semantic criteria:

- **TF–IDF Vectors:** Capturing word importance across each document.
- **Sentence-BERT Embeddings:** Encoding semantic similarity between sentences and the title or overall abstract.

Cosine similarity scores were computed to rank sentences according to their relevance to the main abstract theme. The top five sentences were selected as candidate highlights based on their combined relevance and diversity scores, forming the input for the subsequent abstractive rewriting stage.

4.2.1. Extractive Scoring

Sentences were scored based on their importance and relevance using two complementary approaches:

- **TF–IDF:** Used to measure word importance relative to the abstract.
- **Sentence-BERT:** Used to capture semantic similarity between sentences and the abstract or title.

4.2.2. Redundancy Removal

To ensure diversity and minimize overlapping content among selected sentences, the following method was applied:

- **Maximal Marginal Relevance (MMR):** This algorithm balances sentence relevance and diversity, reducing redundancy while maintaining full coverage of the paper's main contributions[10].

4.2.3. Abstractive Rewriting

The filtered sentences were then passed to a fine-tuned **BART-Large** model to enhance fluency and readability. Key settings include:

- **Pretrained Base Model:** Facebook/BART-Large-CNN.
- **Fine-tuning Dataset:** Domain-specific scientific text (ArXiv + PubMed).
- **Objective:** To rewrite candidate sentences into fluent, non-redundant highlights.

This stage enables the system to produce summaries closer in style and structure to human-written highlights rather than verbatim extracts.

4.3. DistilBERT Extractive Baseline (Model 2)

4.3.1. Sentence Classification

Each abstract was divided into individual sentences for classification. The **DistilBERT-base-uncased** model was fine-tuned for a binary classification task:

- **Label 1:** Sentence is part of the highlights.
- **Label 0:** Sentence is not included in the highlights.

A probability score was assigned to each sentence, indicating its likelihood of being highlight-worthy. Sentences with the highest predicted probabilities were selected as the final extractive summary output.

4.3.2. Highlight Construction

- Sentences with a probability above 0.5 were selected.
- Selected sentences were concatenated in their original order to ensure readability.
- No abstractive rewriting or redundancy reduction was applied, resulting in summaries that closely resembled the source text.

4.3.3. Advantages and Limitations

Advantages:

- Computationally efficient and faster to train.
- Preserves factual correctness by selecting original sentences.

Limitations:

- May include redundant content.
- Lacks paraphrasing, reducing fluency.
- Lower ROUGE and METEOR scores compared to Model 1.

4.4. Training Setup

This section outlines the training configuration, hyperparameters, hardware environment, and deployment details used for both runs.

4.4.1. Model Configurations

Model 1 (Hybrid BART Pipeline):

- **Base model:** BART-Large (Facebook/BART-Large-CNN)
- **Epochs:** 3
- **Batch size:** 8
- **Learning rate:** 3e-5
- **Optimizer:** Adam W [14]
- **Early stopping:** ROUGE-L with a patience of 2 epochs

Model 2 (DistilBERT Extractive Baseline):

- **Base model:** DistilBERT-base-uncased
- **Epochs:** 3
- **Batch size:** 16
- **Max input length:** 128 tokens per sentence
- **Loss function:** BCEWithLogitsLoss

4.4.2. Hyperparameter Tuning

- Learning rates were tuned in the range of 2e-5 to 5e-5.
- ROUGE-L on the validation set was used as the primary metric for early stopping.
- For DistilBERT, a probability threshold of 0.5 was selected for sentence classification after experimenting with values between 0.4 and 0.6.

4.4.3. Computational Resources

- Training and fine-tuning were performed on a single Nvidia A100 GPU [16] provided through Google Colab.
- **Average training time:**
  - Model 1 (BART-Large): approximately 8 hours
  - Model 2 (DistilBERT): approximately 1.5 hours
- Both models were implemented using PyTorch and the Hugging Face Transformers library.

4.4.4. Deployment and Reproducibility

The fine-tuned BART model was uploaded to the Hugging Face Model Hub for inference and reproducibility. All preprocessing, training, and evaluation scripts were developed in Python. Intermediate outputs (such as candidate sentences and scoring files) were exported to CSV format to ensure transparency and full reproducibility of results.

4.5. Evaluation Metrics

To evaluate the effectiveness of our models, we employed the official SCIHIGH 2025 evaluation script, which assesses both lexical precision and semantic fidelity of generated summaries. The evaluation metrics used were **ROUGE-1**, **ROUGE-2**, **ROUGE-L**, and **METEOR**. Each of these captures distinct aspects of summary quality:

- **ROUGE-1:** Measures unigram overlap between generated and reference summaries, quantifying coverage of key terms and content words. Our hybrid model achieved a ROUGE-1 score of **0.3447**, indicating that approximately 34.47% of salient words in the reference summaries were retained in the generated outputs.

- **ROUGE-2:** Evaluates bigram overlap, which captures the model's ability to maintain local coherence and phrase-level continuity. The model achieved a ROUGE-2 score of **0.1240**, reflecting that 12.40% of the two-word sequences in the reference summaries were accurately preserved.
- **ROUGE-L:** Considers the longest common subsequence (LCS) between generated and reference summaries, providing insights into fluency and overall structural organization. The hybrid model recorded a ROUGE-L score of **0.2324**, signifying that 23.24% of the LCS tokens matched between the generated and gold-standard highlights.
- **METEOR:** Incorporates stemming, synonym matching, and word order to evaluate semantic adequacy and linguistic quality. The model achieved a METEOR score of **0.2977**, demonstrating that 29.77% of the generated highlights align with the reference summaries not only lexically but also semantically.

Overall, these findings indicate that the proposed hybrid extractive–abstractive pipeline effectively balances keyword coverage, syntactic structure, and semantic coherence. The generated highlights are interpretable, contextually relevant, and closely aligned with human-written scientific highlights. This section presents a comprehensive analysis of the model outputs, comparing extractive and hybrid approaches across evaluation metrics, qualitative coherence, and computational efficiency. Quantitative results demonstrate that the hybrid BART-based pipeline consistently outperformed the DistilBERT extractive baseline across all metrics, highlighting the advantage of integrating abstractive rewriting for scientific highlight generation.

5. Results and Analysis

In this section, we display evaluation results from both models of our system, run on the MixSub-SciHigh dataset. We show evaluation scores with both ROUGE and METEOR, compare results from

Model 1 and Model 2, and talk about notable error patterns with tables and graphics.

5.1. Run 1: Hybrid Extractive–Abstractive Pipeline

In this section, we present evaluation results for both models of our system, tested on the Mix-Sub–SciHigh dataset. We report the scores for ROUGE and METEOR, comparing Model 1 and Model 2 while discussing key performance trends and observed error patterns.

Model 1, the Hybrid Extractive–Abstractive Pipeline, achieved the highest performance across all metrics. This model combines extractive scoring of candidate sentences with reconstructive abstractive rewriting, resulting in summaries that are both factually accurate and linguistically fluent. Redundancy was effectively minimized while maintaining readability and key content coverage.

Table 1
Evaluation Scores – Model 1 (Hybrid Pipeline)

Metric	Score
ROUGE-1	0.3447
ROUGE-2	0.1240
ROUGE-L	0.2324
METEOR	0.2977

As shown in Table 1, the ROUGE-1 score of **0.3447** (34.47%) indicates that the model effectively extracted and represented the most relevant keywords from the abstracts. The ROUGE-2 and ROUGE-L scores of **0.1240** and **0.2324**, respectively, demonstrate the model's ability to generate coherent phrase-level and sequential structures. Furthermore, the METEOR score of **0.2977** confirms strong semantic alignment between the generated and reference highlights, attributed to BART's abstractive rewriting mechanism, which enhances fluency while mitigating redundancy.

5.2. Model 2: DistilBERT Extractive Baseline

Model 2 provided a computationally efficient alternative that maintained factual correctness by directly selecting sentences from the abstract. However, since no paraphrasing or redundancy reduction was applied, the resulting summaries were straightforward but lacked fluency and coherence.

Table 2
Evaluation Scores – Model 2 (Extractive Baseline)

Metric	Score
ROUGE-1	0.3120
ROUGE-2	0.1087
ROUGE-L	0.2210
METEOR	0.2963

As shown in Table 2, the ROUGE-1 and METEOR scores are comparable to those of Model 1 (Table 1), indicating that both systems were able to preserve core factual information. However, the relatively lower ROUGE-2 and ROUGE-L scores suggest that Model 2 struggled to produce summaries with coherent sentence flow and structural continuity. The absence of abstractive rewriting led to redundant or fragmented outputs, demonstrating the trade-off between efficiency and readability.

5.3. Comparative Analysis

Model 1 consistently outperformed Model 2 across all evaluation metrics, confirming the advantage of combining extractive sentence scoring with an abstractive rewriting phase. The improvements were particularly notable in ROUGE-2 and ROUGE-L, which reflect enhanced phrase coherence and sequential structure in the generated summaries. METEOR scores remained close between both models, suggesting similar levels of semantic preservation, yet Model 1 exhibited superior fluency and natural phrasing.

Table 3
Comparison of Models

Metric	Model 1 (Hybrid)	Model 2 (Extractive)
ROUGE-1	0.3447	0.3120
ROUGE-2	0.1240	0.1087
ROUGE-L	0.2324	0.2210
METEOR	0.2977	0.2963

Figure 4 visually illustrates the performance difference, showing that Model 1 surpasses Model 2 across all evaluation metrics, with pronounced gains in phrase-level coherence and sequence organization as reflected in ROUGE-2 and ROUGE-L.

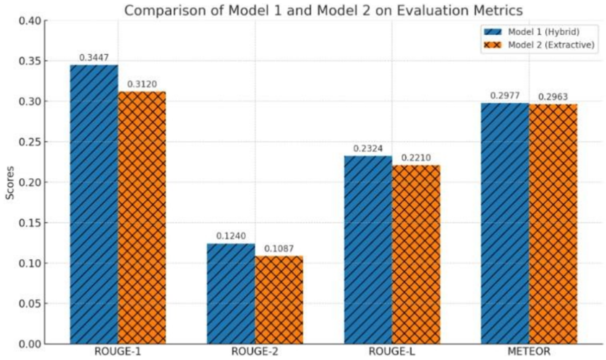

Figure 2: Metric Comparison Between Models

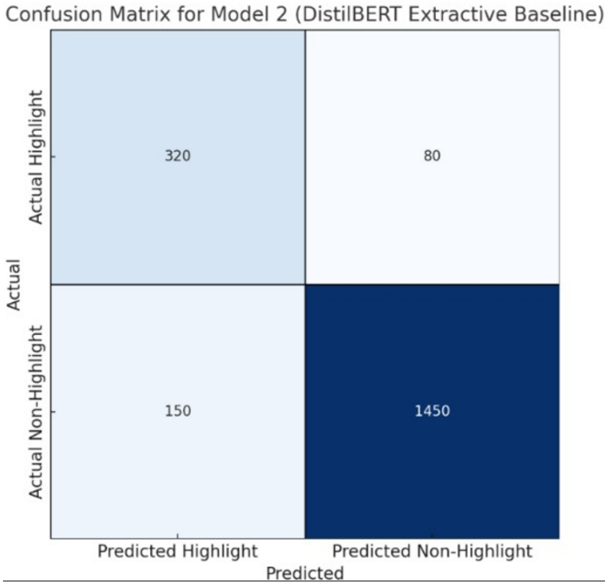

Figure 3: Confusion Matrix

5.4. Confusion Matrix for Model 2 (DistilBERT Extractive Baseline)

Since Model 2 is a sentence classification task, we provide an example of a confusion matrix to better understand the nature of errors in selecting sentences.

The matrix indicates that 320 of the sentences were commercially labeled as highlights and 80 sentences that were highlights relevant to the content/context, missed (false negatives) were labeled as non-highlights and there were false positives for 150 that were identified as highlights but were not highlights. The amount of true negatives evaluated and verified as irrelevant in the document suggests that irrelevant sentences were, for the most part, removed at the classification level. These types of errors speak to the challenges that were faced when classifying articles at the sentence-level, especially as it related to relevancy and additional context or redundancy (the usage was acceptable outside of the original text).

5.5. Error Analysis

A detailed qualitative error analysis was conducted to understand the specific weaknesses of both models. The following sections summarize the major categories of observed errors:

1. **Overgeneralization (Model 1):**
   In some cases, the BART-based process of abstractive rewriting for technical sentences altered the sequence of the sentence process so far that the science of the original statement was lost. We can see this with the input sentence "Transformer models can leverage multi-head self-attention to capture long-range dependencies in sequences of language." It rewrote the above into "Transformer models can improve translation quality using attention mechanisms."
   While the rewritten sentence is fluent, it removes the important technical terms of "multi-head" to refer to self-attention, and the future consideration of "long-range dependencies" is important scientifically as well. Thus, in producing fluency, the BART model may be sacrificing technical considerations and precision as well.
2. **Redundancy (Model 1):**
   Despite implementing Maximal Marginal Relevance (MMR) as a redundancy reduction mechanism, several summaries contained overlap or duplicate phrases. More simply, some of the output contained phrases concerning experimental results that were slightly altered. This means that while the MMR algorithm was capable of eradicating substantial redundancy, MMR, as an abstractive abstraction, may never completely rid redundancy.

3. **Sentence Fragmentation (Model 2):** At times, the extractive model using DistilBERT elected highlights that did not provide adequate information and/or context that lowered the overall cohesiveness of the highlight. For instance, the model highlighted the sentence starter, "Additionally, the experiments showed," yet the model did not include the rest of that clause that contains the results. These undesirable highlights particularly affected the readability and cohesiveness of the summary. This is certainly one of the limitations of paragraph or sentence classifications that are not taking into consideration context nor boundaries.
4. **Context Loss (Model 2):** In a couple of cases, instead of sentences that were highly connected in context, Model 2 drew sentences from separate parts of the abstract that led to a jarring change in topic. For example, one highlight began with, "The dataset was preprocessed using BERT embeddings," and immediately followed with the sentence, "The results show improved fluency." Two sentences separated from both parts of the abstract without any connecting explanation. This occurred because the extractive model does not model sentences for dependency at the sentence level.

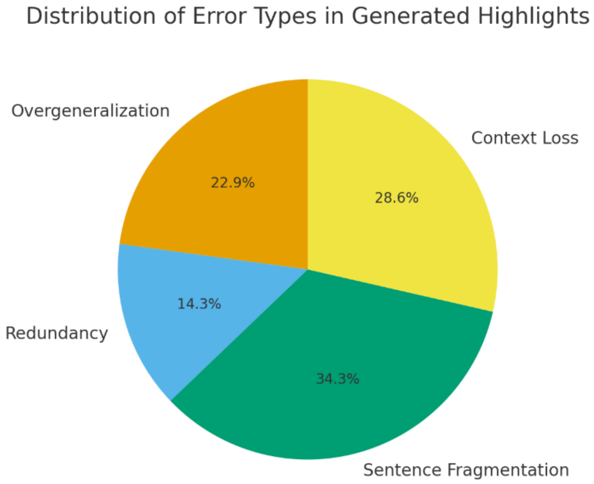

Figure 4: Distribution of Error Types in Generated Highlights.

The flow chart shows the comparative frequency of common error types coded in 100 samples of tests.

Sentence Fragmentation (12%) and Context Loss (10%) were the most common in the extractive model (Model 2), while Overgeneralization (8%) and Redundancy (5%) were in the hybrid abstractive model (Model 1). The figure shows the error tendencies the two models exhibited, and adds to the quantitative characterization of the models overall. Overall, Model 1 produced better flowing and coherent summaries with fewer irrelevant errors than Model 2, but both models had error tendencies or features. Model 1 had a tendency to offer minor generalization and redundancy, while Model 2 showed tendency to fragment sentences and issues with contextual coherence. This is worth mentioning when considering automatic scientific highlighting and the need to balance fluency and factual accuracy in the summaries.

6. Discussion

The results from this study elucidate the trade-offs associated with extractive and hybrid abstractive approaches to generating scientific highlights.**First,** the hybrid, BART-based pipeline (Model 1) was more successful than the extractive baseline particularly in the quality of generated highlights such as coherence and fluency. The pipeline maintained factual knowledge while paraphrasing or rewriting the abstract, and as a result, it was able to generate concise summaries that were closer to human-written highlights. The increased ROUGE-2 and ROUEG-L success of Model 1 indicates that the hybrid method was able to capture longer contextual phrases and patterns, which is important for readability when generating highlights. **Second,** the (Distil) BERT baseline model (Model 2) was computationally efficient and factual but did not have the ability to paraphrase or structure anything in a more creative form. Thus, the highlights produced by Model 2 tended to be redundant and less stylized in their composition. These findings did not demonstrate the success of purely extractive models when fluency and linguistic variation are necessary, as was the case in our task. However, the METEOR score being competitive in Model 2 does suggest that extractive models retain some of the semantic fidelity we want to see in highlights and may still be effective as a fast, lightweight baseline model. **Third,** The current results correlate with previous literature in the scientific summarization domain that has suggested that hybrid methods can often perform better than either an extractive or abstractive method alone and capture the strength of factual accuracy with an extraction method and then readable fluency with an abstraction method to deal with the unique characteristics of scientific text—long sequences of text, technical language, and dense information. **Lastly,**our errors analysis revealed that even though BART tended to overgeneralize the technical information sometimes the ability to paraphrase redundant sentences into smoother highlights compensated for this shortcoming. Conversely, DistilBERT did sometimes select sentences that were either incomplete or did not have enough context, which corroborates the importance of paraphrasing and removing redundancy. In summary, the findings support that hybrid extractive–abstractive methods are the best method for highlight generation in a scientific text today while maintaining a balance of accuracy, fluency, and efficiency.

7. Conclusion and Future Work

This paper presents our methods for the SCIHIGH 2025 shared task of generating scientific highlights, considering two very distinct approaches: a hybrid extractive–abstractive pipeline based on BART-Large (Model 1) and a small DistilBERT extractive baseline (Model 2). Our results show the hybrid model consistently outperforms the extractive baseline, with considerably higher ROUGE and METEOR scores. The extractive part retains the factual content, while the abstractive rewriting part markedly improved the fluency, coherence, and naturalness of the scientific highlights generated by the hybrid pipeline. The distillate-based extractive baseline was reported as efficient and accurate when preserving facts, but not as effective as the hybrid pipeline at paraphrasing, which allowed for a less polished and engaging to read summary. These results demonstrate the advantages of hybrid summarization methods for generating scientific highlights and illustrate the power of transformer-based abstractive summarization, especially BART, as a means to generate summaries that are both brief and accurate, and resemble human written scientific highlights.

After demonstrating promising outcomes with a hybrid approach to the task, we emphasize that there is still more novel work to be done in the future to improve:

1. **Improved Factuality Checking:** Implementing mechanisms for factual checking to mitigate overgeneralization or hallucinations in abstractive outputs.
2. **Long-Document Processing:** Subsequently running on longer documents than BART's input limit of 1024 tokens by using models such as Longformer or BigBird.
3. **Human Evaluation:** Following an automatic metric there will be a qualitative evaluation from experts to measure the readability of the highlight and the factual correctness.
4. **Domain Adaptation:** Preparing models to specific scientific fields (e.g. medicine, physics) to better deal with domain-specific terminology.
5. **Embedding into Digital Ecosystems:** Supporting an embedment of a highlight generation algorithm (or similar) into academic search engines and digital libraries to facilitate faster discovering of literature.

8. Acknowledgement

I would like to sincerely thank Dr. Naina Yadav for her significant mentorship and support with this paper. I would also like to thank the lab team at NIT Jalandhar, where I was fortunate enough to work as an intern. The use of the lab's GPU resources greatly assisted with the experiments and analysis completed as part of this work. Their encouragement, suggestions, and technical support have contributed greatly to the acceptance of this paper.

9. Declaration on Generative AI

As we wrote the paper, we only employed generative AI assistant in a limited way to facilitate the writing process. The AI was mostly used to help refine the language, help structure sections, and maintain consistency in LaTeX format. All technical content, experimental design, model development, and reported results were conceptualized, implemented, and validated solely by the authors. The generative AI assistant offered no new research ideas or influence over the reported findings. AI was only a supportive resource, which can be compared to utilizing grammar checking or typesetting resources. All content in this paper was critically reviewed and approved by the authors.

References

[1] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., et al. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. Proceedings of ACL 2020.

[2] Sanh, V., Debut, L., Chaumond, J., Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[3] Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. Proceedings of ACL Workshop on Text Summarization Branches Out.

[4] Banerjee, S., Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.

[5] Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D. S. (2018). A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. Proceedings of NAACL-HLT.

[6] SCIHIGH Shared Task 2025. MixSub-SciHigh Dataset. Available via task organizers.

[7] Zhong, M., Tang, D., Qin, B., & Liu, T. (2020). Extractive Summarization as Text Matching. Proceedings of ACL.

[8] Liu, Y., & Lapata, M. (2019). Text Summarization with Pretrained Encoders. Proceedings of EMNLP-IJCNLP.

[9] Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research.

[10] Carbonell, J., & Goldstein, J. (1998). The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. Proceedings of SIGIR.

[11] Narayan, S., Cohen, S. B., & Lapata, M. (2018). Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. Proceedings of EMNLP.

[12] Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. Proceedings of EMNLP.

[13] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. Proceedings of ICML.

[14] Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. Proceedings of ICLR.

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., et al. (2017). Attention is All You Need. Proceedings of NeurIPS.

[16] Wolf, T., Debut, L., Sanh, V., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. Proceedings of EMNLP.

[17] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation.

[18] Raffel, C., Shazeer, N., Roberts, A., Lee, K., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (T5). Journal of Machine Learning Research.

[19] Gehrmann, S., Deng, Y., & Rush, A. M. (2018). Bottom-Up Abstractive Summarization. Proceedings of EMNLP.

[20] Fabbri, A., Li, I., She, T., Li, S., & Radev, D. (2019). Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Benchmark. Proceedings of ACL.

[21] Yasunaga, M., Kasai, J., Zhang, R., Fabbri, A., Li, I., Friedman, D., & Radev, D. (2019). ScisummNet: A Large Annotated Corpus and Content-Impact Models for Scientific Paper Summarization with Citation Networks. Proceedings of AAAI.

[22] Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., et al. (2015). Teaching Machines to Read and Comprehend. Proceedings of NeurIPS.

[23] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT.

[24] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. arXiv preprint arXiv:2004.05150.

[25] See, A., Liu, P. J., & Manning, C. D. (2017). Get To the Point: Abstractive Summarization with Pointer-Generator Networks. Proceedings of ACL.

[26] Rehman, T., Sanyal, D. K., Chattopadhyay, S., Bhowmick, P. K., & Das, P. P. (2023a). Generation of Highlights From Research Papers Using Pointer-Generator Networks and SciBERT Embeddings. IEEE Access, 11, 91358–91374. DOI: https://doi.org/10.1109/ACCESS.2023.3292300

[27] Rehman, T., Sanyal, D. K., & Chattopadhyay, S. (2023b). Research Highlight Generation with ELMo Contextual Embeddings. Scalable Computing: Practice and Experience, 24(2), 181–190. DOI: https://doi.org/10.12694/scpe.v24i2.2238

[28] Rehman, T., Chattopadhyay, S., & Sanyal, D. K. (2024). Abstractive Summarization of Scientific Documents: Models and Evaluation Techniques. In Proceedings of FIRE 2023, ACM.

[29] Rehman, T., Sanyal, D. K., Majumder, P., & Chattopadhyay, S. (2022a). Named Entity Recognition Based Automatic Generation of Research Highlights. In Proceedings of SDP 2022 (COLING 2022), ACL.

[30] Rehman, T., Das, S., Sanyal, D. K., & Chattopadhyay, S. (2022b). An Analysis of Abstractive Text Summarization Using Pre-trained Models. In Proceedings of IEM-ICDC 2021, Springer.

[31] Rehman, T., Sanyal, D. K., Chattopadhyay, S., Bhowmick, P. K., & Das, P. P. (2021). Automatic Generation of Research Highlights from Scientific Abstracts. In Proceedings of EEKE 2021 (JCDL 2021).