Scientific Summarization: Techniques and Challenges for Extractive Summarization and Simplification of Scientific Literature

1st Prathamesh Pawar
Khoury College of Computer Sciences
Northeastern University
Boston, United States
pawar.prath@northeastern.edu

2nd Ashwin Sharan
Khoury College of Computer Sciences
Northeastern University
Boston, United States
sharan.a@northeastern.edu

3rd Matthew Chan
Khoury College of Computer Sciences
Northeastern University
United States
chan.matth@northeastern

***Abstract*—We discuss the importance of scientific summarization and simplification in natural language processing (NLP) and the challenges associated with it. It highlights the potential of fine-tuning pretrained language models such as BART and Google-Pegasus on scientific texts for this task. We investigate the effectiveness of different preprocessing techniques and evaluation metrics for scientific summarization and simplification, aiming to propose novel evaluation metrics that are tailored to scientific texts. The experiments are conducted on a large dataset of scientific articles from various domains, and the results provide insights into the opportunities and challenges of using pretrained models for scientific summarization and simplification. The paper concludes that the results are promising further research can be pursued in developing domain based fine tuning of language models to get superior summaries and simplification for better readability.**

***Index Terms*—Scientific summarization, Scientific simplification, Natural language processing (NLP), BART, Pegasus, Fine-tuning, Preprocessing techniques, Evaluation metrics, ROUGE, Domain-specific terminology, Scientific articles, Language models, Extractive summarization.**

I. Introduction

Scientific summarization and simplification are crucial tasks in the field of natural language processing (NLP) that aim to generate concise and understandable summaries of complex scientific texts. These tasks have numerous applications, including assisting researchers in quickly understanding and extracting key information from large volumes of scientific literature, aiding in science communication to non-experts, and facilitating knowledge extraction for downstream tasks such as question answering and information retrieval. Recently, pretrained models, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT-3 (Generative Pre-trained Transformer 3), have demonstrated remarkable performance in a wide range of NLP tasks. Finetuning these pretrained models on scientific texts has shown promising results in scientific summarization and simplification. However, there are still challenges that need to be addressed, such as handling the specific language and structure of scientific texts, dealing with domain-specific terminology, and addressing the readability and simplicity of the generated summaries.

Moreover, the choice of preprocessing techniques and evaluation metrics can significantly impact the performance of finetuned models. Preprocessing techniques, such as sentence splitting, tokenization, and special handling of symbols and equations, can affect the quality of input data, while evaluation metrics, such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy), may not fully capture the quality of scientific summaries and simplifications.

In this paper, we investigate the effectiveness of finetuning pretrained models for scientific summarization and simplification with different types of preprocessing techniques and evaluation metrics. We aim to contribute to the literature by exploring the impact of various preprocessing techniques on the performance of finetuned models, and by proposing novel evaluation metrics that are tailored to scientific texts. Our experiments are conducted on a large dataset of scientific articles from various domains, and our results shed light on the challenges and opportunities in leveraging pretrained models for scientific summarization and simplification.

II. Related Work

The present study focuses on summarizing and simplifying scientific articles using language models. Previous attempts to achieve this goal have been limited, but recent advancements in language models have rekindled interest in this area. In particular, the paper "Making Science Simple" [1] compares various approaches, including the use of BART, which is employed in this study, to address this problem. The article highlights the varying levels of readability and abstractness between two different datasets, which can be utilized to cater to the needs of different applications.

Another study, "Text Summarization with Pretrained Encoders" [3], adopts a similar idea of fine-tuning pretrained models to achieve text summarization. The study builds an extractive model on top of the encoder by stacking several Transformer layers, and for abstractive summarization, different optimizers are adopted for the encoder and decoder to alleviate the mismatch between the two.

An earlier work, ScisummNet [4], aimed to perform scientific article summarization by employing a combination of LSTM (Long Short-term Memory Networks) and GCN (Graph Convolution Network). The study follows a similar preprocessing approach as the present paper, which involves establishing word relations and removing non-usable data.

All the studies mentioned above use the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric [2] for evaluating the performance of their models. ROUGE provides a means of automatically assessing the quality of a summary by comparing it with other (ideal) summaries written by humans, which is an efficient measure of performance for the models presented in this paper.

III. Data Gathering and Preprocessing

- Identifying Relevant Scientific Topics: The first step in creating a list of scientific topics is to identify the areas of research or subjects that are of interest. This can be done by reviewing relevant scientific journals, conference proceedings, textbooks, and other reputable sources in the field. It is important to consider the current trends and advancements in the field and select topics that are relevant and of high interest. The list can include a wide range of topics, such as physics, chemistry, biology, computer science, environmental science, psychology, and many others.
- Researching Wikipedia as a Source: Wikipedia is a widely used online encyclopedia that contains millions of articles on various topics, including scientific subjects. It can be a valuable source of information for researchers as it provides a general overview, historical background, and references on a wide range of topics. However, it is important to note that Wikipedia is a user-edited platform and its content may not always be completely accurate or up-to-date. For the models to work the accuracy of the information hold little sway as we are only trying the increase the size of the scientific vocabulary and word relations to improve the summarizer.
- Extracting Wikipedia Articles: Once the list of scientific topics is finalized, the next step is to extract the Wikipedia articles related to these topics. This can be done using web scraping techniques or using the Wikipedia API, which allows programmatic access to Wikipedia's content.
- Extracting Wikipedia Summaries: In addition to the full articles, Wikipedia also provides summaries for many of its articles, which can be useful for obtaining a concise overview of the topic. The summaries are typically located at the beginning of the articles and provide a brief overview of the main points covered in the article. Extracting the summaries can be done using the same web scraping or API techniques used to extract the full articles.
- Cleaning and Preprocessing the Data: The methodology for data preprocessing in this study employs four distinct approaches, namely Raw, Traditional, Custom, and Combined. Each approach offers unique strengths and benefits, which enable the researchers to tailor the data processing to improve evaluation and achieve higher accuracy rates in their models. The four approaches are as follows:
  - Raw approach: The content is used in its original form to establish a baseline for further ablation research.
  - Traditional approach: The researchers use a range of traditional preprocessing techniques such as stopword removal, punctuation filtering, and tokenization. To perform these tasks, they rely on the Spacy library, known for its effectiveness in handling scientific text and related terminology. Additionally, they use the Term Frequency-Inverse Document Frequency (TFIDF) method to reduce the document's size to 1500 tokens for optimal model performance. They then perform sentence segmentation and rank the importance of each sentence using the TFIDF score. Finally, they retain only the highest ranking sentences that fit under the 1500 token limit, ensuring that only the most relevant and informative content is included in the analysis.
  - Custom approach: In this approach, the researchers remove non-informational sections of the content, such as references, notes, or symbols, before running TFIDF on it.
  - Combined Approach: The researchers use both Traditional and Custom approaches for preprocessing as part of the ablation study to facilitate a better comparison and determine whether it can provide higher accuracy.

  The four preprocessing approaches provide a comprehensive analysis of the data, which leads to more accurate results. The researchers aim to improve the accuracy rates of their models by utilizing these approaches.

IV. Transfer Learning

Transfer learning is the process of using a pretrained model trained on a general task to improve performance on a specific target task or domain with limited data. It allows for efficient adaptation of models to scientific domains, where data may be limited or expensive to acquire. Transfer learning can lead to improved accuracy, efficiency, and generalization in scientific applications.

Transfer learning is a crucial step in scientific summarization due to its ability to leverage knowledge learned from large, diverse datasets in general domains, such as news articles or social media, to improve the summarization of scientific literature. Scientific texts, including research articles, conference papers, and patents, often contain technical jargon, domain-specific language, and complex structures, which can pose challenges for traditional summarization techniques. Additionally, scientific datasets are often limited in size, making it difficult to train robust summarization models from scratch.

- Improved Accuracy: Pretrained models have already learned general language understanding and generation capabilities, which can be beneficial for summarizing scientific texts. Fine-tuning on scientific datasets allows the model to adapt to the specific language and structures commonly found in scientific literature, leading to improved accuracy in generating concise and relevant summaries.
- Efficient Adaptation: Fine-tuning a pretrained model requires less data compared to training a model from scratch. This is particularly advantageous in scientific domains where data may be limited, expensive to acquire, or require domain expertise for annotation. Transfer learning enables efficient adaptation of models to scientific texts, even with limited data, saving time and resources.
- Generalization: Transfer learning helps the model to capture general knowledge and patterns from diverse domains during pretraining, which can improve its ability to generalize to different scientific topics or domains. This allows the model to generate accurate and relevant summaries for a wide range of scientific texts, even those outside of the training dataset.

V. Model Selection

Facebook's BART model and Google's Pegasus model are both highly regarded pretrained language models that show great potential for fine-tuning in the field of scientific summarization. These models have been developed by leading technology companies and have been widely used in natural language processing (NLP) research and applications. There are several reasons why BART and Pegasus are considered to be excellent choices for fine-tuning in scientific summarization tasks.

Firstly, BART and Pegasus are both based on the Transformer architecture, which has been shown to be highly effective in capturing long-range dependencies and contextual relationships in text data. The Transformer architecture allows for efficient processing of sequential data, making it well-suited for tasks such as summarization that require understanding of the context and relationships between different parts of a document.

Secondly, BART and Pegasus are pretrained on large amounts of diverse and high-quality text data, including a wide range of scientific literature. This pretrained knowledge allows the models to capture general language understanding and semantic representations that can be fine-tuned for specific scientific domains. The large-scale training data also helps the models to learn rich representations of text data, including the syntactic, semantic, and contextual information that is essential for generating accurate and coherent summaries.

Thirdly, BART and Pegasus both have encoder-decoder architectures, which make them well-suited for sequence-to-sequence tasks such as summarization. The encoder component of the models is responsible for encoding the input text into a continuous representation, while the decoder component generates the summary based on this representation. This architecture allows for flexible and effective generation of summaries, as the models can learn to encode and decode text in a coherent and meaningful manner.

Furthermore, both BART and Pegasus have been shown to achieve state-of-the-art performance in various NLP tasks, including summarization, on benchmark datasets. Their performance has been demonstrated in scientific literature and in evaluations conducted by the original authors and other researchers.

Finally, BART and Pegasus are open-source models, which means that their architectures, pretrained weights, and fine-tuning techniques are publicly available for researchers and practitioners to use and modify. This allows for further customization and adaptation of the models for specific scientific domains or applications, making them highly versatile and flexible for different use cases.

VI. Methodology

In this experiment, four different variations of data preprocessing techniques will be defined to prepare the scientific articles dataset. Two pretrained models, such as BART and Pegasus, will be chosen for fine-tuning using the dataset variations. Performance of the pretrained models will be evaluated using appropriate metrics and statistical analysis. Results will be interpreted to draw conclusions about the effectiveness of different preprocessing variations for scientific summarization, contributing to the field's understanding and potential improvements.

- Select a scientific dataset, divide it into training and validation sets, and apply four different variations of data preprocessing.
- Train two different pretrained models, such as BART and Pegasus, using the preprocessed training data.
- Evaluate the performance of the trained models using appropriate evaluation metrics. Perform statistical analysis to determine if there are significant differences in performance.
- Analyze and interpret the results, discussing findings and drawing conclusions.
- Choose Simplification Techniques: Select appropriate techniques for simplifying the scientific summary. This may include techniques such as lexical substitution, sentence restructuring, concept simplification, or paraphrasing.
- Evaluate Simplified Summary: Evaluate the quality and effectiveness of the simplified scientific summary. This may include assessing the readability, coherence, and accuracy of the simplified summary.
- Summarize the experiment's findings and their implications for scientific summarization.

VII. Summarization Training

Here are the summarization model training results for pegasus and bart for different forms of preprocessed data. As shown in the figures below.

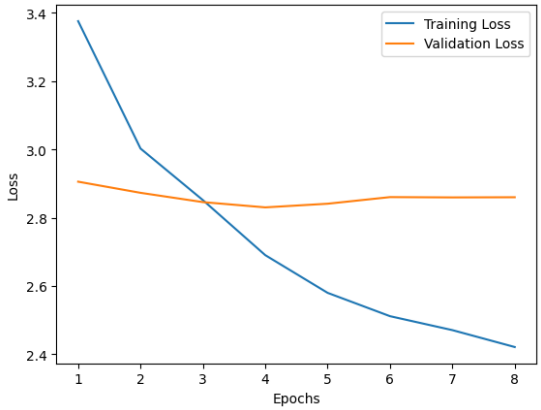

Fig. 1: Results of the bart model when fine tuned on raw data.

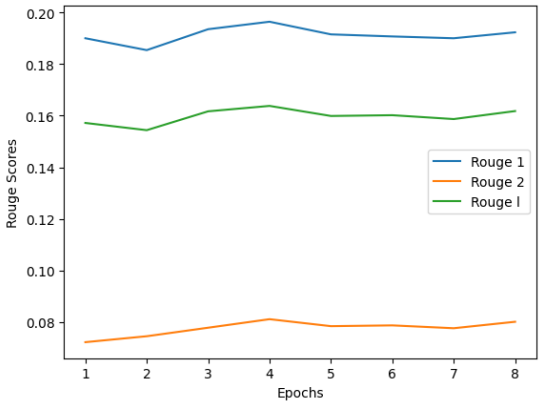

Fig. 2: ROUGE scores of the bart model when fine tuned on raw data.

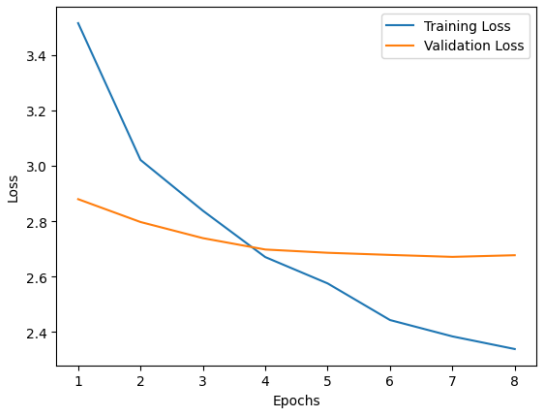

Fig. 3: Results of the bart model when fine tuned on the traditional approach.

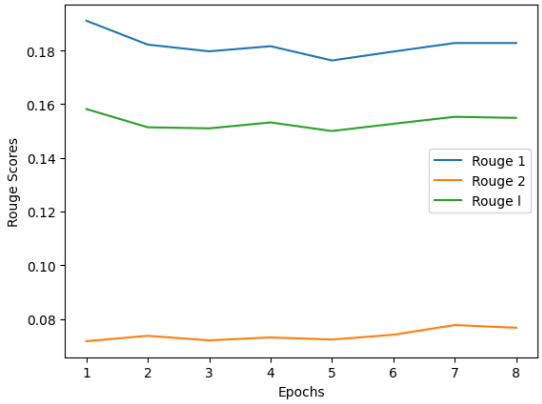

Fig. 4: ROUGE scores of the bart model when fine tuned on the traditional approach.

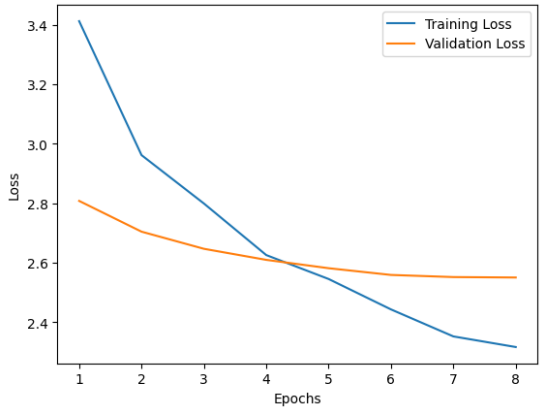

Fig. 5: Results of the bart model when fine tuned on the custom approach.

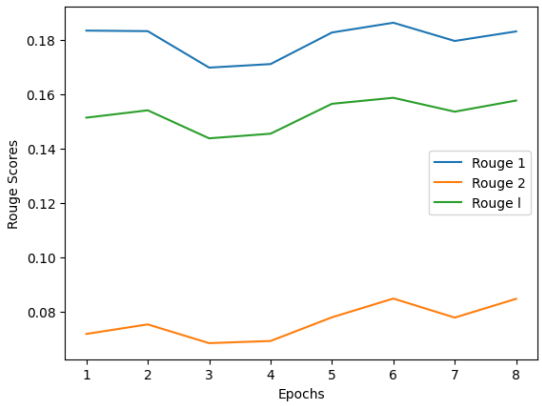

Fig. 6: ROUGE scores of the bart model when fine tuned on the custom approach.

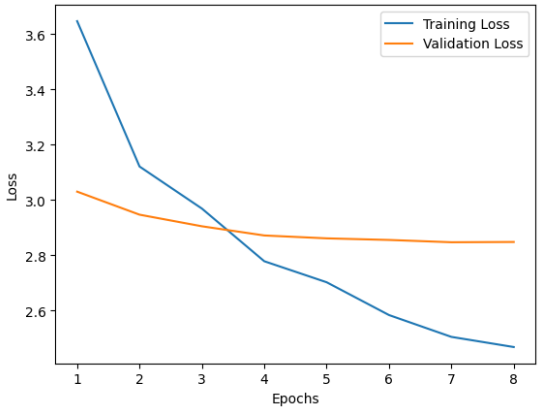

Fig. 7: Results of the bart model when fine tuned on the combined approach.

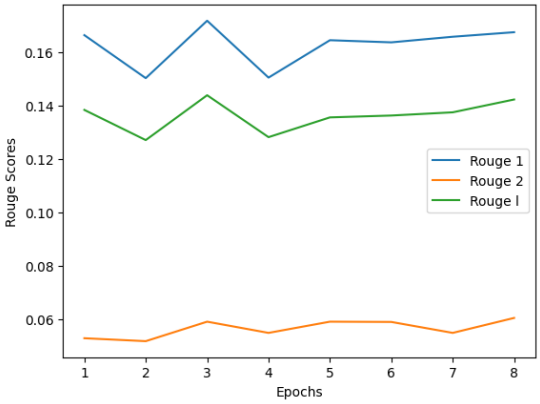

Fig. 8: ROUGE scores of the bart model when fine tuned on the combined approach.

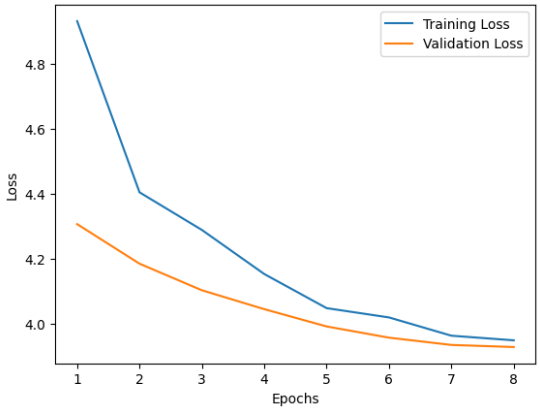

Fig. 9: Results of the pegasus model when fine tuned on the raw content.

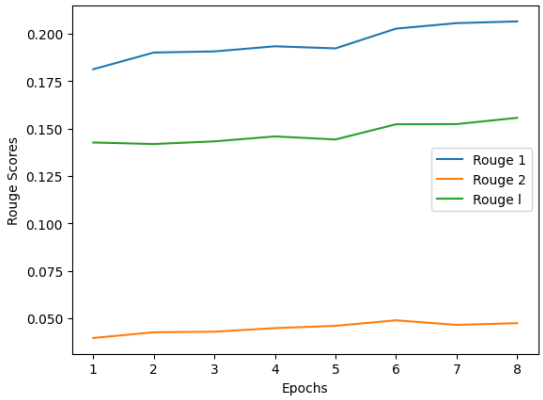

Fig. 10: ROUGE scores of the pegasus model when fine tuned on the raw content.

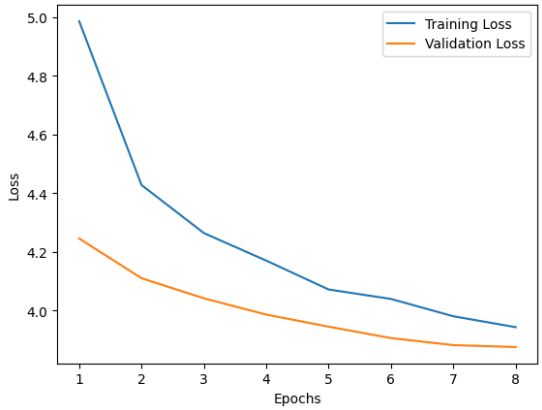

Fig. 11: Results of the pegasus model when fine tuned on the traditional approach.

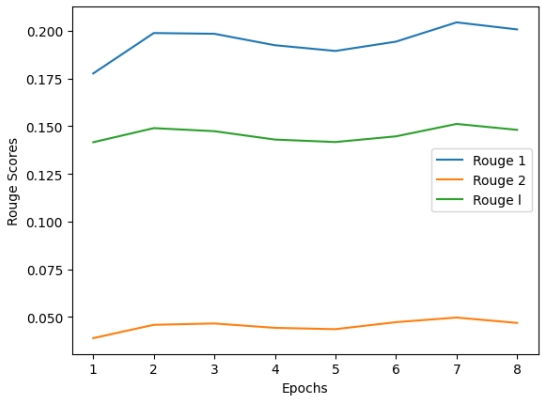

Fig. 12: ROUGE scores of the pegasus model when fine tuned on the traditional approach.

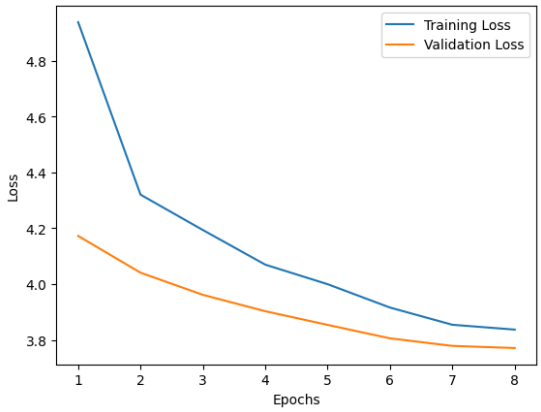

Fig. 13: Results of the pegasus model when fine tuned on the custom approach.

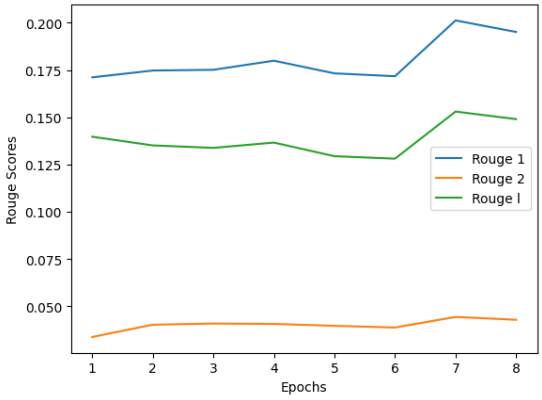

Fig. 14: ROUGE scores of the pegasus model when fine tuned on the custom approach.

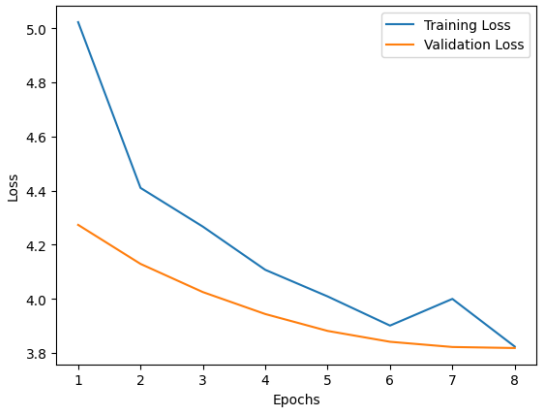

Fig. 15: Results of the pegasus model when fine tuned on the combined approach.

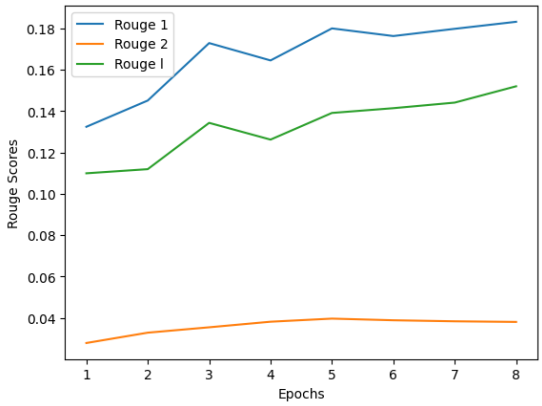

Fig. 16: ROUGE scores of the pegasus model when fine tuned on the combined approach.

VIII. Simplification

After generating the summaries, we wanted to further simplify the language and sentence structures. To ensure that the generated summaries are easy to understand, we try to simplify the summary by using simpler language and sentence structures. We looked into using a decoder architecture that uses a sequence-to-sequence model such as the Transformer decoder or Long Short Term Memory-based decoder. We found that the Transformer decoder seemed to work best using the Wikipedia datasets in keeping the meaning of the summary while simplifying the text. Specifically Hugging Face's Transformers library seemed to have the most available models and support for text simplification tasks. We found Google's T5-base model to work best for our use and this was based on human evaluation of the outputs as well as relevant metrics like ROUGE-2 score. Using the T5-base model with the Transformer decoder achieved the best simplification in our generated summaries.

IX. Evaluation

Chin-Yew Lin proposed a unique metric to measure the quality of the generated summary with respect to reference summary

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of evaluation metrics commonly used in natural language processing (NLP) tasks, such as summarization, to assess the quality of generated summaries compared to reference summaries. There are several variations of ROUGE scores, including ROUGE-1, ROUGE-2 and ROUGE-L each with its own significance:

- ROUGE-1: Also known as unigram ROUGE, measures the overlap of unigrams (single words) between the generated summary and the reference summary. It evaluates the quality of the generated summary in terms of word-for-word matches with the reference summary.
- ROUGE-2: Also known as bigram ROUGE, measures the overlap of bigrams (two consecutive words) between the generated summary and the reference summary. It

captures the quality of the generated summary in terms of consecutive word pairs that match those in the reference summary.

- ROUGE-L: Measures the longest common subsequence (LCS) between the generated summary and the reference summary. It considers the longest contiguous sequence of words that match between the two summaries, which allows for some variation in word order and captures the coherence and fluency of the generated summary.

The significance of these different variations of ROUGE scores lies in their ability to capture different aspects of summary quality, such as word-for-word matches, word order, coherence, fluency, and semantic similarity. They provide a comprehensive evaluation of the generated summary's performance compared to the reference summary, helping researchers and practitioners assess the effectiveness of their summarization models and make informed decisions in their NLP tasks.

X. Results

Average Score Evaluation strategy: Here we ran ten datapoints to evaluate the accuracy of the models by using rouge-1, rouge-2 and rougel on a variation of preprocessing approach's.

Models and Approaches	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Pegasus Base	24.1323	5.6319	22.14926
Pegasus Raw	25.82989	6.44946	23.27936
Pegasus Custom	26.47427	6.02105	23.0815
Pegasus Traditional	25.1203	5.81051	22.5527
Pegasus Combined	23.54458	6.24217	21.936
Bart Base	22.707	3.80962	21.6978
Bart Raw	24.0126	4.6726	23.1023
Bart Custom	23.9828	5.262	22.6811
Bart Traditional	22.104	5.30546	20.5785
Bart Combined	20.33895	4.36807	19.43074

The evaluation results from using the google's t-5 simplification model

Simplification Model	ROUGE-1	ROUGE-2	ROUGE-L
google T-5 base	0.57575	0.4878	0.57575

XI. Conclusion

In conclusion, our experimental findings demonstrate that preprocessing techniques play a crucial role in influencing the performance of the model. Moreover, our results provide compelling evidence that fine-tuning the Pegasus model with custom data leads to significant improvements in the summarization performance compared to the base model. Furthermore, our findings suggest that additional simplification of the summary can facilitate the dissemination of new scientific developments and discoveries to a broader audience, thereby expanding the impact of scientific research in diverse communities. These findings underscore the importance of carefully curating and preprocessing data for fine-tuning and summarization tasks, and highlight the potential of leveraging advanced language models like Pegasus for enhancing scientific communication and knowledge dissemination

References

[1] Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. Making science simple: Corpora for the lay summarisation of scientific literature. *arXiv preprint arXiv:2210.09932*, 2022.

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

[3] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*, 2019.

[4] Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7386–7393, 2019.