News Article Summarization Using Advance NLP Techniques - Team 07

Course: 16:954:577:01 - Statistical Software

Team Members:

Anirudh Gaur (216007356)

Rishik Shekar Salver (217001710)

Xinran Zhao (218009814)

Department of Statistics & Data Science

Rutgers University

New Brunswick, NJ

Supervised by

Prof. Ajita John

Note: Presentation to Cohort is completed (Cohort -3)

1. Introduction

The burgeoning volume of information available on the internet has spurred considerable interest in the research community to develop automated text summarization systems. These systems play a crucial role in condensing vital information from lengthy documents into concise summaries, facilitating more efficient information consumption. The inspiration for the exploration comes from a research paper [1]. The two primary methods of document summarization are classified as Extractive and Abstractive.

Extractive summarization involves the selection and extraction of key sentences or phrases directly from the original document. This method relies on identifying the most relevant information without generating entirely new content. Common techniques in extractive summarization include sentence scoring, where algorithms like TextRank or LexRank assess sentences based on features such as length and word frequency, and graph-based methods like PageRank, which represent sentences as nodes in a graph to determine importance.

On the other hand, Abstractive summarization is a more sophisticated approach, aiming to generate concise summaries by interpreting and rephrasing the content. This method requires a deeper understanding of language semantics. Techniques in abstractive summarization include sequence-to-sequence models, which use neural networks with an encoder-decoder architecture, and attention mechanisms that enhance performance by allowing models to focus on specific parts of the input. Additionally, reinforcement learning and the utilization of pre-trained language models like BERT or GPT have shown success in improving the quality of abstractive summaries.

While extractive summarization offers a straightforward method of condensing texts by extracting key sentences, abstractive summarization seeks to create a more nuanced and human-like summary by understanding and rephrasing the original content. The choice between these methods depends on the specific requirements and complexities of the summarization tasks. The abstract method demands a high level of complexity in natural language processing [2].

2. Dataset & Exploratory Analysis

The CNN/DailyMail Dataset comprises over 300,000 news articles from CNN and the Daily Mail. The dataset is adapted for extractive and abstractive summarization tasks, measured by ROUGE scores. The dataset includes article and highlight strings, with an average token count of 781 for articles and 56 for highlights. Here is the dataset link Link. For computational convenience, only 10% of the train and test data was picked out randomly. There are 28711 lines in train data and 1149 lines in test_data, the summary and article length distribution can also be seen in the below graph.

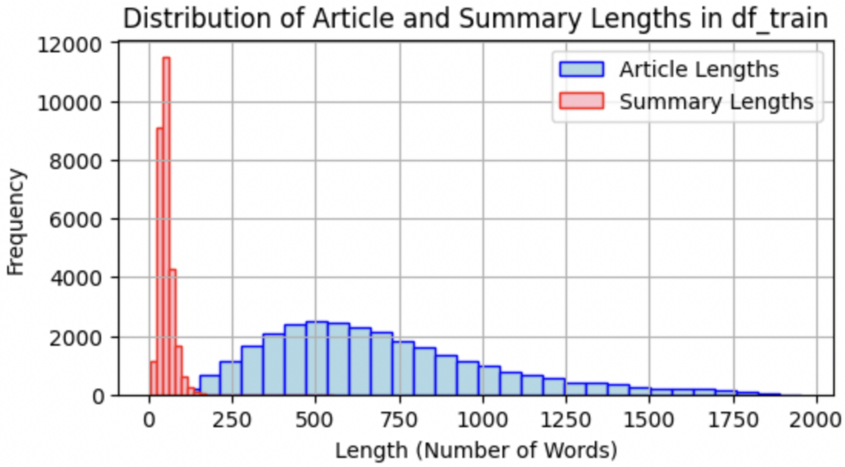


Fig 2.1: Number of words in train and test dataset

In the data preparation phase, Regular Expressions were meticulously applied for data cleaning to enhance the quality of the dataset. This process involved stripping escape characters, reducing repeated characters, and eliminating multiple spaces. Upon cleaning the data, the data was structured for the text summarization model by adding 'sostok' and 'eostok' tokens. These tokens are instrumental in signaling the beginning and end of text sequences, enabling the model to effectively understand and generate coherent summaries.

3. Seq2Seq Baseline Model

The baseline text summarization model adopts a seq2seq architecture, utilizing LSTM cells for both the encoder and decoder components. This dual-component approach involves the processing of input sequences in the encoder with pre-trained word embeddings and the generation of summaries in the decoder using LSTM cells. Employing the Keras Tokenizer facilitates the tokenization of reviews, enabling one-hot encoding in the target variable (Y) and ensuring uniformity in sequence lengths through padding. This architecture, while not the most intricate, serves as a good starting point for our baseline model.

For the latent space and word embeddings, a 300-dimensional latent space and a 200-dimensional embedding space are chosen. The model undergoes training using the RMSprop optimizer and sparse categorical cross-entropy loss function. To mitigate overfitting, early stopping is implemented with a patience of 2 epochs. Model performance is assessed using metrics like ROUGE scores, validating its effectiveness in abstractive text summarization. The training process spans 30 epochs, during which the convergence of the model is monitored for optimal performance.

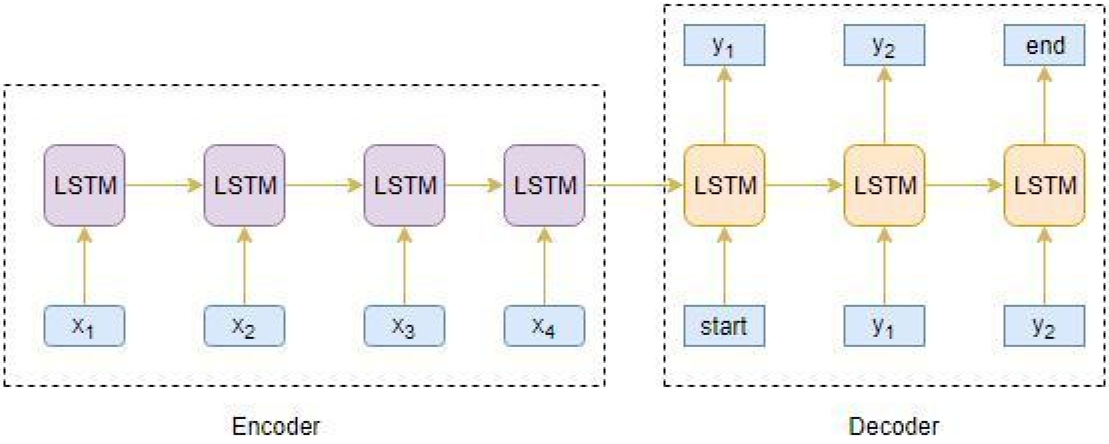


Fig 3.1: Example architecture of a seq2seq model with LSTM for predicting summaries

The baseline model configuration comprises a three-layer LSTM encoder with dropout (0.4) and recurrent dropout (0.4) to prevent overfitting. On the decoder side, a single LSTM layer with dropout (0.4) and recurrent dropout (0.2) is employed. The TimeDistributed dense layer with softmax activation generates probability distributions over the target vocabulary.

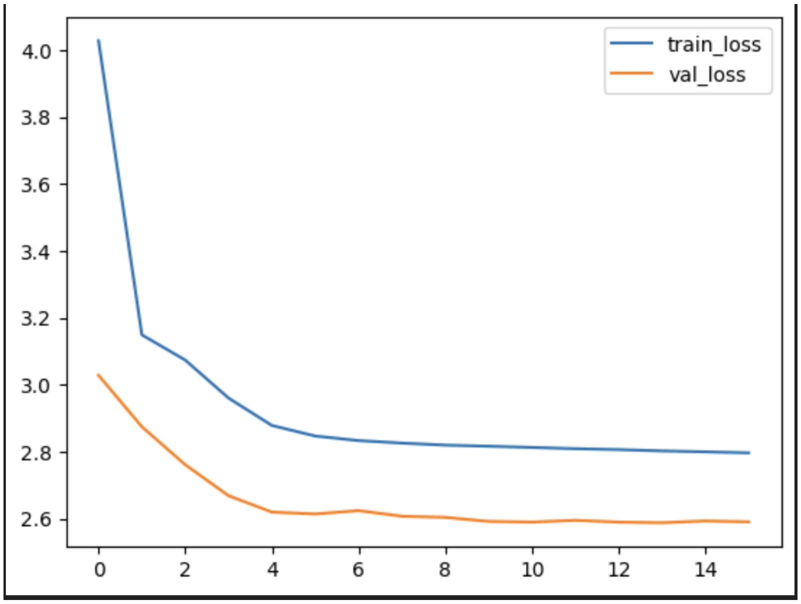


Fig 3.1: Train and validation loss for seq2seq

The graph illustrates the trend of training and validation loss over epochs. Initially, both training and validation loss decrease, indicating the model's learning and adaptation to the training data. However, around the 10th epoch, the reduction in loss stabilizes, suggesting that the model has acquired substantial knowledge from the training data. Training is halted at the 14th epoch to prevent potential overfitting and ensure the model generalizes well to unseen data. This decision strikes a balance between optimizing performance and avoiding unnecessary computational costs associated with extended training.

4. T5 Model

Diving into the next approach for abstractive summarization, a more complex model i.e. T5, was chosen. Text-to-Text Transfer Transformer, is a transformer-based language model developed by Google. It is an encoder-decoder model and is trained using teacher forcing, making it versatile for various language generation tasks. T5's text-to-text framework allows it to be easily adapted for summarization tasks. The model architecture is a standard sort of vanilla encoder-decoder transformer.

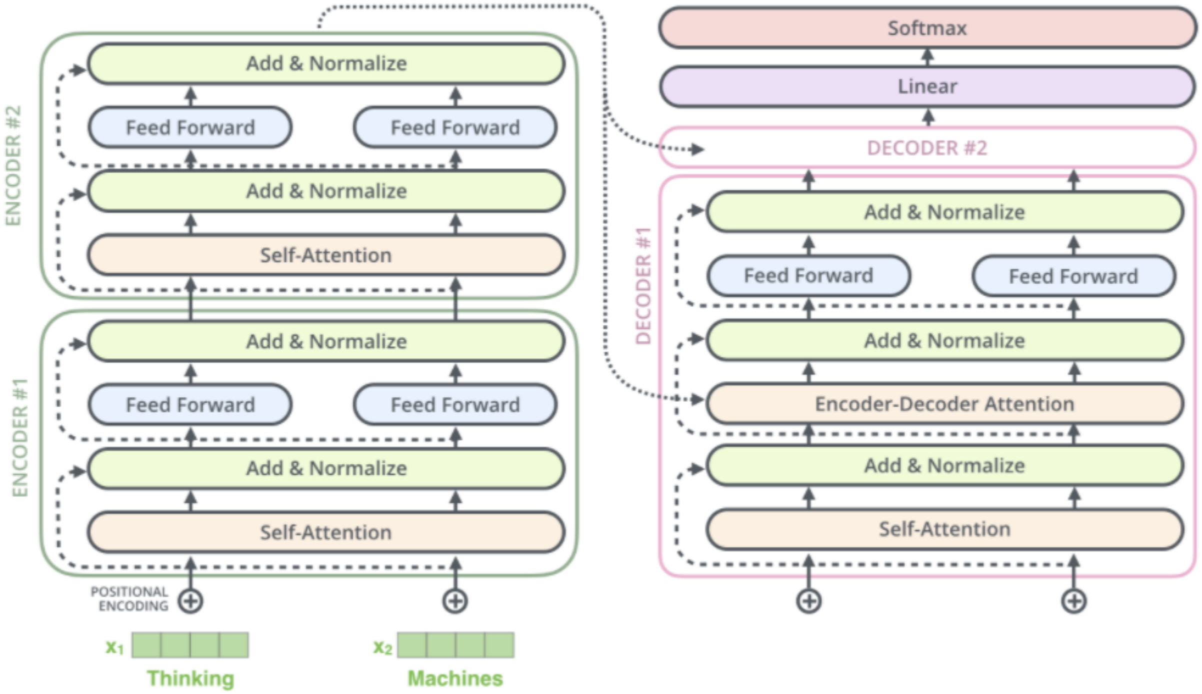


Fig 4.1: Architecture of the T5 model

Initially, the pre-trained T5 base model was used without any fine-tuning to predict the summaries. The model has 24 layers in each encoder and decoder and has around 220M parameters. The model employs relative scalar embeddings, which, unlike traditional word embeddings that capture word meanings, focus on encoding the positional information of tokens within a sequence. In this approach, conventional word embeddings were not utilized, as positional encodings were found to be particularly beneficial for the summarization task.

To process input sequences, the AutoTokenizer from T5 was used, ensuring the data was appropriately formatted for the model. Each input sequence, prefixed with 'summarize:', was systematically fed into the model's encoder. This strategic conditioning helps the T5 model associate the task of generating abstractive summaries with sequences that begin with this specific string.

After the summaries were generated on the test data, the ROUGE score was used as the metric to assess the performance of the predicted summaries.

To further improve the model, the T5 model was fine-tuned to generate summaries that are more concise and capture the nuances. Fine-tuning the T5-base model involved experimenting with various training parameters. A custom dataset structure was developed, optimizing input_ids and labels for seamless integration into the T5 model.

The focus on creating a custom dataset structure was crucial, ensuring optimal preparation of input_ids and labels for effective use in the T5 model. This strategic approach aimed to enhance the model's ability to grasp the nuances of the summarization task. Throughout the training phase, there was a priority on limiting the maximum length of the article to 512 and the summary to 150, selectively excluding longer articles. This deliberate choice aimed to strike a balance between model performance and computational efficiency.

The training was done over 3 epochs, and a maximum length of 512 for the article and 150 for the summary was set, excluding longer articles. Again, the ROUGE score was used as the metric to compare the generated summaries with the gold labels from the trained model.

5. Extractive BERT Summarization

Exploring beyond abstractive models, the extractive summarization approach was employed to evaluate its performance and draw comparisons with the abstractive models. This analysis aimed to compare the strengths and differences between the two approaches.

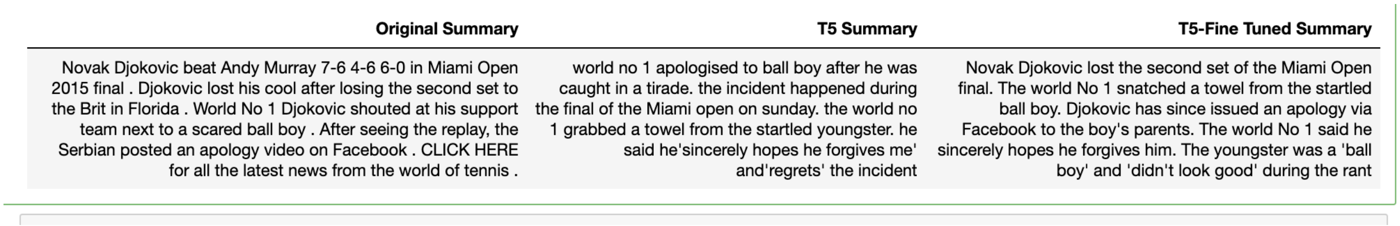
In the BERT-based extractive summarization process, the scoring of sentences is a crucial step that determines their relevance and importance in the generated summary. BERT employs a fine-tuned algorithm to evaluate the significance of each sentence within the context of the entire document. The scoring mechanism involves considering various linguistic and contextual features, such as the relationships between words, entities, and the overall semantic structure.

The BERT model assigns scores to individual sentences based on their contextual importance, emphasizing sentences that encapsulate key information and contribute significantly to the overall meaning of the document. This scoring process is conducted by analyzing the contextual embeddings of words in each sentence, allowing BERT to capture intricate relationships and nuances in the language.

In essence, the sentences are ranked according to their scores, and the top-ranked sentences are selected to form the extractive summary. By leveraging BERT's contextualized embeddings and sophisticated understanding of language, this approach ensures that the generated summary encapsulates the most relevant and informative content from the original document. The scoring mechanism is a testament to BERT's ability to grasp contextual nuances, making it a powerful tool for extractive summarization tasks.

6. Results and Discussion

The visual representation below contrasts ground-truth summaries with summaries generated by both the T-5 model and the fine-tuned version of T-5 applied to one of the articles.

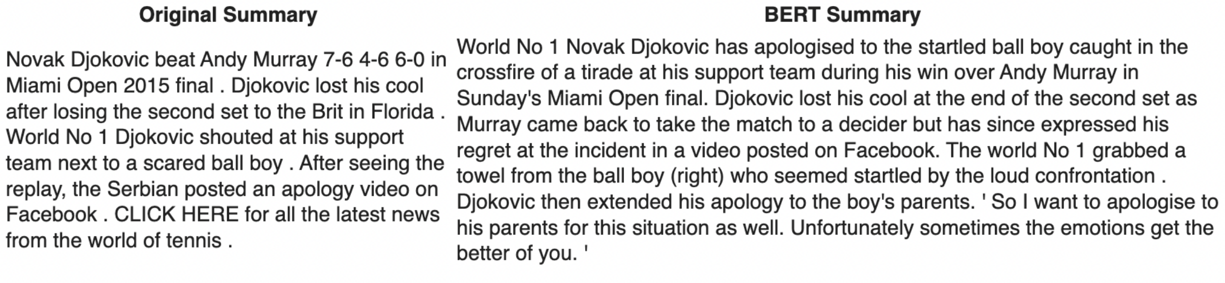


Original Summary	T5 Summary	T5-Fine Tuned Summary
Novak Djokovic beat Andy Murray 7-6 4-6 6-0 in Miami Open 2015 final . Djokovic lost his cool after losing the second set to the Brit in Florida . World No 1 Djokovic shouted at his support team next to a scared ball boy . After seeing the replay, the Serbian posted an apology video on Facebook . CLICK HERE for all the latest news from the world of tennis .	world no 1 apologised to ball boy after he was caught in a tirade. the incident happened during the final of the Miami open on sunday. the world no 1 grabbed a towel from the startled youngster. he said he'sincerely hopes he forgives me' and'regrets' the incident	Novak Djokovic lost the second set of the Miami Open final. The world No 1 snatched a towel from the startled ball boy. Djokovic has since issued an apology via Facebook to the boy's parents. The world No 1 said he sincerely hopes he forgives him. The youngster was a 'ball boy' and 'didn't look good' during the rant

Fig 5.1: Comparing the original summary with T-5 and T-5 fine-tuned

It is evident that while the T-5 model provides a reasonable summary, it falls short in terms of accuracy. The generated summary exhibits issues such as punctuation mistakes, problems with word capitalization, and a failure to capture all essential information.

In contrast, the fine-tuned T-5 model overcomes these limitations, producing summaries that are more refined and accurate. The comparison highlights the impact of fine-tuning on enhancing the summarization capabilities of the T-5 model, addressing issues observed in the original T-5-generated summaries.



Original Summary	BERT Summary
Novak Djokovic beat Andy Murray 7-6 4-6 6-0 in Miami Open 2015 final . Djokovic lost his cool after losing the second set to the Brit in Florida . World No 1 Djokovic shouted at his support team next to a scared ball boy . After seeing the replay, the Serbian posted an apology video on Facebook . CLICK HERE for all the latest news from the world of tennis .	World No 1 Novak Djokovic has apologised to the startled ball boy caught in the crossfire of a tirade at his support team during his win over Andy Murray in Sunday's Miami Open final. Djokovic lost his cool at the end of the second set as Murray came back to take the match to a decider but has since expressed his regret at the incident in a video posted on Facebook. The world No 1 grabbed a towel from the ball boy (right) who seemed startled by the loud confrontation . Djokovic then extended his apology to the boy's parents. ' So I want to apologise to his parents for this situation as well. Unfortunately sometimes the emotions get the better of you. '

Fig 5.2: Comparison of original summary with BERT-generated summary

The above figure is also an example of the same article using BERT-generated summaries, one notable difference not fully captured by BERT is the contextual background and emotional nuances surrounding Novak Djokovic's outburst during the Miami Open 2015. While both the original summary and the BERT-generated summary convey the core information about Djokovic's apology to a startled ball boy, the original summary provides more context on the incident. BERT, while summarizing the main events, may not fully capture the emotional intensity and specific details, potentially leading to a more concise but less comprehensive summary. This emphasizes the need for models to grasp contextual nuances and emotional cues for a more accurate representation of the content.

The subsequent evaluation using ROUGE scores provides a quantitative measure of the improvement achieved through fine-tuning, offering insights into the enhanced performance and effectiveness of the fine-tuned T-5 model in capturing the essence of the articles.

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a set of metrics used to evaluate the quality of summaries by comparing them to reference or ground-truth summaries. It includes precision, recall, and F-measure for unigram, bigram, and longest common subsequence (LCS) matches. Here's what each metric represents:

Precision: The ratio of the number of overlapping words between the generated summary and the reference summary to the total number of words in the generated summary. It measures how many of the generated words are relevant.

Recall: The ratio of the number of overlapping words between the generated summary and the reference summary to the total number of words in the reference summary. It measures how many of the reference words are covered by the generated summary.

F-Measure: The harmonic mean of precision and recall. It provides a balanced measure of the model's ability to generate relevant information while covering the important details present in the reference summary.

Model	**Precision**	**Recall**	**F-Measure**
Seq2seq	0.132	0.0894	0.102

BERT extractive summarization	0.2264	0.5707	0.3091
T-5	0.3585	0.4458	0.3854
T-5 with fine-tuning	0.378	0.4867	0.4128

Table 5.1: ROUGE-1 Scores

Model	**Precision**	**Recall**	**F-Measure**
Seq2seq	-	-	-
BERT extractive summarization	0.0863	0.2111	0.1164
T-5	0.1531	0.1879	0.1635
T-5 with fine-tuning	0.1699	0.2181	0.1850

Table 5.2: ROUGE-2 Scores

Model	**Precision**	**Recall**	**F-Measure**
Seq2seq	0.1305	0.0885	0.1009
BERT extractive summarization	0.1395	0.3510	0.1900
T-5	0.2383	0.2985	0.2569
T-5 with fine-tuning	0.2559	0.3328	0.2806

Table 5.3: ROUGE-L Scores

Interpretation of ROUGE Scores:

1. T-5 Model (Before Fine-Tuning):

   The F-measure for ROUGE-1 (unigram) is 0.3854, indicating that 38.54% of the relevant

words from the ground-truth summary are captured by the T-5 model.
The model shows limitations in capturing precise details, especially for bigrams and longer sequences.

2. T-5 Model (After Fine-Tuning):
   The F-measure improves to 41.28%, highlighting the positive impact of fine-tuning. The model becomes more effective in generating summaries that align with the ground-truth.

3. Seq2Seq Model:
   The F-measure for ROUGE-1 is lower at 10.20%, indicating that the seq2seq model has challenges in capturing relevant information and details from the reference summary.

4. BERT extractive summarization Model:
   The F-measure for ROUGE-1 is 30.91%, suggesting that the BERT model, while better than seq2seq, has room for improvement in capturing details from the reference summary.

These scores serve as quantitative indicators of how well each model performs in comparison to the ground-truth summaries. Higher scores, especially in F-measure, reflect better summarization capabilities. Fine-tuning, as seen in the T-5 model, can significantly enhance the model's performance. The ROUGE-1 score appears higher for the BERT model compared to ROUGE-2 and ROUGE-L. This is because the BERT model returns the relevant sentences from the source article to generate summaries, resulting in a high unigram overlap. However, as the true labels resemble the abstraction summarization process, capturing the nuances from the article and being concise, higher ROUGE-2 and ROUGE-L scores are observed for the T5 model

7. Challenges and Future Prospects

The articles in the dataset were lengthy, often exceeding a thousand words. Compared to many other open-source summarization projects, where the original texts are typically shorter and the resulting summaries are condensed to just one line. This discrepancy in text length added complexity to the summarization tasks. Given the constraints of the computational resources, significant challenges were faced: the necessity to train the model on only 10% of the entire dataset, which might have impacted the ability to produce optimal results. Also relying solely on ROUGE scores may overlook specific nuances and the introduction of diverse metrics such as BLEU, METEOR, and CIDEr could have provided a more comprehensive evaluation. Models like PEGASUS and Bi-LSTM were also explored as they are an abstractive type but due to the limited computational power, the results were not conclusive enough and further fine-tuning on these models was not possible. Exploring these models could have provided insights into diverse methodologies, potentially improving the model's abstraction capabilities. Combining the strengths of both abstractive and extractive methods requires careful integration to maintain coherence and relevance.Hybrid models could have been utilized for reinforcement learning or incorporate extractive summaries as input to the abstractive model, fostering a balanced approach.

For the future prospects, information from multiple documents could be combined and summaries can be generated. Techniques like document clustering, attention mechanisms, and hierarchical structures in the model architecture can be explored to enhance multi-document summarization. Extending the model to handle multiple languages by incorporating language-specific tokenization, pre-trained multilingual embeddings, and language-agnostic can be done which could address the linguistic variations and nuances.

8. Conclusion

The project has delved into diverse models and techniques with each offering unique approaches and meaningful insights from the dataset. From the foundational baseline model which is seq2seq with transformer architecture to advancement into T-5 and T-5 with fine-tuning for abstractive and BERT summarization for extractive summarization, having witnessed why extractive methods are not useful for the case in hand and how further parameter adjustment and working with more data could have helped the modeling further. The comparison of ROUGE scores showed that different models perform differently. This highlights the importance of using more detailed evaluation measures to understand their strengths and weaknesses. The project not only showcased the capabilities of these models but also highlighted the challenges in generating coherent and informative summaries.

The project serves as a foundation step towards what can further be achieved using text summarization which includes but is not limited to multi-document summarization, text summarization which also includes image, multi-language text summarization, hybrid modeling, etc. As the field progresses through these prospects, the ultimate goal remains to develop summarization systems that not only condense information but also capture the essence and context of the underlying content.

9. References

1. N. Patel and N. Mangaokar, "Abstractive vs Extractive Text Summarization (Output based approach) - A Comparative Study," 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 2020, pp. 1-6, doi: 10.1109/INOCON50539.2020.9298416.

2. Gambhir, M., Gupta, V. Recent automatic text summarization techniques: a survey. Artif Intell Rev 47, 1–66 (2017). https://doi.org/10.1007/s10462-016-9475-9

10. Individual Contributions

Anirudh Gaur - Extractive BERT Summarization, Future Prospects
Rishik Shekar Salver - T-5 and T-5 with Fine Tuning, Seq2Seq Baseline Model, Model Evaluation, Future Prospects(Approaches of the Bi-LSTM and Pegasus Model)
Xinran Zhao - Data Preparation, Seq2Seq Baseline Model