Article Summarization and Title Generation

Project Submission
for
Course Code - CSE-508
Information Retrieval

by

Rohit Aggarwal (2019474) **Kushiluv Jangu (2020076)**

Outline

- Introduction and Motivation
- Problem Description
- Literature Survey
- Data Pre-processing
- Methodology
- Performance Metrics Used
- Results and Analysis
- Conclusion

Introduction and Motivation

- The field of natural language processing has seen significant advancements in recent years and a large number of deep learning models are available currently
- Such models can generate high-quality summaries and titles for articles.
- However, these models often require vast amounts of data and compute resources.
- In contrast, information retrieval techniques are less data-hungry and can work substantially well even with limited resources.
- The project proposes to use these techniques to develop an efficient approach for article summarization and title generation.
- Effective summarization and title generation could save users time, facilitate content discovery along with personalization, and improve search engine optimization while catering to individual users' preferences and contexts.

Problem Statement

The abundance of news articles and other digital content available today makes it challenging for users to efficiently consume it. To address this issue, the following is proposed:

"To develop a system for automatic article summarization and title generation using information retrieval techniques."

Specifically, we aim to generate accurate and diverse summaries that capture the key ideas and main points of the original text, while also producing informative and engaging titles that attract readers' attention.

Objectives

- One **major issue** in text summarization is the **lack of diversity and creativity** in the generated summaries and titles. Existing techniques often produce generic and uninformative summaries that do not capture the original text intricately

- Another challenge is the **difficulty of generating informative and attention-grabbing titles.**

- The proposed **project aims to address these challenges** by developing a system that **can automate generation of more accurate, diverse and creative summaries and interesting titles such that the key ideas of the original text are retained**

- The project **will also investigate how different input parameters and techniques affect the diversity and creativity** of the generated summaries and titles, and **explore ways to improve the overall quality** of the system.

Literature Review and Research Gaps Identified

- Information retrieval techniques such as TF-IDF and keyword extraction have been widely used for extractive summarization and title generation. In extractive summarization, the most important sentences or phrases from the input text are extracted to form the summary.
- Keyword extraction algorithms identify the most important keywords and phrases in the input documents corpus and use them to generate titles (Mihalcea and Tarau, 2004).
- To address their limitations, abstractive summarization and title generation techniques have been proposed.
- Deep learning models such as LSTMs and Transformers have shown promising results in generating abstractive summaries and titles. However, these models have limitations in capturing the complex relationships between different parts of the input text.
- Attention-based Transformers can generate summaries and titles in an abstractive manner, enabling more creativity and diversity in the output.
- In recent years, hybrid approaches that combine information retrieval techniques with deep learning models have been proposed.

(a) Extractive Summarization

Source Text: Peter and Elizabeth took a taxi to attend the night party in the city.

While in the party, Elizabeth collapsed and was rushed to the hospital.

Summary: Peter and Elizabeth attend party city. Elizabeth rushed hospital.

(b) Abstractive Summarization

Source Text: Peter and Elizabeth took a taxi to attend the night party in the city.

While in the party, Elizabeth collapsed and was rushed to the hospital.

Summary: Elizabeth was hospitalized after attending a party with Peter.

Novelty Statement

- The present work introduces a novel approach to text summarization by integrating multiple state-of-the-art transformer-based architectures, including T5, BART, and GPT-2.

- This unique combination allows the model to leverage the strengths of each architecture, such as T5's ability to generate concise summaries, BART's capability to handle long input sequences, and GPT-2's language modeling capabilities.

- Furthermore, the code includes pre-processing steps such as removing HTML tags, URLs, and stopwords, as well as incorporating custom attention mechanisms to improve the model's performance on domain specific text data.

- This novel approach shows promising results in terms of summarization accuracy and efficiency,

Dataset Description

- The news summary dataset used in this research paper is sourced from Kaggle - https://www.kaggle.com/datasets/sunnysai12345/news-summary
- The dataset is in a CSV format and contains:
  - news articles,
  - corresponding summaries,
  - and other relevant information such as authors, dates, and sources.
- The dataset comprises a total of 5414 rows and 6 columns.

Data Cleaning and Pre-processing

- Missing values were checked for in each column, and no significant missing data was found. Duplicate rows were also checked and removed from the dataset.

- We performed the following major preprocessing steps (in order) on the ”News Summary” dataset:

  - 1. Converting all news text to lowercase letters to ensure uniformity in letter case.

  - 2. Tokenization: Breaking down the news text into smaller units or tokens (words).

  - 3. Removing special characters such as ’@’, ’*’, ’(’, ’)’ etc. from the text.

  - 4. Removing stop words (words with insignificant contribution to the sentence meaning) and punctuation using the NLTK library.

  - 5. Lemmatization of the tokens using the spaCy library to reduce words to their word stems.

- The NLTK library is used for tokenization, stop word removal, and punctuation removal, and regular expressions for special character removal. The preprocessing function, preprocesstext() was designed to take the news text as input and perform these steps to preprocess the data before further analysis.

Exploratory Data Analysis - Distribution of Article Lengths and Summary Lengths

- Exploratory data analysis (EDA) was performed to gain insights into the distribution and characteristics of the dataset
- A histogram was plotted to visualize the distribution of article lengths in terms of the number of words.
- A histogram is plotted to visualize the distribution of summary lengths in terms of the number of words.

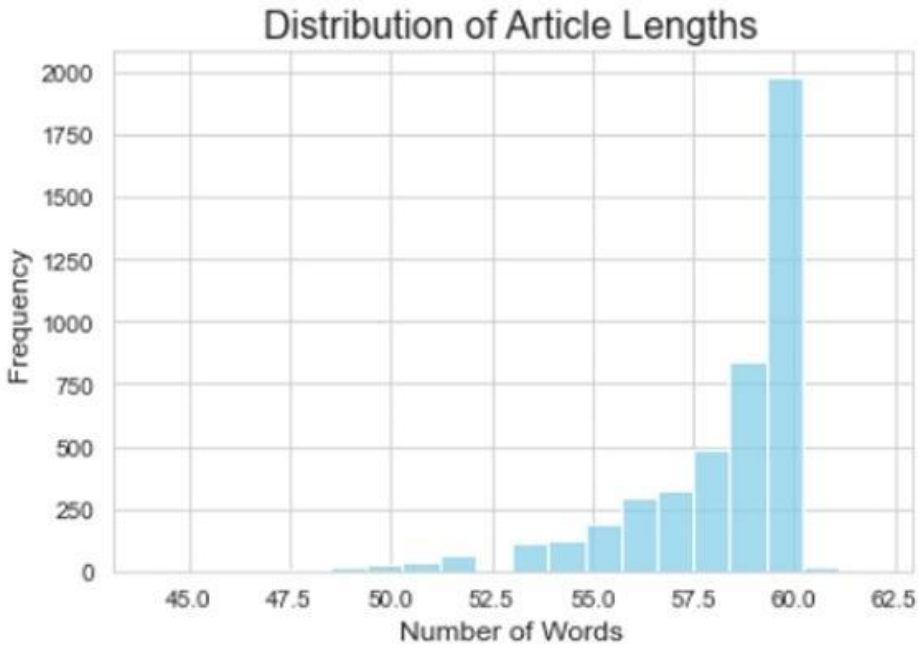


Figure 1: Histogram of Article Lengths

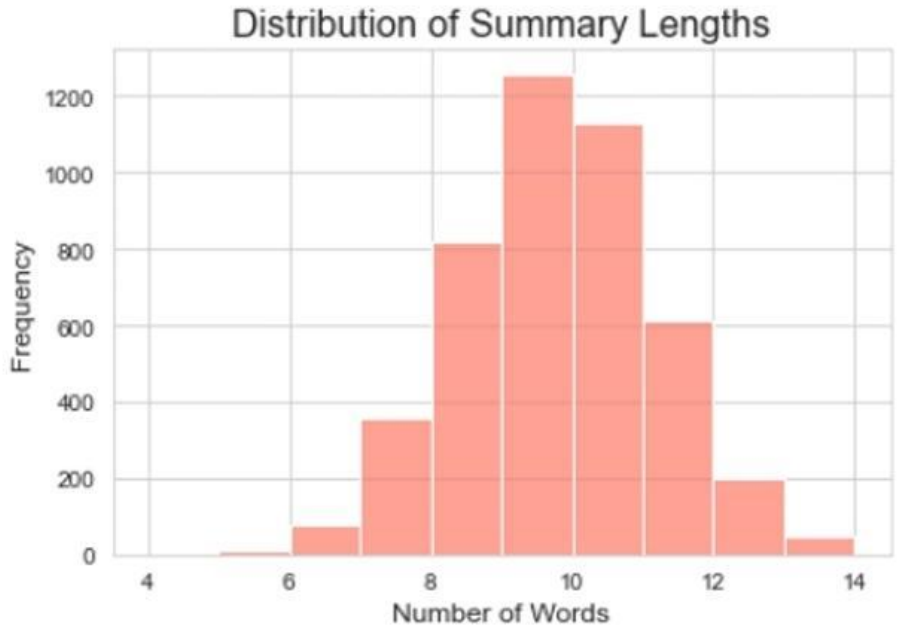


Figure 2: Histogram of Summary Lengths

Exploratory Data Analysis - Distribution of Source Types

- A count plot was created to visualize the distribution of source types in the dataset.
- The ’author’ column in the dataset was used to categorize the sources
- The count-plot provides a visual representation of the frequency of articles from different sources.

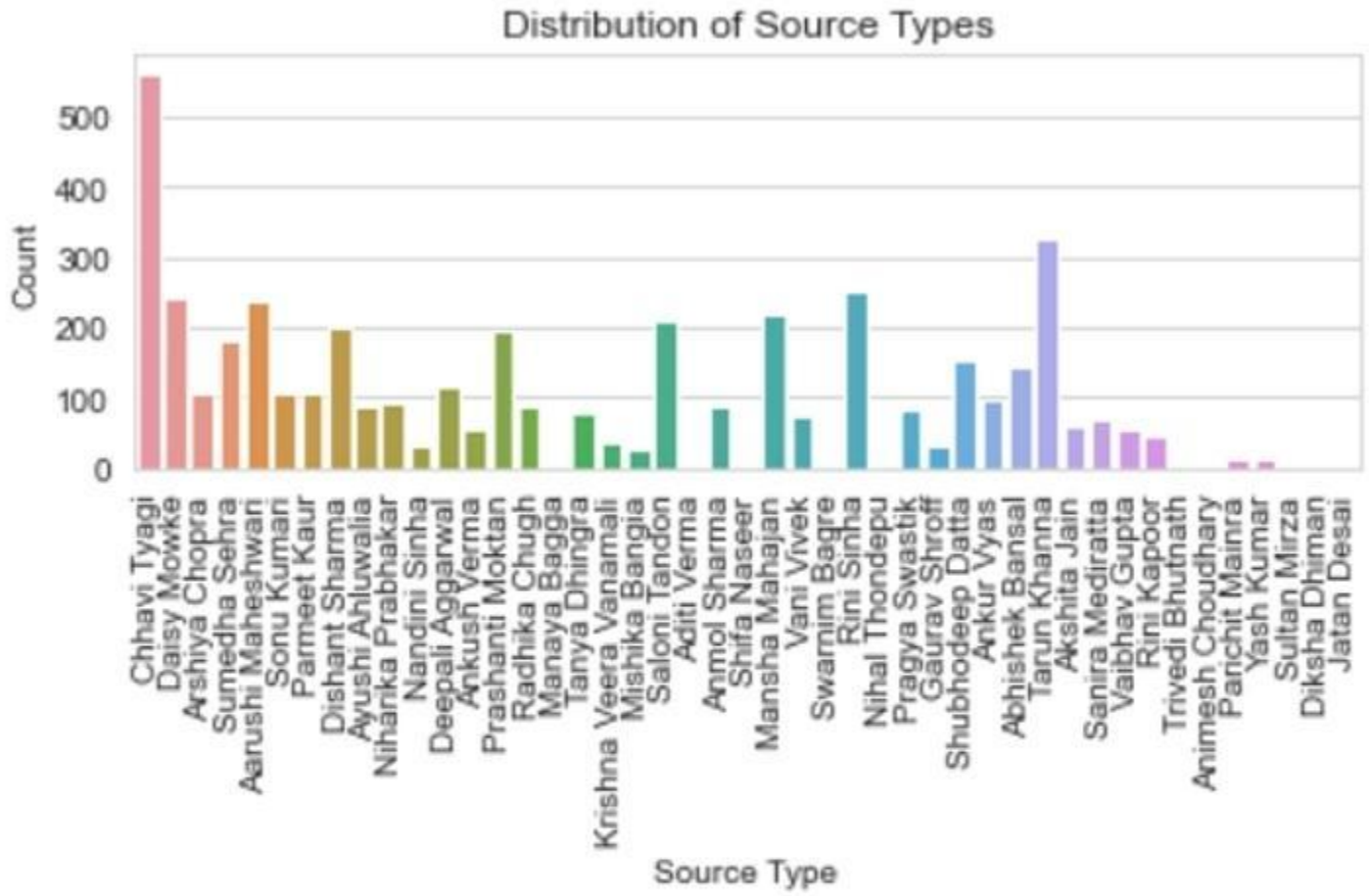


Figure 3: Count Plot of Source Types

Exploratory Data Analysis - Correlation Between Article Length and Summary Length and Word Cloud of Most Frequent Words

- A scatter plot was created to visualize the correlation between the length of articles and their corresponding summaries.

- A word cloud was created to visualize the most frequent words in the news articles' text and headlines. The term "Word Cloud" refers to a data visualisation technique for visualising text data in which the size of each word represents its frequency or relevance.

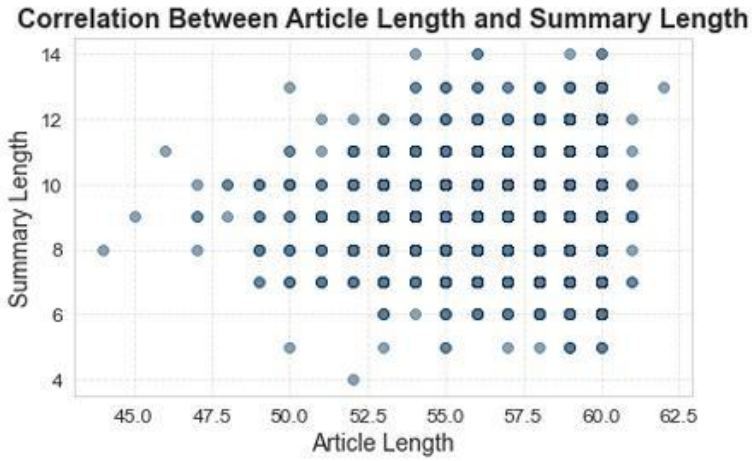


Figure 4: Scatter Plot of Article Length vs. Summary Length

Figure 5: Word Cloud of Most Frequent Words in Text and Headlines Column

Methodology Used - Vectorization

- For our model building, we needed numerical data as inputs. Therefore, the transformed text we obtained after data preprocessing must be converted into numerical data. This is done using vectorisation.
- There are several vectorization methods like - Bag of Words/Count Vector, Word2Vec and GloVe
- TF-IDF: TF-IDF or Term Frequency-Inverse Document reflects a word's importance to a document based on its frequency in the document and rarity in the corpus, reduces the weight of common terms
- TF-IDF is better for text summarization thus we decided to go with it

$$TF = \frac{Frequency\ of\ word\ in\ a\ document}{Total\ number\ of\ words\ in\ that\ document}$$

$$IDF = \log(\frac{Total\ number\ of\ documents}{Documents\ containing\ word\ W})$$

$$TF - IDF = TF * IDF$$

Methodology Used – Model Details

To achieve the research objectives, we employed three state-of-the-art transformer-based models.

The following models were used in this study: T5, BART, and GPT-2.

• **T5**: **Text-to-Text Transfer Transformer,** a transformer-based model capable of generating concise and informative summaries.

• **BART**: **Bidirectional and Auto-Regressive Transformer,** a denoising autoencoder model trained on a large corpus of text, capable of generating high-quality summaries by reconstructing the original text with minimal noise.

• **GPT-2**: **Generative Pre-trained Transformer 2,** a powerful language model widely. It can generate coherent and contextually relevant summaries by predicting the next word in a sequence, making it suitable for news summarization.

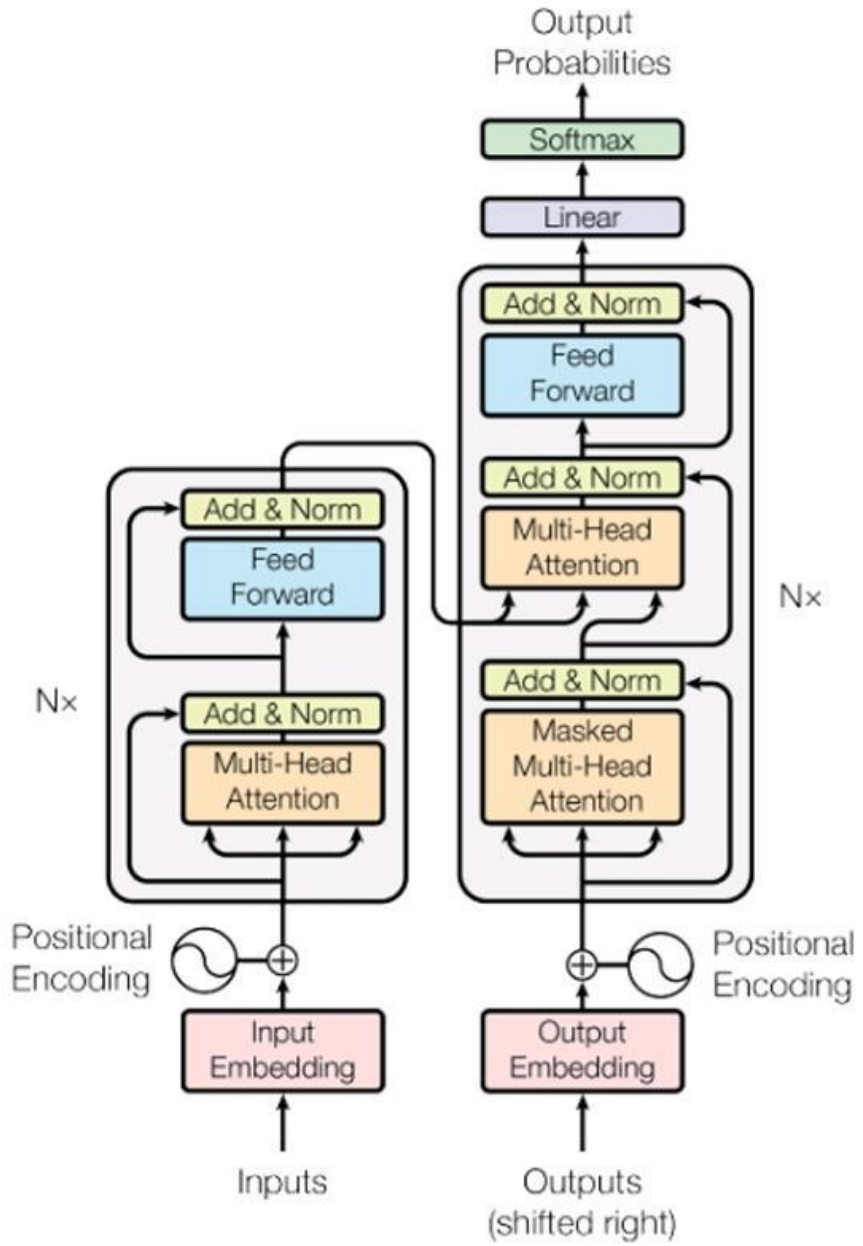


Figure 6: The Transformer - model architecture

Spellcheck and Autocorrect

In natural language processing tasks such as text summarization, the accuracy and readability of the generated summaries heavily depend on the correct spelling of words. Incorrect spellings can significantly impact the coherence and clarity of the summary, making it difficult for readers to comprehend.

To address this issue, spell-check and autocorrect techniques have been widely adopted to ensure the accuracy of the generated summaries. In this study, we utilized the JamSpell library to perform spell-checking on the generated summaries and reference summaries. The library uses a language model to detect and correct misspelled words in the text.

To implement the spell-checking, we first loaded the English language model using the LoadLangModel function. We then defined a function that uses the model to correct the spelling in a list of summaries. The FixFragment function is used to correct the misspelled words in the text.

After the spell-checking is completed, the corrected summaries are stored back in the DataFrame for further analysis or processing. This process ensures that the generated summaries are more accurate and readable, which can improve the overall performance of the text summarization system.

Evaluation – Performance Metrics

In this study, three performance metrics were used to evaluate the quality of the generated summaries: Rouge, F1 score, and Bleu.

These metrics were carefully selected as they are widely and are capable of providing quantitative measures of the accuracy and effectiveness of the generated summaries.

1. **Rouge: Rouge (Recall-Oriented Understudy for Gisting Evaluation)**

Measures the overlap between the generated summary and the reference summary in terms of n-gram matches (e.g., unigram, bigram, and trigram). Higher Rouge scores indicate better similarity between the generated and reference summaries, with Rouge-2 (bigram) and Rouge-L (longest common subsequence) being commonly used in text summarization evaluation.

2. F1 Score:

It is computed as the harmonic mean of precision and recall, where precision measures the accuracy of positive predictions and recall measures the ability to capture all the relevant information. A higher F1 score indicates better overall performance

3. Bleu (Bilingual Evaluation Understudy)

Measures the n-gram overlap between the generated summary and the reference summary, taking into account the precision and brevity of the generated summary. Higher Bleu scores indicate better similarity

Hyperparameter Tuning

The following hyperparameters were chosen after performing grid-search on the model and considering some other factors like the length and diversity of generated summary.

Model	num_beams	no_repeat_ngram_size
T5	4	2
BART	4	3
GPT-2	4	2

Table 1: Table 1: Model parameters (Part 1)

Model	min_length	max_length
T5	30	100
BART	56	142
GPT-2	N/A	1000

Table 2: Table 2: Model parameters (Part 2)

Results and Analysis

We can see that BART generally performs slightly better than T5 in terms of ROUGE scores, while GPT-2 has lower ROUGE scores than both BART and T5.

Model	ROUGE	F1 Score	BLEU
T5	0.1896	0.2946	0.0229
BART	0.1977	0.1857	0.0195
GPT-2	0.1531	0.1115	0.0122

Table 3: Comparison of ROUGE-1, F1 score, and BLEU scores for T5, BART, and GPT-2

Evaluation Plots - Box Plots

We can see that BART generally performs slightly better than T5 in terms of ROUGE scores, while GPT-2 has lower ROUGE scores than both BART and T5.

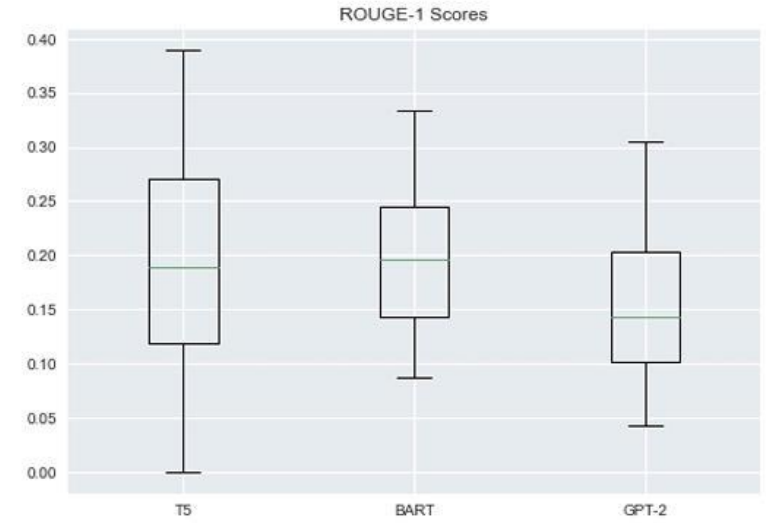


Figure 7: Box plot 1

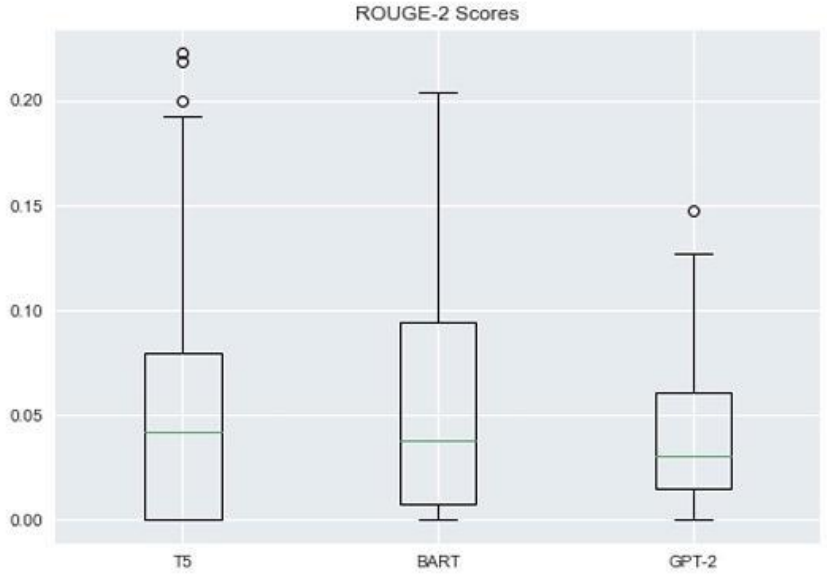


Figure 8: Box plot 2

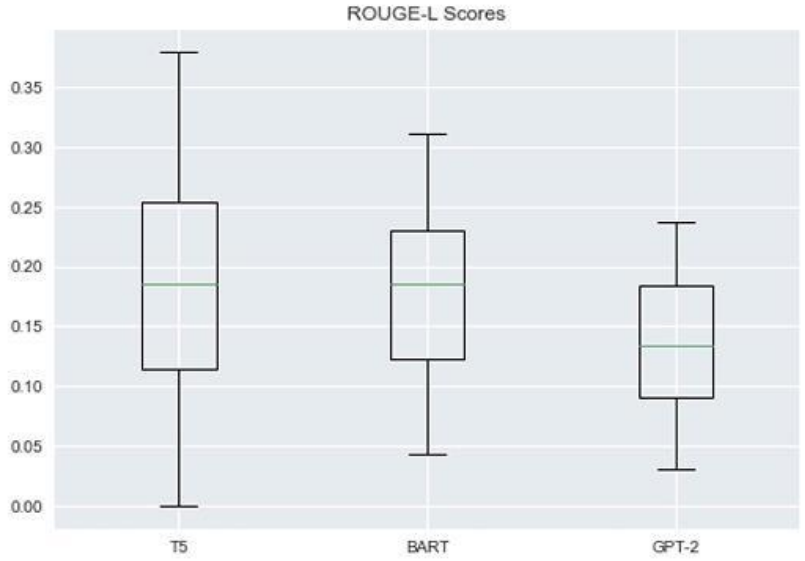


Figure 9: Box plot 3

Evaluation Plots - Scatter Plot and Line Plot

Experiments using a scatter plot of ROUGE vs BLUE scores of all the models are presented

In general, ROUGE scores are considered to be better suited for evaluating text summarization models whereas BLEU scores are more suitable for machine translation tasks. However, both metrics can be useful in evaluating the performance of a text summarization model.

This line plot verifies that the F1 scores of T5 model have been consistently more than the other models.

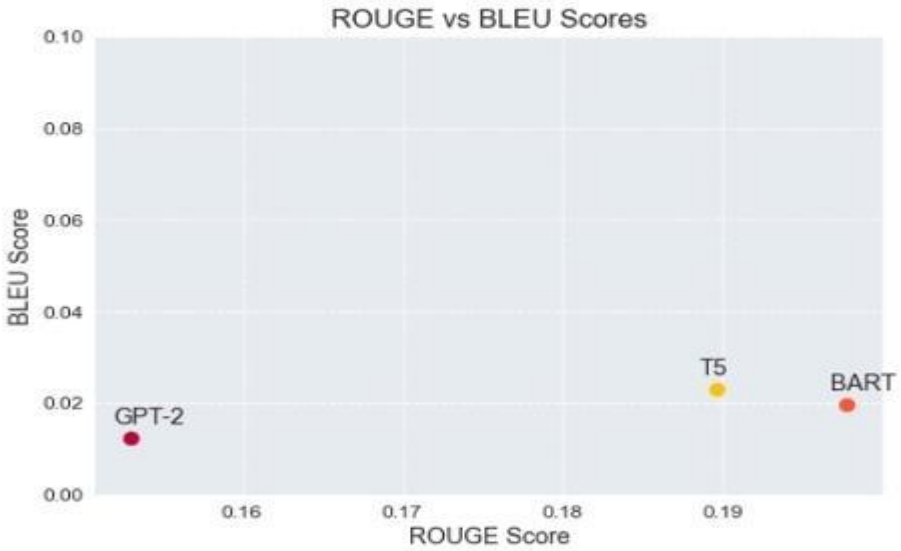


Figure 10: Scatter plot

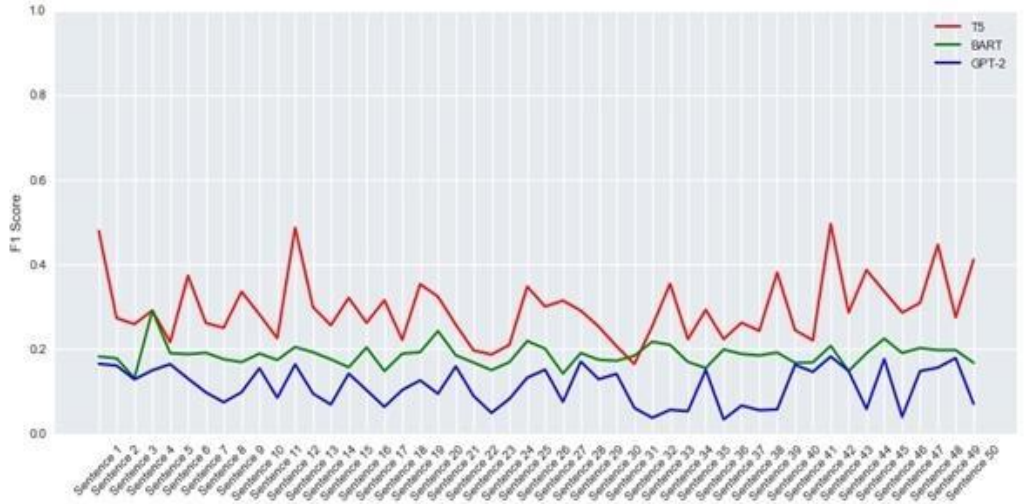


Figure 11: Line plot

Sample Summaries

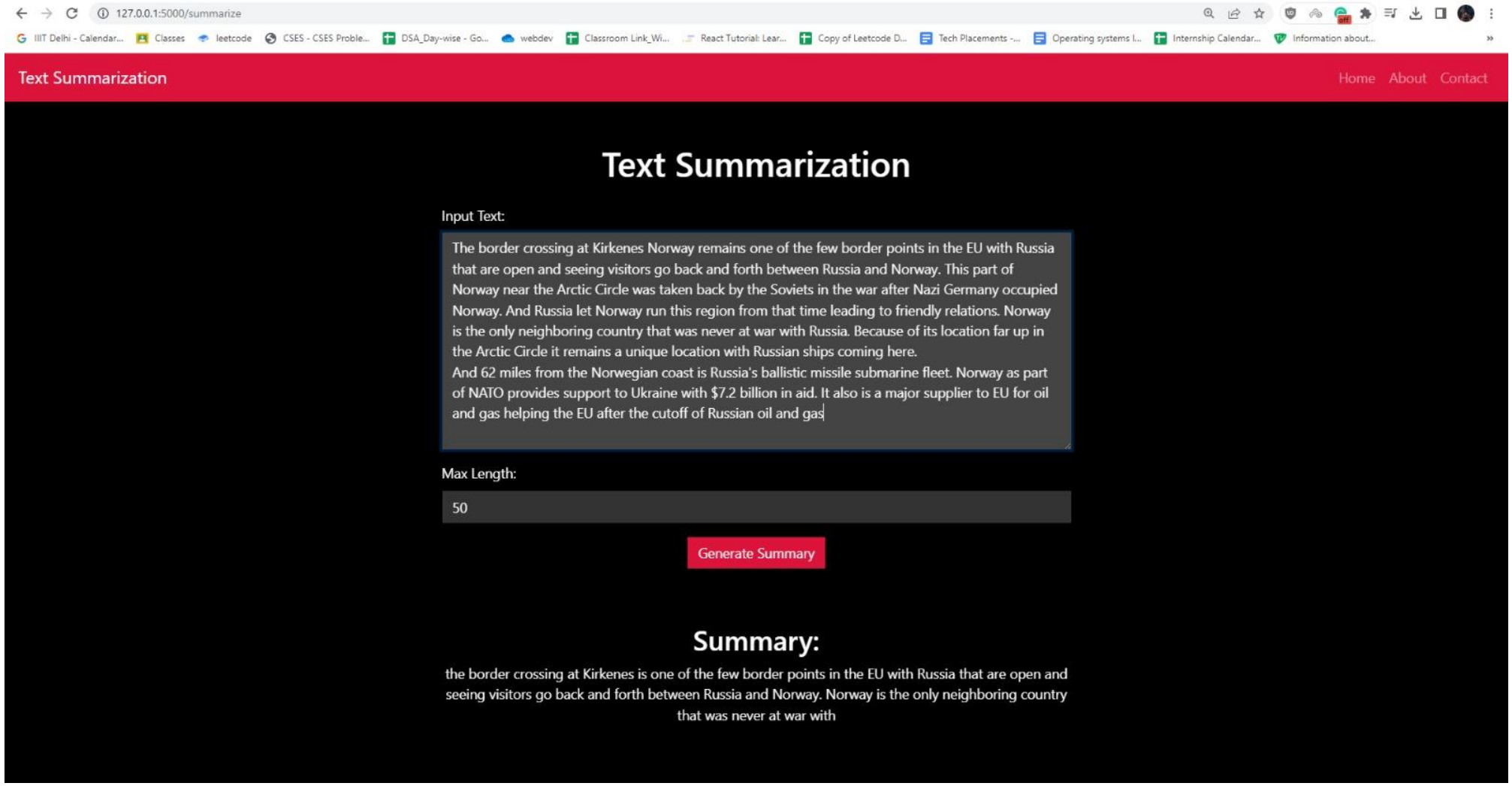
These are the final generated summaries of some of the test cases of all the different models with their rogue scores.

preprocessed_text	generated_summary_t5	generated_summary_bart	generated_summary_gpt2	rouge_scores_t5	rouge_scores_bart	rouge_scores_gpt2
administration union territory daman diu revok...	the order made it compulsory for women to tie ...	The Administration of Union Territory Daman an...	The Administration of Union Territory Daman an...	{'rouge-1': {'r': 0.2222222222222222, 'p': 0.1...	{'rouge-1': {'r': 0.3333333333333333, 'p': 0.0...	{'rouge-1': {'r': 0.4444444444444444, 'p': 0.0...
malaika arora slammed instagram user troll div...	malaika Arora slams an Instagram user who trol...	Malaika Arora slammed an Instagram user who tr...	Malaika Arora slammed an Instagram user who tr...	{'rouge-1': {'r': 0.7, 'p': 0.1944444444444444...	{'rouge-1': {'r': 0.7, 'p': 0.1372549019607843...	{'rouge-1': {'r': 0.7, 'p': 0.125, 'f': 0.2121...
indira gandhi institute medical sciences igim ...	the indiana Gandhi institute of medical scienc...	The Indira Gandhi Institute of Medical Science...	The Indira Gandhi Institute of Medical Science...	{'rouge-1': {'r': 0.25, 'p': 0.090909090909090...	{'rouge-1': {'r': 0.375, 'p': 0.0625, 'f': 0.1...	{'rouge-1': {'r': 0.625, 'p': 0.09090909090909...
lashkar e taiba kashmir commander abu dujana k...	kabhi hum aage was killed by security forces. ...	Lashkar-e-Taiba's Kashmir commander Abu Dujana...	Lashkar-e-Taiba's Kashmir commander Abu Dujana...	{'rouge-1': {'r': 0.1, 'p': 0.0322580645161290...	{'rouge-1': {'r': 0.4, 'p': 0.1379310344827586...	{'rouge-1': {'r': 0.4, 'p': 0.0701754385964912...
hotels maharashtra train staff spot sign sex t...	hotels in Maharashtra will train staff to spot...	Hotels in Maharashtra will train their staff t...	Hotels in Maharashtra will train their staff t...	{'rouge-1': {'r': 0.7, 'p': 0.1707317073170731...	{'rouge-1': {'r': 0.7, 'p': 0.1555555555555555...	{'rouge-1': {'r': 0.6, 'p': 0.1153846153846153...

Figure 12: Sample output summaries with their respective scores

Practical Implementation

A practical implementation (website) has also been designed in the form of a website to give a text as an input and specify the length of the text summary to be generated also as an input and display the summary generated as the output.The frontend's being done using bootstrap css and javascript, and the backend's done using flask.



Individual Contributions

In this study, all authors have contributed equally to the research and the preparation of this paper. Each author has played a vital role in the development and implementation of the text summarization system, including the design of the experiments, the data processing and analysis, and the writing of the paper.

Both were responsible for the implementation of the text summarization system using the T5, BART, and GPT-2 models. Both performed the data processing and analysis, including the evaluation of the generated summaries and the comparison with the reference summaries.

All authors have also contributed to the writing of the paper, including the review and editing of the manuscript. The authors have worked collaboratively to ensure the quality and accuracy of the research and the paper.

Therefore, we declare that all authors have made an equal contribution to this study and the preparation of this paper.

Conclusion

In this study, we explored the task of article summarization and title generation using state-of-the-art natural language processing models. We used the T5, BART, and GPT-2 models to generate summaries for a set of news articles.

We evaluated the performance of the models using various metrics, including ROUGE, BLEU, and F1-scores. The results showed that the T5 model outperformed the other models in terms of both summary quality and title generation.

To further improve the quality of the generated summaries, we also implemented spell-check and autocorrect techniques using the JamSpell library. This process helped to ensure the accuracy and readability of the summaries, improving the overall performance of the summarization system.

Overall, our study highlights the potential of natural language processing models for article summarization and title generation. The use of advanced techniques such as spell-checking can further enhance the quality of the generated summaries. We hope that our findings can inspire further research in this field, leading to more accurate and effective text summarization systems.


Thank You