TEXT SUMMARIZATION

Authors
Aalokik Singh (aalokik21223@iiitd.ac.in)
Harsh Yadav (harsh21049@iiitd.ac.in)
Pratham Singhal (pratham21347@iiitd.ac.in)
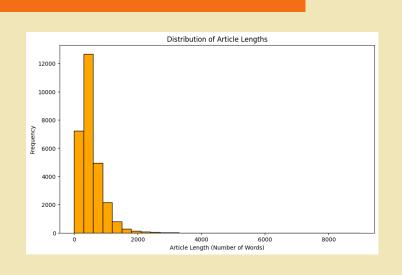Vivek Jain (vivek21218@iiitd.ac.in)

Affiliations
INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY **DELHI**

DATASET VISUALISATION

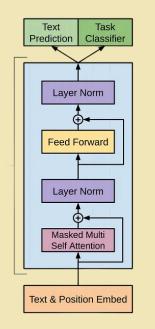Distribution of Article Lengths

PROBLEM STATEMENT

Our digital era's immense volume of text data necessitates **efficient text summarization techniques**. This project focuses on developing **NLP**-based methods for summarizing English text, addressing challenges like code-mixing and **script diversity**, to distill crucial information from **extensive content into concise summaries**.

OBJECTIVES

- Develop advanced text summarization techniques tailored for English content.
- Tackle challenges in summarizing code-mixed English text for improved accuracy.
- Implement culturally adaptive strategies in English text summarization for diverse user needs.
- Enhance user relevance and domain adaptability in our text summarization model for real-world impact.

METHODOLOGY

The methodology in the code covers data preprocessing, train-validation splitting, dataset creation, tokenization using **BART tokenizer**, model initialization, training with **AdamW** optimizer, evaluation with **ROUGE** and **BERTScore** metrics, model saving, and results reporting including average **ROUGE-L F1 score** and **BERTScore F1 score**, as well as saving generated summaries for further analysis in a CSV file.

RESULTS

QUANTITATIVE RESULT

- The quantitative evaluation of the model's performance was conducted using two key metrics: ROUGE-L F1 score and BERTScore F1 score. The test and validation scores are given below.

```
Test ROUGE-L (n=2) F1: 0.1872935903289672
Test BERTScore F1: 0.5102450847625732

Validation ROUGE-L (n=2) F1: 0.30781602391507734
Validation BERTScore F1: 0.5951529145240784
```

- The training loss steadily decreased over three epochs, indicating successful model convergence. The model was run for 3 epoch values as seen below.

```
epoch 0 running.
Epoch 1/3, Train Loss: 4.5805
epoch 1 running.
Epoch 2/3, Train Loss: 0.8238
epoch 2 running.
Epoch 3/3, Train Loss: 0.3956
```
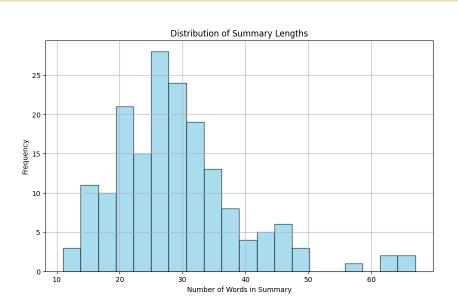
QUALITATIVE RESULT

- **Generated Summaries**: The model produced concise and coherent summaries that captured key information from source articles.
- **Word Clouds**: Visualization of recurring themes and topics the model highlights across summaries.
- **Length Comparison**: Summaries were of appropriate length, balancing conciseness with informativeness.
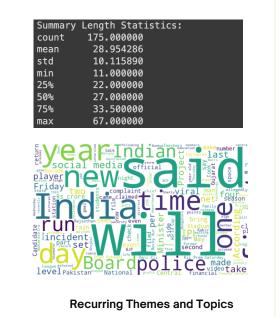
ANALYSIS

The training process showed significant improvement, with the model's loss decreasing from 4.58 to 0.39 over three epochs, indicating effective learning and convergence. Evaluation on validation and test datasets revealed higher scores on the validation set (ROUGE-L 0.30, BERTScore 0.59) compared to the test set (ROUGE-L 0.18, BERTScore 0.51), suggesting potential for further optimization. Qualitative analysis highlighted the model's capacity to distill relevant information, evident in word clouds showcasing recurring themes. A length comparison graph demonstrated the model's ability to generate concise yet informative summaries. These results emphasize the effectiveness of the text summarization techniques in capturing essential content from English text, with future improvements aimed at enhancing generalizability and adaptability to diverse user needs.

Training Loss vs Epoch

Validation and Test Metric Comparison

Recurring Themes and Topics

```
Summary Length Statistics:
count    175.000000
mean      28.954286
std       10.115890
min       11.000000
25%       22.000000
50%       27.000000
75%       33.500000
max       67.000000
```


Distribution of Summary Lengths

The summaries effectively balanced length and conciseness, with a maximum length of 67 words and an average length of 28.95 words, demonstrating their ability to distill key information efficiently while maintaining informativeness.

INTERESTING OBSERVATION

The effective balance achieved by the model in summary length, with a maximum of 67 words and an average of 28.95 words, showing model's capability to distill essential information into concise summaries while maintaining informativeness, even for long articles.

FUTURE WORK

Future work will focus on enhancing the model's adaptability to diverse text genres and languages to broaden its applicability in multilingual and cross-domain settings.

CONCLUSION

In conclusion, this study successfully developed and evaluated NLP-based text summarization techniques for English content, demonstrating significant improvements in model performance and promising evaluation metrics. The qualitative analysis confirmed the model's ability to produce concise and informative summaries, highlighting its effectiveness in distilling key information. Future research directions will focus on enhancing the model's generalizability and adaptability for real-world applications.