Text Summarization of News Articles

Team 3

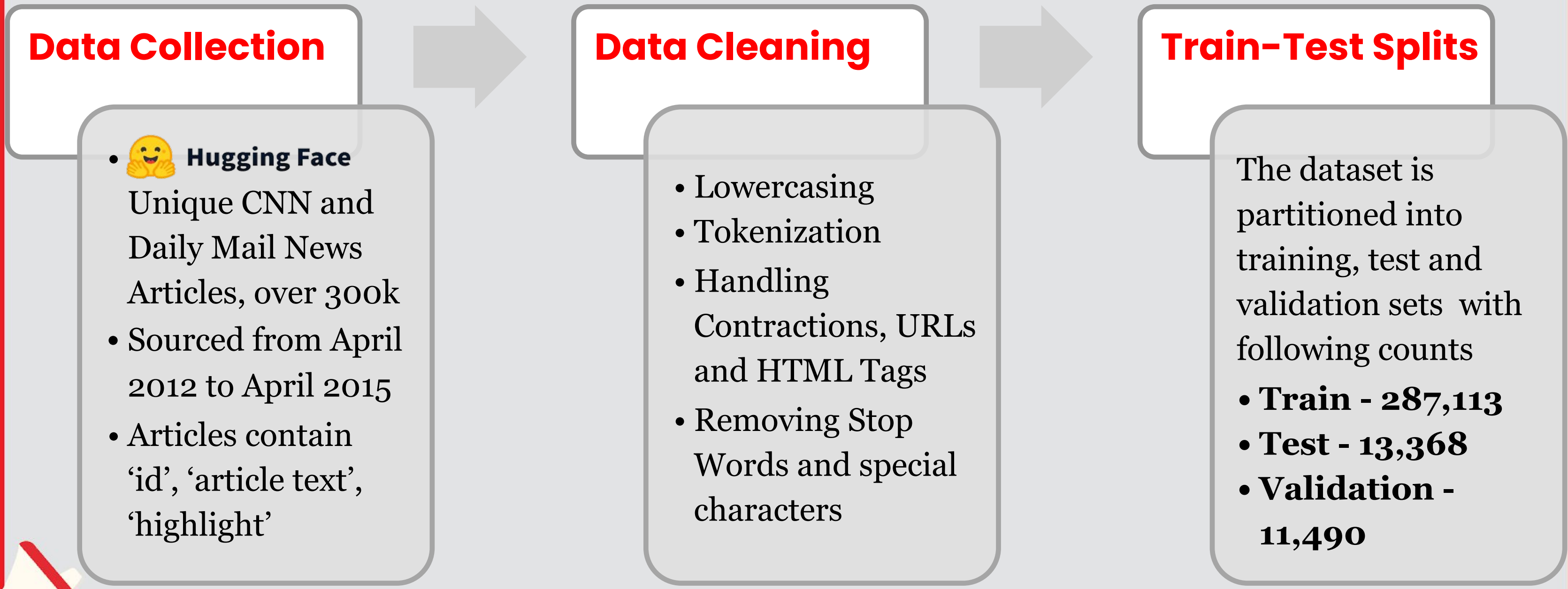
Manisha Damera	md1723
Neha Thonta	nt446
Varun Gandikota	vg411

Cohort Presentation Done

Problem Statement

- Common Experience: Feeling overwhelmed by lengthy reports, seeking only a summary for essential information.
- Condensing large volumes of news articles from CNN/Daily Mail into concise yet informative summaries.
- We aim to improve text summarization by using advanced neural network architectures for better, more informative summaries from lengthy news articles.

Data Preparation



Data Collection

- 🤗 Hugging Face Unique CNN and Daily Mail News Articles, over 300k
- Sourced from April 2012 to April 2015
- Articles contain ‘id’, ‘article text’, ‘highlight’

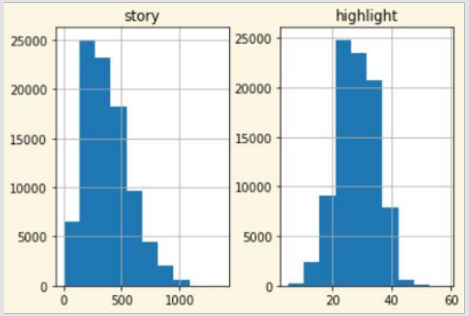
Data Cleaning

- Lowercasing
- Tokenization
- Handling Contractions, URLs and HTML Tags
- Removing Stop Words and special characters

Train-Test Splits

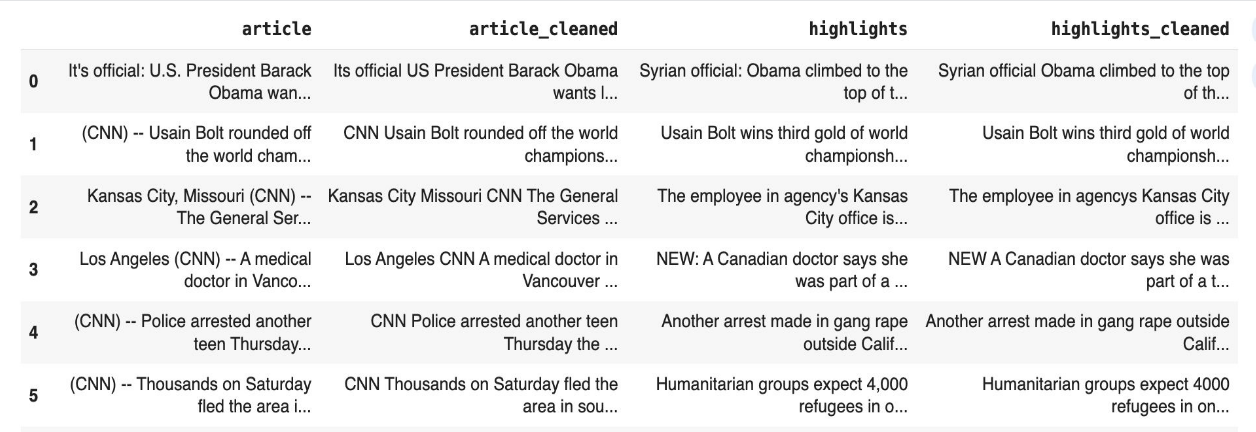
The dataset is partitioned into training, test and validation sets with following counts

- **Train - 287,113**
- **Test - 13,368**
- **Validation - 11,490**



← Word Count Distribution

Pre-processed Text



	article	article_cleaned	highlights	highlights_cleaned
0	It's official: U.S. President Barack Obama wan...	Its official US President Barack Obama wants l...	Syrian official: Obama climbed to the top of t...	Syrian official Obama climbed to the top of th...
1	(CNN) -- Usain Bolt rounded off the world cham...	CNN Usain Bolt rounded off the world champions...	Usain Bolt wins third gold of world championsh...	Usain Bolt wins third gold of world championsh...
2	Kansas City, Missouri (CNN) -- The General Ser...	Kansas City Missouri CNN The General Services ...	The employee in agency's Kansas City office is...	The employee in agencys Kansas City office is ...
3	Los Angeles (CNN) -- A medical doctor in Vanco...	Los Angeles CNN A medical doctor in Vancouver ...	NEW: A Canadian doctor says she was part of a ...	NEW A Canadian doctor says she was part of a t...
4	(CNN) -- Police arrested another teen Thursday...	CNN Police arrested another teen Thursday the ...	Another arrest made in gang rape outside Calif...	Another arrest made in gang rape outside Calif...
5	(CNN) -- Thousands on Saturday fled the area i...	CNN Thousands on Saturday fled the area in sou...	Humanitarian groups expect 4,000 refugees in o...	Humanitarian groups expect 4000 refugees in on...

Experiments

01 Seq2Seq

02 Seq2Seq+Attention

03 Bert2Bert
https://huggingface.co/bert-base-uncased

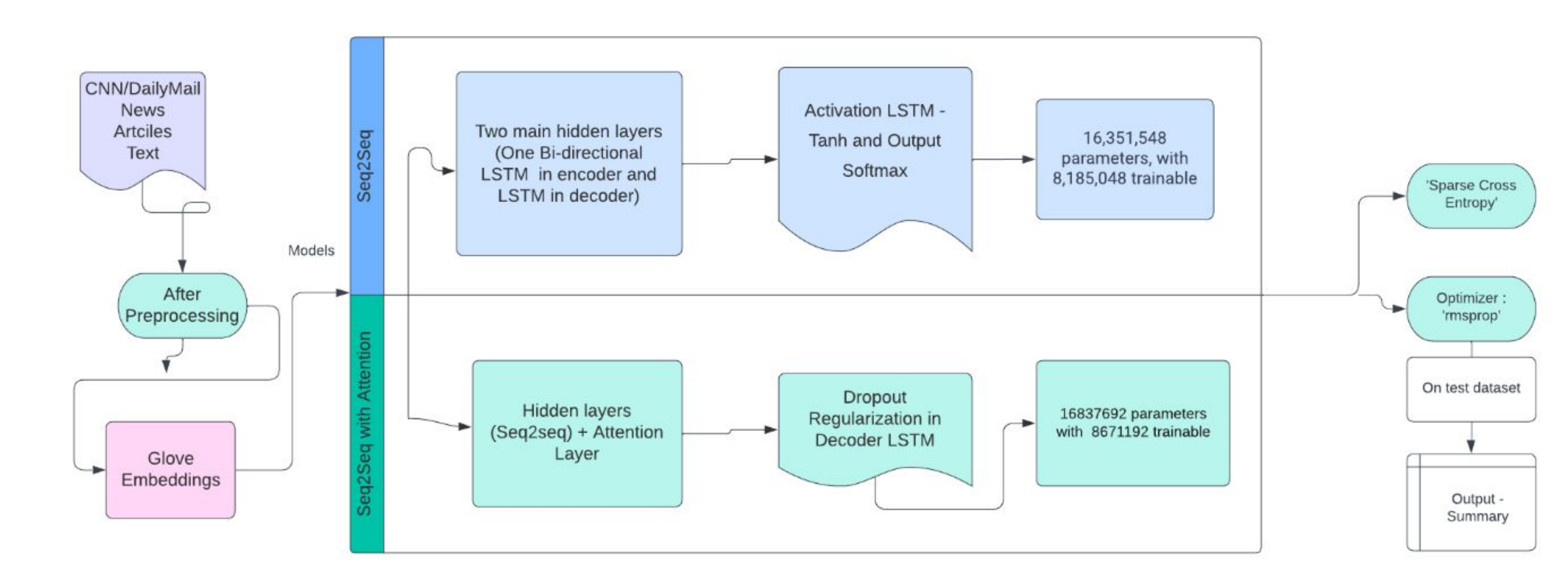
04 T5
https://huggingface.co/docs/transformers/tasks/summarization

05 Pegasus
https://huggingface.co/google/pegasus-large

06 BART
https://huggingface.co/facebook/bart-large-cnn

Baseline Models

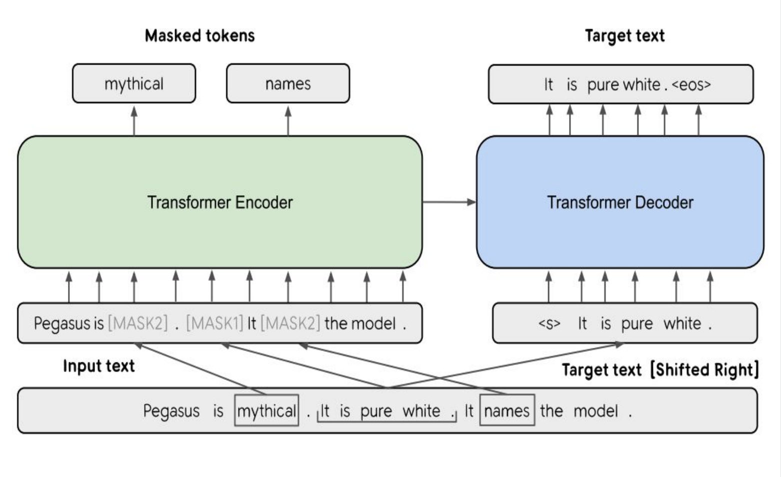
Seq2Seq and Seq2Seq + attention


CNN/DailyMail News Artciles Text
After Preprocessing
Glove Embeddings
Models
Seq2Seq
Seq2Seq with Attention
Two main hidden layers (One Bi-directional LSTM in encoder and LSTM in decoder)
Activation LSTM - Tanh and Output Softmax
16,351,548 parameters, with 8,185,048 trainable
Hidden layers (Seq2seq) + Attention Layer
Dropout Regularization in Decoder LSTM
16837692 parameters with 8671192 trainable
'Sparse Cross Entropy'
Optimizer : 'rmsprop'
On test dataset
Output - Summary

Pre-trained Models:

Pegasus -

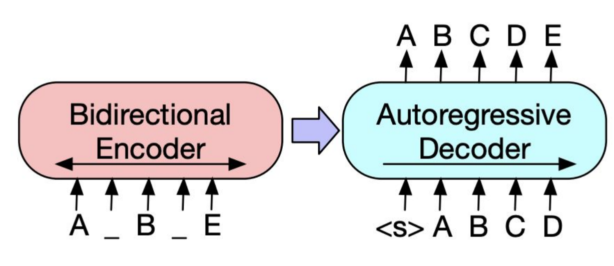
- Transformer based model for Abstractive Summarization.
- Pre-trained on two self-supervised objective functions MLM & GSG.
  MLM - Masked Language Model
  GSG - Gap Sentences Generation
- It has transformer encoders and decoders with 16 layers.

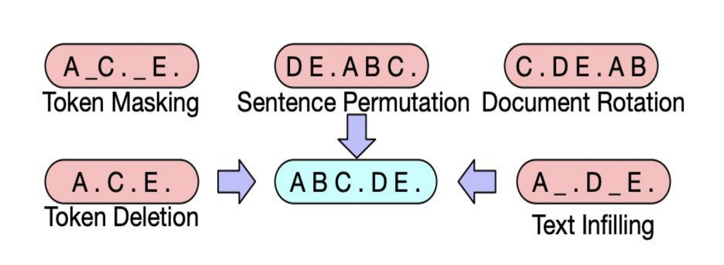

Masked tokens
Target text
mythical
names
It is pure white . <eos>
Transformer Encoder
Transformer Decoder
Pegasus is [MASK2] . [MASK1] It [MASK2] the model .
<s> It is pure white .
Input text
Target text [Shifted Right]
Pegasus is mythical . It is pure white . It names the model .

Pre-trained Models:

BART -

- Pre-trained method that combines Bidirectional and Autoregressive transformers.
- Pre-training has 2 phases. It corrupts the text during pre-training process. Seq2Seq model to reconstruct the original text.
- It has 6 layers for base model and 12 layers for large model.

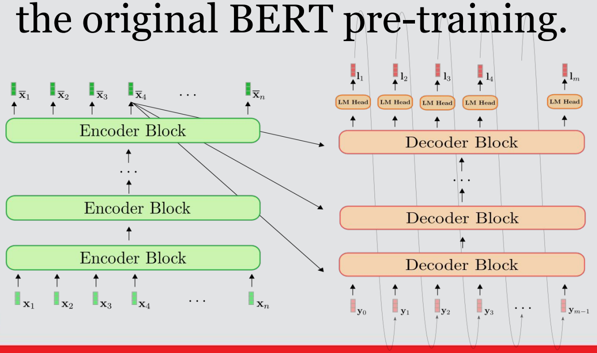




Fine-Tuning the Pre-Trained

Bert2Bert -

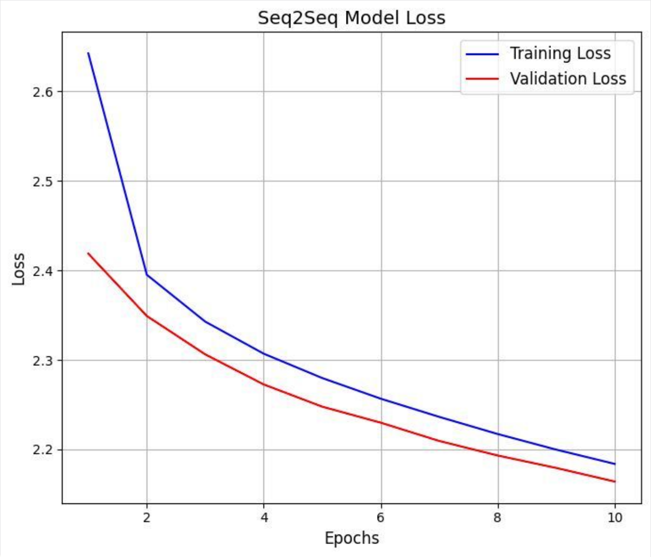
- Both the encoder and decoder are Bert-Base-Uncased Bert Models.
- The decoder needs cross-attention layers and uses auto regressive generation.
- bert2bert saves 45% computation cost of the original BERT pre-training.

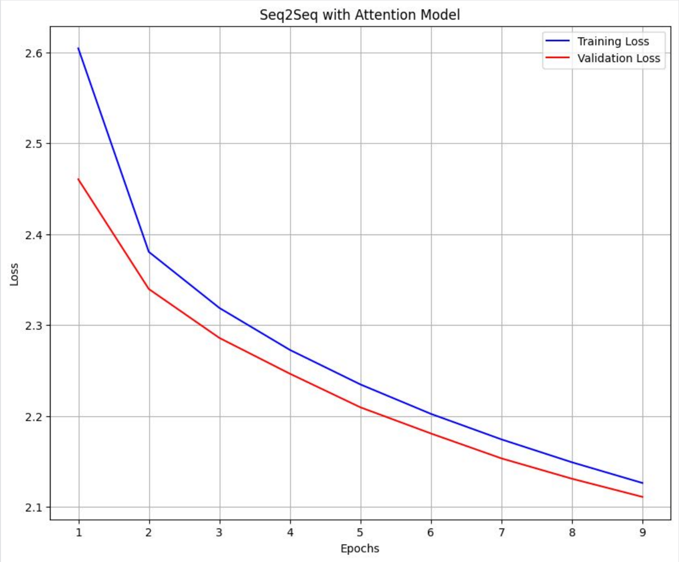

$\bar{x}_1$ $\bar{x}_2$ $\bar{x}_3$ $\bar{x}_4$ ... $\bar{x}_n$
Encoder Block
...
Encoder Block
Encoder Block
x_1 x_2 x_3 x_4 ... x_n
l_1 l_2 l_3 l_4 l_m
LM Head
LM Head
LM Head
LM Head
LM Head
Decoder Block
...
Decoder Block
Decoder Block
y_0 y_1 y_2 y_3 ... y_{m-1}

T5 Text to Text Summarizer

- Parameters - 222M
- The 't5-small' variant strikes a balance between performance and resource usage
- Utilized T5ForConditionalGeneration class

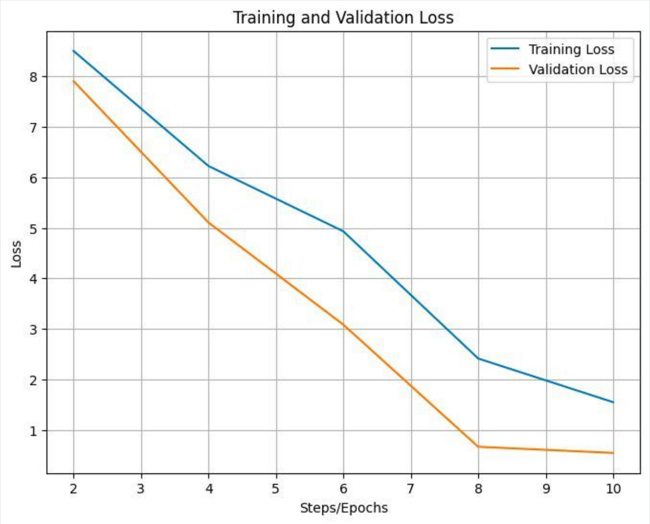
Results


Seq2Seq Model Loss
Training Loss
Validation Loss
Loss
2.6
2.5
2.4
2.3
2.2
2
4
6
8
10
Epochs

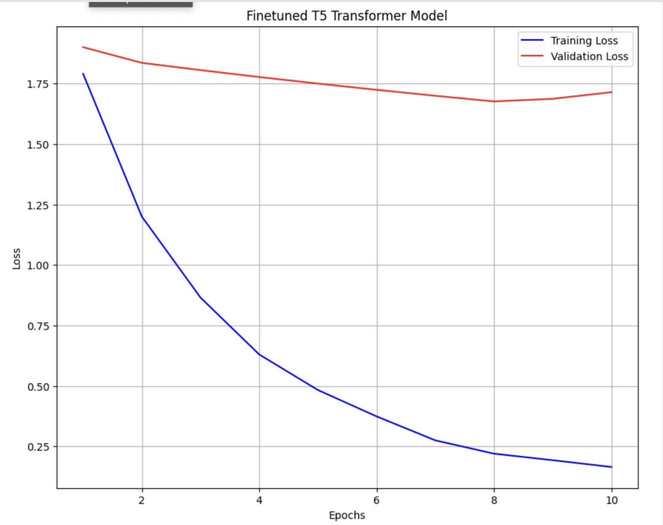

Seq2Seq with Attention Model
Training Loss
Validation Loss
Loss
2.6
2.5
2.4
2.3
2.2
2.1
1
2
3
4
5
6
7
8
9
Epochs

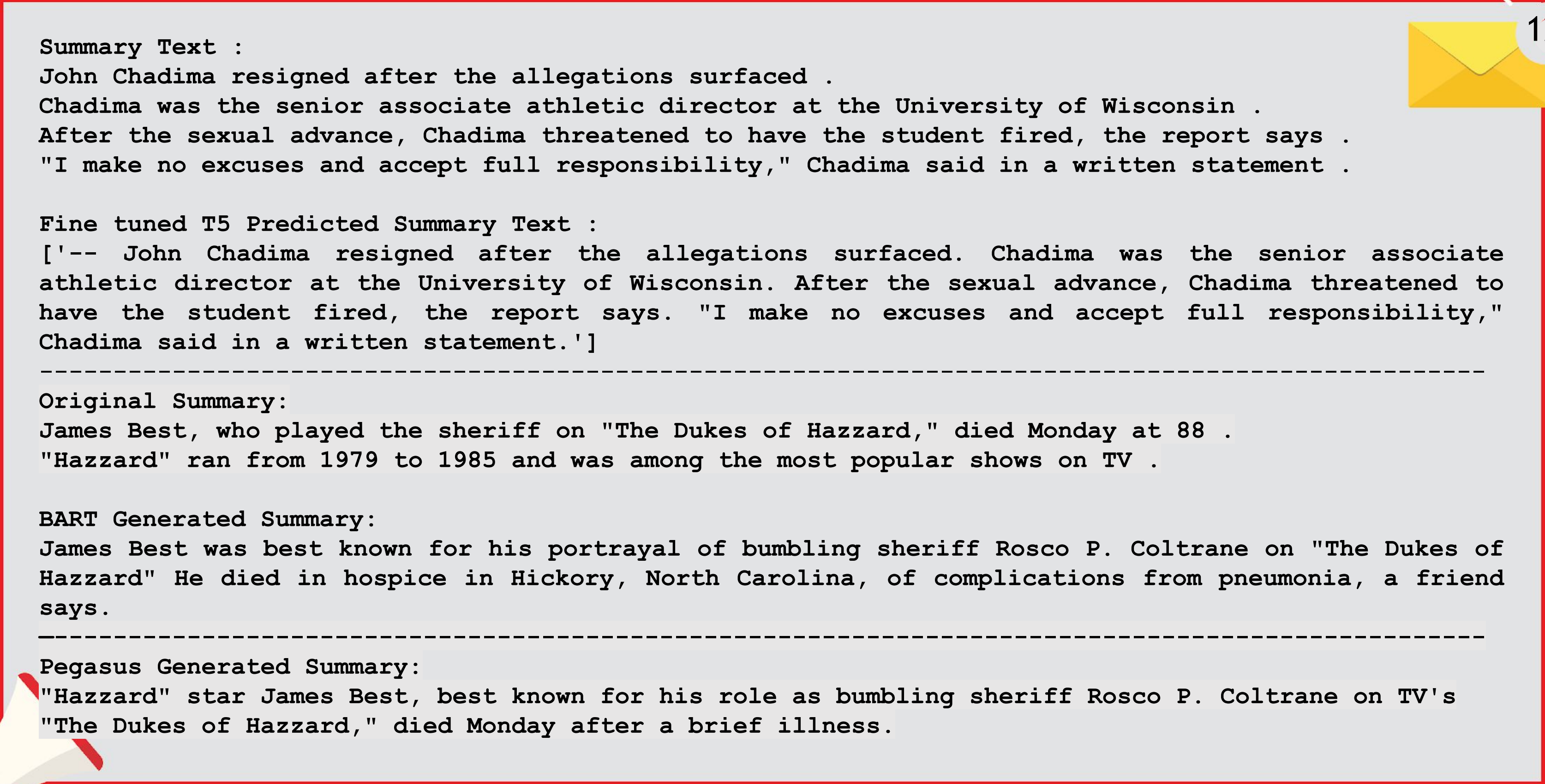
Results

Fine-tuned Bert2Bert Model



Fine-tuned T5 Model





```
Summary Text :
John Chadima resigned after the allegations surfaced .
Chadima was the senior associate athletic director at the University of Wisconsin .
After the sexual advance, Chadima threatened to have the student fired, the report says .
"I make no excuses and accept full responsibility," Chadima said in a written statement .

Fine tuned T5 Predicted Summary Text :
['-- John Chadima resigned after the allegations surfaced. Chadima was the senior associate
athletic director at the University of Wisconsin. After the sexual advance, Chadima threatened to
have the student fired, the report says. "I make no excuses and accept full responsibility,"
Chadima said in a written statement.']
--------------------------------------------------------------------------------------------------
Original Summary:
James Best, who played the sheriff on "The Dukes of Hazzard," died Monday at 88 .
"Hazzard" ran from 1979 to 1985 and was among the most popular shows on TV .

BART Generated Summary:
James Best was best known for his portrayal of bumbling sheriff Rosco P. Coltrane on "The Dukes of
Hazzard" He died in hospice in Hickory, North Carolina, of complications from pneumonia, a friend
says.
--------------------------------------------------------------------------------------------------
Pegasus Generated Summary:
"Hazzard" star James Best, best known for his role as bumbling sheriff Rosco P. Coltrane on TV's
"The Dukes of Hazzard," died Monday after a brief illness.
```

Results

Model	Validation Loss	Rouge Scores
Seq2Seq	2.1549	0.08
Seq2Seq with Attention	2.11	0.03
BART	-	0.34
Pegasus	-	0.41
Bert2Bert Fine Tuned	0.7483	0.0001
T5 Fine Tuned	1.67	0.30

Challenges

- Few Dependency library versions.
- Full training for fine tuning couldn’t be done on our systems due to memory error.
- We couldn’t successfully perform our baseline models on CPU, GPUs.

Conclusion

Insights

“Fine Tuned T5 and pre-trained BART, Pegasus show significant results in text summarization, by high ROUGE score of 0.41. This highlights growing impact of transformers and LLMs in NLP”

Improvements

“Experiment implementation using advanced Transformer models such as GPT-3, and Finetune other BERT models. Also improve rouge scores using different optimizers”

Reflections

“This project offered valuable insights, revealing remarkable evolution from fundamental baselines to sophisticated transformers in the area of text summarization”

CNN
Thank you
GUE CND NES DE RFORNV DUERIL NENK CNEFF POMIN
The Cantdut
The igume
Ohe Conip
Celaupot lonc diner
Gatlewer you ceah Haatiuy
Thank as ith, can ihcredey yud
Can tl hetks vil verh. rill lhalrs