Text Summarization Techniques: Exploring Seq2Seq, BERT, and T5 Models

(DS680 001 NLP)

Himaja Kinthada (sk3355)
Siddhant Soam (ss4688)
Siddharth Pradhan (sp2882)
Tanmay Agarwal (ta424)

Problem Statement: Generating Headlines from News Articles using abstractive text summarization

Introduction:

- News articles contain a wealth of information but can be lengthy and time-consuming to read thoroughly.
- The need for concise and informative headlines is crucial to quickly grasp the essence of an article.

The Challenge:

- Headline generation from news articles poses a significant challenge due to the need to condense information while retaining key details.
- Creating headlines that are informative, engaging, and capture the essence of the article's content accurately

Proposed Solution:

- Utilizing advanced Natural Language Processing (NLP) techniques for data preprocessing and preparation.

- Exploring the following NLP models to generate headlines from the articles:
  1) Seq2seq
  2) T5-base
  3) BERT (Text Summarization with Pretrained Encoders)

- Exploring novel methodologies to extract salient information and generate coherent and informative headlines.

DATASET DESCRIPTION

Link : https://www.kaggle.com/datasets/sunnysai12345/news-summary

news_summary.csv

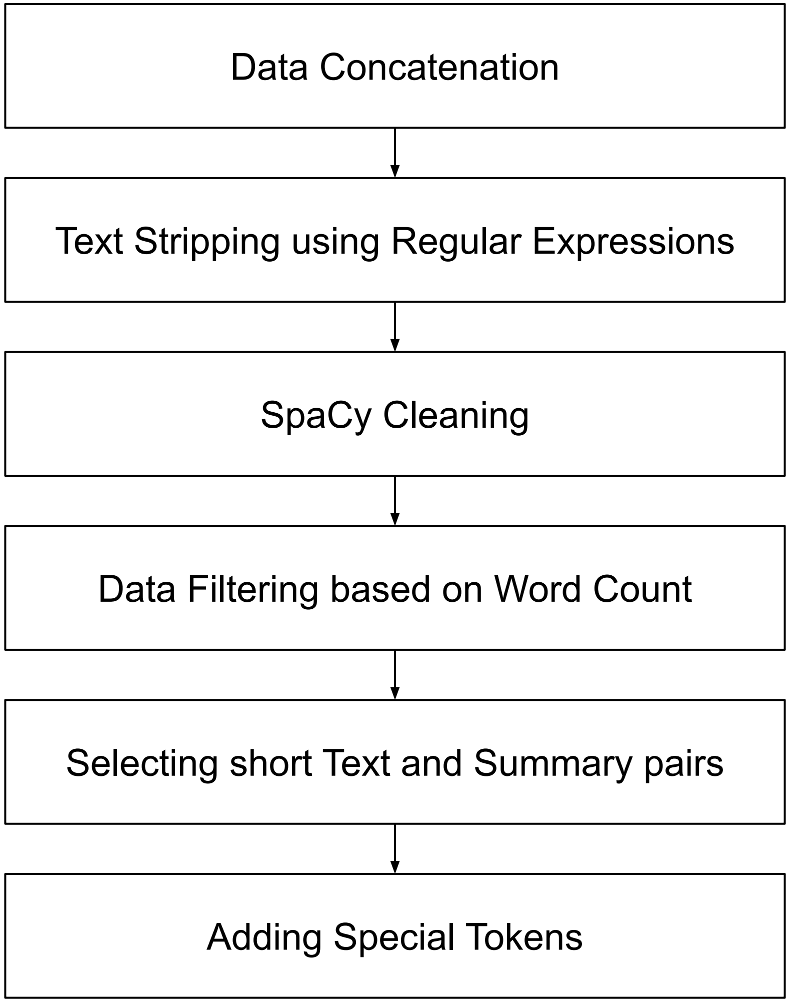
Columns	Description
Author	Author Name
Date	Publish Date
Headlines	News Headline
read_more	Link for Headline
text	Summary
ctext	Complete news text

news_summary_more.csv

Columns	Description
Headlines	News Headline
Text	Summary

- Contains news articles collected from Inshorts, Hindu, Indian times, and Guardian between February and August 2017.
- It consists of 102,274 data points consisting of news summary and headlines.

DATA PREPROCESSING



Data Concatenation

Text Stripping using Regular Expressions

SpaCy Cleaning

Data Filtering based on Word Count

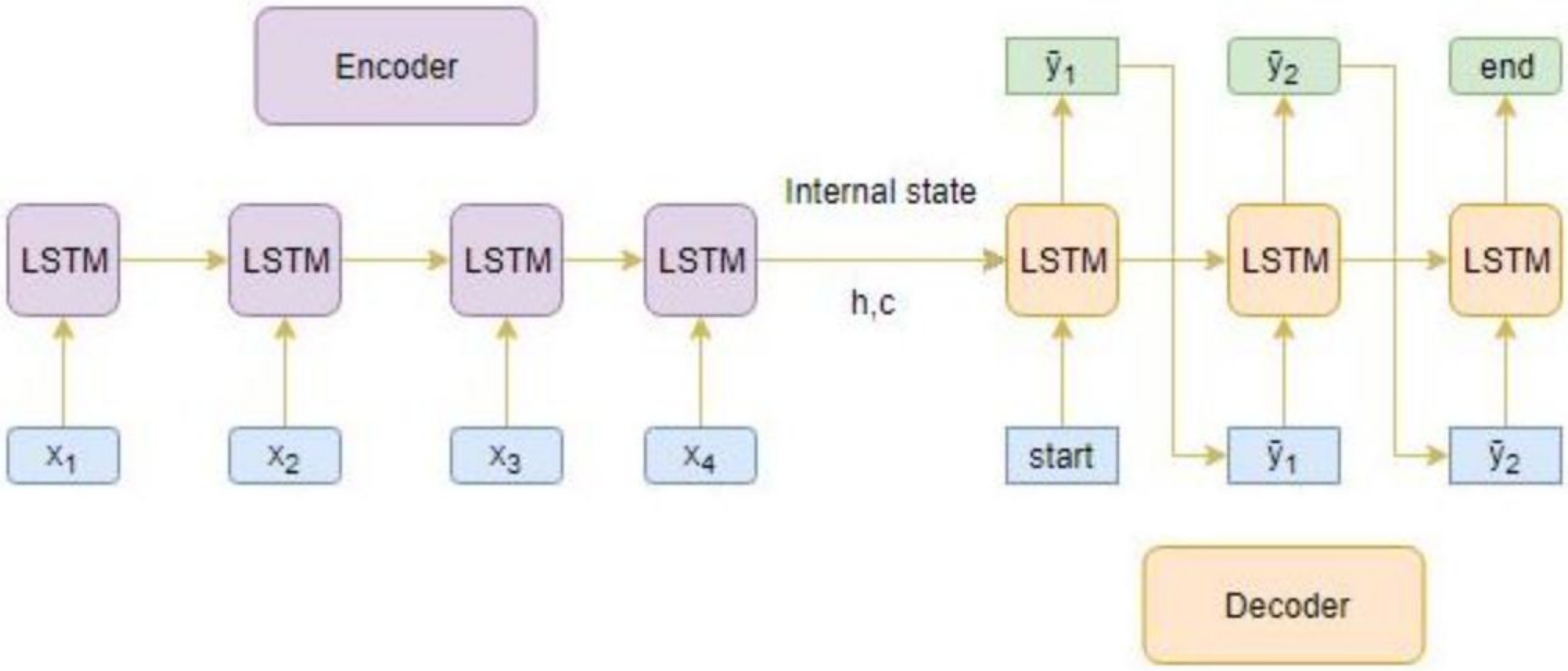
Selecting short Text and Summary pairs

Adding Special Tokens

DATA PREPROCESSING(cont.)

- Training data size = 70000
  Validation data size = 10000
  Test size = 1000
- Rare words identification using tokenization.
- Calculation of number of words in the vocab.
- Removing headlines which has only <START> and <END> tokens (empty headlines).
- Label data preparation by converting words into integers.
- Ensuring all sequences have the same length.

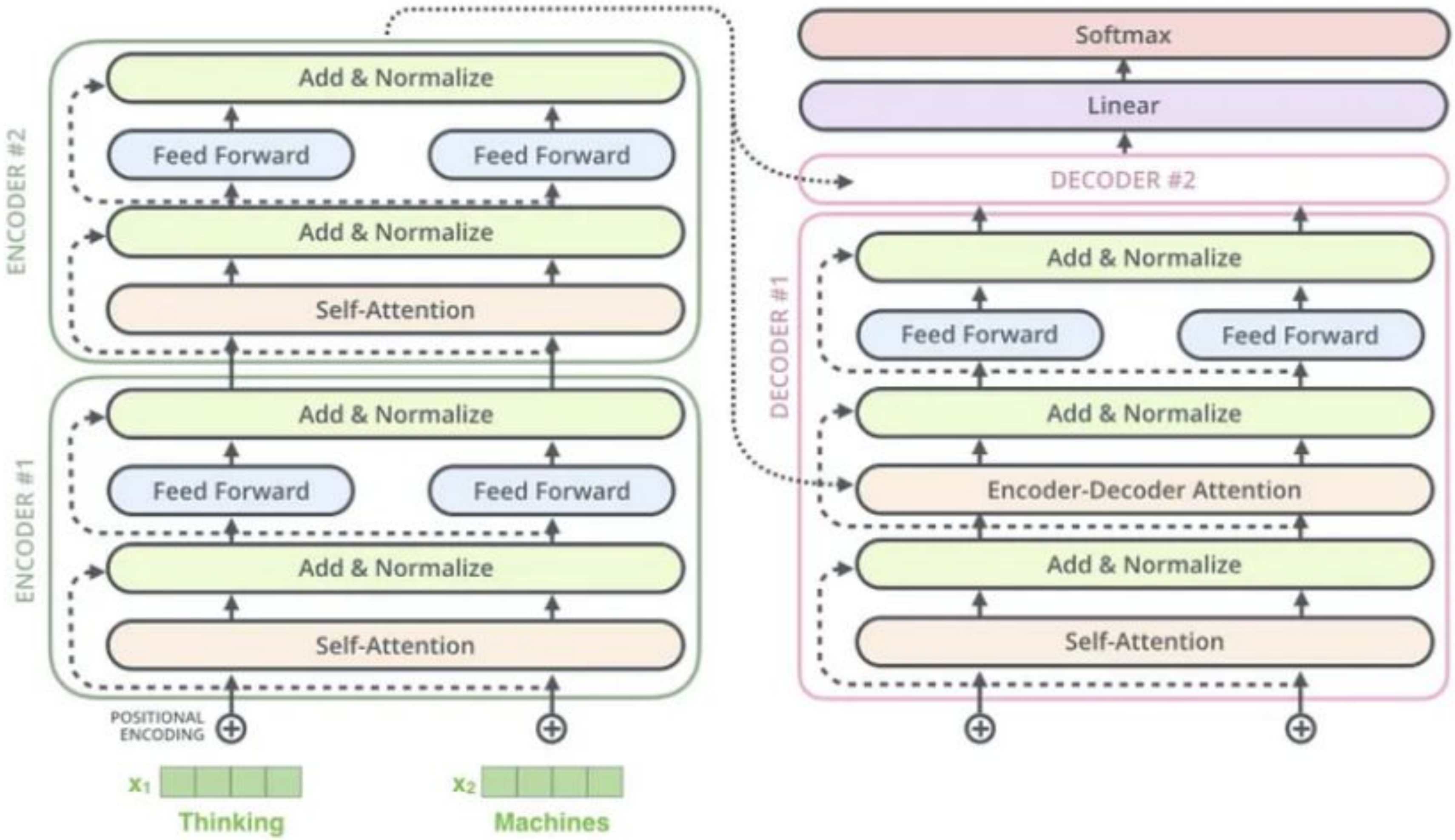
Seq2Seq (Sequence To Sequence) Model Architecture



Overview of Seq2Seq model Training

- Total Number of Parameters: 18 Millions (approx)
- A sequence of integers representing text(news) with size 100 is passed to the encoder.
- A sequence of integers representing words in target summary is passed with flexible size.
- The model is trained using sample of 80K rows of text(news) and ran for 50 epoch with batch size of 128.
- Train Loss: 3.20, Validation Loss = 3.65
- Training Time: 3 hours (approx) on GPU P100

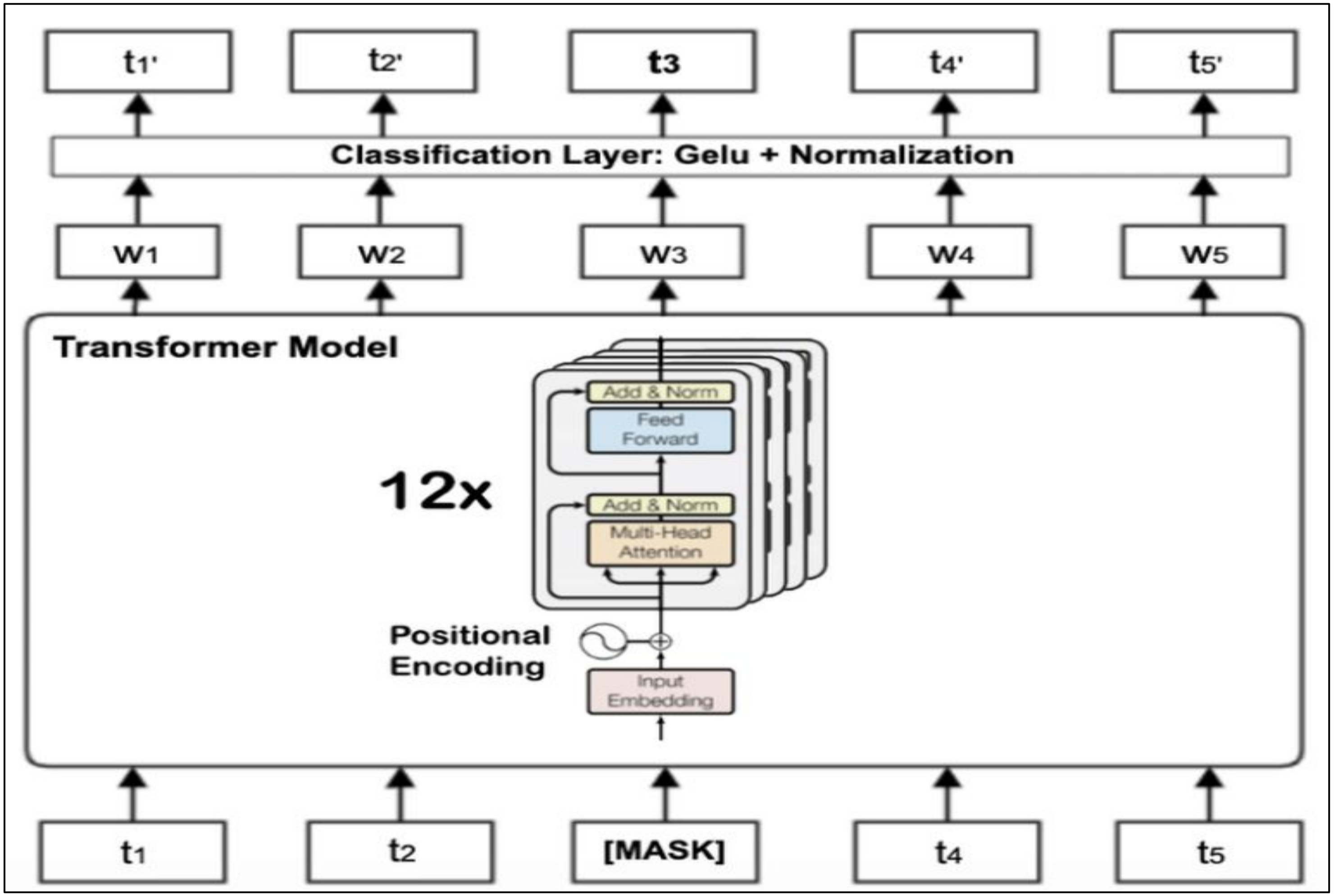
T5 Base Model Architecture



T5 model (cont.)

- Number of parameters: 220 million
- Requires input data in the form of a data frame consists of 2 columns names as “source_text” and “targte_text”.
- It is pretrained, but we did an additional training with 10K rows and ran for 5 epochs.
- Train loss : 0.6599 , val loss : 1.3475
- To generate summaries, we need to add “summarize: ” keyword at the beginning of the text.
- Generated headlines of 100 articles to save execution time.
- Notebook took 43 mins to run on GPU P100

BERT Architecture



BERT model(cont.)

- Bidirectional Encoder Representations from Transformers
- 110M parameters, with 12 transformer layers, with hidden size as 768 and 12 attention heads
- Loaded the pretrained model from the ACL paper: Text Summarization with Pretrained Encoders
- The encoder is fine tuned twice, first on the extractive summarization task & then on the abstractive summarization task, that's why they called it “BertSumExtAbs”.
- As per the paper, this approach enhances the quality of abstractive summarization.
- Notebook took 1 hour to execute, which includes setting up the environment, downloading the model and the generation task, on GPU P100.

Evaluation Metrics

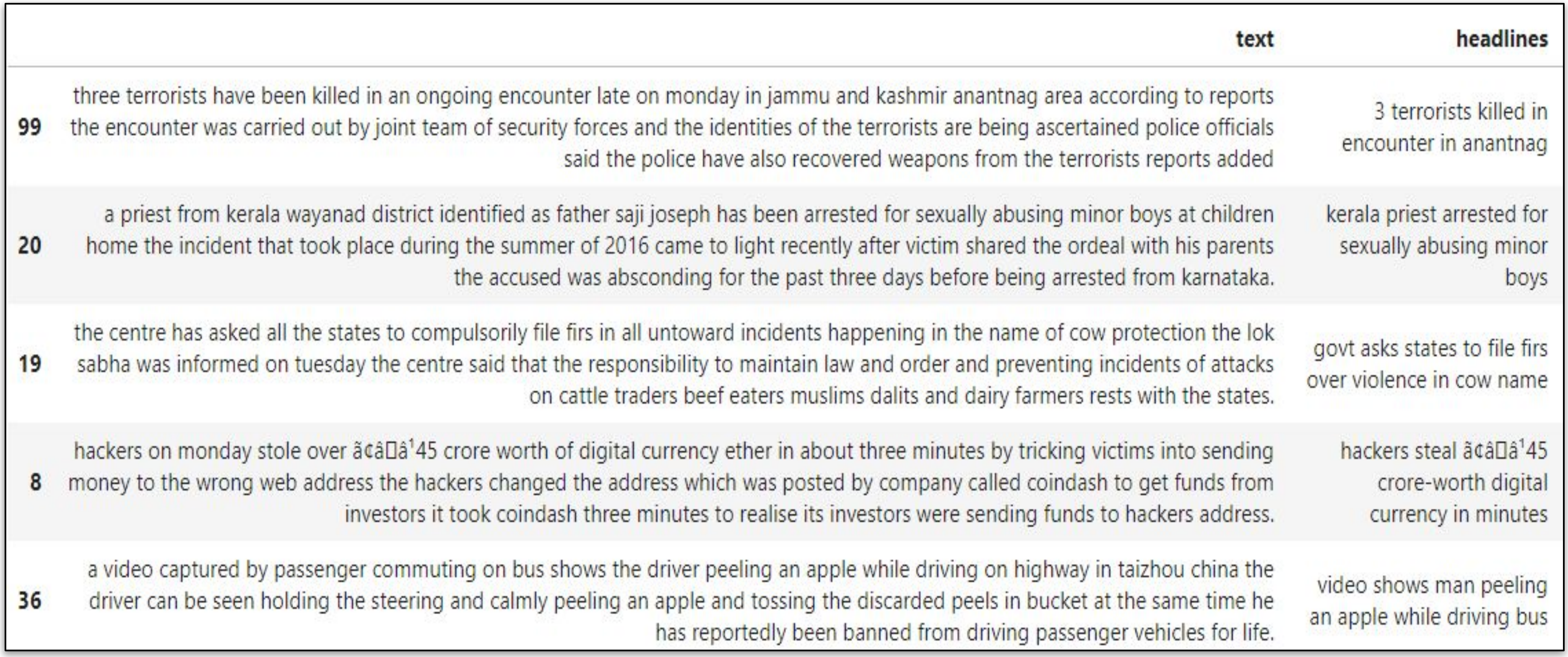
ROUGE Scores:

- ROUGE-1 : Measures overlap of unigrams (single words)
- ROUGE-2 : Evaluates the overlap of bigrams (pairs of words).
- ROUGE-L : Considers the longest common subsequence between the generated and reference text.

BLEU Score:

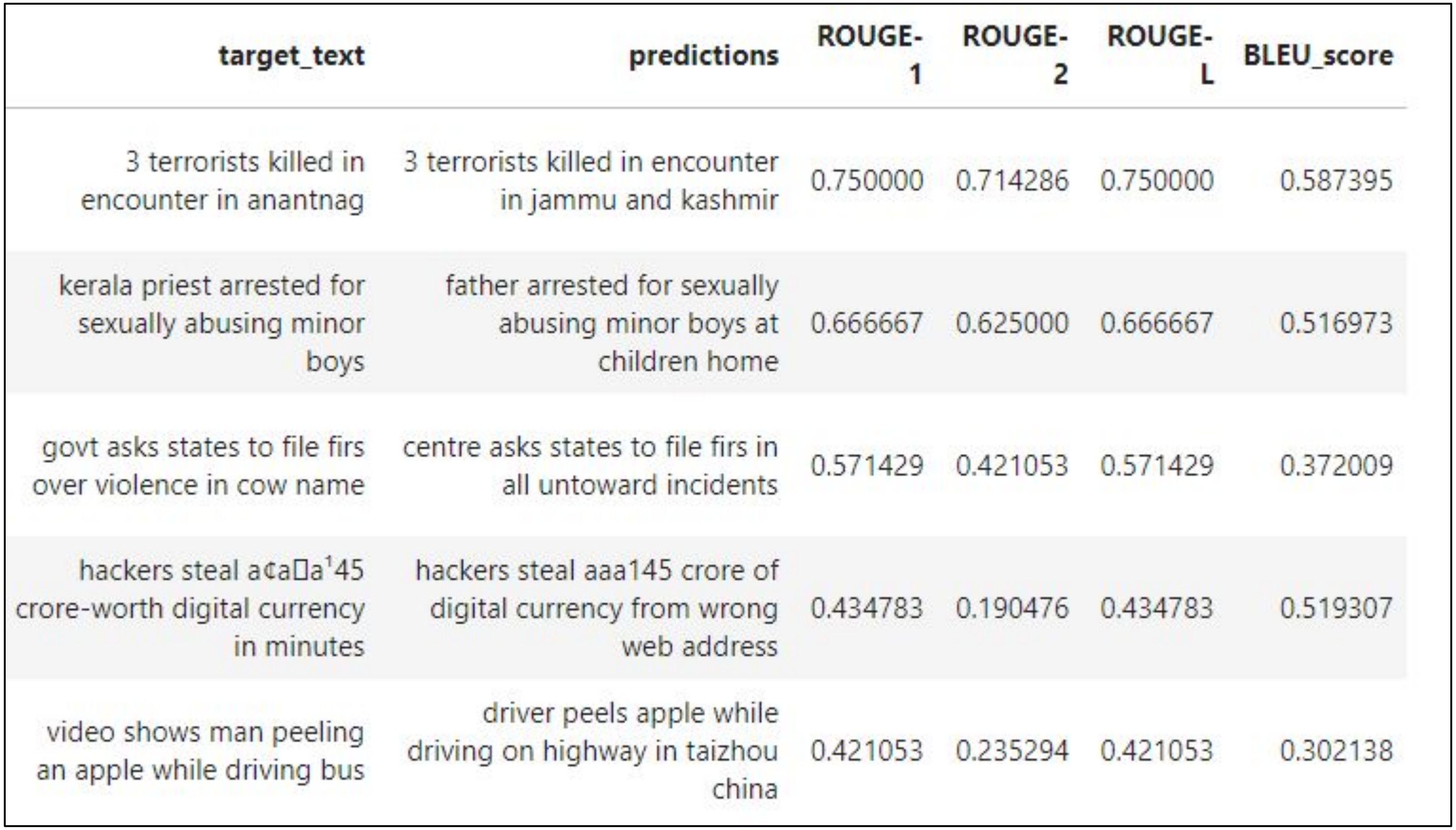
- Bilingual Evaluation Understudy
- Originally designed for machine translation evaluation but also used in text summarization.
- Measures the n-gram overlap (up to a certain length) between the generated text and reference text.

Sample of the test set used for generating summaries:



	text	headlines
99	three terrorists have been killed in an ongoing encounter late on monday in jammu and kashmir anantnag area according to reports the encounter was carried out by joint team of security forces and the identities of the terrorists are being ascertained police officials said the police have also recovered weapons from the terrorists reports added	3 terrorists killed in encounter in anantnag
20	a priest from kerala wayanad district identified as father saji joseph has been arrested for sexually abusing minor boys at children home the incident that took place during the summer of 2016 came to light recently after victim shared the ordeal with his parents the accused was absconding for the past three days before being arrested from karnataka.	kerala priest arrested for sexually abusing minor boys
19	the centre has asked all the states to compulsorily file firs in all untoward incidents happening in the name of cow protection the lok sabha was informed on tuesday the centre said that the responsibility to maintain law and order and preventing incidents of attacks on cattle traders beef eaters muslims dalits and dairy farmers rests with the states.	govt asks states to file firs over violence in cow name
8	hackers on monday stole over ã¢â□â¹45 crore worth of digital currency ether in about three minutes by tricking victims into sending money to the wrong web address the hackers changed the address which was posted by company called coindash to get funds from investors it took coindash three minutes to realise its investors were sending funds to hackers address.	hackers steal ã¢â□â¹45 crore-worth digital currency in minutes
36	a video captured by passenger commuting on bus shows the driver peeling an apple while driving on highway in taizhou china the driver can be seen holding the steering and calmly peeling an apple and tossing the discarded peels in bucket at the same time he has reportedly been banned from driving passenger vehicles for life.	video shows man peeling an apple while driving bus

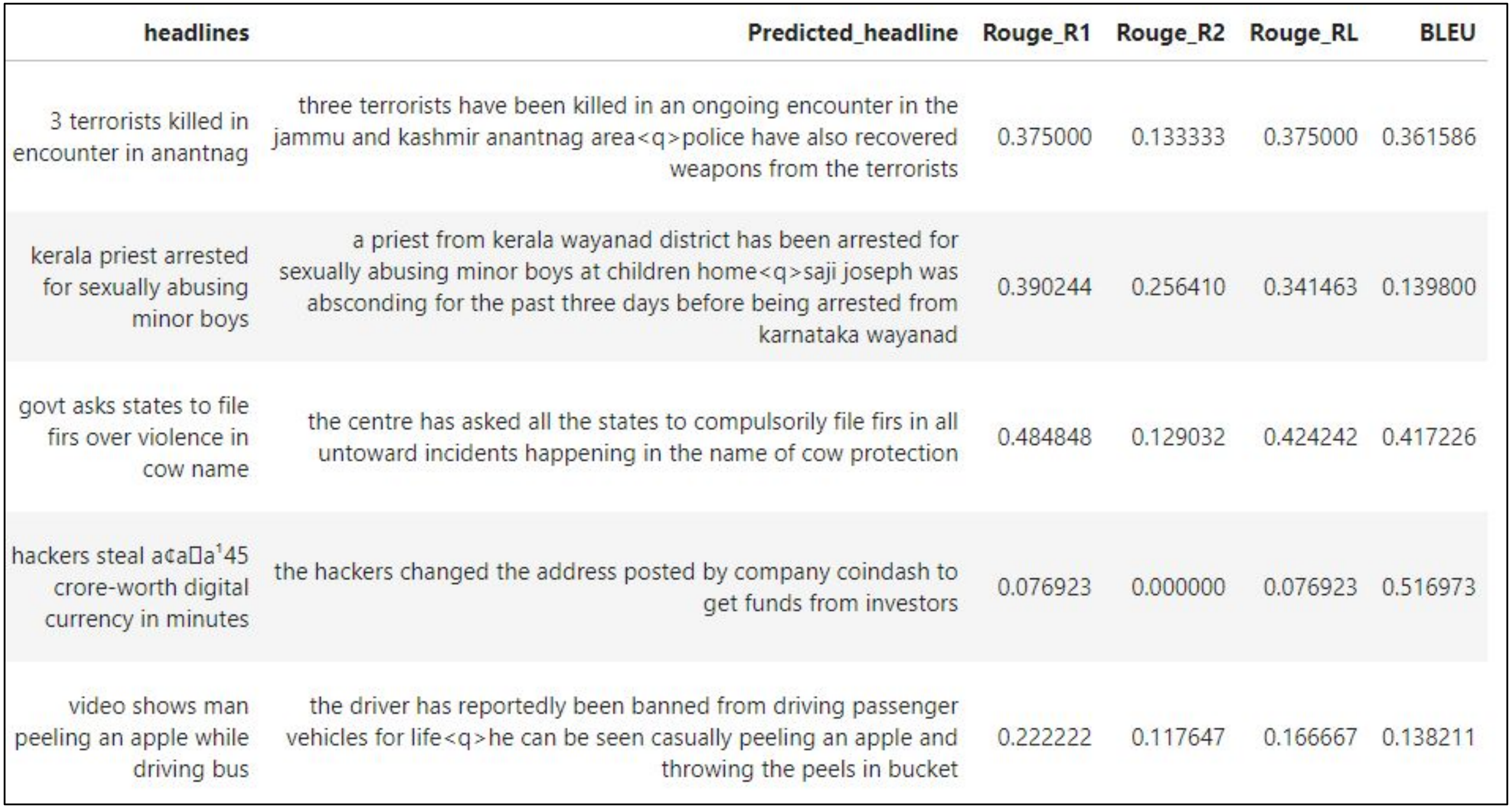
T5 generated summaries



target_text	predictions	ROUGE-1	ROUGE-2	ROUGE-L	BLEU_score
3 terrorists killed in encounter in anantnag	3 terrorists killed in encounter in jammu and kashmir	0.750000	0.714286	0.750000	0.587395
kerala priest arrested for sexually abusing minor boys	father arrested for sexually abusing minor boys at children home	0.666667	0.625000	0.666667	0.516973
govt asks states to file firs over violence in cow name	centre asks states to file firs in all untoward incidents	0.571429	0.421053	0.571429	0.372009
hackers steal a¢a□a^{1}45 crore-worth digital currency in minutes	hackers steal aaa145 crore of digital currency from wrong web address	0.434783	0.190476	0.434783	0.519307
video shows man peeling an apple while driving bus	driver peels apple while driving on highway in taizhou china	0.421053	0.235294	0.421053	0.302138

Eval metric	Rouge-1	Rouge-2	Rouge-L	BLEU
Avg Score	0.537	0.139	0.486	0.472

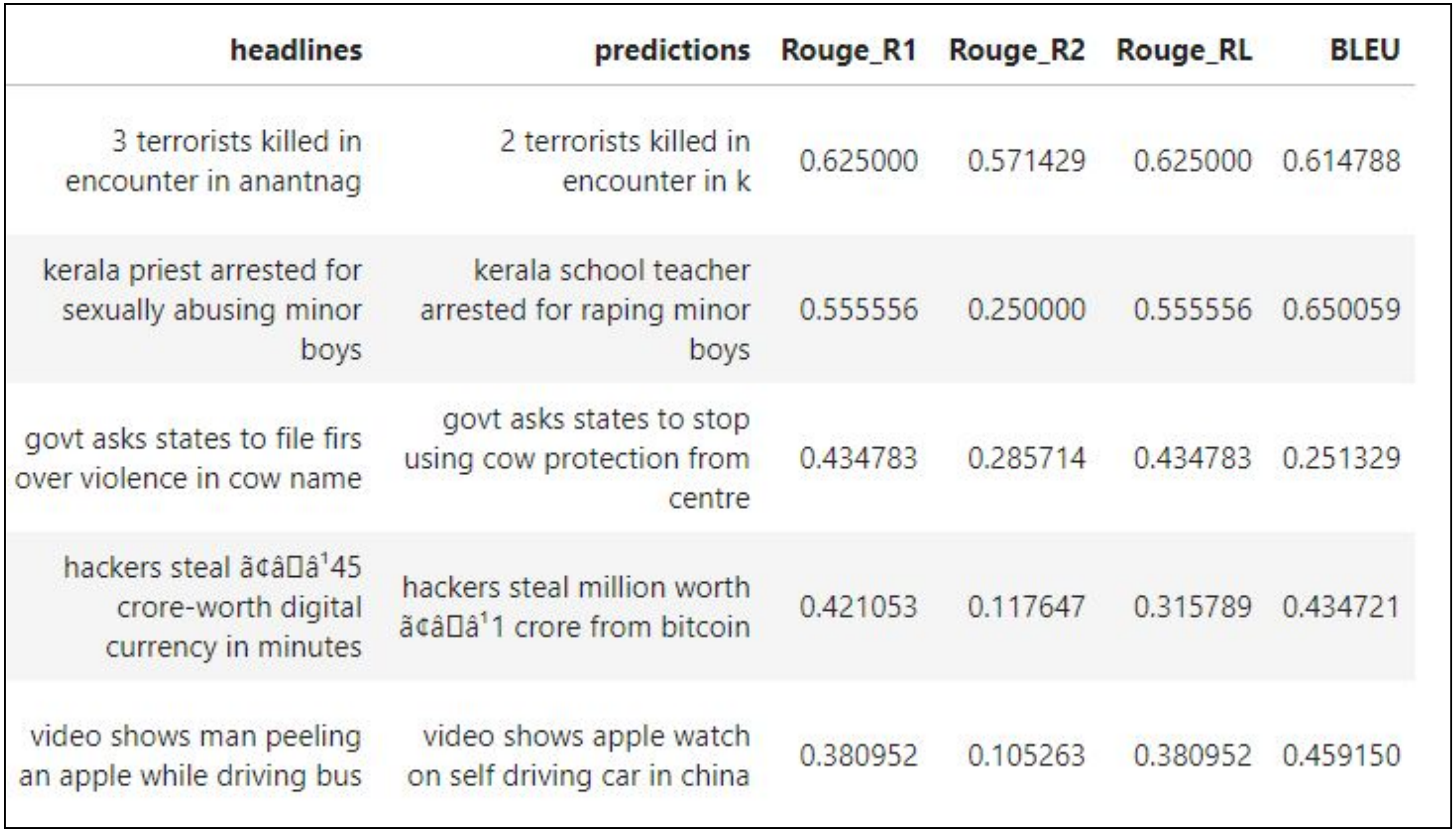
Bert generated summaries



headlines	Predicted_headline	Rouge_R1	Rouge_R2	Rouge_RL	BLEU
3 terrorists killed in encounter in anantnag	three terrorists have been killed in an ongoing encounter in the jammu and kashmir anantnag area<q>police have also recovered weapons from the terrorists	0.375000	0.133333	0.375000	0.361586
kerala priest arrested for sexually abusing minor boys	a priest from kerala wayanad district has been arrested for sexually abusing minor boys at children home<q>saji joseph was absconding for the past three days before being arrested from karnataka wayanad	0.390244	0.256410	0.341463	0.139800
govt asks states to file firs over violence in cow name	the centre has asked all the states to compulsorily file firs in all untoward incidents happening in the name of cow protection	0.484848	0.129032	0.424242	0.417226
hackers steal a¢a□a¹45 crore-worth digital currency in minutes	the hackers changed the address posted by company coindash to get funds from investors	0.076923	0.000000	0.076923	0.516973
video shows man peeling an apple while driving bus	the driver has reportedly been banned from driving passenger vehicles for life<q>he can be seen casually peeling an apple and throwing the peels in bucket	0.222222	0.117647	0.166667	0.138211

Eval metric	Rouge-1	Rouge-2	Rouge-L	BLEU
Avg Score	0.213	0.043	0.198	0.091

Seq2seq generated summaries



headlines	predictions	Rouge_R1	Rouge_R2	Rouge_RL	BLEU
3 terrorists killed in encounter in anantnag	2 terrorists killed in encounter in k	0.625000	0.571429	0.625000	0.614788
kerala priest arrested for sexually abusing minor boys	kerala school teacher arrested for raping minor boys	0.555556	0.250000	0.555556	0.650059
govt asks states to file firs over violence in cow name	govt asks states to stop using cow protection from centre	0.434783	0.285714	0.434783	0.251329
hackers steal ã¢â□â¹45 crore-worth digital currency in minutes	hackers steal million worth ã¢â□â¹1 crore from bitcoin	0.421053	0.117647	0.315789	0.434721
video shows man peeling an apple while driving bus	video shows apple watch on self driving car in china	0.380952	0.105263	0.380952	0.459150

Eval metric	Rouge-1	Rouge-2	Rouge-L	BLEU
Avg Score	0.084	0.007	0.081	0.100

Challenges Faced

- Limited resources hindered our ability to train the Seq2seq model effectively.

- The pre-trained BERT model generating longer headlines which couldn’t generate better rouge and bleu scores.

Thank You

(Q&A)