Using NLP to Analyze News Articles

Team 6
Himanshee
Karnik Ketan Kalani
Lakshmi Satvika Nekkanti
Saumya Sinha

Background and Executive Summary

- Over the last decade, the number of news pieces published each day has skyrocketed. Internet is flooded with news articles and all organizations also have thier own news portals.
- Manually processing this data is a time-consuming and labor-intensive task.
- Hence this project aims to summarize and classify news articles (text data) for user convenience.
- It will enable users to access information of their interests quickly and effectively.

Project Requirements

Functional Requirements	• *Libraries such as Hugging Face Transformers, Pytorch, TensorFlow, Tokenisers, Numpy, Pandas etc.* • *Google Colab Pro Noteooks for faster executions.* • *GitHub/Huggingface repository for version control.*
AI Requirements	• *Models for classification task: Bert , Roberta and GPT-2.* • *Models for summarisation task: T5 and GPt-2.*

Data Requirements	• *Classification task*
	  ◦ *BBC News Articles : Source- Huggingface Libraray, Size-, categories- 4*
	  ◦ *Ag news : Source - Huggingface Librara, Size - ,Categories -4*
	• *Summarization task*
	  ◦ *Cnn_Daily Mails: Source - tensorflow dataset, features-4, Size- 1.29GB*

Team Organisation

- *Asana: Project management tool*
- *Gantt Charts: Planning, coordinating, and tracking project activities*
- *PERT Charts: Visual representation of task dependencies, critical path, estimating project durations*
- *Team Meetings: Zoom and Google Meet*
- *File and notebook Storage: Google Drive*

Data Description

Datasets:

- **AG news-**
  - **H**ugging Face
  - 127601 rows
  - World,Sports,Business,and Science/Tech
- **BBC News-**
  - Kaggle
  - 1624 rows
  - World,Sports,Business,and Science/Tech
- **Cnn_Dailymail news**
  - KerasDatset hub/Hugging face
  - size of 1.29 gb (298603)

Raw BBC dataset sample:

categoryfilenametitlecontent																
business0(	from $63(	which is n	benefited	and less u	AOL	had has m	000 subscr	the compa	which is c	helped by	a sharp cc	when the	TimeWarr	up 27% fr	while reve	meeting o
business0(	so it will h	said Mosc	we will fight them where the rule of law exists under the international arbitration clauses of the credit." Rosneft officials were unavailable for comment.													
business0(	the airline	BA's chief	said the re	and it exp	BA last ye	it increase	while the	further be	BA warne	it said sale	the total r	BA chairm	however	operating	he said. Si	Mr Edding
business0(	but has ye	while Perr	the move	Pernod - a	which has	one of Sc	but lost o	Havana Cl	Courvoisi	Stolichnay	having ea	while Allied has improved the performance of its fast-f				
business0(	figures sh	the data s	suggesting	and we wi	said econ	said Paul	said Rick E	said Ken Mayland, president of ClearView Economics. That means there are a limited number of ne								
business0(	which atte	is unlikely	India's fin	lashed ou	the World	India's fin	argued tha	given Indi	which acc	he said th	he told the	Mr	a prolong	a total of	000 tonne	so as to a
business01	cancellati	depending	including I	which has	passenger	only sche	airlines wi	if a flight i	all passen	airlines do	security al	arguing th	we among	Ryanair de	national e	she said. I
business01	the Italian	has report	up from 3	the compa	and is falli	well-know	have cont	however	and have	the compa	turnaroun	sued Mor	its former	to return	a mass of legal action	
business01	up from a	which rate	lifted India	more cash	buoying th	said Bhan	which star	Egypt and	and one level below Russia.							
business01	said Mr M	looking to	said Gartn	added	says Cara	now its ag	insurance	Ms Diemo	the averag	leaving "a	Ms Diemo	is "scary"	call centre	with nearl	which in t	but that is
business02	reports cl	formerly f	will expos	the Sunda	the report	the group	including t	Hard Rock	which pro	bars and h	including	which mal	that busin	but said a	roadside s	the Mecca
business02	000 in Nov	or 10.8%	as governi	said Frank	he said. Tl	where the	the goverr	paid for b	has sparke	notably a	have dem	but hiring	economis	said Isabelle Kronawitter at Hypo		

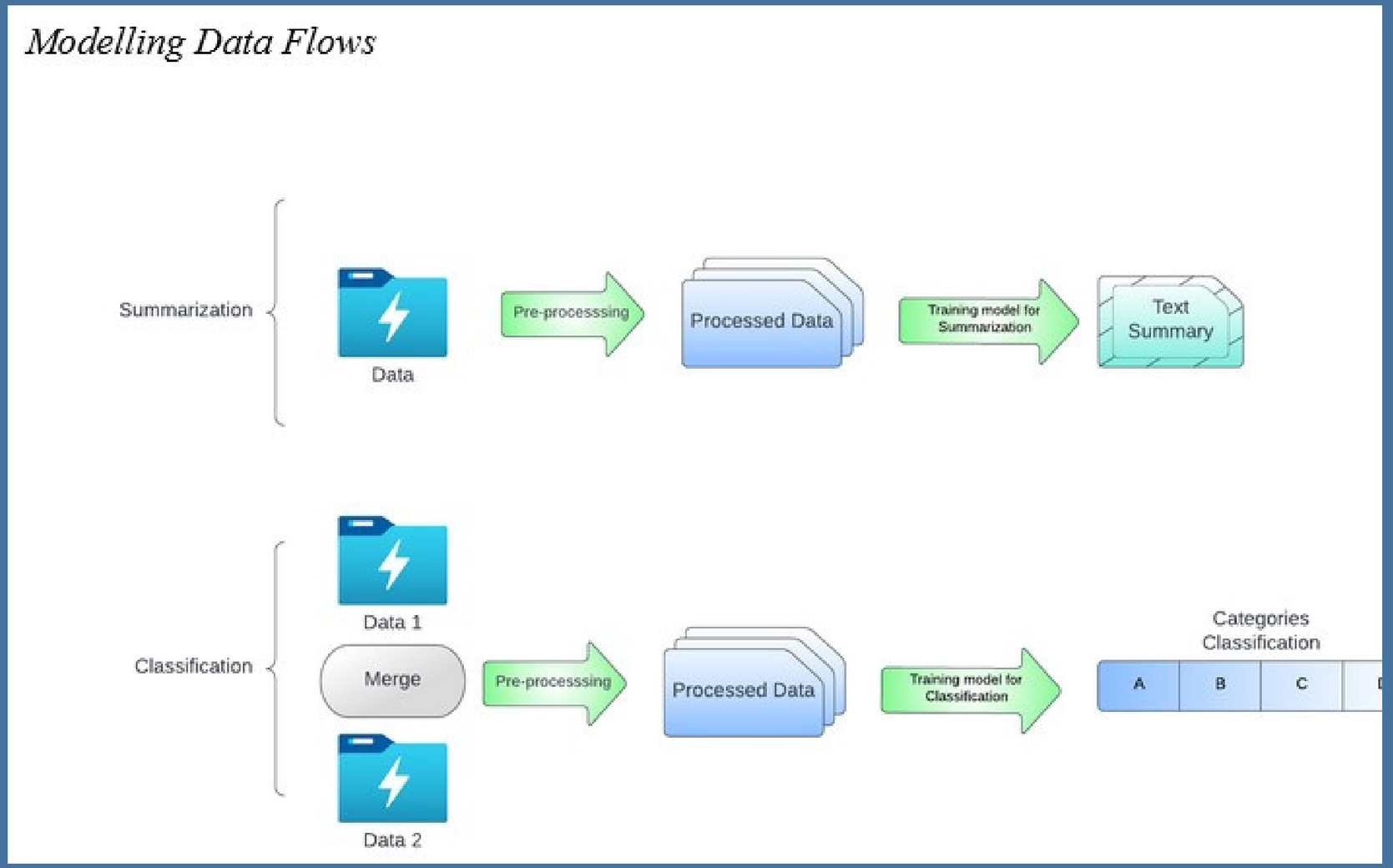
Raw AG news dataset sample:

	Class Index	Title	Description
0	3	Wall St. Bears Claw Back Into the Black (Reuters)	Reuters - Short-sellers, Wall Street's dwindli...
1	3	Carlyle Looks Toward Commercial Aerospace (Reu...	Reuters - Private investment firm Carlyle Grou...
2	3	Oil and Economy Cloud Stocks' Outlook (Reuters)	Reuters - Soaring crude prices plus worries\ab...

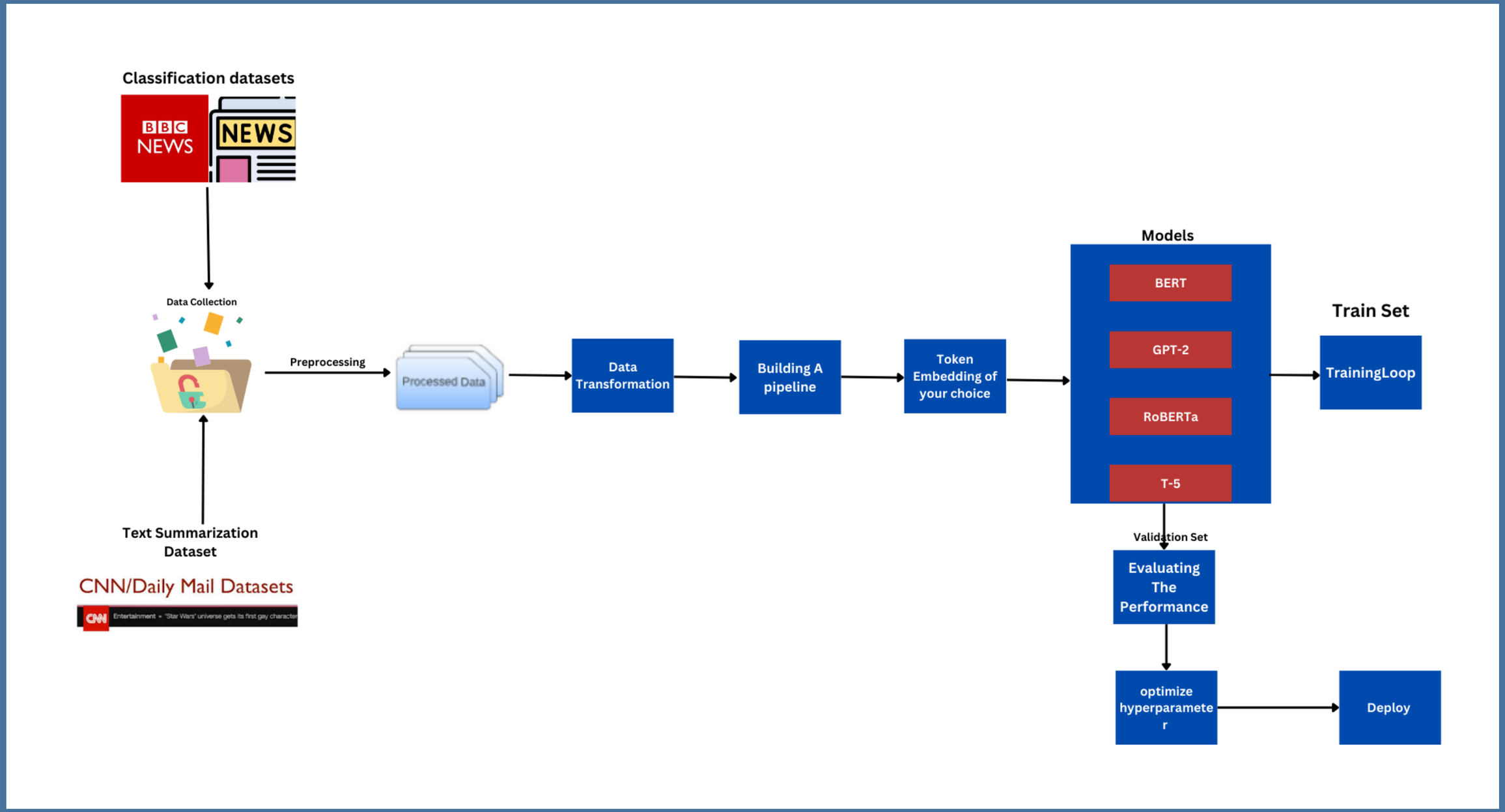
Raw Cnn_dailymail dataset sample:

Unnamed: 0	article	highlights	id	publisher
0	b"Ever noticed how plane seats appear to be ge...	b'Experts question if packed out planes are p...	b'92c514c913c0bdfe25341af9fd72b29db544099b'	b'dm'
1	b"A drunk teenage boy had to be rescued by sec...	b"Drunk teenage boy climbed into lion enclosur...	b'2003841c7dc0e7c5b1a248f9cd536d727f27a45a'	b'dm'
2	b"Dougie Freedman is on the verge of agreeing ...	b"Nottingham Forest are close to extending Dou...	b'91b7d2311527f5c2b63a65ca98d21d9c92485149'	b'dm'
3	b"Liverpool target Neto is also wanted by PSG ...	b'Fiorentina goalkeeper Neto has been linked w...	b'caabf9cbdf96eb1410295a673e953d304391bfbb'	b'dm'

Project Dataflow



Project Architecture



Resource Requirements

Platform	Usage	Cost
Google Drive	To store and share data between teammates	Free upto 15 GB
Google Colab	Developing an environment for the team	Free
Python	To build, train, and analyze NLP models	Open Source
Google Colab Pro	Project Deployment Developing an environment for the team	Free 5 GB (chargeable onwards with $0.0023 per$10 per month
Platform	Usage	Cost
Grammerly	For Documentation	Free
Draw.io	For Diagrams	Free
Google Meet	For Team Meetings	Free
Google Drive	To store and share data between teammates	Free upto 15 GB
Asana	Project planning	Free

Hardware & Software Requirements

Type	Resource	Specifications	Cost
Hardware	GPU	8 - GB	$300 - higher
	CPU	Inter CORE i7- higher	$250 - higher
	RAM	16GB - higher	$50 - higher
Software	OS	Windows 10 or Mac Os	$30 - higher

Data Preparation

Cnn_daily Mail dataset

- *Total features - unnamed:0, id, article , highlights and publisher*
- *Feature types - Object and integer*

Removed unwanted features

Checked for Null, missing and duplicates

Removed Punctuations, html tags

Removes special characters

Data Preparation

BBC and AGnews dataset

- *Total features: "Class Index" - [1,2,3,4] , Title, Description*
- *Class Index[1] refers to world news*
- *Class Index[2] refers to sports*
- *Class Index [3] refers to Business*
- *Class Index[4] refers to Sci/Tech*

Changing the column names

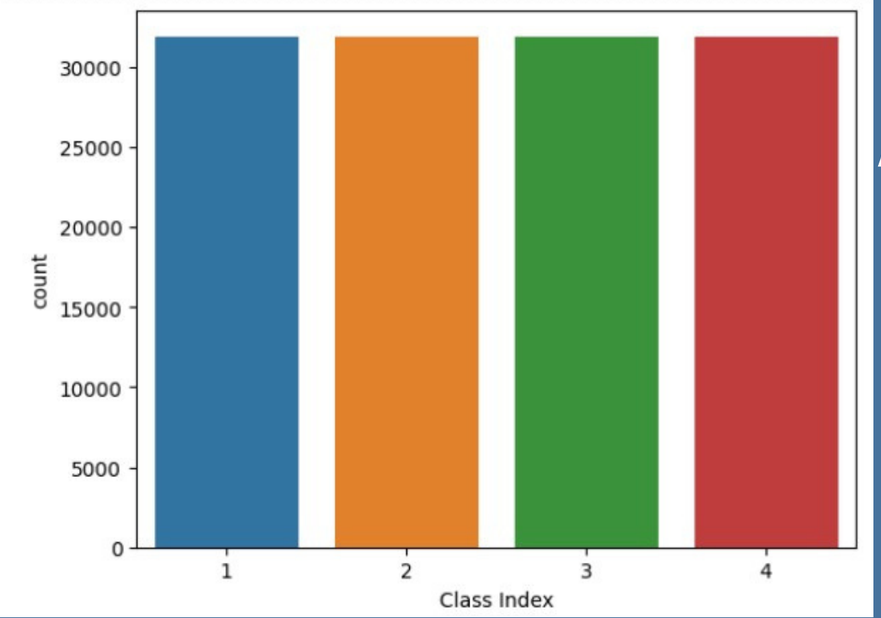
Concatenation of the dataset

Replacing words with numbers such as 1,2,3,4 to describe the Class Index

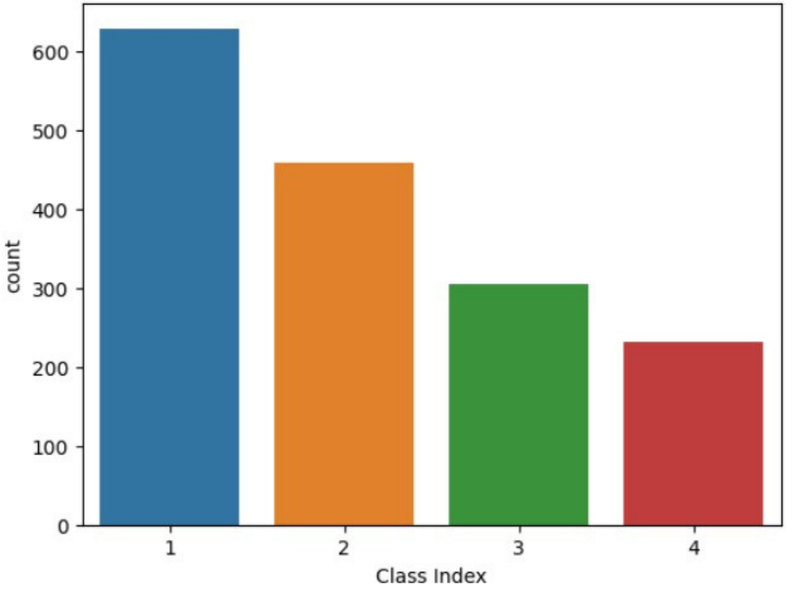
Removing Null Values

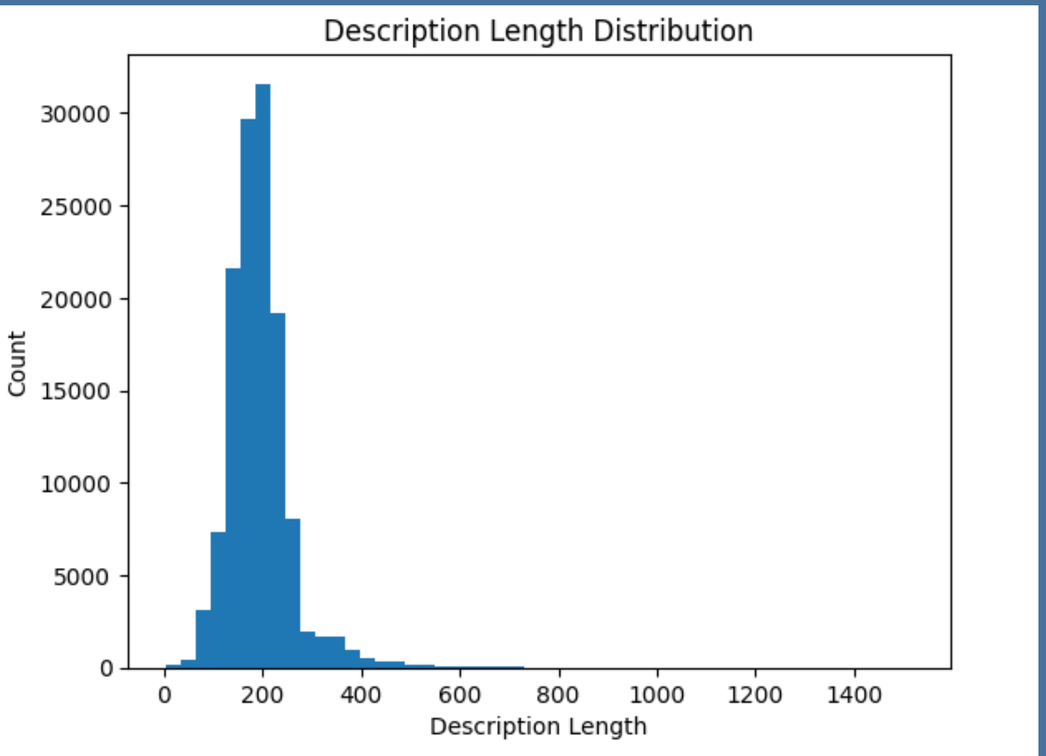
Data Statistics AG news & BBC News

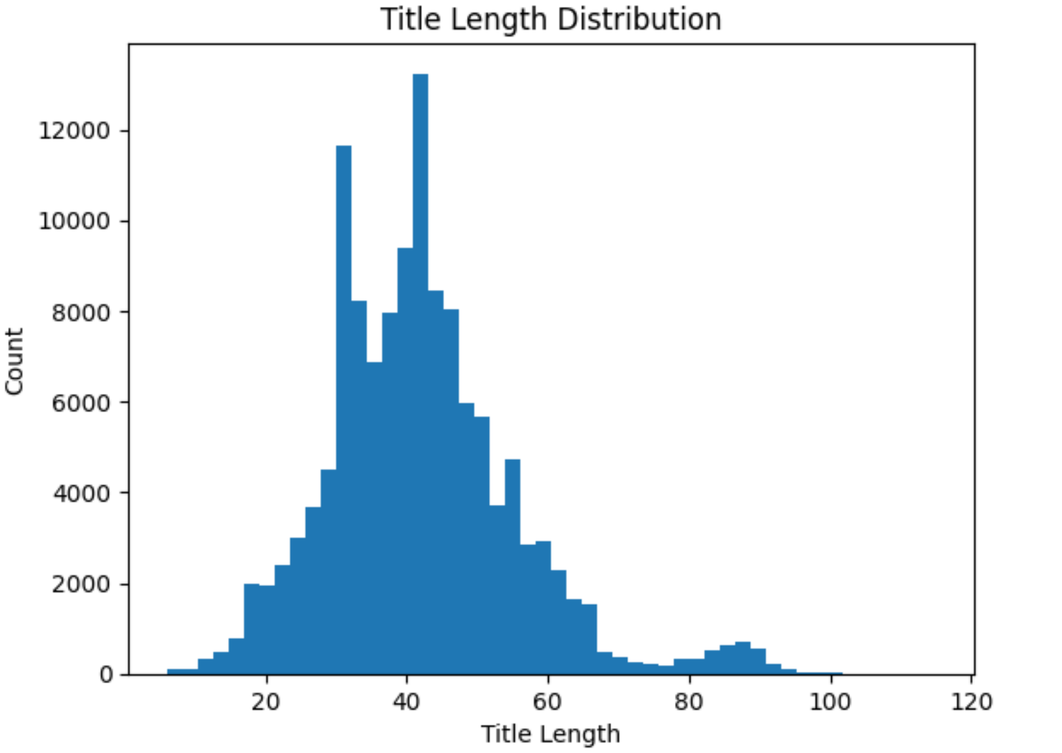
Bar chart showing class distribution in AG news dataset:



Bar chart showing class distribution in BBC dataset:

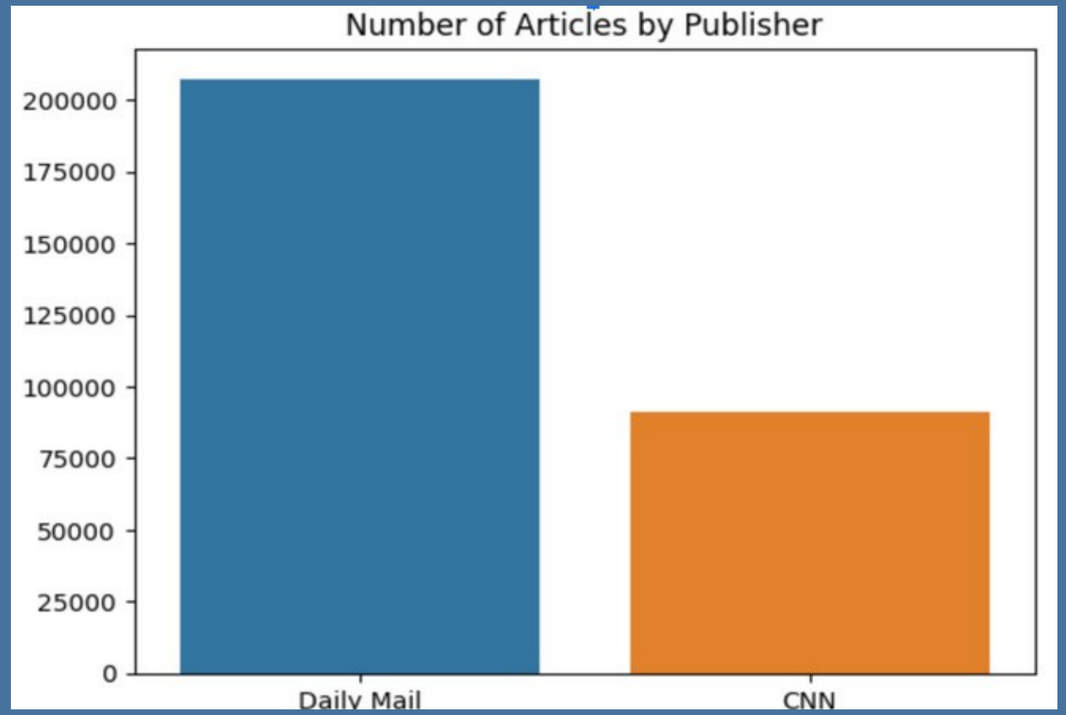




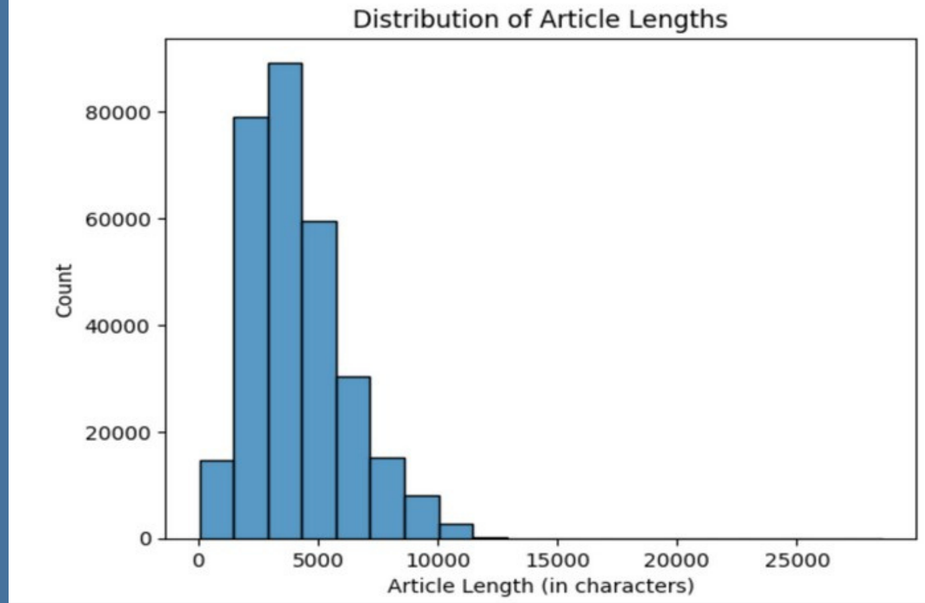


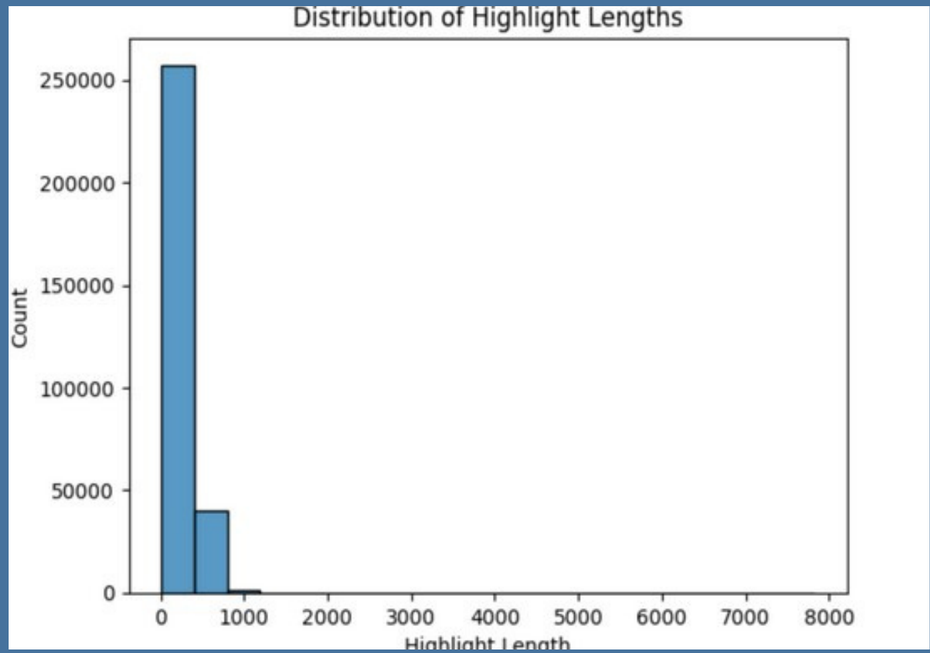
Data Statistics

Cnn_Daily mail dataset



Visualizing the distribution of Article Lengths





Data Transformaions

Cnn_daily Mail dataset

- combining the separated data
  - *Train set -(287113, 5)*
  - *test set - (11490, 5)*
  - *combined set - (298603, 5)*

- Transforming into Squad data set format

```
'>> Article: blyth spartans will learn their fa cup thirdround fate on monday ' and manager t
'>> Highlights: blyth spartans are through to the fa cup third roundnthe northern league side

'>> Article: published 0917 est 19 december 2013 updated 1014 est 19 december 2013 a priest s
'>> Highlights: priest from lxkken denmark says children should not be taught elves are linke
```

Data Transformaions

- Adding Prefic "Summarize" in fron t of Article for pre training T5 model.

summarize: ever noticed how plane s...	experts question if packed out planes are putt...
summarize: a drunk teenage boy had ...	drunk teenage boy climbed into lion enclosure ...
summarize: dougie freedman is on th...	nottingham forest are close to extending dougi...

- Tokenization using Hugging face Library.

Tokenized data:

```
[[443, 32, 109, 108, 1167, 739, 887, 43, 3, 13973, 21, 4038, 1075, 16, 3, 17, 208, 1506, 21, 12
[[443, 32, 747, 2275, 172, 29, 23, 4365, 23, 9193, 160, 15157, 1418, 19, 29, 17, 3, 24092, 12,
[[3, 7, 10405, 3, 8637, 49, 9668, 141, 118, 671, 8873, 53, 25039, 481, 437, 3105, 29, 354, 1316
[[1157, 6123, 6098, 1666, 7, 872, 629, 581, 8, 1576, 13, 251, 30, 3, 4915, 1413, 3822, 257, 209
[[29733, 2615, 8838, 12, 11, 49, 40, 12041, 16, 3, 76, 15, 89, 9, 4192, 5533, 29, 6983, 1846, 4
[[396, 4517, 12346, 7250, 16, 689, 28, 1100, 1614, 3055, 30, 1687, 10736, 29, 4998, 867, 3, 937
```

Data Transformaions

- *Daatset Split:*
  - *Train set - (90444)*
  - *Validation set (19381)*
  - *Test set: (19382)*
- *Combined Dataset:129207*
- *Stop Words Removed*
- *Tokenization*

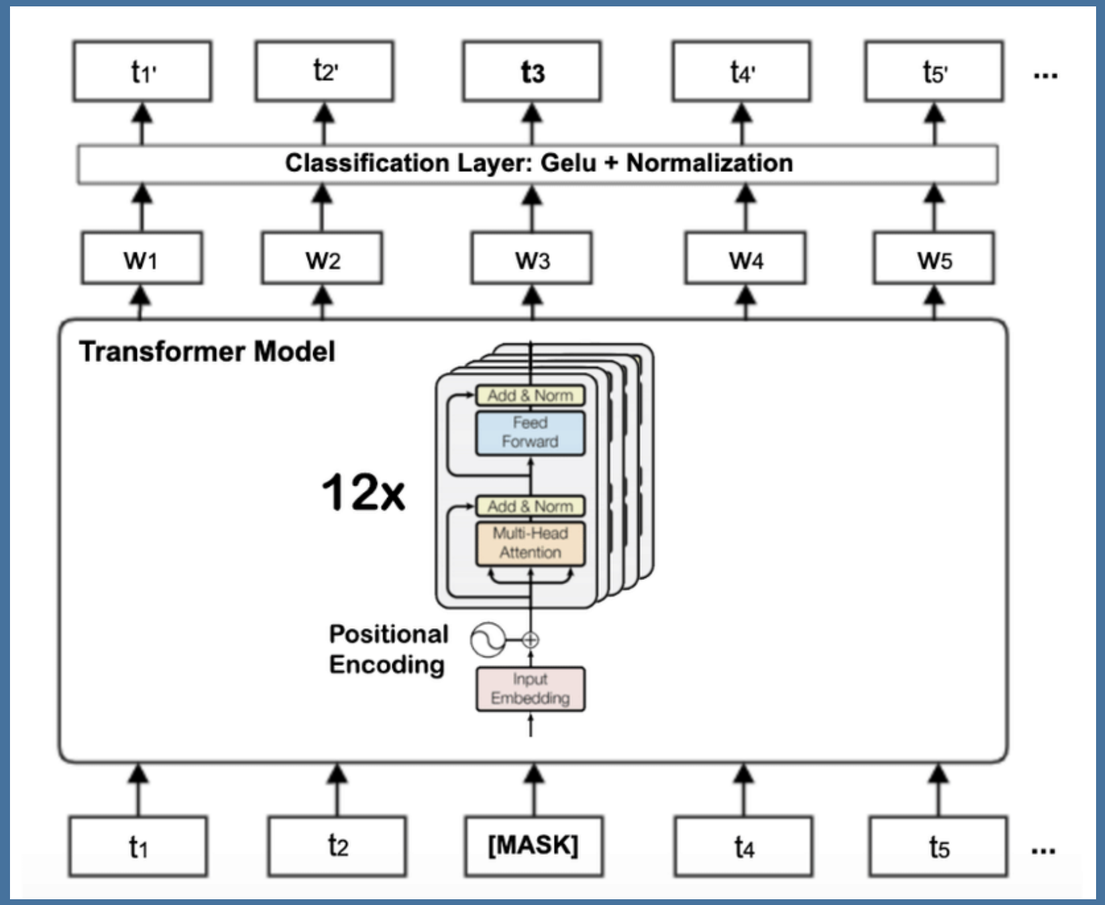
Tokenized classification dataset.

	Description
0	[101, 12174, 11372, 2012, 2149, 2865, 5016, 20...
1	[101, 1996, 5608, 1997, 7861, 14479, 14782, 28...
2	[101, 2329, 13095, 2038, 11248, 2152, 4762, 75...
3	[101, 6661, 1999, 2866, 8974, 1998, 2833, 3813...
4	[101, 2900, 1005, 1055, 4610, 17170, 14050, 20...
5	[101, 2634, 102]
6	[101, 2250, 5467, 2040, 2024, 4039, 2000, 2604...

Project machine learning models

BERT & RoBERTa

- Transformer Architecture: BERT is based on the
- Encoder-Decoder Structure: Only uses encoder block.
- Coming to Text classification is used to predict the words based on their contextual meaning.
- Uses mlm and nsp which is helpful for news article classification.
- Supports transfer learning. Attention blocks help in focusing on important words and relationships which is important for classification.
- As Bert and Roberta are both trained on a large corpus of data they tend to learn the general language patterns well and also the relationships between them which helps in better classification.
- Roberta uses a dynamic masking technique
- Roberta only uses masked language modeling and removes the next sentence prediction unlike Bert.



GPT2 Architecture

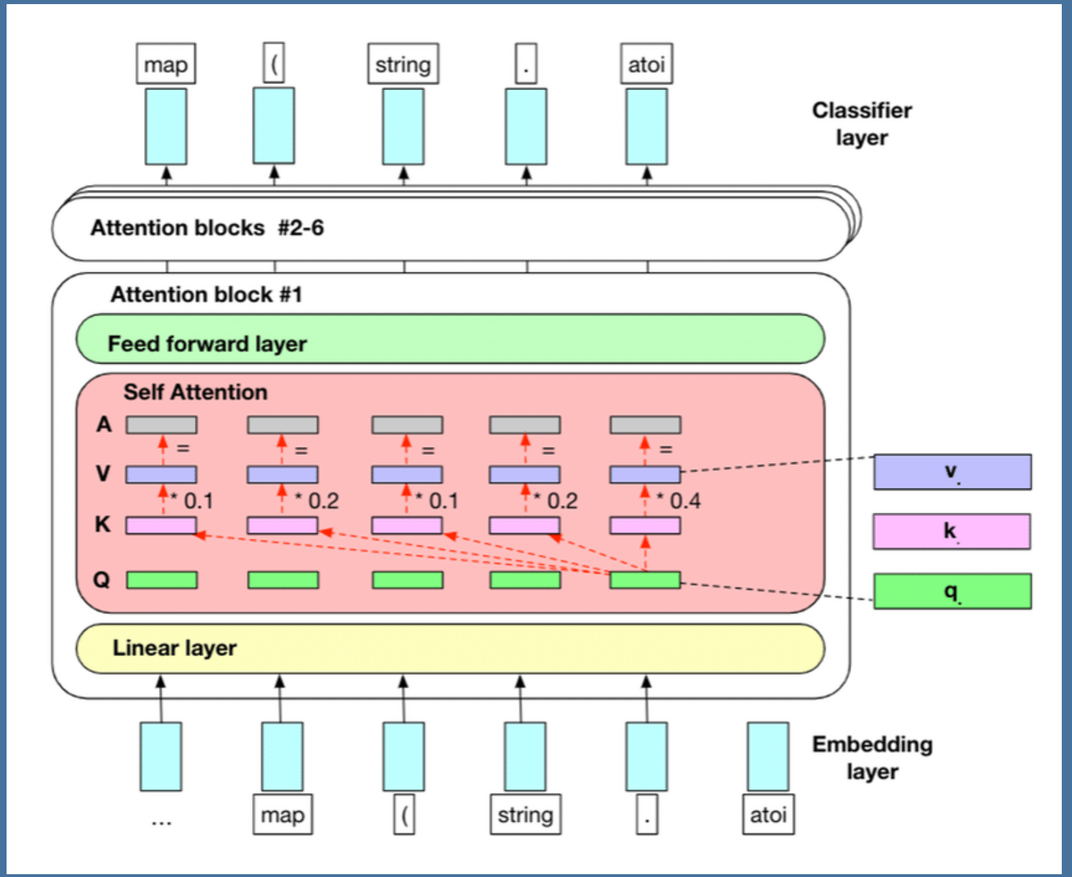
Transformer Architecture: Processes and understands text at various levels (words, sentences, paragraphs), capturing hierarchical relationships.

Encoder-Decoder Structure: Learns input text representations and generates coherent summaries or classifications.

Attention Mechanism: Focuses on relevant text parts, capturing dependencies and relationships for accurate summaries.

Self-Attention: Weighs word/sentence importance based on relevance, capturing long-range dependencies and context.

Multi-Head Attention: Captures different aspects of text simultaneously, enhancing understanding and summarization.



GPT2

Pros:

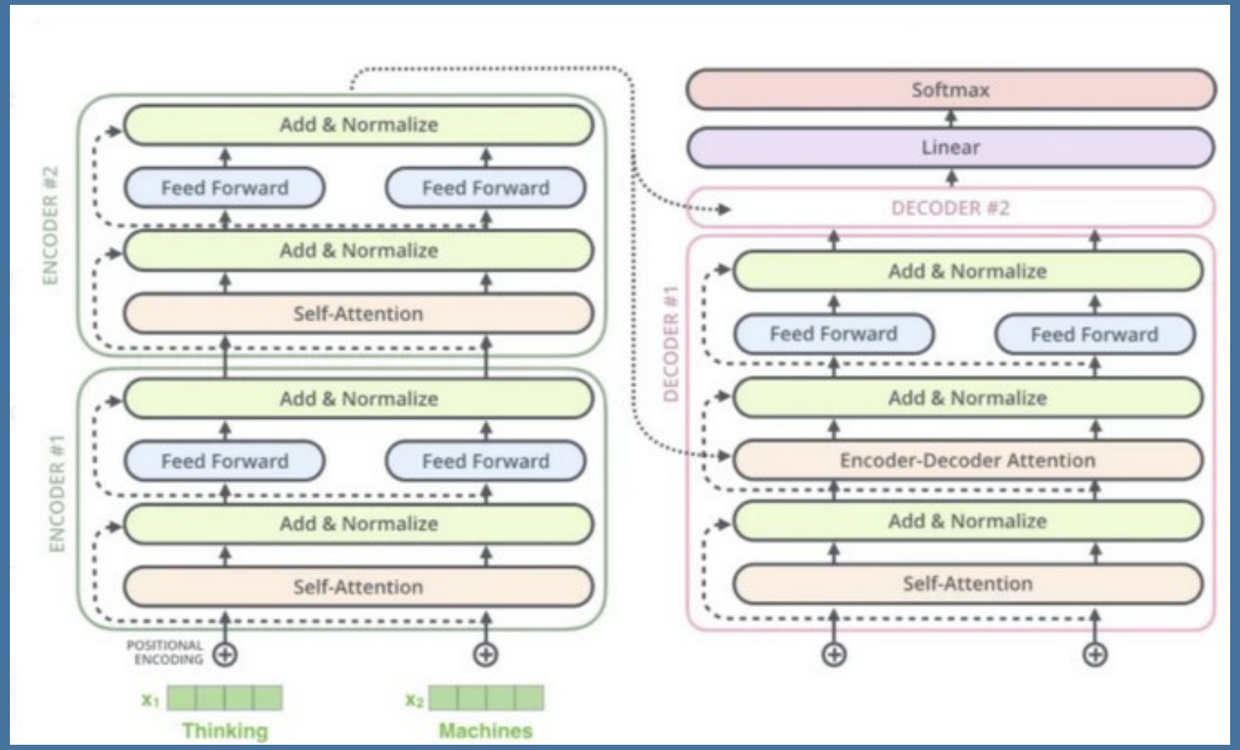
- Extensive diverse training data.
- Coherent and contextually relevant text generation.
- Remarkable performance in summarization and classification.
- Fine-tuning for news articles.
- Captures hierarchical relationships for effective understanding and summarization.
- Generates high-quality informative summaries.

Cons:

- Summaries can be redundant or lack consciousness
- May occasionally produce summaries that are too verbose
- Large model size High computational requirements

T5:

- T5 is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks and proposed in the year 2019.
- **Pros:**
  - T5 is trained on a large volume of text including news articles , intensive pre-training aids in understanding the content's semantics and syntactic structure.
  - Transfer learning capabilities.
  - versatile
- **Cons:**
  - Computationally expensive
  - Interpretability



Prototype


Enter Text:
Sally Forrest, an actress-dancer who graced the silver screen throughout the '40s and '50s in M GM musicals and films such as the 1956 noir While the City Sleeps died on March 15 at her home in Beverly Hills, California. Forrest, whose birth name was Katherine Feeney, was 86 and had lo ng battled cancer. Her publicist, Judith Goffin, announced the news Thursday. Scroll down for v ideo. Actress: Sally Forrest was in the 1951 Ida Lupino-directed film 'Hard, Fast and Beautiful ' (left) and the 1956 Fritz Lang movie 'While the City Sleeps' A San Diego native, Forrest beca me a protege of Hollywood trailblazer Ida Lupino, who cast her in starring roles in films inclu ding the critical and commercial success Not Wanted, Never Fear and Hard, Fast and Beautiful. S ome of Forrest's other film credits included Bannerline, Son of Sinbad, and Excuse My Dust, acc ording to her iMDB\xc2\xa0page. The page also indicates Forrest was in multiple Climax! and Raw hide television episodes. Forrest appeared as herself in an episode of The Ed Sullivan Show and three episodes of The Dinah Shore Chevy Show, her iMDB page says. She also starred in a Broadw ay production of The Seven Year Itch. City News Service reported that other stage credits inclu ded As You Like It, No, No, Nanette and Damn Yankees. Forrest married writer-producer Milo Fran k in 1951. He died in 2004. She is survived by her niece, Sharon Durham, and nephews, Michael a nd Mark Feeney. Career: A San Diego native, Forrest became a protege of Hollywood trailblazer I da Lupino, who cast her in starring roles in films
Process
Summary:
Sally Forrest, an actress-dancer who graced the silver screen throughout the '40s and '50s in M GM musicals and films died on March 15.Forrest, whose birth name was Katherine Feeney, had long battled cancer. San Diego native, Forrest became a protege of Hollywood trailblazer Ida Lupino , who cast her in starring roles in films.
Category:
Entertainment

Model Evaluation

For evaluating the performance of the models the following metrics will be used:

Task	Evaluation Methods
Classification	Accuracy, Precision, Recall, Roc and Auc curve
Summarization	Rouge 1, Rouge2, Rouge L and Rouge N

Future Work

- Train and Develop the proposed models for summarization and categorization.
- Integrate the best model with UI.
- Deploy the web application
- Faster computing, Facilitate more compute units.

Thank You