News Article Analysis using NLP

ALL YOUR NEWSLETTERS IN ONE PLACE

2023

Team 6
Himanshee
Karnik Ketan Kalani
Lakshmi Satvika Nekkanti
Saumya Sinha

PROJECT MOTIVATION AND OBJECTIVES

Motivation:

- The surge in available news and online text data necessitates efficient processing.
- LLMs offer a solution for automating summarization, classification, and NER in news articles.

Background:

- LLMs, with their capabilities in text generation and summarization, pave the way for innovative NLP applications.

PROJECT MOTIVATION AND OBJECTIVES

Objectives:

- Utilize LLMs for text summarization, classification, and NER.
- Fine-tune models on custom datasets and evaluate performance.
- Develop a real-time news analysis application for summarizing and categorizing articles.

Deliverables:

- Final Report
- Presentation slides
- Fine-tuned models
- Fully functional Web Application

LITERATURE SURVEY AND TECHNOLOGY SURVEY

Sr No.	Title	Authors	Summary	Challenges and Results
1)	Abstractive Text Summarization Using T5 Architecture	Ramesh, G. S., Manyam, N. V., Mandula, N. V., Myana, N. P., Macha, N. S., & Reddy, N. G. S	Explore NLP-based text summarization using the T5 transformer model, employing a text-to-text approach with an encoder-decoder structure to automatically extract relevant information while preserving the essence of content.	Compares T5 with an attention-based seq-to-seq approach, finding that T5-based summarization outperforms the attention-based approach in tests, supported by evaluation metrics including recall, precision, F-1 score, and Rouge.
2)	Financial News Analytics Using Fine-Tuned Llama 2 GPT Model	Bohdan M. Pavlyshenko	Unsupervised extractive and abstractive summarization methods are used in the study on the COVID-19 Dataset	Main challenges cost and time in fine-tuning large models, paper’s approach demonstrates the effectiveness of LLAMA-2 in generating accurate and sentiment-rich summaries by comparing the fine-tuned model's results with existing dataset summaries.
3)	Pre Training with Extractive Gap Sentences for Abstractive Summarization	Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu	The paper introduces Pegasus, a transformer-based encoder-decoder model, pre-trained on extensive text corpora for abstractive summarization, focusing on generating sentences from documents, achieving state-of-the-art performance across 12 summarization tasks.	Main challenges include optimizing gap sentence ratios, MLM effects, and vocabulary impact. PegasusBase with 223M parameters showed efficiency in pre-training, paving the way for scaling Pegasus Large model, which exhibited superior performance.
4)	How to Fine-Tune BERT for Text Classification?	Chi Sun, Xipeng Qiu, Yige Xu, Xuanjing Huang	Compared BERT to conventional machine learning models	BERT has outperformed numerous ML models including Logistic Regression , Linear SVC , Multinomial NB , Ridge Classifier , and Passive Aggressive Classifier in different areas of text classification.

LITERATURE SURVEY AND TECHNOLOGY SURVEY

Sr No.	Title	Authors	Summary	Challenges and Results
5)	A Survey on Text Classification Algorithms: From Text to Predictions	Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A.	Delves into sentiment analysis , text classification , topic labelling, news classification, question and answering tasks and compares diffrenert models	Challenges were evaluating diverse NLP tasks by employing precision, recall, F1 score, etc. In the context of text classification, Bert wmerge as top performers, showcasing the significance of attention models in enhancing seq-to-seq architectures.
6)	RoBERTa: A Robustly Optimized BERT Pretraining Approach	Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov	The RoBERTa model, an evolution of BERT with modifications such as extended training data and longer pre-training, achieves superior performance on diverse nlp benchmarks, including GLUE, RACE, and SQuAD.	The challenges addressed in enhancing RoBERTa's performance involved optimizing pre-training periods and batch sizes, resulting in a notable score of 4/9 on GLUE tasks and an impressive 88.5 on the public GLUE.
7)	A Named Entity Recognition Method For Chinese Winter Sports News Based On RoBERTa-WWM,	P. Liu, Y. Cao	Developed a named entity recognition model based on Roberta, enhancing winter sports terminology precision through WWM pre-training and achieving notable performance metrics.	Challenges include incorporating specific winter sports terms, and the results showcase the superiority of the Roberta-WWM-BiLSTM-CRF model with Precision 0.9134, Recall 0.9428, and F1 score 0.9279 over BERT-BiLSTM-CRF and BiLSTM-CRF models.
8)	CoVShorts: News Summarization Application Based on Deep NLP Transformers for SARS-CoV-2	Hunar Batra; Akansha Jain; Gargi Bisht; Khushi Srivastava; Meenakshi Bharadwaj; Deepali Bajaj; Urmil Bharti	Employed diverse models, including BERT, GPT-2, GPT, XLNET, BART, and T5, on a vast COVID-19 public media dataset comprising over 350,000 data points, distinguishing between extractive models like GPT-2, BERT, and XL Net, and abstractive models such as T5 and BART.	GPT-2 outperforming XL Net among, exhibiting higher ROUGE-2 and ROUGE-L scores of 0.354 and 0.364 compared to 0.343 and 0.352, respectively. Additionally, BART demonstrated superior performance over T5.

DATA COLLECTION AND PRE-PROCESSING

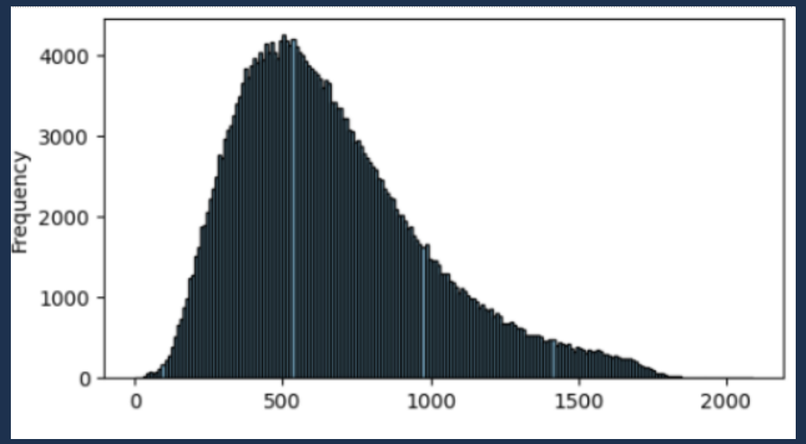
CNN_Dailymail Dataset

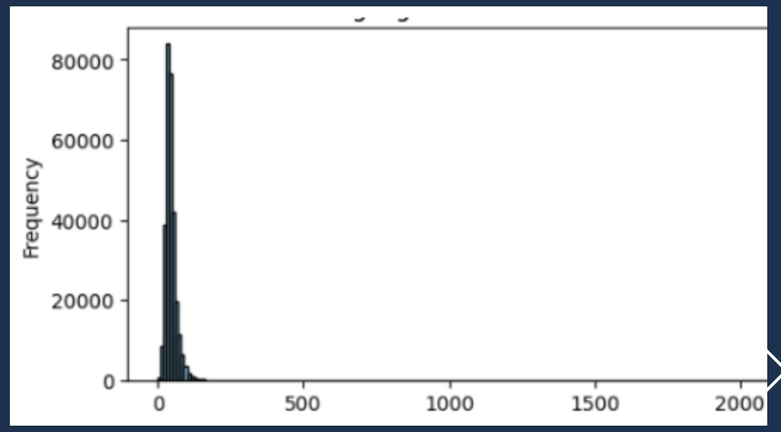
- Source: TensorFlow
- Size: 1.29 GB (298603 rows)
- Features:
  - articles
  - highlights
  - publishers
  - Id
  - unnamed

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'

Word frequency of Article and Highlights





DATA PREPARATION OF CNN-DAILY MAIL

Before and after Removing Duplicates

Before removing duplicates:

```
article       298603
highlights    298603
dtype: int64
```

After removing duplicates:

```
article       295503
highlights    295503
dtype: int64
```

Tokenized Data

Tokenized data:

```
[[443, 32, 109, 108, 1167, 739, 887, 43, 3, 13973, 21, 4038, 1075, 16, 3
[[443, 32, 747, 2275, 172, 29, 23, 4365, 23, 9193, 160, 15157, 1418, 19,
[[3, 7, 10405, 3, 8637, 49, 9668, 141, 118, 671, 8873, 53, 25039, 481, 4
[[1157, 6123, 6098, 1666, 7, 872, 629, 581, 8, 1576, 13, 251, 30, 3, 491
[[29733, 2615, 8838, 12, 11, 49, 40, 12041, 16, 3, 76, 15, 89, 9, 4192,
[[396, 4517, 12346, 7250, 16, 689, 28, 1100, 1614, 3055, 30, 1687, 10736
```

Train, Test and Val split of Dataset- split ratio:70/15/15

```
atasetDict({
    train: Dataset({
        features: ['article', 'highlights', '__index_level_0__']
        num_rows: 206852
    })
    val: Dataset({
        features: ['article', 'highlights', '__index_level_0__']
        num_rows: 44326
    })
    test: Dataset({
        features: ['article', 'highlights', '__index_level_0__']
        num_rows: 44325
    })
```

Data Pre-Processing and preparation steps:

- Removing duplicates
- handling missing values
- Lower casing
- Removing punctuations and numbers
- Removing Unicode character such as \xe2\x80\x99, \xc2\xa0, b, etc.
- Features selection
- Tokenization
- Adding prefix-”summarize”

DATA COLLECTION AND PRE-PROCESSING CLASSIFICATION

AG news and BBC news

- AG news:
  - Source: Hugging Face
  - 127600 rows
  - Belonging Categories
    - World - 0
    - Sports -1
    - Business - 2
    - Science/Tech -3
- BBC News:
  - Source : Kaggle
  - Initially 2225 rows
  - Belonging Categories:
    - World-0
    - Sports-1
    - Business-2
    - Science/Tech-3
    - Entertainment-4

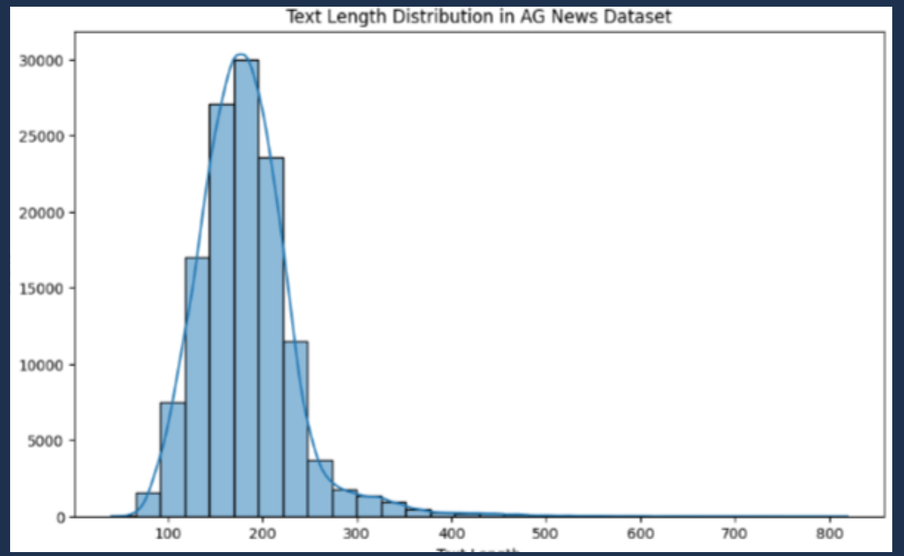
Raw Dataset of Ag news

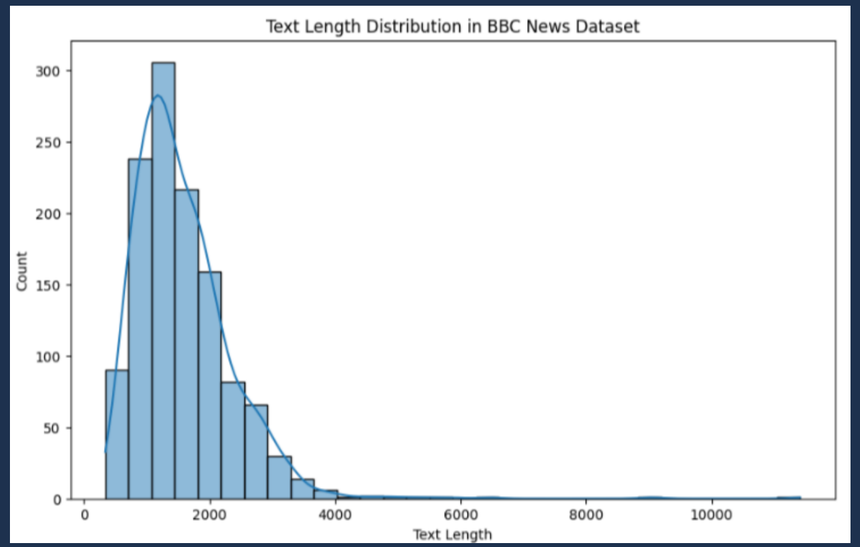
```
Combined CSV file saved successfully.
```

	text	label
127595	Terror is below radar in Russia It took 2 days...	0
127596	Air Canada confirms order for 45 Embraer jets ...	2
127597	White House Shifts Its Focus on Climate The ad...	3
127598	Netflix stock plummets on buzz about Amazon.co...	2
127599	Phillies to interview Russell for vacant manag...	1

Raw Dataset of BBC News

	Unnamed: 0	Description	Class Inde
0	NaN	worldcom ex-boss launches defence lawyers defe...	
1	NaN	german 2 confidence slides german 2 confidence...	
2	NaN	bbc poll indicates economic gloom citizens in ...	
3	NaN	lifestyle governs mobile choice faster bett...	
4	NaN	enron bosses in $168m payout eighteen former e...	



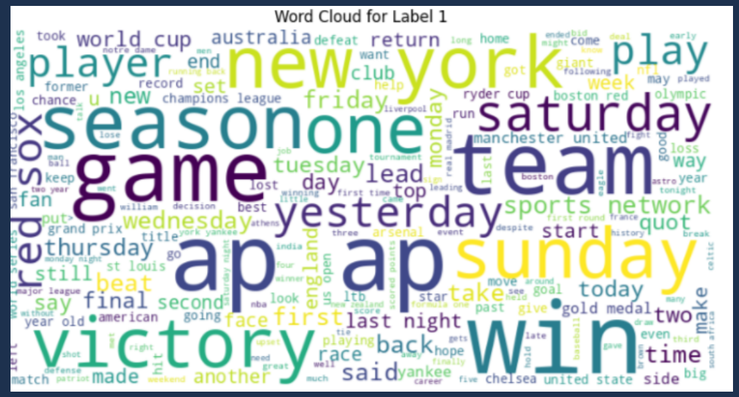


DATA PREPARATION OF AG-NEWS AND BBC NEWS

BBC news

AG news

After preprocessinng and concatenation



	text	label	text_length
0	amd debuts dual core opteron processor amd new...	3	174
1	woods suspension upheld reuters reuters major ...	1	188
2	bush reform may blue states seeing red preside...	2	222
3	halt science decline schools britain run leadi...	3	124
4	gerrard leaves practice london england sports ...	1	185

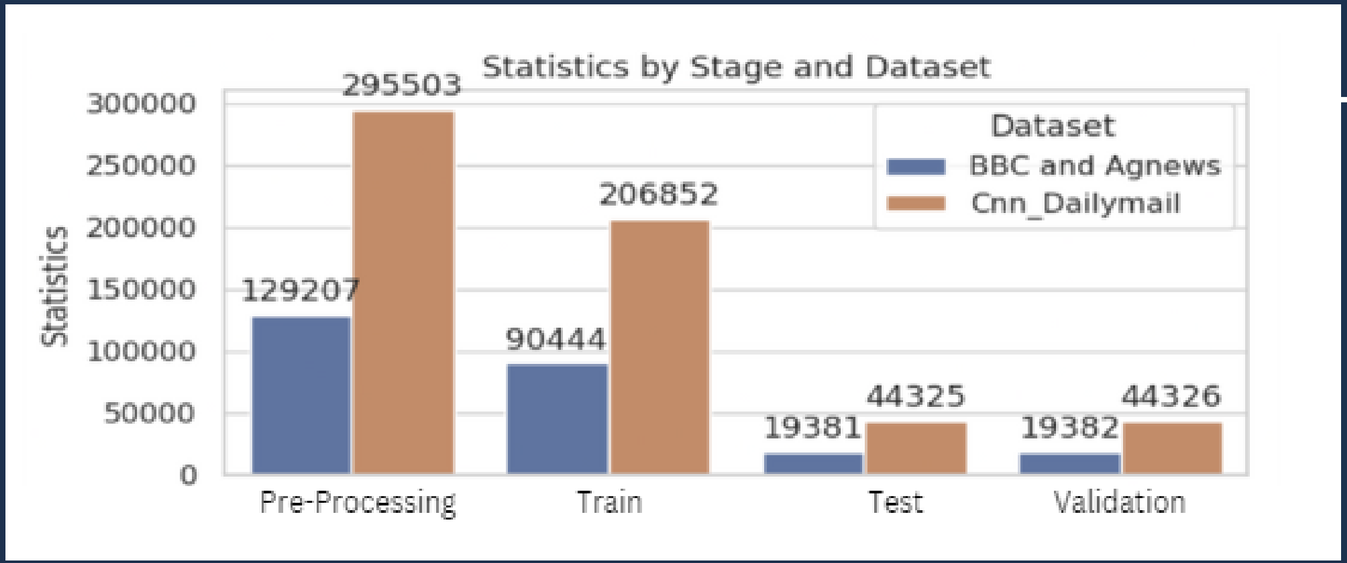
Train ,Test val split was done in *70%-15%-15%*

```
Training data shape: (90171, 2)
Validation data shape: (19323, 2)
Testing data shape: (19323, 2)
```

Data Pre-processing and Preparation steps:

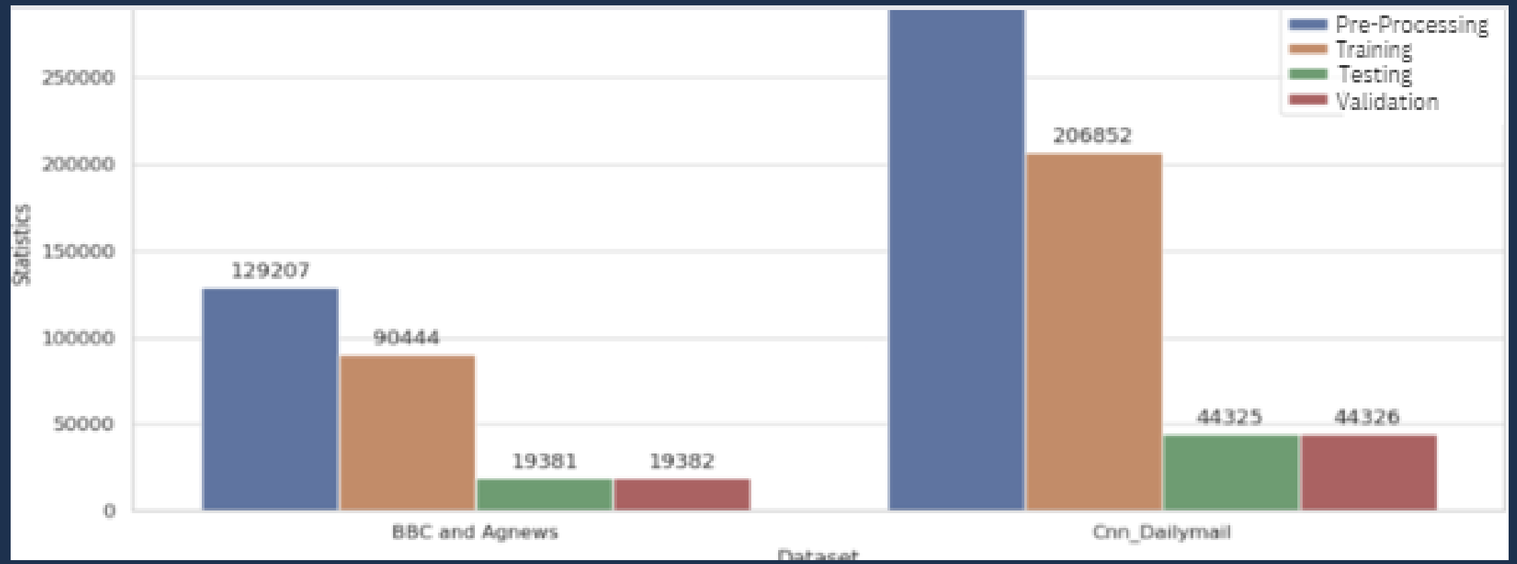
- Checking for duplicates and missing values
- Removing unnecessary columns
- Text was converted to lower case
- Extra whitespaces were removed
- Stop words were removed
- Html tags, special characters,Punctuations were removed.
- Changing the column names from “Description” and “class index” to “Text” and “label”

DATA ANALYSIS

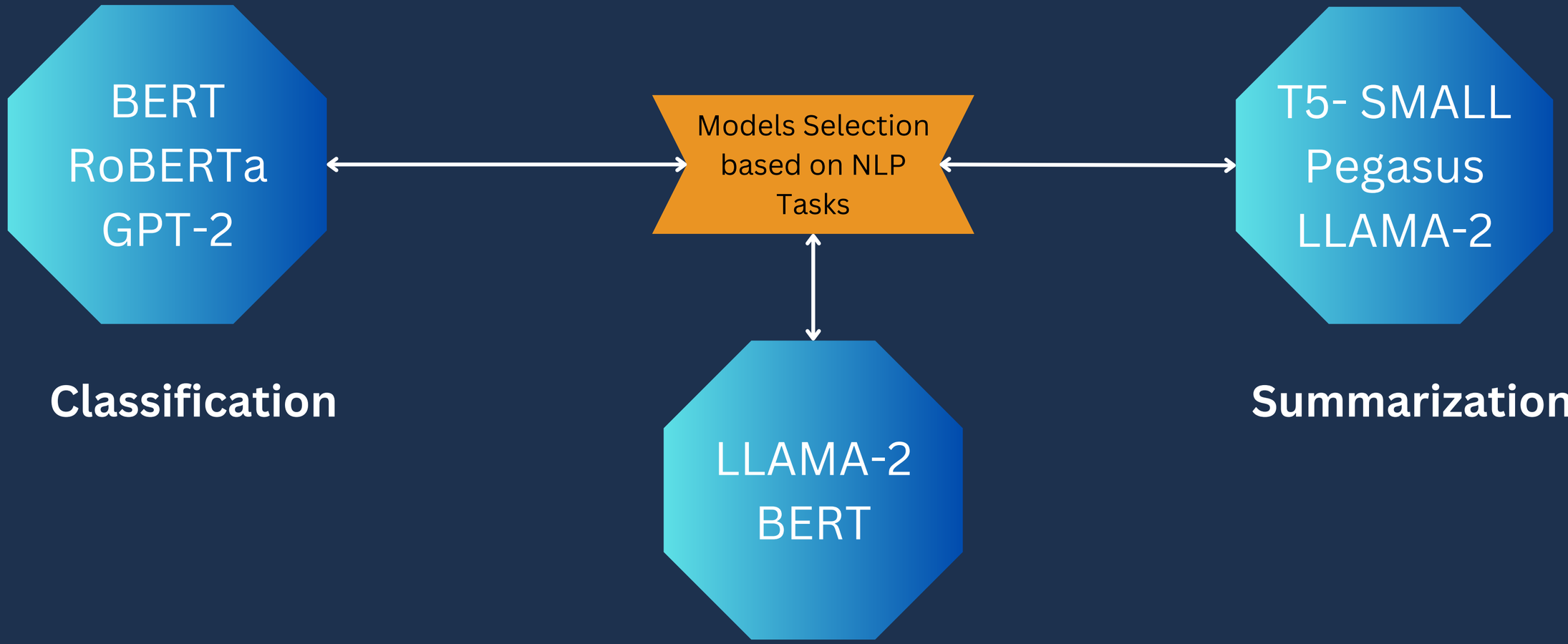


Row Count Over the Stages by Dataset

Dataset Statistics Over the Stages



MODEL SELECTION



MODEL COMPARISION

Model	Size	Source	Why?	Pre training approach
LLAMA-2	7 B	Meta	Competitor of GPT, Open Source	Diverse mix of data
PEGASUS	223 M	Google AI	Specifically designed for abstractive summarization	Pre-trained on C4 and HugeNews corpus
T5	60 M	Google	Unified Text-to-Text Framework, more range of NLP tasks	C4 (Colossal CleanCrawledCorpus)
GPT-2	124 M	Open AI	GPT series revolutionized NLP Human like text processing capabilities	Web Text dataset
BERT	110 M	Google	Bi-directional approach hence good at capturing context	Book corpus,English Wikipedia
ROBERTA	125 M	Meta	Improved version of BERT	Book corpus 11k unpublished books, wikipedia, cc news,Openwebtext,Stories

MODEL EVALUATION

Named Entity Recognition (NER)

Hyperparameters:

- Learning Rate =1e-4,
- Epochs=2
- Batch_size=8
- Max input length =1024
- Max target length = 128
- Optimizer= AdamW

MODELS	RECALL	PRECISION	F1
LLAMA-2 (Baseline)	0.6923	0.6284	0.7180
LLAMA-2(Fine-tuned)	0.7643	0.7582	0.7992
BERT (Baseline)	0.6923	0.6284	0.7180
BERT (Fine-Tuned)	0.7762	0.8236	0.7492

MODEL EVALUATION OF CLASSIFICATION

Models	Accuracy	F1-score	Precision	Recall
Bert Baseline	0.25	-	-	-
Bert Fine-Tuned	0.93	per class: [0.9319627 0.97993411 0.8934665 0.90387225]	per class: [0.95338248 0.96881169 0.88517831 0.90321235]	per class: [0.91148425 0.99131489 0.90191136 0.90453311]
RoBERTa Baseline	0.25	-	-	-
RoBERTa Fine-Tuned	0.92	F1-score per class: [0.92005103 0.97525938 0.88921192 0.89848722]	Precision per class: [0.94434745 0.96353243 0.87926078 0.89745727]	per class: [0.89697347 0.9872753 0.89939088 0.89951953]
Gpt-2 Baseline	0.23	-	-	-
Gpt-2 Fine-Tuned	0.92	per class: [0.9319627 0.95993411 0.8854665 0.88387225	per class: [0.93434745 0.96353243 0.87926078 0.89745727]	per class: [0.92148425 0.97131489 0.90191136 0.91453311]

MODEL EVALUATION OF CLASSIFICATION

Hyperparameters:

- Learning Rate =1e-5,beta_1=0.9, beta_2=0.999, epsilon=1e-07
- Epochs=3
- Batch_size=8
- Max input length =512
- Optimizer= AdamW

Input article

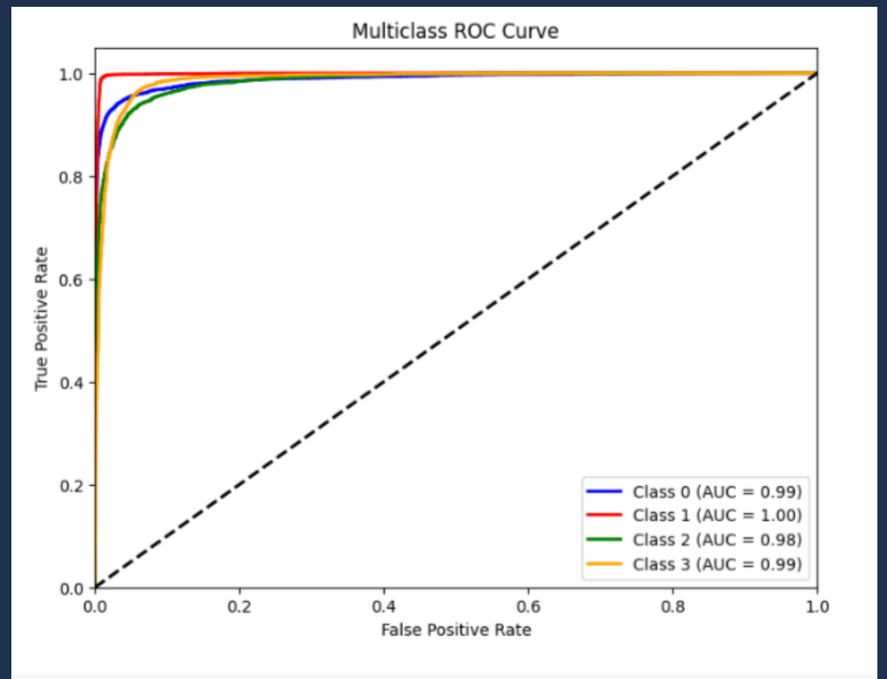
```
('Last night, the underdog team staged a stunning comeback,
 scoring three goals in the final minutes to secure a thrilling
 victory over the reigning champions in a nail-biting soccer match.')
```

Generated Output

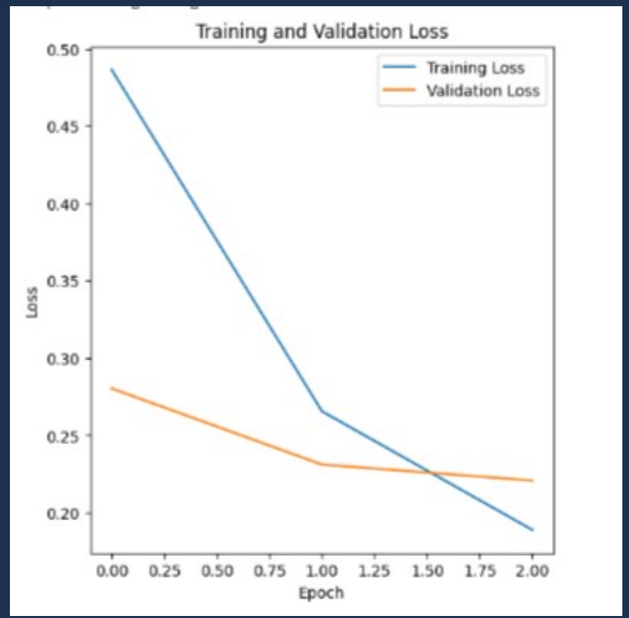
```
np.argmax(probs[0])
```

```
1
```

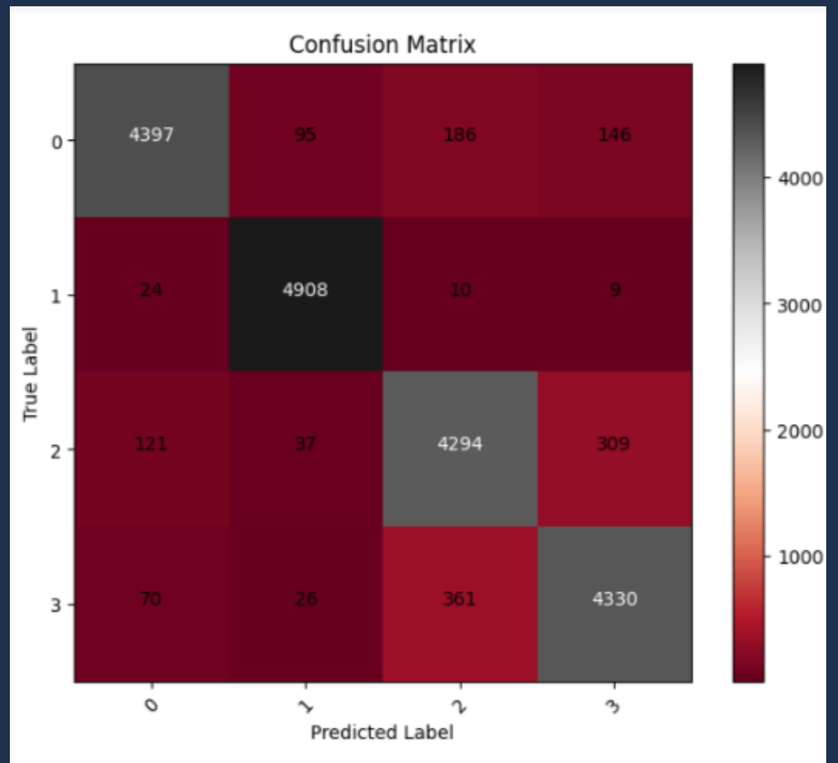
Roc -Auc Curve of Bert



Train vs val loss of Bert

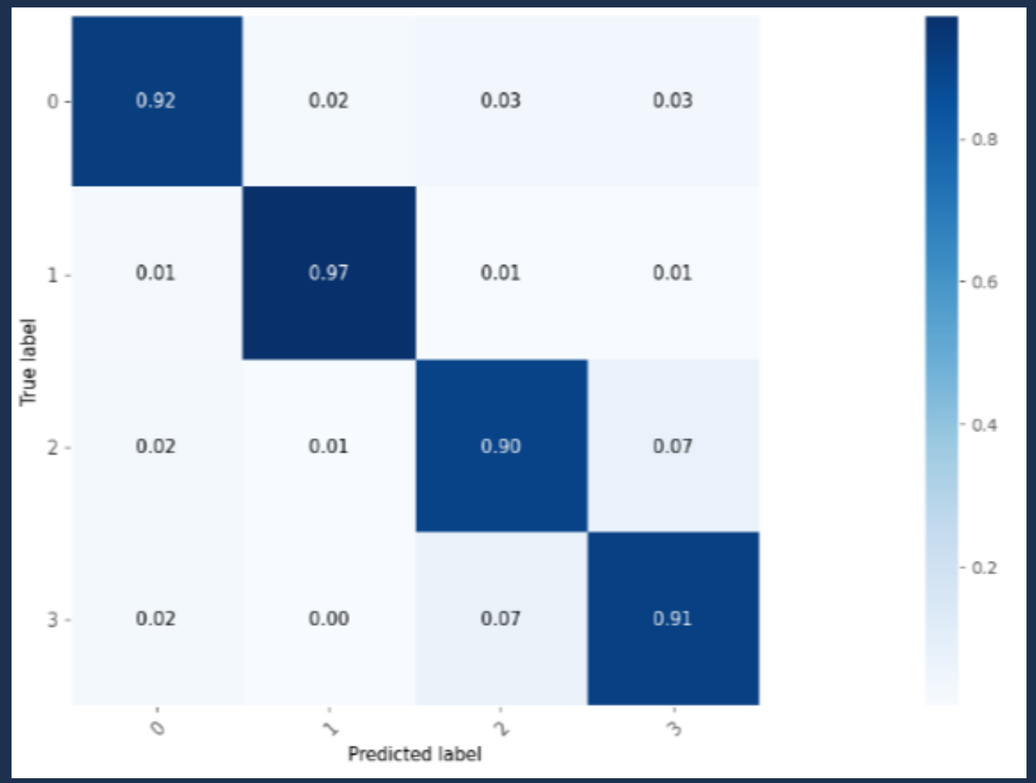


Confusion matrix of Bert

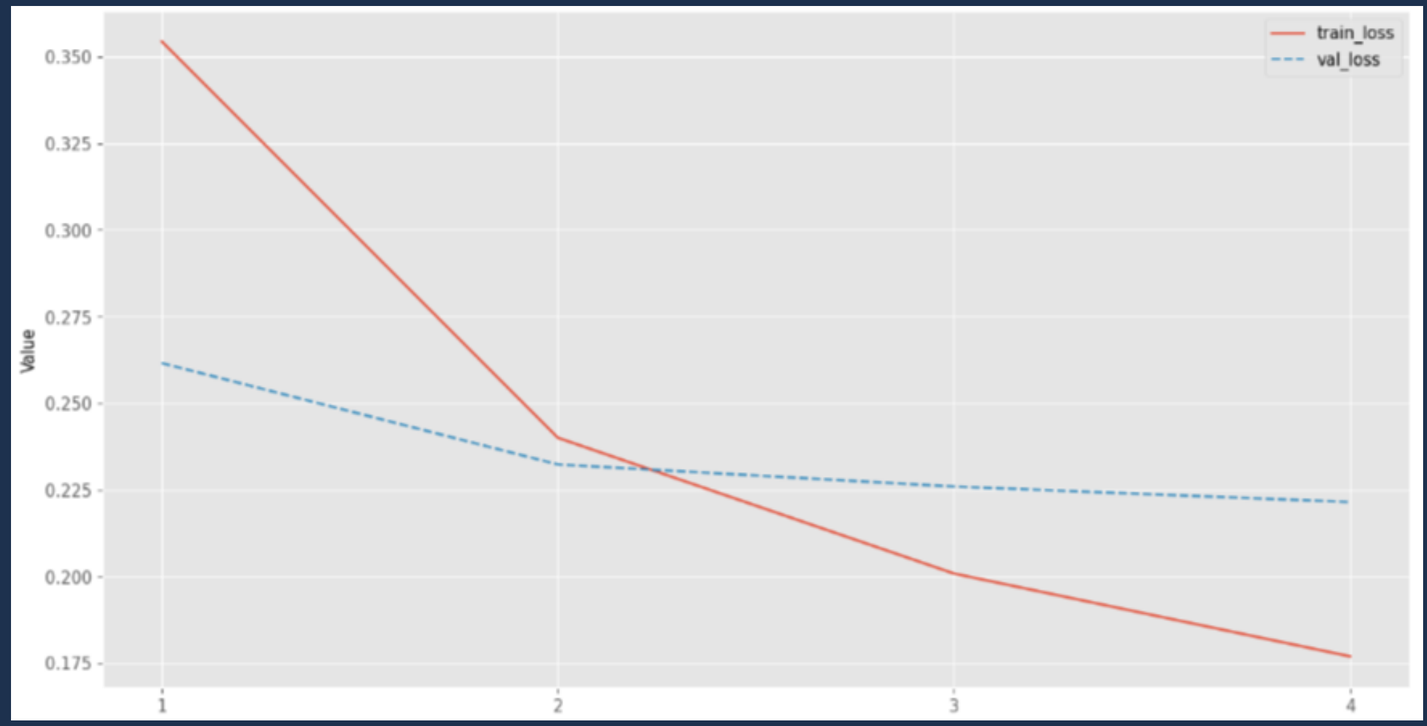


MODEL EVALUATION OF CLASSIFICATION

Confusion Matrix of GPT-2

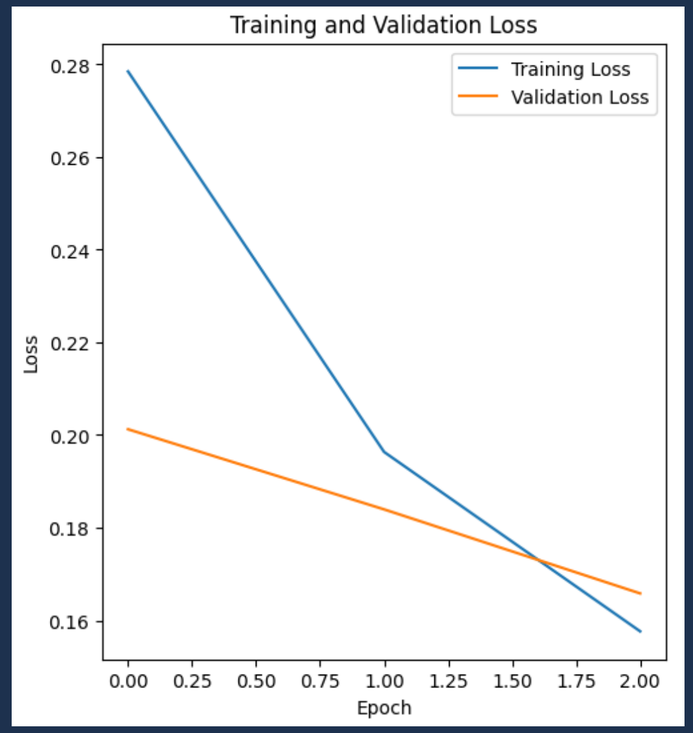


Training vs Val loss of GPT-2

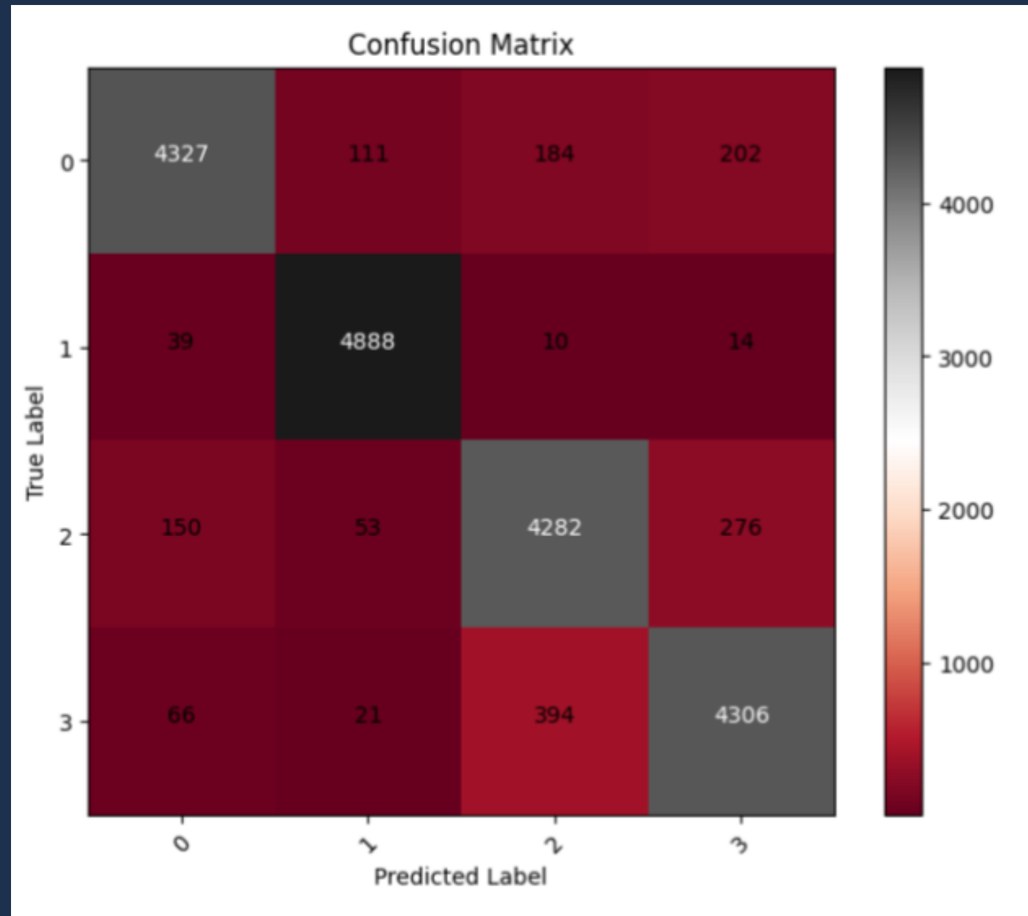


MODEL EVALUATION OF CLASSIFICATION

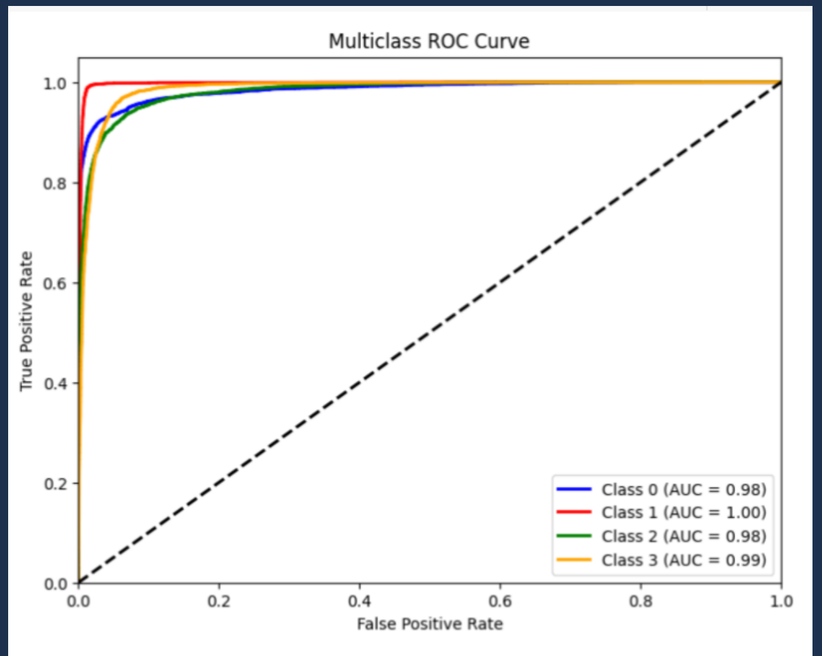
Train vs val loss of RoBERTa



Confusion Matrix of RoBERTa



Roc-Auc curve for RoBERTa



MODEL EVALUATION

Summarization Task

Hyperparameters:

- Learning Rate =3e-5
- Epochs=3
- Batch_size=8
- Max input length =1024
- Max target length = 128
- Optimizer= AdamW

MODELS	ROUGE - L	ROUGE - 1	ROUGE - 2
T5-small (baseline)	0.16	0.21	0.07
T5-small (finetuned)	0.21	0.30	0.12
Pegasus (baseline)	0.21	0.29	0.13
Pegasus (finetuned)	0.27	0.36	0.16
LLAMA-2 (baseline)	0.24	0.28	0.14
LLAMA - 2 (finetuned)	0.26	0.25	0.20

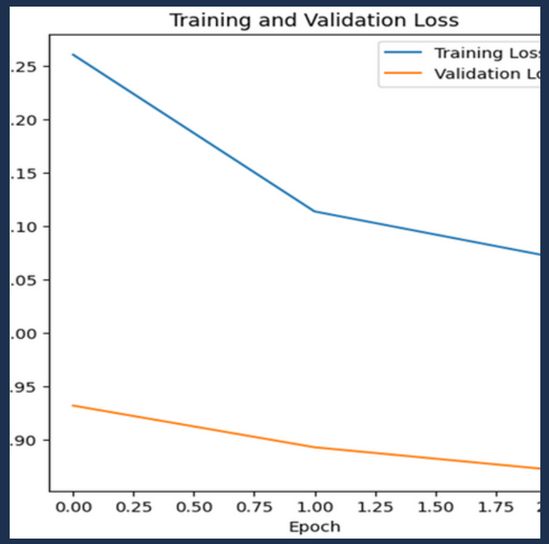
Optimised Result

Learning_rate=5e-5 , Epochs=5, batch_size=16, warmup_steps=5000

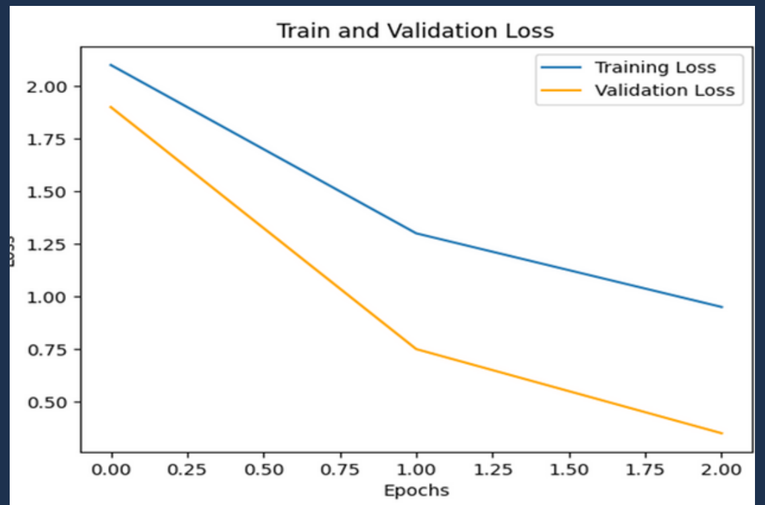
Pegasus	**0.30**	**0.41**	**0.18**

MODEL EVALUATION

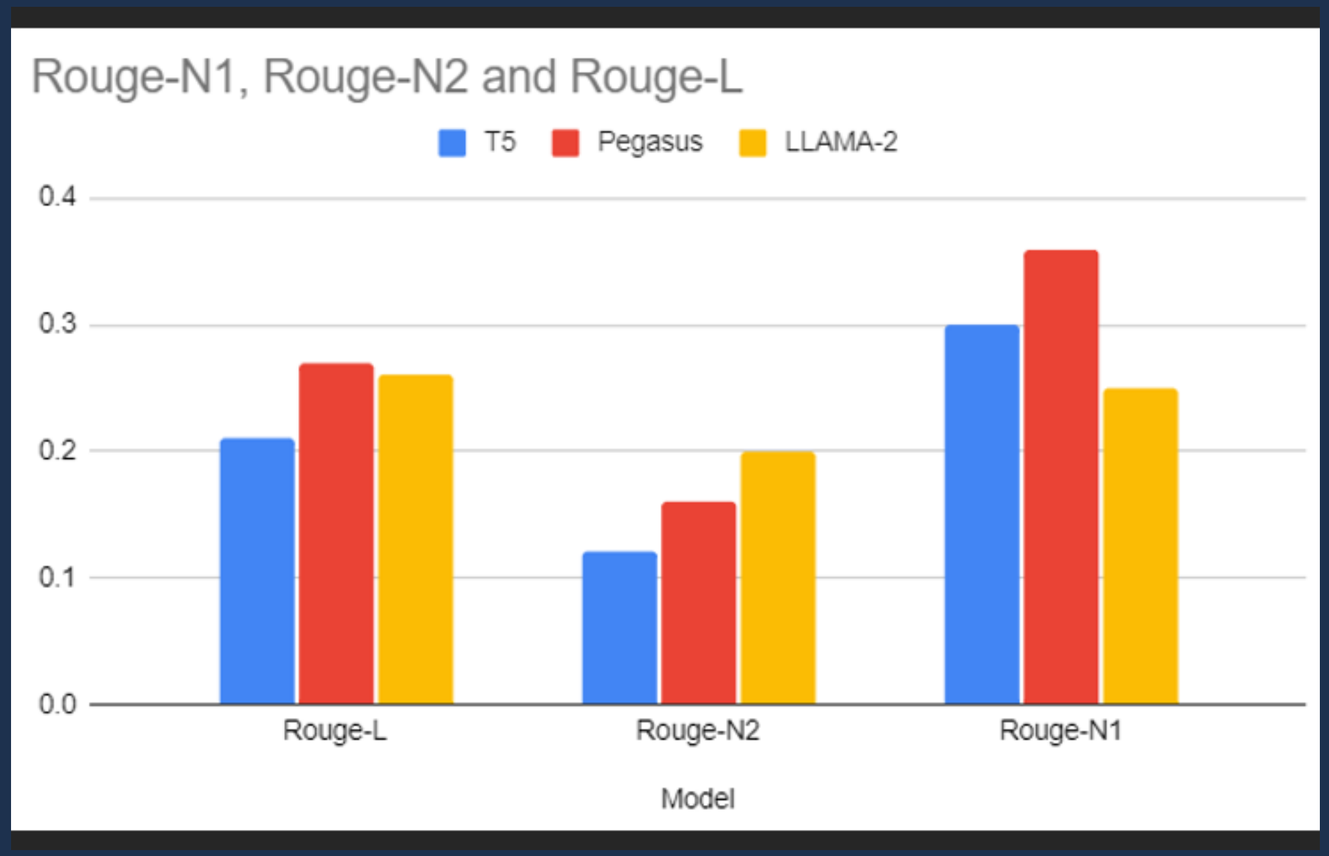
Pegasus: Training and Validation Loss



LLAMA-2 :Training and Validation Loss



Comparison between Summarization Models based on Rouge score



Article Sample

ahead of this weekends premier league action sportsmail will be providing you with all you need to know about every fixtu
with team news provisional squads betting odds and opta stats here is all the information you need for manchester uniteds
home clash with aston villa manchester united vs aston villa old trafford team news manchester united striker robin van
persie is out of manchester uniteds barclays premier league match against aston villa with an ankle injury leftback luke
has recovered from a hamstring problem but centreback chris smalling is a big doubt due to illness robin van persie score
against leicester city earlier in the season but he is ruled out of the aston villa game defender jonny evans sits out th
fourth match of his sixgame ban for spitting at newcastle striker papiss cisse provisional squad de gea lindegaard valdes
shaw blackett rojo smalling jones mcnair rafael pereira young blind fellaini herrera valencia di maria januzaj mata falca
rooney wilson aston villa skipper ron vlaar could return for aston villa after finally shaking off a calf injury for the
to manchester united the defender has been out since february and he will be joined in the squad by joe cole after a
hamstring issue while jores okore knee is also fit ron vlaar trains ahead of his expected return to the aston villa squad
manchester united kieran richardson and philippe senderos both calf are in training but could still miss out while there
doubts surrounding aly cissokho groin and nathan baker knee and united loanee tom cleverley is ineligible provisional squ
guzan given hutton lowton okore vlaar clark cissokho kinsella delph sinclair gil westwood sanchez grealish cole nzogbia
agbonlahor benteke weimann kickoffsaturday pm odds subject to change manchester united draw aston villa refereeroger east
managerslouis van gaal manchester united tim sherwood aston villa headtohead league recordmanchester united wins draws as
villa wins key match stats supplied by opta manchester united have premier league wins against aston villa the jointmost
any team against any opponent in the history of the competition along with man utd v everton aston villa have won just on
the last premier league games against manchester united l d w wayne rooney has scored goals against aston villa his
jointhighest tally against any opponent in premier league history along with newcastle united aston villa are the only si
in the premier league this season yet to see a substitute score or assist a single one of their goals radamel falcao scor
manchester uniteds equaliser in their draw at aston villa in december rooney has scored eight goals in his last six premi
league games at old trafford against the villans netting four braces villa have seen six players sent off in the premier
league this season two more than they had in any previous premier league campaign rooney has scored goals at old trafford
only alan shearer with at st james park two for blackburn and thierry henry with at highbury have scored more goals on on
single ground in premier league history the red devils have won of their last premier league matches at old trafford d l
just eight occasions this season have a side used two or more players that are under in a premier league game seven of th
have been by manchester united arsenal the other aston villa have scored just two goals in the final half an hour in the
premier league this season man utd have scored more in the th minute or later

Reference summary

robin van persie ruled out with ankle injury for manchester unitednchris smalling a doubt but luke shaw back from hamstring complaintnron vlaar could make return to aston villa squad following calf injurynjoe cole and jores okore have also regained fitness for villans

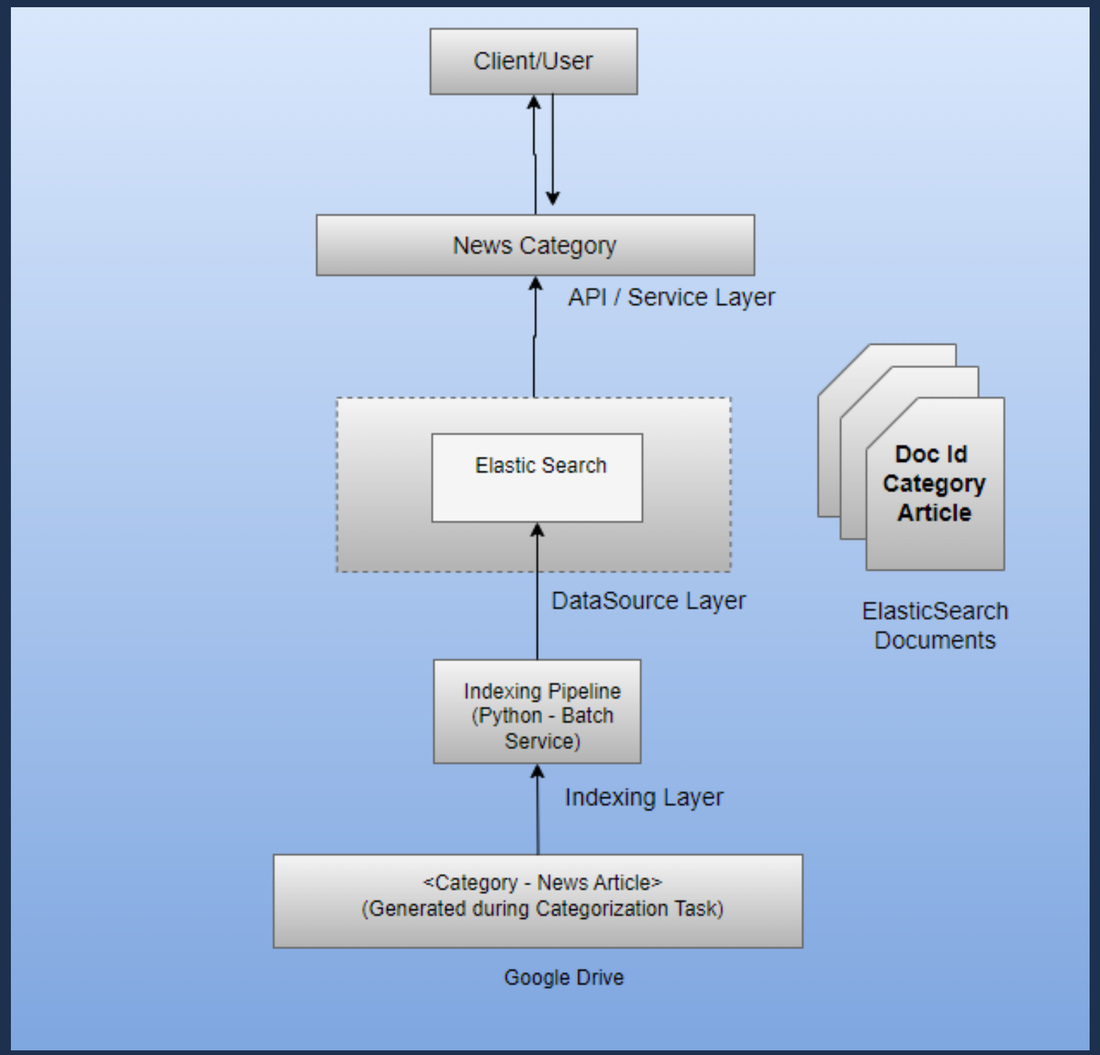
Generated summary

robin van persie is out of manchester uniteds premier league match against aston villanluke shaw has recovered from a hamstring problem but chris smalling is a big doubtnron vlaar could return for aston villa after finally shaking off a calf injury

MODEL PERFORMANCE CONCLUSION

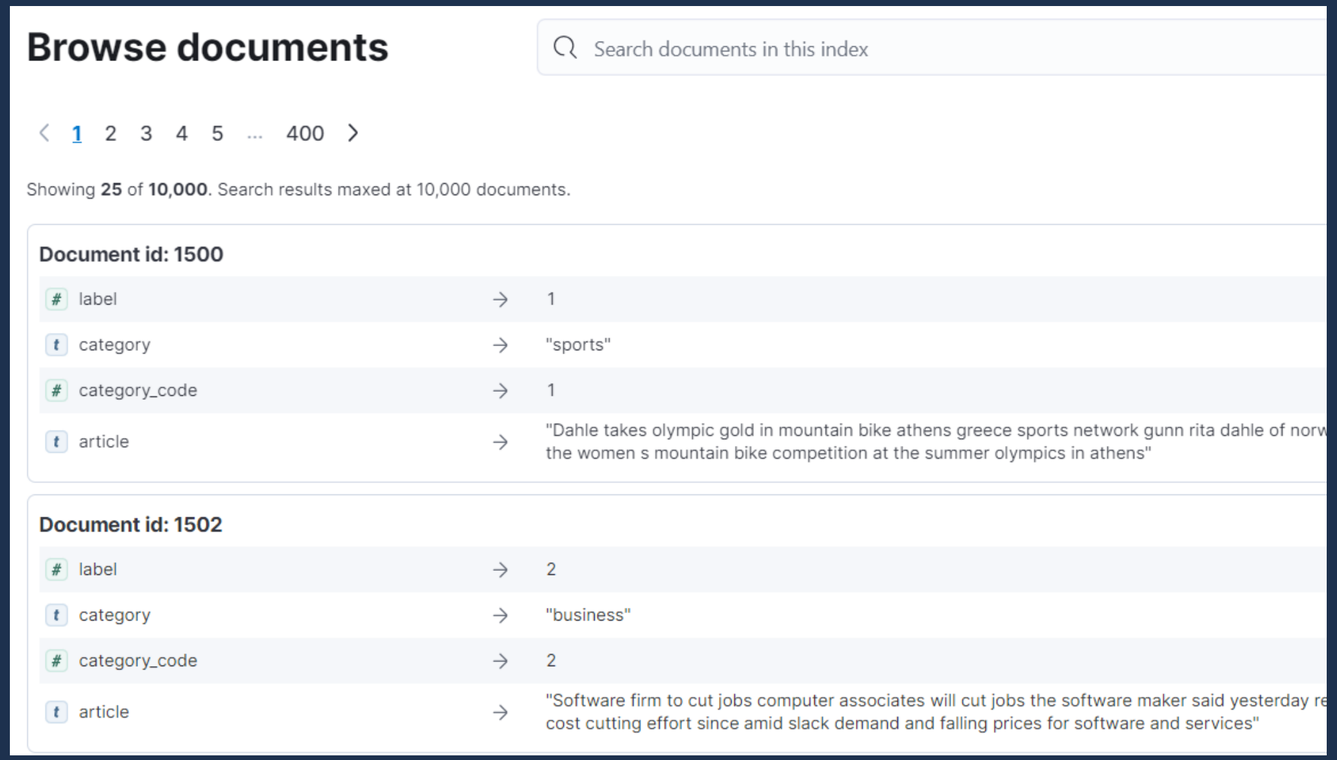
- Bert has performed slightly better than RoBERTa and Gpt-2 in the classification task.
- All three models when coming to precision, and F1-score have each performed well for different categories.
- Pegasus has given good results for Rouge 1 Rouge -L score but LLAMA performed well in terms of ROUGE-2 score for summarization.
- LLAMA-2 is seen to be the second-best-performing model.
- Despite performing poorly, T-5 small has a faster training time.
- For NER, LLAMA-2 performed better than BERT, however it was a close race.

ELASTICSEARCH FUNCTIONALITY



- Search Index is built which will store the categories generated in the Categorization task.
- Search provides following functionalities:
  - Given an input category, return all news associated news articles with the selected category.
- Indexing process provides the following functionalities:
  - Index the categories along with News Article generated during Categorization task.

ELASTICSEARCH FUNCTIONALITY



Authorization: "<API_KEY>"

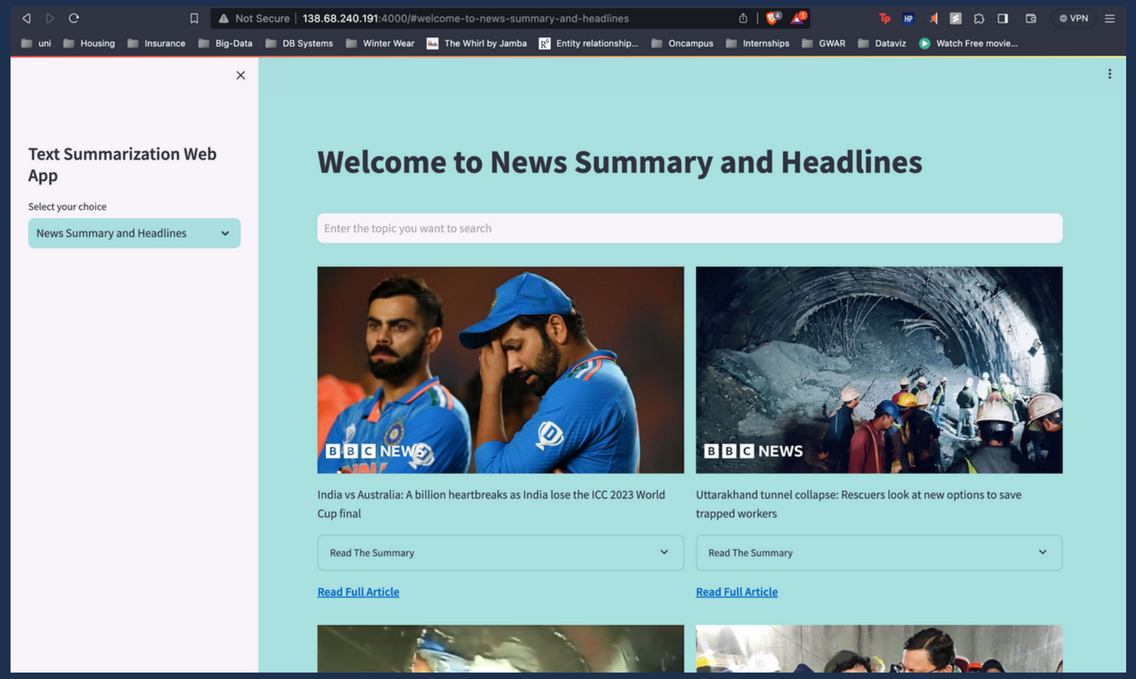
Elastic Search URL: https://e8206b505dc648f78d24b894fef165b1.us-central1.gcp.cloud.es.io:443/search-news-article-data-298b/_search?q=category:<CATEGORY>&from=0&size=1000&_source=article"
]

JSON document structure

WEB APPLICATION

Web Application is designed using StreamLit

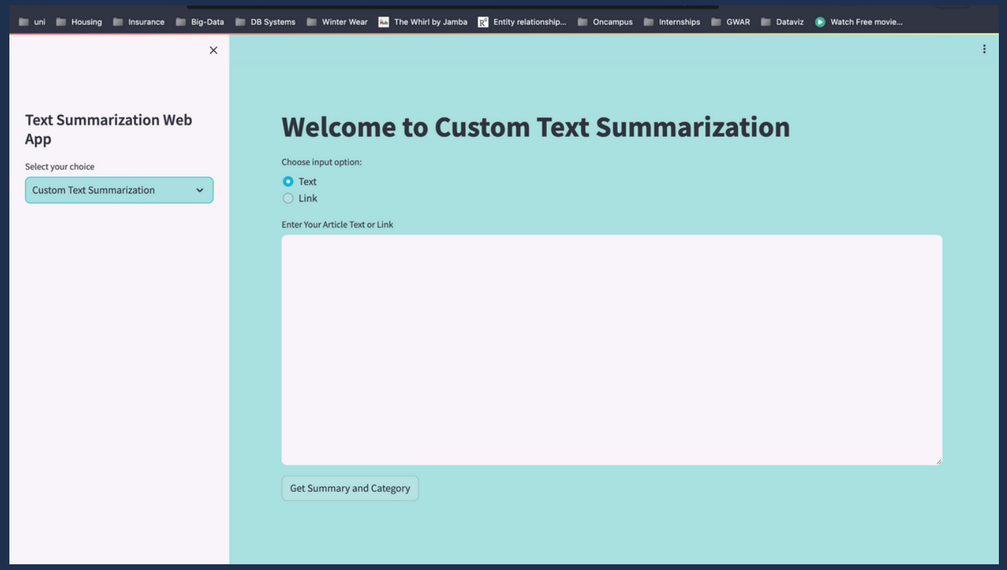
- Easy To Use
- Built-in Widgets
- Easy Deployment with Hosting Services

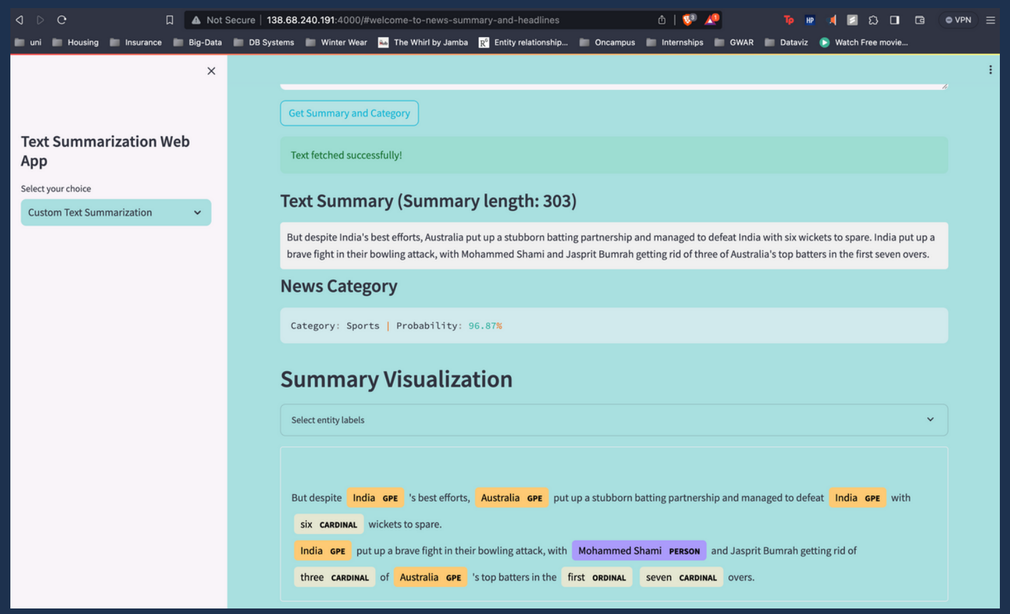


GUI Sequence 1 Flow -
News Summary and Headlines

- Retrieve Recent BBC Articles
- Summarize Retrieved Articles
- Display Summaries on Screen

WEB APPLICATION

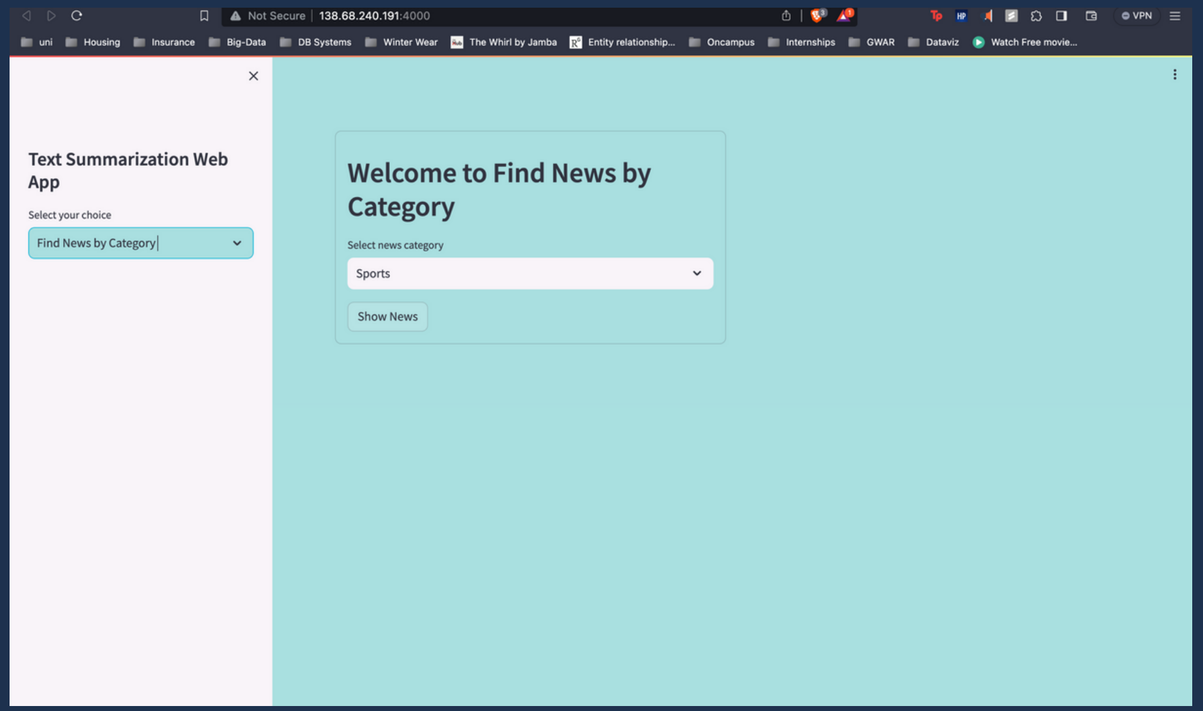




GUI Sequence 2 Flow - Custom Text Summarization

- User-provided or article-fetched text
- Models generate output
- Article summary, category, and entities displayed

WEB APPLICATION



GUI Sequence 3 Flow - Custom Text Summarization

- User Chooses Category
- GCP-Based Elastic Search Activation
- Display of Category-Specific Articles

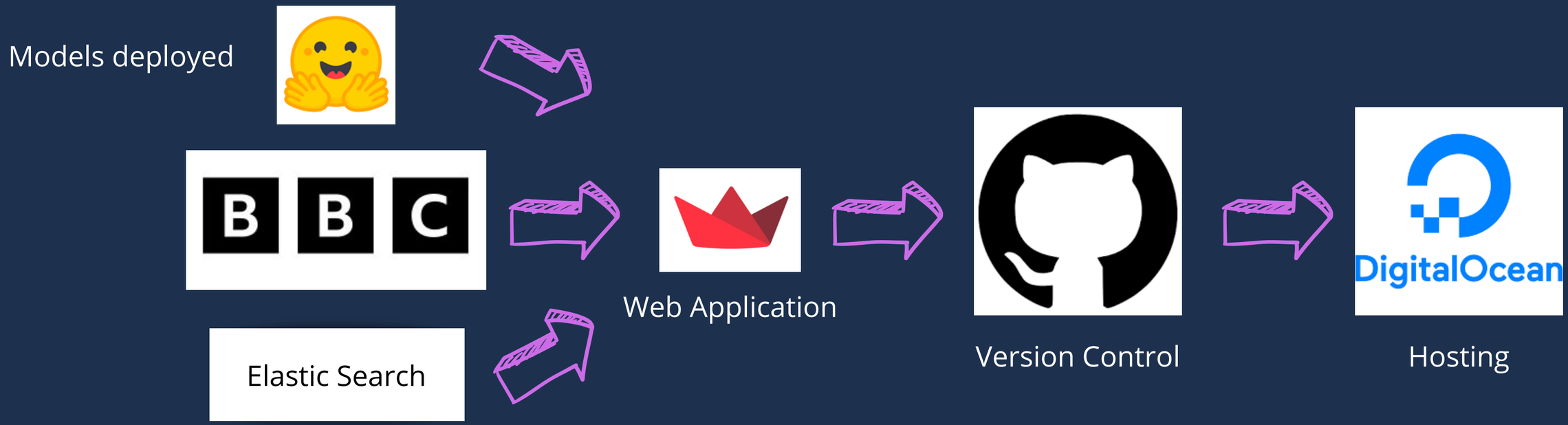
WEB APPLICATION DEPLOYMENT AND ARCHITECHTURE

Deployment is done on Digital Ocean and the code is pushed from Github

Advantages:

1) User-Friendly Interface 2) Cost-Effective Hosting

3) GitHub Integration



SYSTEM DEMO

Application Link: http://138.68.240.191:4000

CHALLENGES

- Training Time: Training large language models requires significant computational power and time. Training models on vast amounts of data can take days or even weeks.
- GPU/TPU Usage: To train and fine-tune models effectively, high-performance GPUs or TPUs are required, which can be expensive and resource-intensive.
- Latency: Deploying large language models for real-time applications can be challenging due to high inference latency. Serving predictions quickly often requires dedicated hardware or optimized model architectures.
- Cost: Utilizing cloud resources for training and inference need substantial costs, particularly when dealing with large models and datasets

APPLICATIONS

- Scholar Works: Summarizing Research papers
- News Website: To analyze news articles
- Legal Documents: Everyone has to go through some legal documents on a daily basis like Job offer letters, leasing agreements, and terms and conditions but do not have the time to read.
- Sentiment Analysis: Once the text is analyzed, it is easier to provide sentiment analysis on reviews or complaints.
- Independent News portals: Now a days every institution or organization have their own news portal, hence it can be used by SJSU for SJSU newsletters as well.

THANKYOU!