

Group 6's (AASD 4010 - Deep Learning I) Project Report

Topic: News Classification and Summarization

Team members:

1. Fatima Aghapourasl
2. Kareem Albeetar
3. Chun Lam Cheung
4. Penelope De Freitas
5. Eltigani Hamadelniel Elfatih Hamadelniel
6. Yi Liu
7. Jeffrey Ching Lok Ng
8. Ivy Tsai

Table of Contents

List of Figures.....	3
List of Tables.....	4
1.0 Problem Background and Motivation.....	5
2.0 Methodology.....	6
2.1 The General Pipeline.....	7
2.2 Project Approach.....	7
2.3 About the Dataset.....	8
2.3.1 Classification Dataset.....	8
2.3.2 Summarization Dataset.....	9
2.4 Training and Testing Environments.....	9
2.5 Data Preprocessing and EDA.....	9
2.5.1 Data Preprocessing (Classification).....	9
2.5.2 Exploratory Data Analysis (EDA).....	10
2.5.3 Data Preprocessing (Summarization).....	10
2.5.4 Exploratory Data Analysis (EDA).....	11
2.6 About the Models.....	11
2.6.1 RoBERTa.....	12
2.6.2 GPT-2.....	12
2.6.3 T5.....	12
2.6.4 BART.....	12
2.6.5 Longformer.....	13
2.7 Evaluation Metrics.....	13
2.7.1 Rouge.....	13
2.7.2 Additional Metrics.....	13
3.0 Results and Discussion.....	14
3.1 Classification Results.....	14
3.2 General Summarizer Model Results.....	23
3.3 Specialized Summarizer Models.....	25
3.3.1 Entertainment Summarization.....	25
3.3.2.1 More Exploration from the model.....	32
3.3.3 Business Summarization.....	32
3.3.4 Sports Summarization.....	36
4.0 Conclusions.....	40
5.0 References.....	41

List of Figures

- [Figure 1. Project Pipeline](#)
- [Figure 2. Article Content Acquisition Pipeline](#)
- [Figure 3. The Distribution of Categories in the Classification Dataset](#)
- [Figure 4. Word Count of the Text in the Classification Dataset](#)
- [Figure 4. Word Count of the Text in the Classification Dataset](#)
- [Figure 5. The Distribution of Categories in the Summarization Dataset](#)
- [Figure 6. Hyperparameters for DistilRoBERTa \(Original 18k Sample\)](#)
- [Figure 7. Training and Validation results](#)
- [Figure 8. Performance on Validation Set](#)
- [Figure 9. Performance on Summary Set](#)
- [Figure 10. Hyperparameters for DistilBERT \(Original 18k Sample\)](#)
- [Figure 11. Training and Validation results](#)
- [Figure 12. Training Results for DistilBERT \(Original 18k Sample\)](#)
- [Figure 13. Performance on Summary Set](#)
- [Figure 14. Hyperparameters for DistilRoBERTa \(Full Data\)](#)
- [Figure 15. Training and Validation results](#)
- [Figure 16. Performance on Summary Set](#)
- [Figure 17. Hyperparameters for DistilRoBERTa \(Full Data\)](#)
- [Figure 18. Training history for DistilRoBERTa \(Full Data\)](#)
- [Figure 19. Performance on Test Set for GPT-2 \(Full Data\)](#)
- [Figure 19. Performance on Test Set for GPT-2 \(Full Data\)](#)
- [Figure 20. Longformer \(Froze Layers\) Training Results](#)
- [Figure 21. Longformer Training Results](#)
- [Figure 22. T5-Base Fine-tuned Hyperparameters](#)
- [Figure 23. T5-Base Training Results](#)
- [Figure 24. Summaries Comparison for General Summarizer Models](#)
- [Figure 25. Hyperparameters of T5-base](#)
- [Figure 26. T5-base training](#)
- [Figure 27. Training Framework for the T5 Model](#)
- [Figure 28. Training and Validation for the BART model](#)
- [Figure 29. Training Framework for the Bart Model](#)
- [Figure 30. Fine-tuned Model Summarization](#)
- [Figure 31. Bart Base Model Summarization](#)
- [Figure 32. High Length_penalty VS Low_penalty](#)
- [Figure 33. Training Framework for BART \(Business Summarization\)](#)
- [Figure 35. Training Framework for T5 \(Business Summarization\)](#)
- [Figure 36. Training and Validation Results for T5 \(Business Summarization\)](#)
- [Figure 37. Training Parameters for BART \(Sports Summarization\)](#)
- [Figure 38. Training Parameters for T5 \(Sports Summarization\)](#)
- [Figure 39. Training and Validation Results for T5 \(Sports Summarization\)](#)

List of Tables

- [Table 1. Overall Classification Training Results of all Models](#)
- [Table 2. Example Summarization Comparison between T5-base Fine-tuned model, and handwritten summary \(Entertainment Summarization\)](#)
- [Table 3. Example Summarization Comparison between BART-base with Fine-tuned model \(Sports Summarization\)](#)
- [Table 4. Example Summarization Comparison between BART-base with T5 model \(Sports Summarization\)](#)

1.0 Problem Background and Motivation

In the modern information landscape, the escalating volume of news articles poses a significant challenge to effective information consumption. This project emerges from the recognition that classifying and summarizing news articles can bestow substantial benefits upon various stakeholders. The motivations driving our pursuit of this endeavor are multifaceted and underscore the transformative potential for users, decision-makers, and industries alike.

Addressing Information Overload

The ubiquity of digital content has ushered in an era of information overload, where the sheer volume of available data can be overwhelming. Our focus on classifying and summarizing news articles aims to empower users to navigate vast amounts of information more effectively. By distilling crucial information into concise summaries, we endeavor to provide users with a streamlined and accessible means of obtaining the most pertinent details.

Enhancing Time Efficiency

Recognizing the constraints on individuals' time, our project seeks to streamline the news consumption process. Summarized news content offers a time-efficient solution, allowing users to quickly grasp essential information without delving into lengthy articles. This approach aligns with the contemporary need for expeditious access to news updates.

Personalization for Tailored Experiences

The project also delves into the realm of personalization, acknowledging the diverse preferences and interests of news consumers. Through effective classification, this project holds the potential to deliver personalized news content, tailored to individual preferences. This not only enhances the user experience but also ensures that readers are exposed to information most relevant to their interests.

Empowering Decision-Makers

Decision-makers across various domains heavily rely on news sources to make well-informed choices. Our project recognizes the pivotal role of summarized and categorized news in expediting decision-making processes. By distilling critical information, decision-makers can stay abreast of developments, fostering agility in responding to evolving scenarios.

Facilitating Trend Monitoring

Summarizing and classifying news data extend beyond individual benefits to encompass broader implications for trend monitoring. Businesses and policymakers can leverage these summaries to monitor trends, patterns, and emerging issues over time. This proactive approach equips stakeholders with valuable insights for strategic planning and adaptation.

In essence, our project endeavors to address the challenges posed by information overload, time constraints, and the diverse needs of users and decision-makers. By harnessing the power of classification and summarization, we aim to create a robust solution that not only eases the news consumption experience but also facilitates informed decision-making and trend monitoring for a wide array of stakeholders.

Specifically, we harnessed the power of large language models in the realm of deep learning to enhance the classification and summarization processes applied to news articles. These models played a pivotal role in categorizing articles into distinct genres, including general news, entertainment, business, and sports. By employing large language models, we aimed to bolster the precision and efficiency of our classification and summarization efforts, thereby providing a more nuanced and insightful understanding of diverse news content across various domains.

2.0 Methodology

In the dynamic world of journalism, it is also important to highlight the differences in writing styles that are strategies employed across the various branches of news categories to both captivate the audience but also communicate the message effectively. Each sub-branch or category of news such as hard news, features, opinion pieces, sports, economics, finance, and entertainment pieces requires a distinct approach to engage readers and fulfill the intended purpose of the piece. For instance, hard news articles that focus on communicating facts are often straightforward and concise. On the other hand, entertainment and human interest articles lean towards a narrative-driven approach, one that uses descriptive language to evoke an emotion and connection between the reader and the content.

Therefore it is important to keep in mind the importance of writing styles when developing our approach. To tackle this we have decided to go with a two-staged approach that features classification and summarization. Where in the first section we focus on identifying the particular category or content features based on its content and writing style. Based on the derived category we then employ the use of a fine-tuned model to produce a summary tailored to embody that particular style of writing reducing the article to its key components while still maintaining the cohesive flow crafted by the journalist.

2.1 The General Pipeline

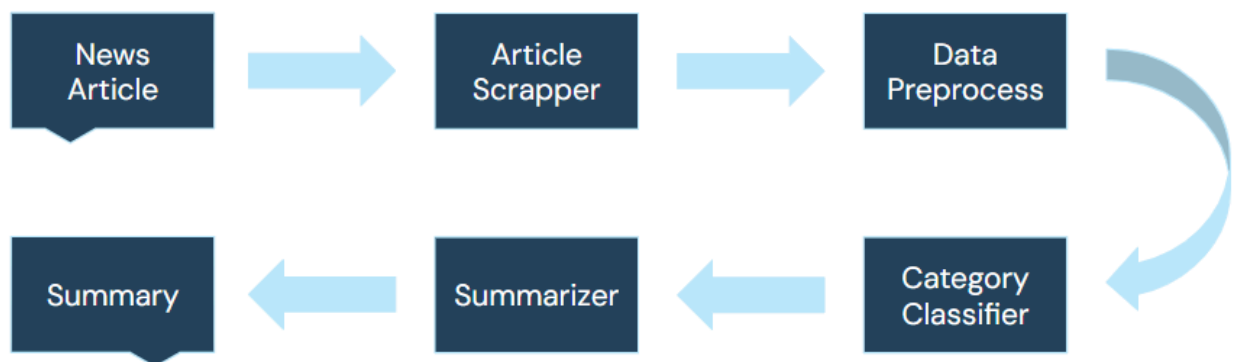


Figure 1. Project Pipeline

Based on our defined approach we curated a six-step pipeline that can be observed in the above figure. When given an **input article website link X** we use a crawling algorithm to extract the textual content of the data, reducing it to its bare-bone structure by eliminating the presence of any formatting such as emojis, hyperlinks, HTML tags, stripping white spaces and so on. This now cleaned text is fed into our classification neural network to identify the category of the article given its content. This is a vital point in our pipeline as based on this derived category a summarization model is selected that has been fine-tuned on a dataset of articles for that class, this summarizer would **then generate a summary Y** that is tailored to the particular writing style and pattern seen within articles of this category serving as our final output. Therefore it is pivotal that the classifier featured within the pipeline yields a very low misclassification rate to ensure the right summarizer is selected for the article generating the best quality summary.

2.2 Project Approach

Attacking this project, our team was split into pairs, with each pair focusing on developing a particular component of our pipeline with the focus of exploring the performance and capabilities of various large language models on both the classification of news articles and generating summarizations for them. We pointed our focus to the top three categories with the largest number of samples for both our downstream tasks to ensure our models were able to see a wide spread of samples for a particular style of writing and generate higher-quality summaries.

2.3 About the Dataset

2.3.1 Classification Dataset

The dataset used to train the classification component of our pipeline was retrieved from Kaggle. It features 210k news headlines from 2012 to 2022 saved in a JSON format from TheHuffingPost, a prominent online news and opinion platform that features articles covering a diverse set of topics including politics, entertainment, culture, and lifestyle.

Regarding features, each record in the dataset consists of the following attributes:

- **Category:** The category in which the article was published.
- **Headline:** The headline of the news article.
- **Authors:** The list of authors who contribute to the article.
- **Link:** The URL link to the original news article.
- **Short Description:** An abstract of the article's content.
- **Date:** The publication date for the article.

The target variable for this dataset is the category, this dataset contains a total of 42 news categories including politics, wellness, entertainment, travel, style & beauty, parenting, and more. We later added two additional datasets to try and make the classifier more diverse which added 8K more data points.

Article Content Acquisition

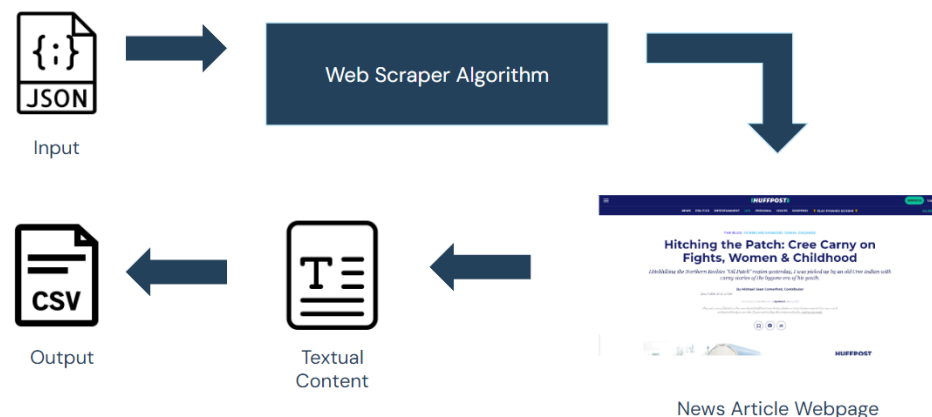


Figure 2. Article Content Acquisition Pipeline

To acquire the articles for the classification component of our pipeline we first read the JSON file into a pandas data frame for easier data manipulation. Next, we leveraged the BeautifulSoup library to create a web scraping algorithm that would select each record within our data frame iteratively sending a GET request using the website URL link, retrieve its

HTML DOM, and extract the textual content for the article through its class ID and storing it back into our data frame. Finally, the data frame was exported as a clean CSV file with 193,000 records after addressing broken links due to archived news articles.

2.3.2 Summarization Dataset

Our Summarization dataset was acquired by merging two data sources from Kaggle HuggingFace and GitHub respectively. These datasets feature 5500 scrapped articles from various news platforms namely: The New York Times, CNN, Business Insider, and Breitbart.

Regarding features in the summarization dataset, each record consisted of the following attributes:

- **Category:** The category in which the article was published.
- **Article:** The textual content of the news articles.
- **Human Summary:** Hand-written summary of the articles.

The target variable in this dataset is the human summaries.

2.4 Training and Testing Environments

Throughout the training and testing phases, platforms like Kaggle and Google Colab were leveraged for their supplementary resources. These platforms provided an enriched computational environment, facilitating the execution of complex tasks and the utilization of additional computing resources. The use of Kaggle and Google Colab enhanced the efficiency of model development and evaluation, offering collaborative and accessible environments that streamlined the experimentation process.

2.5 Data Preprocessing and EDA

2.5.1 Data Preprocessing (Classification)

For the classification dataset, we initially gathered data from multiple sources. To ensure uniformity, we standardized column names and labels across all datasets. This involved addressing discrepancies such as lowercase labels and variations in category names, like "sports" versus "Sports." Once these issues were rectified, articles not categorized as business, sports, or entertainment were relabeled as "others."

Subsequently, we merged all datasets into a single data frame and began preprocessing the text data. This entailed removing unwanted characters and converting all text to lowercase. Leveraging the NLTK library, we then eliminated stopwords and performed lemmatization on the article content.

During this process, we identified instances of articles with minimal word count. Upon investigation, it was discovered that some entries contained only NaN values, prompting their removal. Finally, any duplicate rows were also eliminated from the dataset.

2.5.2 Exploratory Data Analysis (EDA)

The classification dataset exhibited significant imbalance, with a majority of articles falling under the "other" category. Given the dataset's considerable size, our approach involved downsampling both the "other" and "entertainment" categories to approximately 6500 data points each. To address the diverse range of topics within the "other" category, we opted for stratified random sampling. This ensured a representative selection across various genres such as health, politics, food, and more.

```
category
OTHER      163283
ENTERTAINMENT  17158
BUSINESS    6851
SPORTS      5953
Name: count, dtype: int64
```

Figure 3. The Distribution of Categories in the Classification Dataset

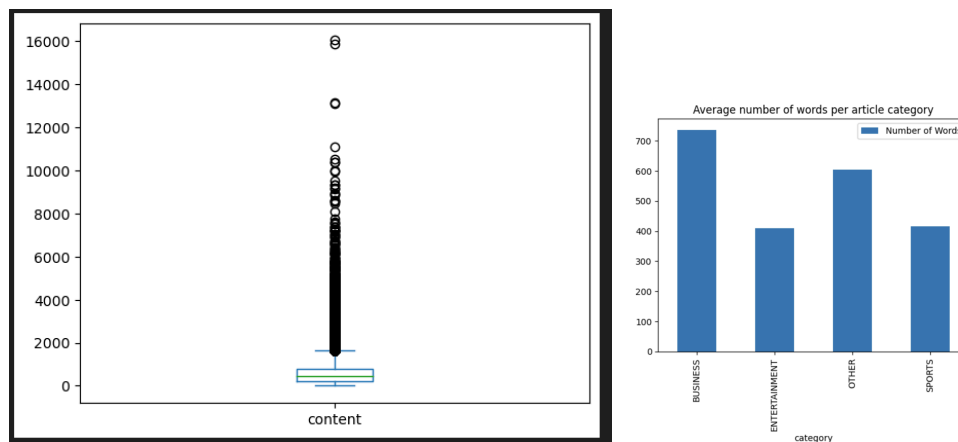


Figure 4. Word Count of the Text in the Classification Dataset

2.5.3 Data Preprocessing (Summarization)

For the Summarization dataset, we gathered data from multiple sources to create a comprehensive dataset including Kaggle, Hugging Face and dataset on other websites. We've standardized column names and ensured that all column names across different datasets were consistent.

We've chosen to retain punctuation marks within the text data. This decision likely stems from the understanding that punctuation can convey important semantic and syntactic information in natural language text, especially in summarization tasks where maintaining the original structure and flow of sentences is crucial for generating accurate summaries.

We're providing flexibility to members who will be training models on this dataset. By not pre-processing or removing punctuation from the text data, allowing the members to decide how they want to handle punctuation during their model training process. They can choose to keep the punctuation as is, remove it, or apply other text processing techniques based on the requirements and characteristics of their specific models.

2.5.4 Exploratory Data Analysis (EDA)

We identified that our dataset contains more than 10 categories, with a significant portion of articles being not our main preference. This could potentially lead to biases in model training and affect the performance of the classifier, especially if certain classes are underrepresented.



File_path	
accidents	4
architecture	4
art	2
business	1228
crime	110
entertainment	925
environment	1
health	2
law	41
lifestyle	78
politics	1158
science	25
sport	1021
sports	30
tech	802
technology	18
dtype:	int64

Figure 5. The Distribution of Categories in the Summarization Dataset

2.6 About the Models

For this project, our team experimented with a diverse set of large language models. The following subsection provides a brief description of each of the models that were utilized.

2.6.1 RoBERTa

The RoBERTa model, an extension of the BERT architecture, has demonstrated remarkable efficacy in natural language processing tasks, including classification and summarization. By optimizing the pre-training process with a larger dataset and removing the next sentence prediction task, RoBERTa achieves enhanced contextual understanding and representations. Its robust transformer-based design allows it to excel in classification tasks by capturing intricate relationships within textual inputs and assigning accurate labels. In the realm of summarization, RoBERTa's proficiency lies in its ability to comprehend and distill essential information from lengthy documents, providing coherent and contextually relevant summaries.

2.6.2 GPT-2

The GPT-2 model, developed by OpenAI, has emerged as a pivotal transformer-based architecture renowned for its remarkable language generation capabilities. For classification tasks, GPT-2 can be fine-tuned to discern and categorize textual inputs effectively. In the realm of summarization, its autoregressive nature allows it to generate concise and coherent summaries by extracting salient information from lengthy documents. GPT-2's versatility and proficiency in both creative text generation and practical natural language processing tasks underscore its significance as a robust and adaptable model in the field of deep learning.

2.6.3 T5

The T5 (Text-To-Text Transfer Transformer) model, developed by Google, has emerged as a transformative architecture in natural language processing, excelling in both classification and summarization tasks. In classification tasks, T5 maps input texts to predefined labels, leveraging the same architecture that is fine-tuned for summarization. For summarization, T5 stands out by framing the task as generating a concise summary from the input text, showcasing its flexibility in handling diverse document lengths. The model's ability to unify different NLP tasks under a consistent paradigm underscores its elegance and effectiveness, positioning T5 as a prominent and adaptable player in the landscape of transformer-based models for natural language understanding and generation.

2.6.4 BART

Text summarization is an essential task in natural language processing (NLP), aiming to condense long pieces of text into shorter versions that capture the most critical information. The Bidirectional and Auto-Regressive Transformers (BART) model has emerged as a potent framework for this task due to its pretraining on a diverse range of text and its ability to understand and reproduce text effectively.

2.6.5 Longformer

The Longformer model introduces an innovative approach to handling long-range dependencies in sequential data, particularly beneficial for tasks involving extensive document-level information. It extends the transformer's capabilities by incorporating a novel global-local attention mechanism, allowing it to efficiently process vast sequences with significantly reduced computational complexity. This design is particularly advantageous for tasks such as document classification and summarization, where contextual understanding across extensive textual spans is crucial.

2.7 Evaluation Metrics

2.7.1 Rouge

The **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) metric was employed in this study since it serves as a crucial tool in the evaluation of text summarization. ROUGE's utility in summarization lies in its ability to quantitatively gauge the informativeness, coherence, and conciseness of machine-generated summaries, thereby facilitating advancements in the development of robust and contextually relevant summarization models.

Employing various metrics such as ROUGE-N (measuring overlap of n-grams), ROUGE-L (calculating the longest common subsequence), and ROUGE-W (evaluating word overlap), the metric provides a comprehensive evaluation of the effectiveness of summarization models.

2.7.2 Additional Metrics

Accuracy was another metric that was employed because it holds paramount importance in the training and testing phases within the domains of text classification or summarization. During training, accuracy serves as a fundamental gauge of a model's ability to correctly predict categories or generate summaries, reflecting the overall effectiveness of the learning process. In testing, accuracy becomes a crucial benchmark for evaluating the model's generalization capabilities to new, unseen data.

We complement accuracy with other metrics such as the **F-measure** and **precision**. These metrics become imperative during training to fine-tune models for optimal performance and guide the selection of appropriate thresholds. Precision, represents the ratio of true positives to the sum of true positives and false positives, whereas, the f-measure combines precision and recall, and brings a balanced assessment by considering both false positives and false negatives. In testing, the F-measure and precision offer a comprehensive understanding of a model's capability to make accurate and relevant predictions or generate precise summaries.

In addition to appraising the models' performance, **inference analyses** were undertaken by contrasting the original text articles with both manually crafted summaries and summaries generated by the models. This comparative scrutiny enabled us to discern the efficacy of the automated summarization process in comparison to human-generated summaries. It served as a crucial step in evaluating the models' ability to capture the essence of the content and generate concise, coherent summaries that align with the nuanced understanding demonstrated by human summarizers.

3.0 Results and Discussion

3.1 Classification Results

At first we decided to tackle this problem by testing out two models, DistilBERT and DistilRoBERTa. We trained these models on the original 18k dataset and had a separate dataset for validation, which we will call Summary Set. This Summary Set came from the dataset we were going to use to train our summarizers, and contained articles from many different websites.

After receiving feedback from the presentation day, we wanted to try out different architecture models to see how they would perform in classification tasks. The two architectures we chose were encoder only, and decoder only. For the encoder only we decided to use DistilRoBERTa and DistilBERT. For the decoder only, we wanted to use XLNet, but unfortunately, the GPU frequently ran out of memory while training, even with batch size at 1. We finally decided to use GPT-2 instead to at least try out a decoder-only model, even though GPT-2 is used more for text generation. We also moved on to the full dataset as described above.

Model	Test Accuracy	Time to Train (Not including validation time)
DistilBERT (Original 18k Sample)	0.89	33 Min (4 Epochs)
DistilRoBERTa (Original 18k Sample)	0.90	33 Min (4 Epochs)
DistilRoBERTa (Full Data)	0.91	33 Min (3 Epochs)
GPT-2 (Full Data)	0.90	40 Min (3 Epochs)

Table 1. Overall Classification Training Results of all Models

When we first started this project we trained only on the original 18k dataset from Huffington post, as we were unsure how long it would take to train. Since it was already starting to

overfit at the 4th epoch, we didn't tune the number of epochs. Instead, we tuned the learning rate and batch size.

DistilRoBERTa (Original 18k Sample)

After testing various configurations, we had this configuration as our best.

```
In [22]: training_args = TrainingArguments(  
    report_to="wandb",  
    output_dir="DistilroBERTa",  
    learning_rate=2e-5,  
    per_device_train_batch_size=8,  
    per_device_eval_batch_size=8,  
    num_train_epochs=4,  
    weight_decay=0.01,  
    evaluation_strategy="epoch",  
    save_strategy="epoch",  
    load_best_model_at_end=True,  
    push_to_hub=False,  
)
```

Figure 6. Hyperparameters for DistilRoBERTa (Original 18k Sample)

With the following results after training

 [8332/8332 33:15, Epoch 4/4]

Epoch	Training Loss	Validation Loss	Accuracy
1	0.413800	0.359183	0.903296
2	0.340300	0.380372	0.904916
3	0.266600	0.402998	0.903836
4	0.245400	0.422403	0.900054

Figure 7. Training and Validation results

We can see that validation loss was increasing after every epoch showing some signs of overfitting.

On the validation set, we see that while it formed well in identifying Business, Entertainment, and Sports, it performed a lot worse on Others. It often misclassified something that should belong to others as business. We think the reason for this is that since business is such a broad topic, it can often overlap with other topics. One example is that if two tech companies are performing a merger, let's say these two companies are Google and OpenAI. The article may mention financial aspects, regulators from different countries, and how the tech of AI

may change. While this article might be classified as Technology (looking at how AI would change), it could also be thought of as Business (The financials of both companies and might have a little mix of World News (Maybe EU regulators are already looking to block the merger and what kinds of argument they will use).

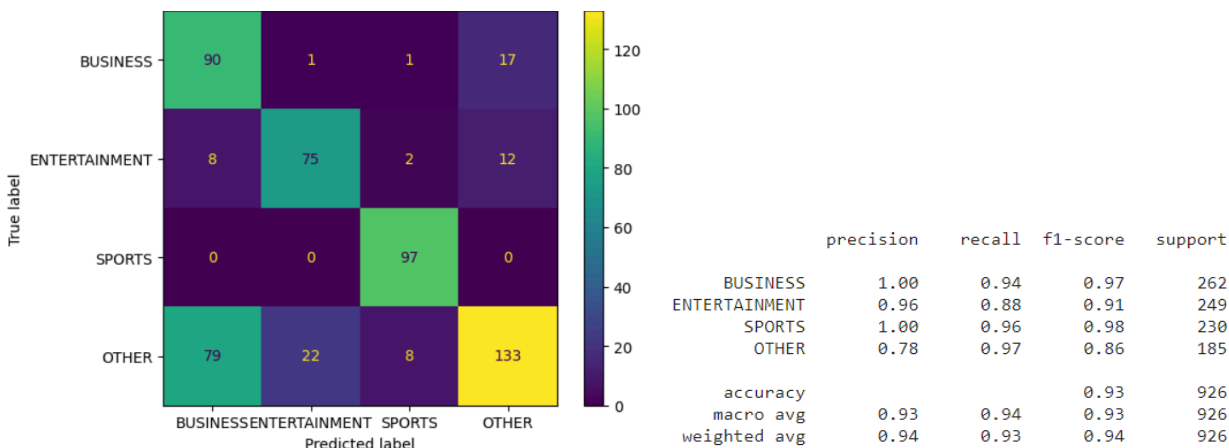


Figure 8. Performance on Validation Set

After training DistilRoBERTa, we decided to also check on a separate dataset with data from another website in which we can see that it performed a lot worse at classifying the 'other' category.

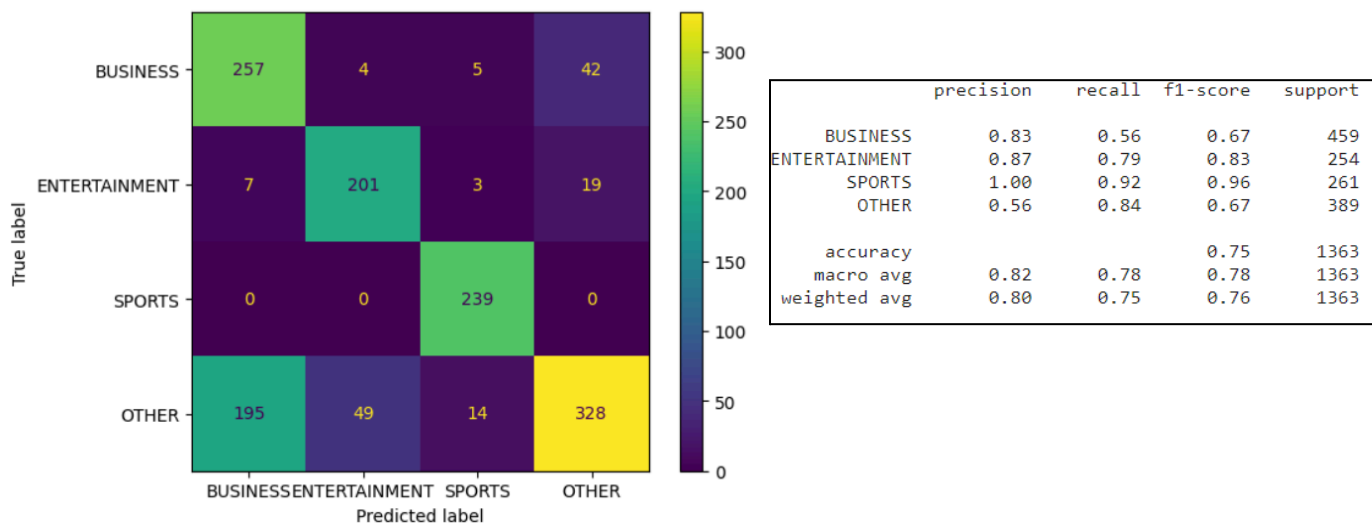


Figure 9. Performance on Summary Set


DistilBERT (Original 18k Sample)

The following was our best result for DistilBERT.

```
In [19]: training_args = TrainingArguments(  
    output_dir="DistilBERT",  
    learning_rate=2e-5,  
    per_device_train_batch_size=8,  
    per_device_eval_batch_size=8,  
    num_train_epochs=4,  
    weight_decay=0.01,  
    evaluation_strategy="epoch",  
    save_strategy="epoch",  
    load_best_model_at_end=True,  
    push_to_hub=False,  
)
```

Figure 10. Hyperparameters for DistilBERT (Original 18k Sample)

As a result we were able to get the following during training.



Epoch	Training Loss	Validation Loss	Accuracy
1	0.425800	0.388695	0.883306
2	0.340700	0.396263	0.887628
3	0.262600	0.445465	0.891410
4	0.237200	0.470516	0.893571

Figure 11. Training and Validation results

	precision	recall	f1-score	support
BUSINESS	0.81	0.60	0.69	860
ENTERTAINMENT	0.84	0.82	0.83	477
SPORTS	1.00	0.92	0.96	555
OTHER	0.61	0.82	0.70	833
accuracy			0.77	2725
macro avg	0.82	0.79	0.79	2725
weighted avg	0.79	0.77	0.77	2725

Figure 12. Training Results for DistilBERT (Original 18k Sample)

On the separate summary set, the model performed quite poorly, similar to DistilRoBERTa, we see that it is often misclassifying other articles as business.

	precision	recall	f1-score	support
BUSINESS	0.80	0.61	0.69	419
ENTERTAINMENT	0.84	0.83	0.83	239
SPORTS	1.00	0.91	0.96	280
OTHER	0.63	0.81	0.71	425
accuracy			0.78	1363
macro avg	0.82	0.79	0.80	1363
weighted avg	0.79	0.78	0.78	1363

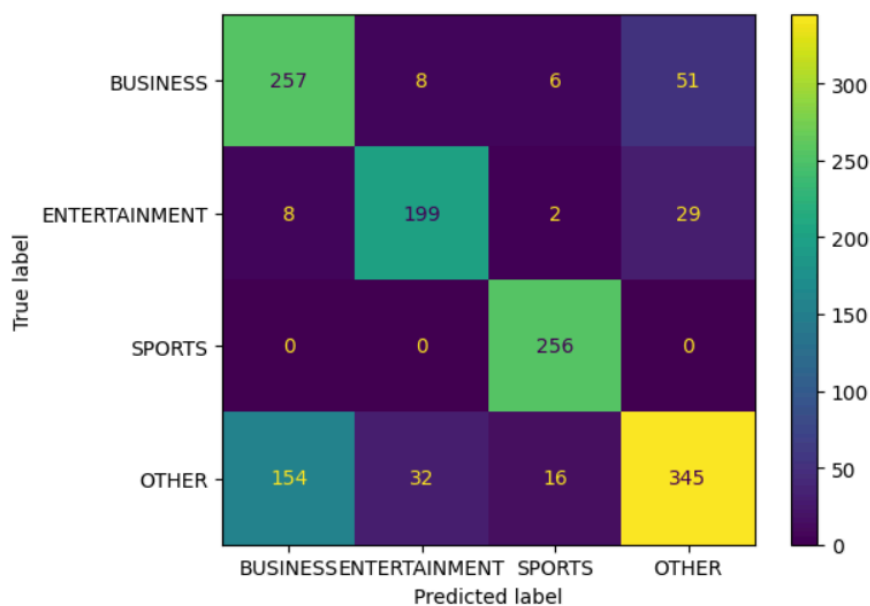


Figure 13. Performance on Summary Set

DistilRoBERTa (Full Data)

We started training this model following the feedback we were given. We also added early stopping to this model as we saw in the previous models that it is prone to overfitting.

In [20]:

```
training_args = TrainingArguments(
    report_to="wandb",
    output_dir="DistilroBERTa-cat_df",
    learning_rate=2e-5,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=4,
    weight_decay=0.01,
    evaluation_strategy="epoch",
    save_strategy="epoch",
    load_best_model_at_end=True,
    push_to_hub=False,
)
```

In [22]:

```
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_data['train'],
    eval_dataset=tokenized_data['test'],
    tokenizer=tokenizer,
    data_collator=data_collator,
    compute_metrics=compute_metrics,
    callbacks=[early_stop]
)
```

Figure 14. Hyperparameters for DistilRoBERTa (Full Data)

During training it did an early stop and only trained on 3 epochs. Additionally, overall with the increase in the data, the validation accuracy went up a bit more.

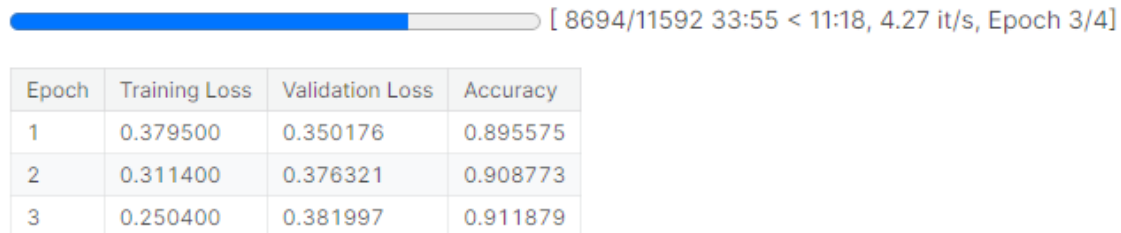


Figure 15. Training and Validation results

Overall with the increase in the data the validation performed better at classifying everything, with only a slight decrease in performance on Sports. It has improved a lot more at classifying others but still has trouble with it sometimes. It seems that overall it is quite good at differentiating between our 3 main topics, but still confuses it with others.

	precision	recall	f1-score	support
BUSINESS	0.87	0.92	0.89	230
ENTERTAINMENT	0.92	0.90	0.91	252
SPORTS	0.93	0.96	0.94	241
OTHER	0.77	0.71	0.74	203
accuracy			0.88	926
macro avg	0.87	0.87	0.87	926
weighted avg	0.88	0.88	0.88	926

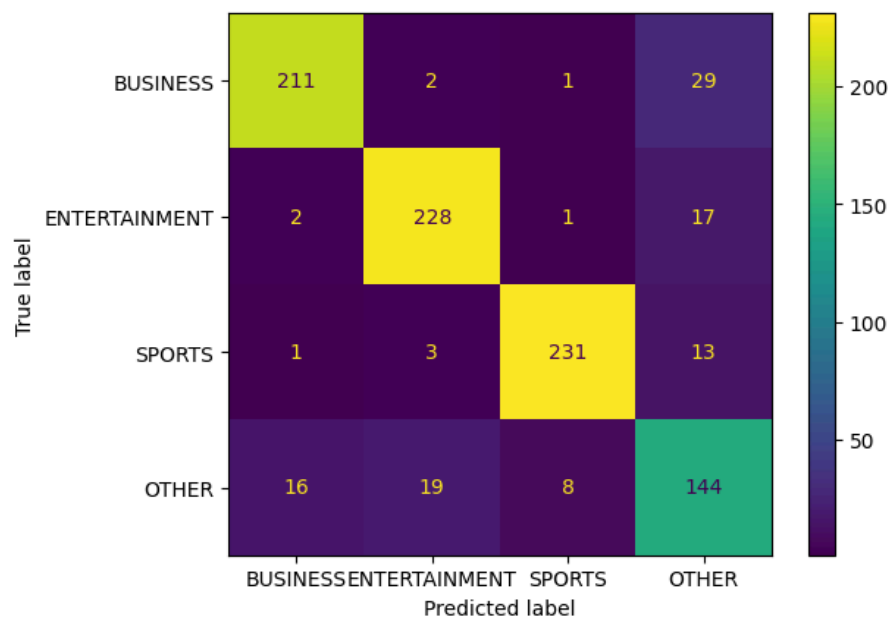


Figure 16. Performance on Summary Set

GPT-2 (Full Data)

After failing numerous times trying to train XLNet, we decided to try GPT-2 for text classification instead, as it is also a decoder-only model.

```
EPOCHS = 3
model = SimpleGPT2SequenceClassifier(hidden_size=768, num_classes=4, max_seq_len=512, gpt_model_name='distilbert/distilgpt2')
LR = 1e-5

train(model, df_train, df_val, LR, EPOCHS)
torch.save(model.state_dict(), "gpt2-text-classifier-model-E4.pt")
```

Figure 17. Hyperparameters for DistilRoBERTa (Full Data)

Unfortunately, due to time constraints and running out of GPU resources, we were not able to experiment too much with GPT-2 as a text classifier.

Here are the results after training

```
100%|██████████| 10301/10301 [13:10<00:00, 13.02it/s]

Epochs: 1 | Train Loss: 0.227 | Train Accuracy: 0.846 | Val Loss: 0.175 | Val Accuracy: 0.875

100%|██████████| 10301/10301 [13:10<00:00, 13.03it/s]

Epochs: 2 | Train Loss: 0.113 | Train Accuracy: 0.922 | Val Loss: 0.170 | Val Accuracy: 0.886

100%|██████████| 10301/10301 [13:11<00:00, 13.02it/s]

Epochs: 3 | Train Loss: 0.038 | Train Accuracy: 0.974 | Val Loss: 0.225 | Val Accuracy: 0.899
```

Figure 18. Training history for DistilRoBERTa (Full Data)

While GPT-2 overall performed only slightly worse than DistilRoBERTa, we see that it was able to increase the accuracy on classifying other by a lot, as well as increases in sports and business, with only a slight decrease of accuracy in entertainment,

	precision	recall	f1-score	support
0	0.90	0.88	0.89	689
1	0.88	0.95	0.91	686
2	0.95	0.96	0.95	596
3	0.85	0.77	0.81	605
accuracy			0.89	2576
macro avg	0.89	0.89	0.89	2576
weighted avg	0.89	0.89	0.89	2576

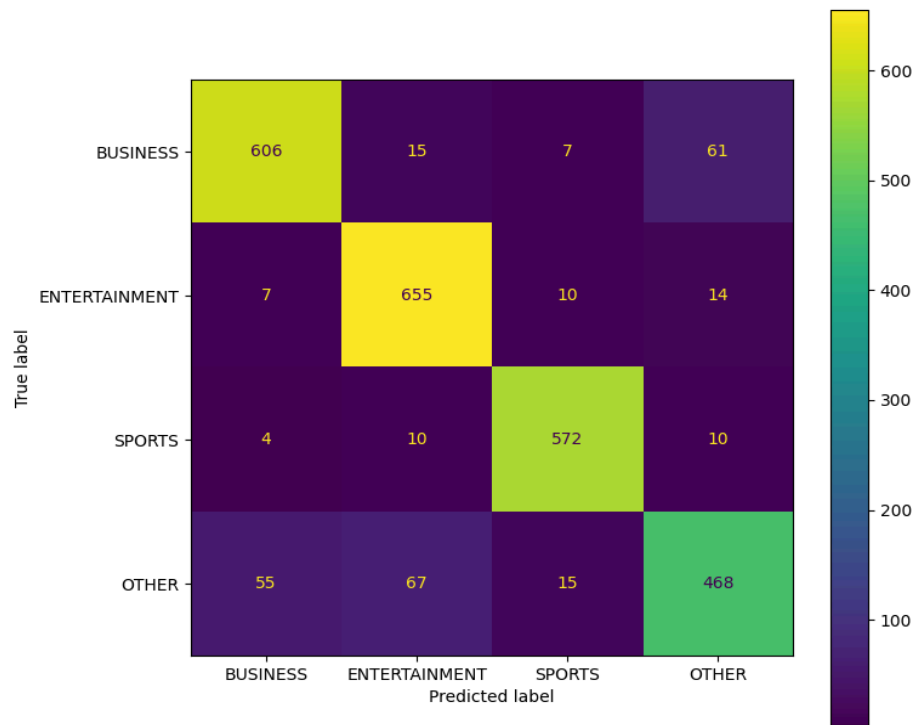


Figure 19. Performance on Test Set for GPT-2 (Full Data)

Overall, both DistilRoBERTa and GPT-2 performed well in classifying the news articles. We should note that since we aim to have a fine-tuned summarizer on the topic, we care more about making sure we don't have high false positives on our three chosen categories. This means we want to see higher precision overall in those three categories. Overall, GPT-2 (average precision on 3 categories was 91%) was able to beat DistilRoBERTa (average precision on 3 categories was 90.67%) by only 0.33%. Thus we would be using GPT-2 for our model.

If we were to do this task again, we feel like we would be able to accomplish a lot more, as we would have been able to do more training with the free GPU resource we had, instead of spending a lot of time fixing bugs. We would also like to spend more time fine-tuning GPT-2,

and also try to figure out a way to train an XLNet model. We think overall for our first deep learning project, the plan of attack went well as we got to experiment with a lot of different models. But for next time, now that we have much more experience, we think that we would have been able to spend more GPU resources on fine-tuning and training models, instead of debugging.

3.2 General Summarizer Model Results

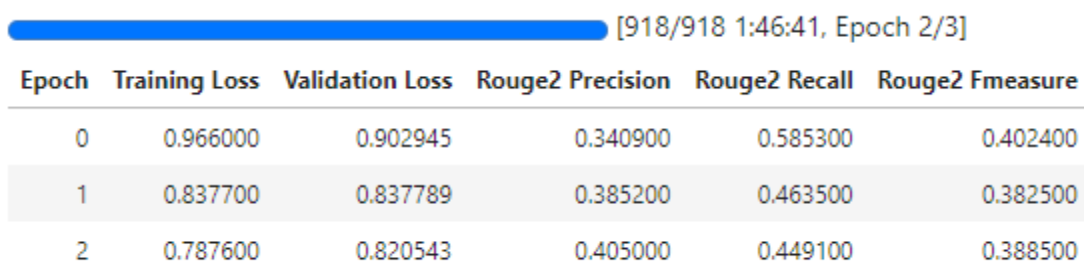
Our general summarizer had two purposes. One was to ensure that we still had a summarizer model for any news that was classified as other. The other part was to serve as a baseline for the other more specialized model to perform better in.

For general summarization, we compared Longformer and T5. While Longformer seemed like an exciting new model we wanted to use, this model crashed a lot when we were training. Longformer took much more memory, had way more parameters, and took much longer to train. This made it so that we were not able to experiment with Longformer as much as other models. As a result, T5 has better-generated summarization; therefore T5 was used in all specialized summarizer models(business, sports, and entertainment).

Longformer

As mentioned above, Longformer often takes a very long time to train, even on GPUs. With our limited resources, we were unfortunately unable to experiment too much with longformer. We did try to increase epoch size a few times, but the GPU would run out of memory, and would often happen only when it was almost done training. It also took a long time to debug as many errors came up while training, resulting in a lot of loss models.

One way we tried to lower the training time was all the layers except the last one, but it did not affect the training time at all. Unfortunately, even with an epoch of only 3, it would still take around 3 hours to train (Validation took around 15 mins each time), if it didn't crash.



Epoch	Training Loss	Validation Loss	Rouge2 Precision	Rouge2 Recall	Rouge2 Fmeasure
0	0.966000	0.902945	0.340900	0.585300	0.402400
1	0.837700	0.837789	0.385200	0.463500	0.382500
2	0.787600	0.820543	0.405000	0.449100	0.388500

Figure 20. Longformer (Froze Layers) Training Results

[918/918 1:42:30, Epoch 2/3]

Epoch	Training Loss	Validation Loss	Rouge2 Precision	Rouge2 Recall	Rouge2 Fmeasure
0	0.240900	0.223211	0.573400	0.492400	0.521800
1	0.184100	0.208663	0.596600	0.546000	0.562700
2	0.131200	0.208519	0.615800	0.583600	0.592600

Figure 21. Longformer Training Results

T5-base

The T5-base model was fine-tuned with the hyperparameters shown below in Figure 21. Although the T5-base model was not as large as Longformer, it still took about 30 minutes to train for 1 epoch. Therefore, only 3 epochs were trained; for future fine-tuning, more epochs should be considered for better performance. Additionally, to decrease the training time, freezing some of the layers during training should be considered as well.

Hyperparameters	
Prefix added to each input text	"summarization: "
Model	t5-base
Tokenizer	t5-base
Data collator	DataCollatorForSeq2Seq
max_length	150
batch_size	4
Learning rate	2e-5
Epoch	3
weight_decay	0.01

Figure 22. T5-Base Fine-tuned Hyperparameters

[2943/2943 37:20, Epoch 3/3]

Epoch	Training Loss	Validation Loss
1	1.211700	0.705703
2	0.672900	0.678294
3	0.656700	0.671496

ROUGE Scores:
ROUGE-N F1 Score: 0.4519763007101942
ROUGE-L F1 Score: 0.4450328824264304

Figure 23. T5-Base Training Results

Here is an example of the generated summaries from both models being compared with the human-written summary.

Hand Written Summary	T5-Base Summary	LongFormer Summary
But Ms Short said the effect of the parallel coalition would be to undermine the UN. She said only the UN had the "moral authority" to lead the relief work. The US was "very bad at coordinating with anyone" and India had its own problems, Ms Short said. Ms Short said the countries involved could not boast good records on their response to major disasters. Former Cabinet minister Clare Short has criticized the US-led tsunami aid coalition, saying the UN should be leading efforts.	Former Cabinet minister Clare Short has criticized the US-led tsunami aid coalition, saying the UN should be leading efforts. President Bush has announced that an alliance of the US, India, Australia and Japan will <u>co-ordinate</u> a humanitarian drive. I think this initiative from America to set up four countries claiming to <u>co-ordinate</u> sounds like yet another attempt to undermine the UN when it is the best system we have got	Ms Short also said that the US was very bad at coordination with anyone' and India Had its Own problems, Mr Short said. "But MsShort said The effect of a parallel coalition will be to undercut the UN," she said. Former cabinet minister Clare short has criticized THE US-Led tsunami aid Coalition, sayingThe UN should being leading efforts .

Figure 24. Summaries Comparison for General Summarizer Models

One of the difficulties we faced was the extended training time due to working with the entire summarization dataset containing about 6,000 examples. In conclusion, although Longformer had lower losses and better ROUGE scores than T5-base when reading the actual generated summary, the T5-base model made more sense. Additionally, the T5-base took less time to train, had more consistent outcomes (no crashes), and had faster inferences.

3.3 Specialized Summarizer Models

3.3.1 Entertainment Summarization

For the entertainment category we deployed the use of T5 and BART transformer models. Explored how their difference in architecture impacted training times and performance on downstream tasks in particular the summarization of entertainment articles.

In general, we found that the training times taken by BART were a lot longer than T5 which could be attributed to its bidirectional and auto-regressive nature. We found through our testing that BART's sequence-to-sequence architecture offered more freedom and flexibility when fine-tuning due to its prompt engineering capabilities that can be tailored to the specific use case and task.


Hyperparameter Fine-tuning

T5-base

```
training_args = TrainingArguments(  
    output_dir=OUT_DIR,  
    num_train_epochs=EPOCHS,  
    per_device_train_batch_size=BATCH_SIZE,  
    per_device_eval_batch_size=BATCH_SIZE,  
    warmup_steps=50,  
    weight_decay=0.01,  
    logging_dir=OUT_DIR,  
    logging_steps=10,  
    evaluation_strategy='steps',  
    eval_steps=200,  
    save_strategy='epoch',  
    save_total_limit=2,  
    report_to='tensorboard',  
    learning_rate=0.0003,  
    dataloader_num_workers=4  
)
```

Figure 25. Hyperparameters of T5-base

During our experimentation and testing, we tried various combinations of hyperparameters and ultimately found that the combination observed in the figure above yielded the best results. It is important to note that the batch size used was 4 and this was due to computational limitations. Any other batch size used would often lead to CUDA out-of-memory errors when training. Furthermore, due to our limited samples for our training dataset (832 samples), we found that a smaller number of warm-up steps helped to improve our performance in summarization and reduced training times. Additionally, we tried two learning rates 0.0001 and 0.0003 and we found that the higher learning rate improved the performance of the model when generating summaries.

 [1040/1040 27:20, Epoch 10/10]

Step	Training Loss	Validation Loss	Rouge1	Rouge2	RougeL	Gen Len
200	0.478800	0.234253	0.915700	0.852700	0.901800	191.010800
400	0.309000	0.200994	0.925700	0.866800	0.913300	191.204300
600	0.391700	0.185056	0.931900	0.878400	0.920400	191.204300
800	0.284900	0.175317	0.935700	0.885200	0.925000	191.204300
1000	0.273700	0.173083	0.937600	0.889600	0.927700	191.204300

 [12/12 00:06]

```
{'eval_loss': 0.17307457327842712,
 'eval_rouge1': 0.9376,
 'eval_rouge2': 0.8891,
 'eval_rougeL': 0.9275,
 'eval_gen_len': 191.2043,
 'eval_runtime': 23.5104,
 'eval_samples_per_second': 3.956,
 'eval_steps_per_second': 0.51,
 'epoch': 10.0}
```

Figure 26. T5-base training

Our validation and testing results indicated that our fine-tuned model was able to generate summaries with a high degree of unigram overlap between the generated output and the hand-written summary as observed by the higher Rouge 1 score. In contrast, the T5 Model had a lower degree of bigram overlap between the generated summary and the hand-written one which can be seen in the slightly lower score for the Rouge 2 metric as opposed to the Rouge 1 and Rouge L scores. We hypothesize this to be due to T5 using synonyms and thus fewer pairs match between the output and the hand-written summary. We also observed that the model performed extremely well in the Rouge L metric showcasing that the generated outputs were able to maintain the key points highlighted in the hand-written summary as it had a large degree of overlap in the longest common subsequences (LCS) between the generated output and the hand-written summary.

Hand-written Summary	T5 Base Model	T5 Fine-tuned
<p>Belle & Sebastian have been named the best Scottish band of all time after a three month-long public poll. Scottish magazine The List recently compiled a list of the top 50 Scottish bands of all time, but left the final decision to the public. Scottish bands from earlier musical eras also made it into the final list, including 1970s tartan boy band the Bay City Rollers and goth favourites the Jesus and Mary Chain. Scottish-based band Snow Patrol, who finished 14th in the vote and have been nominated for a pair of Brit Awards, were among the performers who covered well-known Scottish pop songs at the party on Wednesday night. BBC Radio Scotland presenter Vic Galloway, who has been involved in the project, said it had been "great fun" to look back at Scotland's musical heritage and take note of up-and-coming Scottish acts.</p>	<p>Belle & Sebastian have been named the best Scottish band of all time. the group beat Travis and Idlewild into second and third place respectively. other bands from earlier musical eras also made it into the final list.</p>	<p>best Scottish band' Belle & Sebastian have been named the best Scottish band of all time after a three month-long public poll. Scottish magazine The List recently compiled a list of the top 50 Scottish bands of all time, but left the final decision to the public. Scottish-based band Snow Patrol, who finished 14th in the vote and have been nominated for a pair of Brit Awards, were among the performers who covered well-known Scottish pop songs at the party on Wednesday night. Scottish band Snow Patrol, who finished 14th in the vote and have been nominated for a pair of Brit Awards, were among the performers who covered well-known Scottish pop songs at the party on Wednesday night. Scottish magazine The List recently compiled a list of the top 50 Scottish bands of all time, but left the final decision to the public.</p>

Table 2. Example Summarization Comparison between T5-base Fine-tuned model, and handwritten summary (Entertainment Summarization)

We provided both the base model and fine-tuned model in the same sample to observe the difference in the generated outputs and the handwritten summary for that sample. The base model generated a significantly shorter focus only on the top 3 finalist bands from the poll. In contrast, our fine-tuned model provided greater detail and had a summary with approximately the same length as the hand-written one. It can be observed that the fine-tuned model's output very closely resembled the hand-written summary and often used the same phrases present in the hand-written summary with no use of synonyms. Furthermore, our fine-tuned model output repeated two phrases twice in its generated summary.

	Metric	Score
0	ROUGE-1 Precision	0.811509
1	ROUGE-1 Recall	0.418409
2	ROUGE-1 F-measure	0.549271
3	ROUGE-2 Precision	0.661468
4	ROUGE-2 Recall	0.332101
5	ROUGE-2 F-measure	0.439809

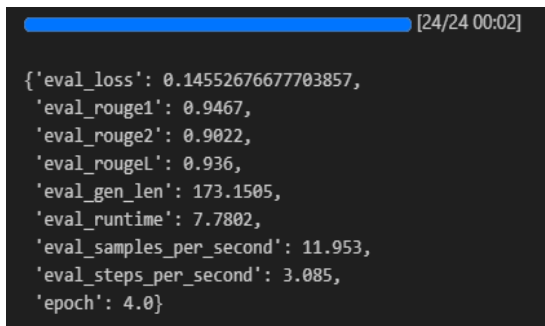
Figure 27. Training Framework for the T5 Model

We ran further testing inference on the fine-tuned T5 model picking 10 random samples from the test dataset, generating summaries for each sample, and calculating the average precision, recall, and F measure of the Rouge 1 and Rouge metrics across all 10 samples. Based on the scores observed in Figure 35 we can see that the model-produced summaries with a high degree of unigram overlaps between the hand-written summaries and the generated ones which is reflected in the high precision of the Rouge 1 metric. However, we can see that the model performed poorly in the Rouge 2 metrics, this could be from the issue we have seen earlier with the model repeating phrases in the generated summaries and therefore missing out on several unique bigrams that are present in the reference hand-crafted summaries thus explaining the lower scores for the Rouge 2 metrics.

Bart

Throughout our experimentation and rigorous testing process, we explored diverse configurations of hyperparameters for our BART model. After extensive trials, we identified that the parameter setup depicted in the figure above produced the most optimal outcomes. Notably, we constrained the batch size to 4, primarily due to computational constraints. Additionally, given the relatively modest size of our training dataset comprising 832 samples, we determined that a reduced number of warm-up steps was advantageous. This strategy not only enhanced the performance of our summarization tasks but also contributed to shorter training durations, effectively optimizing our workflow.

Training our BART model for 4 epochs yielded higher performance compared to the T5-base model. Despite the T5-base model's robust capabilities, our experimentation revealed that extended training epochs with BART led to more impressive results across various evaluation metrics.



[832/832 05:43, Epoch 4/4]						
Step	Training Loss	Validation Loss	Rouge1	Rouge2	RougeL	Gen Len
200	0.486300	0.195726	0.927000	0.870600	0.915700	173.150500
400	0.373800	0.180362	0.932600	0.876400	0.920200	173.150500
600	0.114400	0.162766	0.939300	0.888200	0.927300	173.150500
800	0.078500	0.145852	0.946600	0.901900	0.936000	173.150500

Figure 28. Training and Validation for the Bart Model

	Metric	Score
0	ROUGE-1 Precision	0.734517
1	ROUGE-1 Recall	0.398924
2	ROUGE-1 F-measure	0.516232
3	ROUGE-2 Precision	0.544366
4	ROUGE-2 Recall	0.293116
5	ROUGE-2 F-measure	0.380161

Figure 29. Training Framework for the Bart Model

The validation and testing of our fine-tuned model showed strong performance in generating summaries, particularly in terms of unigram overlap with hand-written summaries, as indicated by a higher Rouge 1 score. In comparison, the Bart Model had lower bigram overlap with hand-written summaries, reflected in a slightly lower Rouge 2 score. However, the evaluation scores are influenced by the generation function, which incorporates several parameters such as `num_beams`, `length_penalty`, and `early_stopping`. We attribute a significant role to `length_penalty` in effectively managing the varying lengths of articles within our dataset. Given the diverse lengths of articles in our dataset, calculating a specific value for the generated summary may not adequately capture performance variations across different cases.

Generated result

First Content 71% match	Second Content 88% match
<p>Superhero clichs</p> <p>The 45-year-old actor and stand up comedian played an alien who came to Earth to protect his planet, becoming a Bollywood star to blend₁</p> <p>in. But</p> <p>despite high anticipation for the film, which was directed by Oscar winner Chlo Zhao, it received a relatively poor 47% rating on Rotten Tomatoes and achieved the lowest score of any Marvel Cinematic Universe film in audience survey Cinema₁</p> <p>Score. Empire said at the time of release Chlo Zhao's entry into the superhero</p> <p>world is assured, ambitious and told on a dizzyingly cosmic scale - but even it can't escape the clichs of superhero₁</p> <p>storytelling. The</p> <p>Guardian's Peter Bradshaw wrote There are some nice touches and an attractive new diversity worn lightly, but this is an underpowered and uncertain₄</p> <p>film. Nanjiani and his</p> <p>wife, Emily V Gordon, were Oscar nominees in 2018 for their romantic comedy The Big Sick, in which he starred alongside Zoe₁</p> <p>Kazan. He's one of many stars who have spoken out about dealing with bad reviews in film and TV. Melissa McCarthy said in an interview that she confronted a critic at the Toronto Film Festival who had previously criticised her for her appearance in 2014 film₁</p> <p>Tammy. Jennifer Lawrence said in Variety that she gets defensive after reading negative reviews of her work. You're so in the zone, you put your whole soul and body, you move to shoot a₁</p> <p>movie, and you</p> <p>then love it, obviously because you wouldn't be there if you didn't love₁</p> <p>it, and then</p> <p>people just destroy₁</p> <p>it, she explained.</p>	<p>The 45-year-old actor and stand up comedian played an alien who came to Earth to protect his planet, becoming a Bollywood star to blend₁</p> <p>inBut</p> <p>despite high anticipation for the film, which was directed by Oscar winner Chlo Zhao, it received a relatively poor 47% rating on Rotten Tomatoes and achieved the lowest score of any Marvel Cinematic Universe film in audience survey Cinema₁</p> <p>ScoreThe comic</p> <p>world is assured, ambitious and told on a dizzyingly cosmic scale - but even it can't escape the clichs of superhero₁</p> <p>storytellingThe</p> <p>Guardian's Peter Bradshaw wrote There are some nice touches and an attractive new diversity worn lightly, but this is an underpowered and uncertain₄</p> <p>filmShe</p> <p>said in an interview that she confronted a critic at the Toronto Film Festival who had previously criticised her for her appearance in 2014 film₁</p> <p>TammyJiani and his</p> <p>wife, Emily V Gordon, were Oscar nominees in 2018 for their romantic comedy The Big Sick, in which he starred alongside Zoe₁</p> <p>KazanShe was so in the zone that</p> <p>you put your whole soul and body, you move to shoot a₁</p> <p>movie and you</p> <p>then love it, obviously because you wouldn't be there if you didn't love₁</p> <p>it and then</p> <p>people just destroy₁</p> <p>it she explained</p>

Figure 30. Fine-tuned Model Summarization

According to the result of the similarity comparison between the original article and the generated result from the bart-base fine-tuned model.

First Content 99% match	Second Content 99% match
<p>Superhero clichs The 45-year-old actor and stand up comedian played an alien who came to Earth to protect his planet, becoming a Bollywood star to blend in. But despite high anticipation for the film, which was directed by Oscar winner Chlo Zhao, it received a relatively poor 47% rating on Rotten Tomatoes and achieved the lowest score of any Marvel Cinematic Universe film in audience survey Cinema Score. Empire said at the time of release₂</p> <p>Chlo</p> <p>Zhao's entry into the superhero world is assured, ambitious and told on a dizzyingly cosmic scale - but even it can't escape the clichs of superhero storytelling. The Guardian's Peter Bradshaw wrote There are some nice touches and an attractive new diversity worn lightly, but this is an underpowered and uncertain film. Nanjiani and his wife, Emily V Gordon, were Oscar nominees in 2018 for their romantic comedy The Big Sick, in which he starred alongside Zoe Kazan. He's one of many stars who have spoken out about dealing with bad reviews in film and TV. Melissa McCarthy said in an interview that she confronted a critic at the Toronto Film Festival who had previously criticised her for her appearance in 2014 film Tammy. Jennifer Lawrence said in Variety that she gets defensive after reading negative reviews of her work. You're so in the zone, you put your whole soul and body, you move to shoot a movie, and you then love it, obviously because you wouldn't be there if you didn't love₁</p> <p>it, and then</p> <p>people just destroy it, she explained.₃</p>	<p>Superhero clichs The 45-year-old actor and stand up comedian played an alien who came to Earth to protect his planet, becoming a Bollywood star to blend in. But despite high anticipation for the film, which was directed by Oscar winner Chlo Zhao, it received a relatively poor 47% rating on Rotten Tomatoes and achieved the lowest score of any Marvel Cinematic Universe film in audience survey Cinema Score. Empire said at the time of release₂</p> <p>Chloek</p> <p>Zhao's entry into the superhero world is assured, ambitious and told on a dizzyingly cosmic scale - but even it can't escape the clichs of superhero storytelling. The Guardian's Peter Bradshaw wrote There are some nice touches and an attractive new diversity worn lightly, but this is an underpowered and uncertain film. Nanjiani and his wife, Emily V Gordon, were Oscar nominees in 2018 for their romantic comedy The Big Sick, in which he starred alongside Zoe Kazan. He's one of many stars who have spoken out about dealing with bad reviews in film and TV. Melissa McCarthy said in an interview that she confronted a critic at the Toronto Film Festival who had previously criticised her for her appearance in 2014 film Tammy. Jennifer Lawrence said in Variety that she gets defensive after reading negative reviews of her work. You're so in the zone, you put your whole soul and body, you move to shoot a movie, and you then love it, obviously because you wouldn't be there if you didn't love₁</p> <p>it. And then</p> <p>people just destroy it, she explained.₃</p>

Figure 31. Bart Base Model Summarization

The following result is the similarity comparison between the original article and the generated result from the bart-base model. There are no summarizations because they are the same

3.3.2.1 More Exploration from the model

First Content 97% match	Second Content 44% match
<p>The 45-year-old actor and stand up comedian played an alien who came to Earth to protect his planet, becoming a Bollywood star to blend inBut despite high anticipation for the film, which was directed by Oscar winner Chlo Zhao, it received a relatively poor 47% rating on Rotten Tomatoes and achieved the lowest score of any Marvel Cinematic Universe film in audience survey</p> <p>Cinema₁</p> <p>ScoreShe</p> <p>said in an interview that she confronted a critic at the Toronto Film Festival who had previously criticised her for her appearance in 2014 film₂</p> <p>Tammy</p>	<p>The 45-year-old actor and stand up comedian played an alien who came to Earth to protect his planet, becoming a Bollywood star to blend inBut despite high anticipation for the film, which was directed by Oscar winner Chlo Zhao, it received a relatively poor 47% rating on Rotten Tomatoes and achieved the lowest score of any Marvel Cinematic Universe film in audience survey</p> <p>Cinema₁</p> <p>ScoreThe comic world is assured, ambitious and told on a dizzyingly cosmic scale - but even it can't escape the clichés of superhero storytellingThe Guardian's Peter Bradshaw wrote There are some nice touches and an attractive new diversity worn lightly, but this is an underpowered and uncertain filmShe said in an interview that she confronted a critic at the Toronto Film Festival who had previously criticised her for her appearance in 2014 film₂</p> <p>TammyJiani and his wife, Emily V Gordon, were Oscar nominees in 2018 for their romantic comedy The Big Sick, in which he starred alongside Zoe KazanShe was so in the zone that you put your whole soul and body, you move to shoot a movie and you then love it, obviously because you wouldn't be there if you didn't love it and then people just destroy it she explained</p>

Figure 32. High Length_penalty VS Low_penalty

In the generation function, the parameter length penalty and early stopping affect the most compared to the result given. When a high length_penalty is applied, the resulting summary tends to be shorter. This is because the model prioritizes focusing more on the main points or targets mentioned in the first paragraph. The length_penalty parameter influences the balance between generating longer, more detailed summaries versus shorter, more concise ones. With a high length_penalty, the model tends to favor brevity and may emphasize capturing the essential information from the initial paragraph rather than elaborating on additional details.

3.3.3 Business Summarization

In the realm of business summarization for this project, we employed two distinct models: BART and T5. This dual-model strategy enhanced the robustness of our summarization efforts, providing a comprehensive perspective on the efficiency and effectiveness of different architectures in the context of business-oriented text summarization.

The ‘facebook/bart-base’ Model

The Facebook/bart-base model was trained on an available Google Colab V100 GPU to utilize its computational efficiency. It can significantly speed up training times for complex neural networks. Hyperparameters were carefully selected to optimize training outcomes:

Hyperparameter	
device	cuda if torch.cuda.is_available() else cpu
model_name	facebook/bart-base
learning_rate	0.0002
weight_decay	0.01
epsilon	0.0
num_warmup_steps	50
num_training_steps	len(train_dataloader) * 10
early_stopping_rounds	2
accumulation_steps	20
batch_size	4
max_length	512
optimizer	AdamW
scheduler	get_linear_schedule_with_warmup
scaler	GradScaler
clip_grad_norm_	clip_grad_norm_
epoch	10

Figure 33. Training Framework for BART (Business Summarization)

Facebook/bart-base model Optimization and Results

Throughout the project, the model underwent several optimization cycles, with fine-tuning of the learning rate and an increase in the number of training epochs, leading to enhanced summarization performance.

Learning Rate of 5e-5: At this intermediate learning rate, the training losses across epochs plateaued around 0.66 to 0.64, denoting a consistent performance.

Learning Rate of 2e-5: With this lower learning rate, the model's training loss improved significantly, reaching a loss near 0.7, suggesting an enhanced ability to capture the nuances of the business texts.

Learning Rate of 2e-4: A decision was made to revert to the initial learning rate of 2e-4. When coupled with an increased number of epochs from 3 to 10, this learning rate resulted in the best performance, achieving the lowest training loss and thereby indicating the most effective summarization capability.

The final configuration of a learning rate of 2e-4 over 10 epochs led to a training loss of approximately 0.5074, which was determined to be the optimal outcome of our training efforts.

Facebook/bart-base model Rouge Metric Assessment: The evaluation results, measured by the Rouge metric, were as follows:

Downloading builder script:		
[5]:		
	Metric	Score
0	ROUGE-1 Precision	0.688022
1	ROUGE-1 Recall	0.413103
2	ROUGE-1 F-measure	0.513980
3	ROUGE-2 Precision	0.486871
4	ROUGE-2 Recall	0.296194
5	ROUGE-2 F-measure	0.366684

Figure 34. Training Framework for T5 (Business Summarization)

These scores illustrate the model's competency in capturing key terms and phrases from the original texts, albeit with room for refinement in encompassing the entirety of the source content.

The T5-base Model

The T5-base model underwent training on a Google Colab platform, utilizing an A-100 GPU to expedite computational efficiency. The selection of specific training parameters aimed to harness the full potential of the model, optimizing its performance during the training process. The strategic use of the A-100 GPU not only facilitated faster computations but also contributed to enhancing the overall training efficiency of the T5-base model, allowing for more comprehensive and effective model exploration.

```

training_args = TrainingArguments(
    output_dir=OUT_DIR,
    num_train_epochs=EPOCHS,
    per_device_train_batch_size=BATCH_SIZE,
    per_device_eval_batch_size=BATCH_SIZE,
    warmup_steps=50,
    weight_decay=0.01,
    logging_dir=OUT_DIR,
    logging_steps=10,
    evaluation_strategy='steps',
    eval_steps=200,
    save_strategy='epoch',
    save_total_limit=2,
    report_to='tensorboard',
    learning_rate=0.0003,
    dataloader_num_workers=4
)

```

Figure 35. Training Framework for T5 (Business Summarization)

Moreover, small batch sizes of 4 were used, a maximum of 10 epochs for a maximum length of 512 for tokenized word sequences.

[2460/2460 1:02:53, Epoch 10/10]						
Step	Training Loss	Validation Loss	Rouge1	Rouge2	RougeL	Gen Len
200	0.132500	0.309195	0.915000	0.858400	0.899200	216.910600
400	0.079500	0.320744	0.917700	0.862500	0.902000	216.910600
600	0.051400	0.331799	0.921100	0.868900	0.905200	216.910600
800	0.064900	0.326689	0.922800	0.873200	0.907200	216.910600
1000	0.064200	0.339646	0.925500	0.877600	0.909700	216.910600
1200	0.052800	0.342000	0.926700	0.880800	0.911400	216.910600
1400	0.025800	0.351525	0.927800	0.882100	0.911900	216.910600
1600	0.045100	0.358839	0.928200	0.883700	0.912600	216.910600
1800	0.029900	0.364549	0.928400	0.883800	0.912900	216.910600
2000	0.020700	0.367140	0.929800	0.885800	0.914200	216.910600
2200	0.018500	0.370888	0.929400	0.886100	0.914000	216.910600
2400	0.023400	0.371883	0.929600	0.886300	0.914200	216.910600

```

{'eval_loss': 0.37215113639831543,
 'eval_rouge1': 0.9295,
 'eval_rouge2': 0.8863,
 'eval_rougeL': 0.9142,
 'eval_gen_len': 216.9106,
 'eval_runtime': 64.5482,
 'eval_samples_per_second': 3.811,
 'eval_steps_per_second': 0.961,
 'epoch': 10.0}

```

Figure 36. Training and Validation Results for T5 (Business Summarization)

Overall, the model demonstrated positive training trends, with decreasing loss values and improving Rouge scores, indicating enhanced summarization performance over the training iterations.

It appears that convergence begins to take place around the 2000th training step. Beyond this point, the training and validation loss values remain relatively steady, with only minor fluctuations. The stabilized loss values indicate that the model has reached a point where further training iterations are yielding diminishing returns, suggesting that the model has learned the underlying patterns in the data.

BART vs. T5 for Business Summarization: A comparative analysis between BART and T5 models revealed that T5 outperformed BART. This conclusion is drawn from the observed higher Rouge scores and lower training losses exhibited by the T5 model. Higher Rouge scores, encompassing measures like Rouge1, Rouge2, and RougeL, reflect more accurate and relevant summarization. Additionally, lower training losses signify a better convergence of the T5 model during the training process. The findings suggest that T5 is more effective in distilling essential information from business-related text, showcasing its superior capabilities in generating high-quality summaries when compared to BART.

3.3.4 Sports Summarization

To further explore the sports news dataset, we used two pre-trained models: BART-base and T5-base. We aimed to compare the performance of two models in generating summarizations for sports new articles. In our dataset, we observed a wide range of distribution of the article lengths and summary lengths. This brought challenges to the settings for the hyperparameters and related thresholds. During the training attempts, we observed variations of the rouge scores of the prediction based on the same input articles and brought that to the comparison.

The BART-base Model

The optimizer was initialized with a learning rate of $2e-5$ and weight decay of 0.01 to optimize the model parameters during training. Additionally, a learning rate scheduler was configured to adjust the learning rate linearly from 0 during the warm-up phase, and then linearly decrease it over the remaining training steps. Early stopping is implemented to prevent overfitting by halting training if there's no improvement in performance for a specified number of consecutive epochs. The best ROUGE score achieved during training is tracked to monitor performance, starting with an initial value of -1 to ensure any obtained score is considered an improvement. The choice of max length was based on the observation of the data distribution in our dataset. We chose to undersample to prevent overfitting. Finally, current_round is initialized to track the current training round, serving as an epoch counter. Together, these components establish a comprehensive training environment for the model, facilitating effective learning and performance monitoring.

Hyperparameters	Model	Max_length	Epoch	Batch size	Learning rate	Weight_decay
BART-base	BART-base	155	3	4	$2e-5$	0.01

Figure 37. Training Parameters for BART (Sports Summarization)

The following table shows an example of the summarization comparison between the base model and the fine-tuned model.

Base (88% match)	Fine-Tuned (97% match)
<p>isinbayeva claims new world best pole vaulter yelena isinbayeva broke her own indoor world record by clearing 489 metres in lievin on saturday it was the russians 12th world record of her career and came just a few days after she cleared 488m at the norwich union grand prix in birmingham the olympic champion went on to attempt 505m at a meeting on france but failed to clear that height in the mens 60m former Olympic 100m champion maurice greene could only finish second to leonard scott it was greenes second consecutive defeat at the hands of his fellow american who also won in Birmingham last week i ran my race perfectly said sc</p>	<p>isinbayeva claims new world best pole vaulter yelena isinbayeva broke her own indoor world record by clearing 489 metres in lievin on saturday it was the russians 12th world record of her career and came just a few days after she cleared 488m at the norwich union grand prix in birmingham the olympic champion went on to attempt 505m at the meeting on france but failed to clear that height in the mens 60m former olympic 100m champion maurice greene could only finish second to leonard scott it was greenes second consecutive defeat at the hands of his fellow american who also won in birmingham last week i ran my race perfectly said scott who won in 646secs his best time indoors i am happy even if i know that maurice is a long way from being at his peak at the start of the season</p>

Table 3. Example Summarization Comparison between BART-base with Fine-tuned model (Sports Summarization)

For the model evaluation:

- ROUGE-1 : 0.6059322033898306
- ROUGE-2: 0.7489361702127659

The T5-base Model

In the T5 model, the hyperparameter tuning was done with *TrainingArguments* from the Hugging Face Transformers library. The following figure shows the parameters that govern various aspects of the training process, such as the directory for output and logging, the number of training epochs (8), batch sizes set as 4 for both training and evaluation, settings for the learning rate scheduler including warm-up steps and weight decay, logging intervals, evaluation strategy, and checkpoint saving strategy. These settings enable control over the training procedure and operate optimization of training efficiency and performance.

```

training_args = TrainingArguments(
    output_dir=OUT_DIR,
    num_train_epochs=EPOCHS,
    per_device_train_batch_size=BATCH_SIZE,
    per_device_eval_batch_size=BATCH_SIZE,
    warmup_steps=50,
    weight_decay=0.01,
    logging_dir=OUT_DIR,
    logging_steps=10,
    evaluation_strategy='steps',
    eval_steps=200,
    save_strategy='epoch',
    save_total_limit=2,
    report_to='tensorboard',
    learning_rate=0.0002,
    dataloader_num_workers=4
)

```

Figure 38. Training Parameters for T5 (Sports Summarization)

The following log table represents the training progress of a text summarization model over multiple steps or epochs. Each row corresponds to a specific step during training, with associated metrics recorded for analysis. A decrease in training loss as expected indicates the model learned to fit the training data better. Then it went up could indicate a slight overfitting was occurring at the last 400 steps. However, the constantly decreased validation loss, in this case, suggests that the model is learning to generalize well to unseen data despite the increase in training loss. This scenario indicates that the model is effectively capturing the underlying patterns in the data and is not solely memorizing the training examples. The increase in rouge score also supports that the model is improving. The overall rouge scores: rouge-1: 0.8491, and rouge-2: 0.7449, give fair scores on how well the generated summaries align with the reference summaries in terms of word overlap at both the unigram and bigram levels.

[1896/1896 11:56, Epoch 8/8]						
Step	Training Loss	Validation Loss	Rouge1	Rouge2	RougeL	Gen Len
200	1.042700	1.029591	0.773100	0.635300	0.732600	185.452800
400	0.787100	0.919268	0.786200	0.653900	0.751000	185.452800
600	0.862000	0.842314	0.799100	0.668300	0.765100	185.452800
800	0.551000	0.771776	0.813400	0.687100	0.782400	185.452800
1000	0.392800	0.732438	0.825000	0.704700	0.797200	185.452800
1200	0.467600	0.687911	0.834300	0.718300	0.808600	185.452800
1400	0.366600	0.658650	0.840400	0.729900	0.817300	185.452800
1600	0.420300	0.639894	0.845500	0.739400	0.824100	185.452800
1800	0.512100	0.628516	0.849200	0.745100	0.829000	185.452800

```

[27/27 00:02]
{
  'eval_loss': 0.6282171010971069,
  'eval_rouge1': 0.8491,
  'eval_rouge2': 0.7449,
  'eval_rougeL': 0.8286,
  'eval_gen_len': 185.4528,
  'eval_runtime': 4.6893,
  'eval_samples_per_second': 22.605,
  'eval_steps_per_second': 5.758,
  'epoch': 8.0
}

```

Figure 39. Training and Validation Results for T5 (Sports Summarization)

The BART-base Model vs. T5-base Model for Sports News

When looking at them side by side, T5 came out on top compared to BART. This was evident from the higher Rouge scores and lower training losses seen in the T5 model. These Rouge scores indicate that T5's summaries were more precise and on point. Also, the lower training losses suggest that T5's training process went more smoothly. In summary, T5 is better at extracting important info from sports-related news articles and making top-notch summaries compared to BART.

Implications

Both BART-base and T5 models are powerful transformer architectures with their strengths and applications. Due to the wide distribution of our dataset and its unbalances, the accuracy for both models doesn't appear to be close to perfect. With the same input article, the generated summaries with two models performed differently as an example was provided below. For our dataset, the T5-base model performs better than the BART-base model in terms of the rouge scores; however, in readability, both of the models deliver good results even though different models have their advantages. Therefore, the choice of model for the task we were given is subjective to the dataset that was used for training.

BART (71% match)	T5 (91% match)
Kewell 26 had hoped to play against Fulham last weekend and to be ready for the Champions League tie against Bayer Leverkusen on 22 FebruaryMandic continued in the Sydney Morning Herald Harrys struggling a bit but the club are doing everything they can to get him rightHarry was desperate to be fit for the Fulham game as part of his plan to play himself in for the Leverkusen match hes had scans which dont show anything substantial but despite his best efforts at the moment he simply cant shake off the discomfort and get back on the park mandic	kewell waits as injury drags on liverpool winger harry kewell is struggling to recover from his achilles tendinitis problem and may not recover until march claims his agent kewell 26 had hoped to play against fulham last weekend and to be ready for the champions league tie against bayer leverkusen on 22 february but bernie mandic said hes had scans which dont show anything substantial but despite his best efforts at the moment he simply cant shake off the discomfort and get back on the park mandic continued in the sydney morning herald harrys struggling a bit but the club are doing everything they can to get him right harry was desperate to be fit for the fulham game as part of his plan to play himself in for the Leverkusen match kewell has not played since december 19 and misses out on international duty this week with australia facing south africa in durban on wednesday

Table 4. Example Summarization Comparison between BART-base with T5 model (Sports Summarization)

4.0 Conclusions

The classification and summarization of news data have emerged as invaluable tools in information management, catering to the need for staying well-informed in today's dynamic world. This report delves into the design decisions and scientific experiments behind the inner workings of such tools and underscores their crucial role in leveraging data for actionable insights.

Our findings indicate that while there is a level of parity among the models for classification tasks, the T5 model shows a slight edge in summarization tasks, particularly in the business news category. This suggests that the choice of model architecture and training regimen can be crucial to the quality of the summarization output.

As we conclude this project, we recognize the dynamic nature of deep learning and NLP technologies and their continuous evolution. The success of our project serves as a testament to the potential of these technologies to transform the way we interact with the ever-growing expanse of digital news.

Future work in this domain can expand upon our methodologies to include real-time news processing and summarization, catering to the instantaneous nature of news dissemination in the digital era. The ultimate goal remains to deliver succinct, relevant, and personalized news content to end-users, thereby enriching the landscape of information consumption in our digitally-driven society.

In closing, the utilization of pre-trained deep learning models forms a robust foundation for this project, promising a positive impact on information management and, consequently, on the broader societal landscape.

5.0 References

- BART*. (n.d.). Huggingface.Co. Retrieved February 10, 2024, from https://huggingface.co/docs/transformers/en/model_doc/bart
- DataScience_FP*. (n.d.). Retrieved February 10, 2024, from https://github.com/reddzzz/DataScience_FP
- facebook/bart-base · Hugging Face*. (n.d.). Huggingface.Co. Retrieved February 10, 2024, from <https://huggingface.co/facebook/bart-base>
- Harinatha, S. R. K., Tasara, B. T., & Qomariyah, N. N. (2021). Evaluating Extractive Summarization Techniques on News Articles. *2021 International Seminar on Intelligent Technology and Its Applications (ISITIA)*.
- Issa, A. (2023, January 26). *Transformer, GPT-3, GPT-J, T5 and BERT*. - Ali Issa. Medium. <https://aliissa99.medium.com/transformer-gpt-3-gpt-j-t5-and-bert-4cf8915dd86f>
- Jorge, L. (2023, March 30). *RoBERTa vs. GPT: A comprehensive comparison of state-of-the-art language models, with expert insights from CronJ*. Medium. <https://medium.com/@livajorge7/roberta-vs-86ee82a44969>
- Longformer*. (n.d.). Huggingface.Co. Retrieved February 10, 2024, from https://huggingface.co/docs/transformers/model_doc/longformer
- Misra, R. (2022). *News Category Dataset* [Data set].
- multi_news · Datasets at Hugging Face*. (n.d.). Huggingface.Co. Retrieved February 10, 2024, from https://huggingface.co/datasets/multi_news/viewer/default/train
- News Summarization*. (2022). [Data set].
- T5*. (n.d.). Huggingface.Co. Retrieved February 10, 2024, from https://huggingface.co/docs/transformers/model_doc/t5