ECE1786 FINAL REPORT

SUM(Text)

December 13, 2022

Yichen Zhang 1008653967 | Zijie Zhao 1003521478

Total Words Count: 1977/2000
Penalty: 0%

Table of Contents

Permissions

Team Member: **Yichen Zhang**

permission to post video: wait till see video
permission to post final report: yes
permission to post source code: yes

Team Member: **Zijie Zhao**

permission to post video: wait till see video
permission to post final report: yes
permission to post source code: yes

1. Introduction

When you open news sites, do you just start reading every news article? Probably not. People typically glance at the short news summary and then read more details if interested. Short, concise and informative summaries can help to select preferred news articles more efficiently and accurately. However, manual text summarization is a time-expensive task. The automation of the summarization task has been gaining increasing popularity.

This project aims to implement and experiment with various models that can make automatic text summarization specifically on news articles to provide concise and accurate news summaries to people.

2. Illustration / Figure

The figure below illustrates the overall project idea. Details are provided in the following sections.

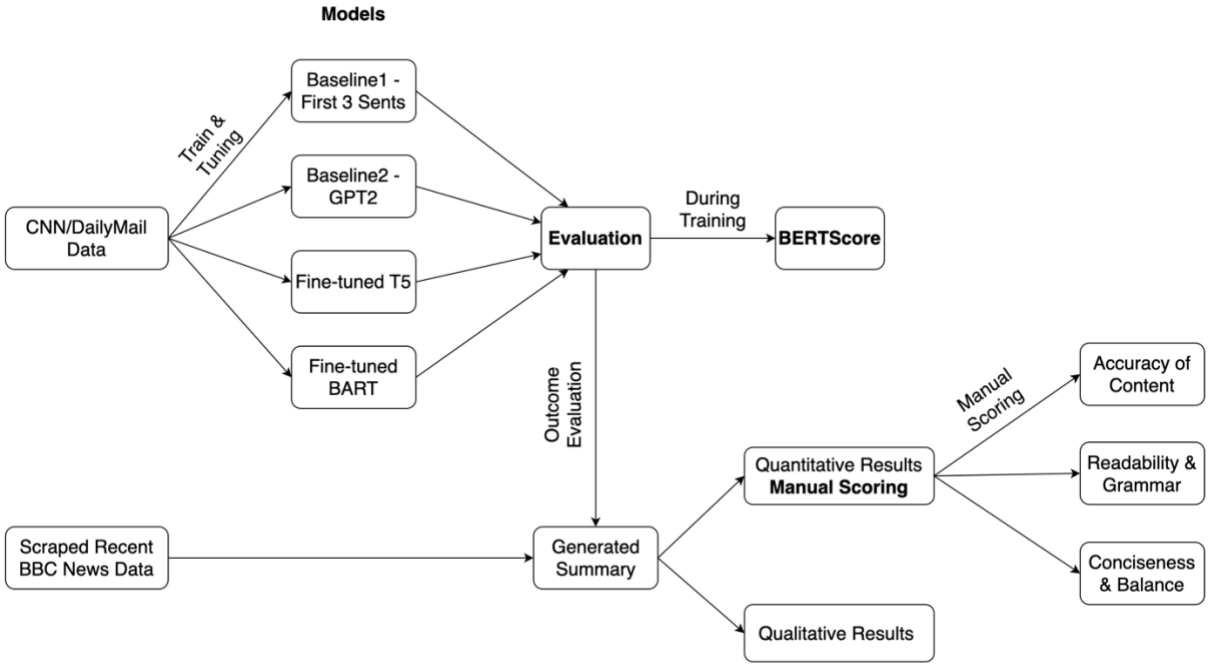

Figure 1. Project Illustration

3. Background & Related Work

The amount of text data available from various sources has exploded in the big data era. This volume of text is an inestimable source of information and knowledge which needs to be effectively summarized to be useful. This increasing availability of documents has demanded exhaustive research in the NLP area for automatic text summarization.

In general, there are two different approaches for automatic text summarization: extraction and abstraction, as shown in Figure 2 below.

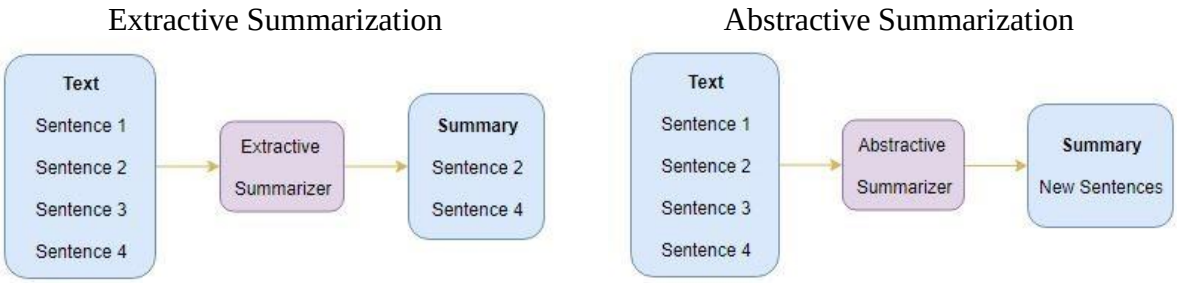

Figure 2. Major Summarization Methods Illustration

Recent studies have applied deep learning in extractive summarization. Yong Zhang[1] proposed a document summarization framework based on convolutional neural networks to learn sentence features and perform sentence ranking jointly using a CNN model for sentence ranking.

Abstractive summarization methods aim at producing summaries by interpreting the text using advanced natural language techniques to generate a new shorter text. An example is Liu et al.[2], whose work proposes an adversarial framework to jointly train a generative model and a discriminative model. In the framework, a generative model takes the original text as input and generates the summary using reinforcement learning to optimize the generator for a highly rewarded summary. Further, a discriminator model tries to distinguish the ground truth summaries from the generated summaries by the generator.

4. Data & Data Processing

The data for this project contains two main parts. The first part contains datasets which are used to train the models. The datasets are subsets from the CNN/DailyMail dataset by taking portions from the original training, validation, and test dataset. The amount for the training, validation, and test dataset is shown in Figure 3 below. For each instance, an input article and a reference summary would be used in the training process.

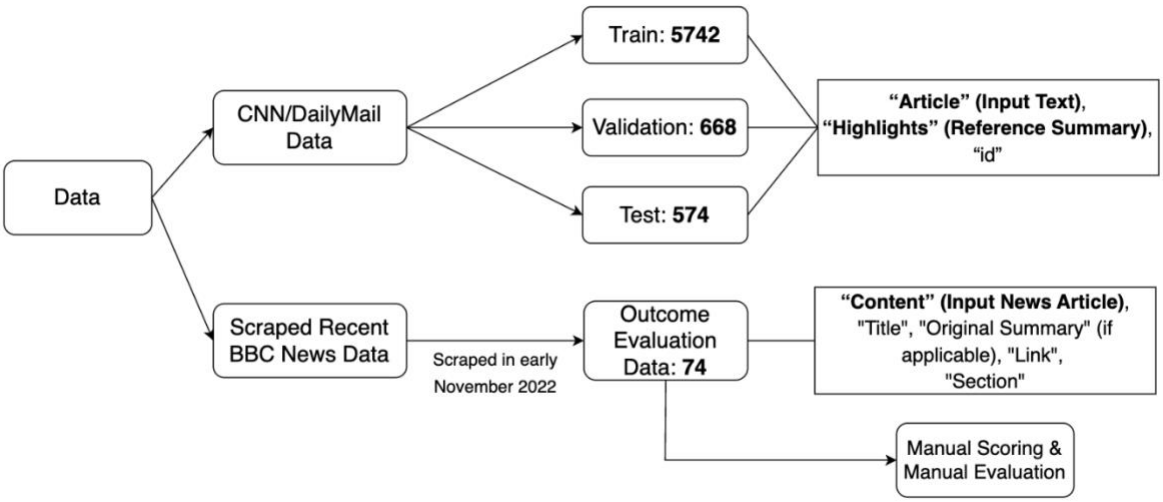

Figure 3. Data Illustration

The other part of the data contains scraped recent news data from BBC. The intention of using this data is to evaluate the performance of the models. There are 74 data instances as shown in Figure 2 and the data contains BBC news articles with topics within the business, technology, science & environment, world, stories, entertainment & arts. The summarization results on the recent news dataset will be evaluated manually by the team members on three dimensions (Accuracy of content, Readability & Grammar, and Conciseness & Balance). Each of the criteria will have a score from 1 to 3 based on the results. Details are covered in section 7 of the report.

The team applied regular expressions to clean the data. Special characters as well as some punctuations are substituted or removed. Figure 4 below shows an example of the cleaned data.

id	text	reference_summary
24521a2abb2e1f5e34e6824e0f9e56904a2b0e88	washington doctors removed five small polyps from president bush s colon on saturday, and none appeared worrisome, a white house spokesman said. the polyps were removed and sent to the national naval medical center in bethesda, maryland, for routine microscopic examination, spokesman scott stanzel said. results are expected in two to three days. all were small, less than a centimeter half an inch in diameter, he said. bush is in good humor, stanzel said, and will resume his activities at camp david. during the procedure vice president dick cheney assumed presidential power. bush reclaimed presidential power at 9 21 a.m. after about two hours. doctors used monitored anesthesia care, stanzel said, so the president was asleep, but not as deeply unconscious as with a true general anesthetic. he spoke to first lady laura bush who is in midland, texas, celebrating her mother s birthday before and after the procedure, stanzel said. afterward, the president played with his scottish terriers, barney and miss beazley, stanzel said. he planned to have lunch at camp david and have briefings with national security adviser stephen hadley and white house chief of staff josh bolten, and planned to take a bicycle ride saturday afternoon. cheney, meanwhile, spent the morning at his home on maryland s eastern shore, reading and playing with his dogs, stanzel said. nothing occurred that required him to take official action as president before bush reclaimed presidential power. the procedure was supervised by dr. richard tubb, bush s physician, and conducted by a multidisciplinary team from the national naval medical center in bethesda, maryland, the white house said. bush s last colonoscopy was in june 2002, and no abnormalities were found, white house spokesman tony snow said. the president s doctor had recommended a repeat procedure in about five years. a colonoscopy is the most sensitive test for colon cancer, rectal cancer and polyps, small clumps of cells that can become cancerous, according to the mayo clinic. small polyps may be removed during the procedure. snow said on friday that bush had polyps removed during colonoscopies before becoming president. snow himself is undergoing chemotherapy for cancer that began in his colon and spread to his liver. watch snow talk about bush s procedure and his own colon cancer . the president wants to encourage everybody to use surveillance, snow said. the american cancer society recommends that people without high risk factors or symptoms begin getting screened for signs of colorectal cancer at age 50. e mail to a friend .	five small polyps found during procedure none worrisome, spokesman says . president reclaims powers transferred to vice president . bush undergoes routine colonoscopy at camp david .

Figure 4. Cleaned Data Example

5. Architecture & Software

The best models the team has built are a fine-tuned T5-base model and a fine-tuned BART-base model. The reason for choosing these two models is that both models are sequence-to-sequence transformer models and are well-known models which can perform text generation tasks including summarization tasks.

T5 is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks and for which each task is converted into a text-to-text format [3].

BART is a denoising autoencoder for pretraining sequence-to-sequence models. BART is trained by (1) corrupting text with an arbitrary noising function, and (2) learning a model to reconstruct the original text. It uses a standard Tranformer-based neural machine translation architecture which, despite its simplicity, can be seen as generalizing BERT (due to the bidirectional encoder), GPT (with the left-to-right decoder), and many other more recent pretraining schemes [4].

Figure 5 below shows the training process of the models. The team has used pre-trained T5 and BART models, fine-tuned them, and trained them to perform news article summarization tasks.

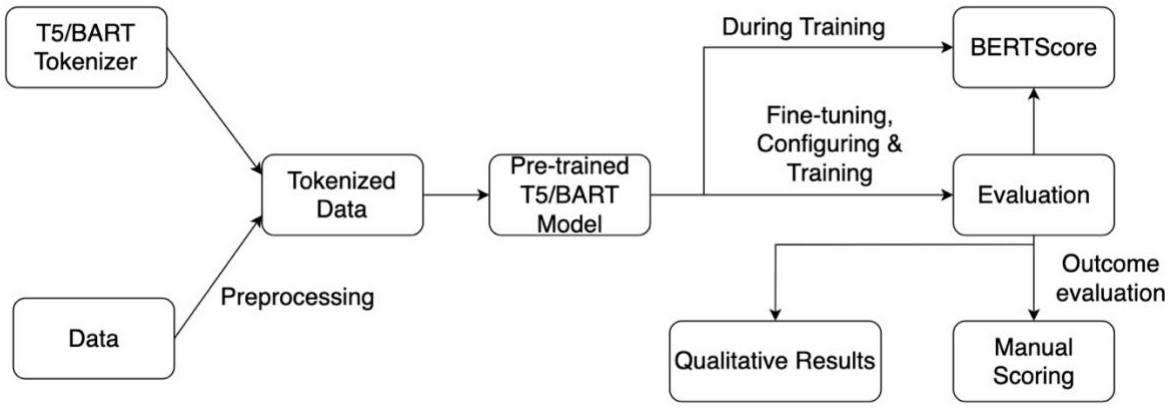

Figure 5. Models Illustration

Table 1 below presents the details of the model configurations.

	T5	BART
Configuration	"d_ff": 3072, "d_kv": 64, "d_model": 768, "decoder_start_token_id": 0, "dense_act_fn": "relu", "dropout_rate": 0.1, "eos_token_id": 1, "feed_forward_proj": "relu", "initializer_factor": 1.0, "is_encoder_decoder": true, "is_gated_act": false, "layer_norm_epsilon": 1e-06, "model_type": "t5", "n_positions": 512, "num_decoder_layers": 12, "num_heads": 12, "num_layers": 12, "output_past": true, "pad_token_id": 0, "relative_attention_max_distance": 128, "relative_attention_num_buckets": 32,	"activation_dropout": 0.1, "activation_function": "gelu", "add_bias_logits": false, "add_final_layer_norm": false, "architectures": ["BartForConditionalGeneration"], "attention_dropout": 0.1, "bos_token_id": 0, "classif_dropout": 0.1, "classifier_dropout": 0.0, "d_model": 768, "decoder_attention_heads": 12, "decoder_ffn_dim": 3072, "decoder_layerdrop": 0.0, "decoder_layers": 6, "decoder_start_token_id": 2, "dropout": 0.1, "early_stopping": true, "encoder_attention_heads": 12, "encoder_ffn_dim": 3072, "encoder_layerdrop": 0.0, "encoder_layers": 6, "eos_token_id": 2, "forced_bos_token_id": 0, "forced_eos_token_id": 2, "gradient_checkpointing": false,
# of Parameters	About 222 million	About 140 million

Table 1. Model Key Information

6. Baseline Model

The team has implemented two baseline models, which are the extractive summarization baseline model and the abstractive summarization baseline model. The extractive summarization baseline model uses the method of taking the first three sentences from the input article as the summary. It makes sense and the results are reasonable as the "topic" sentences tend to appear at the beginning of news articles.

The abstractive summarization model is a finetuned GPT-2. The team leveraged the pre-trained GPT2 model transformer with a language modelling head on top and finetuned the model on the summarization task. The team modified the input data structure as "article" + "<summarize>" +

"summary", and targets are the same as inputs, making the model learn the summary after processing the <summarize> token.

7. Quantitative Results

The metric the team used during the model-training stage is BERTScore, which was introduced by Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi [5]. BERTScore is an automatic metric that can measure the semantical similarity between the reference summary and the generated summary. Figure 6 shows the details of how BERTScore works. It uses pre-trained BERT contextual embedding and computes similarity scores (F1, Precision, and Recall) as shown in Figure 6. A higher BERTScore value indicates higher semantical similarity scores between the reference summary and the generated summary, which is desirable. Monitoring BERTScore during the training process allows the team to detect unexpected behaviours. For example, when the model is generating outputs that are unrelated to the input text, the BERTScore is expected to be low. BERTScore is crucial for the training process and can provide some insights in terms of models' performance.

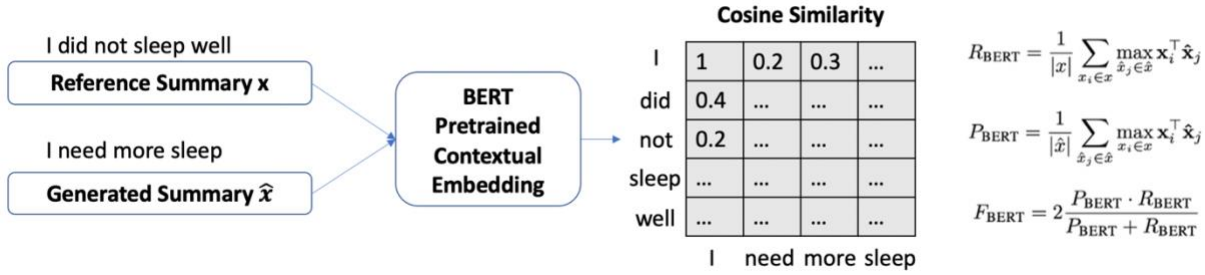

Figure 6. BERTScore Illustration

Table 2 below shows the results of the test dataset for each model in terms of BERTScore. It can be observed that T5 and BART models outperform baseline models in terms of BERTScore F1, Precision, and Recall. Although BERTScore is helpful during the training process, it cannot evaluate the models' performance in some aspects such as grammar correctness and conciseness of the generated summaries. Therefore, the team introduce manual scoring to evaluate the models' performance.

	BERTScore - Mean			**BERTScore - Median**		
Model	F1	Precision	Recall	F1	Precision	Recall
Baseline-First 3 Sentences	0.855	0.841	0.870	0.854	0.839	**0.871**
Baseline-GPT2	0.864	0.869	0.858	0.863	0.869	0.856
T5	0.870	0.878	0.851	0.870	0.883	0.857
BART	**0.875**	**0.891**	**0.860**	**0.874**	**0.892**	0.859

Table 2. Model performance on the test dataset

As mentioned in section 3, the team has scraped some recent news articles from BBC to evaluate the outcome of the project. The team members had given manual scores based on human judgement to the generated summaries. Figure 7 below shows how the manual scoring process

was conducted. Figure 7 also introduces the three dimensions that the team evaluates the performance of the models as well as the scoring criteria.

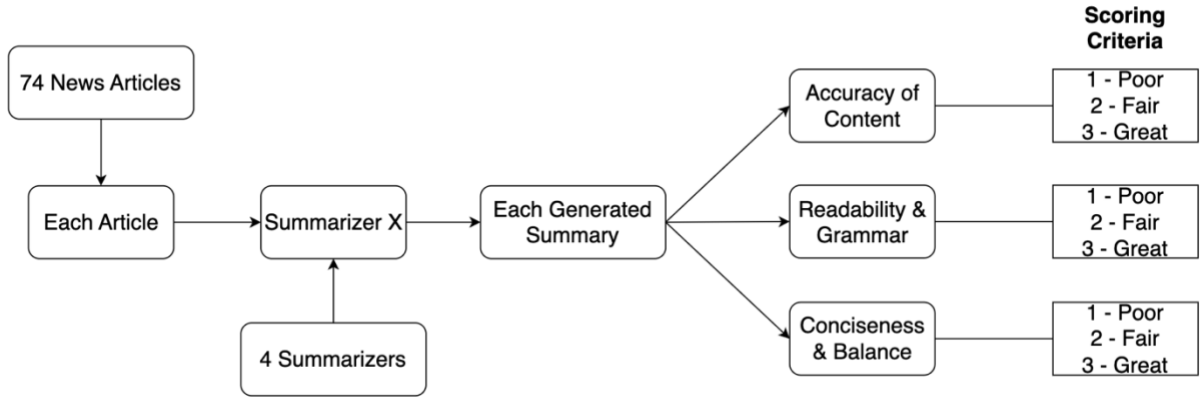

Figure 7. Manual Scoring Illustration

Both team members have contributed to the manual scoring process. Figure 8 shows a portion of the manual scoring table.

Article	Baseline_First3				Baseline_GPT2				T5				BART			
	Output	Accuracy of content	Readability & Grammatical Correctness	Conciseness & Balance	Output	Accuracy of content	Readability & Grammatical Correctness	Conciseness & Balance	Output	Accuracy of content	Readability & Grammatical Correctness	Conciseness & Balance	Output	Accuracy of content	Readability & Grammatical Correctness	Conciseness & Balance
The takeov	the takeove	2	3	2	UK governr	2	2	3	UK firm's de	3	3	2	Nexperia m	2	3	3
Elon Musk h	elon musk h	3	3	2	Elon Musk	2	2	3	Elon Musk s	3	2	3	Elon Musk t	3	3	3
Chancellor	chancellor j	3	3	3	Chancellor	2	3	2	Chancellor	3	3	3	UK chancell	3	3	3
The first sig	the first sigr	2	3	2	Workers at	2	3	3	More than	2	3	3	A number o	2	3	2
UK-based fo	uk based fo	2	3	2	UK-based f	2	3	3	Deliveroo is	2	3	3	UK-based fo	2	3	2

Figure 8. Manual Scoring Table

The team managed to obtain Table 3 shown below which shows the average manual scoring.

Criteria: 1-poor, 2-fair, 3-great	Accuracy of content	Readability & Grammar	Conciseness & Balance
Baseline-First3	**2.64**	**2.99**	2.05
Baseline-GPT2	2.01	2.70	2.96
T5	2.42	2.85	**2.99**
BART	2.34	2.81	2.92

Table 3. Average Manual Scoring Outcome

According to the results, taking the first three sentences as the summary works great in terms of the accuracy of content and readability. However, it performs significantly worse than other models in terms of conciseness, which is crucial considering the purpose of the project. For the three abstractive summarization models, GPT2 performs significantly worse than T5 and BART in terms of accuracy of content and readability. Overall speaking, T5 and BART have a good balance in terms of accuracy of content, readability, and conciseness. T5 and BART are, therefore, considered the most successful models based on their performance.

8. Qualitative Results

Here is a news example shown in Figure 9 that mainly talked about after Elon Musk bought Twitter, he demanded employees at Twitter to fully commit to working and follow some strict new rules or they would be fired.

After summarization, we can see that the first 3 sentences captured the main idea of the news, but it's not concise enough. GPT-2 produced concise and linguistically correct sentences, but the contents were not accurate if taking a careful look.

T5 and BART summarized in a concise manner, and the contents were accurate and readable. They produced high-quality results.

Elon Musk has told Twitter staff that they must commit to working "long hours at high intensity" or else leave the company, according to reports.In an email to staff, the social media firm's new owner said workers should agree to the pledge if they wanted to stay, the Washington Post reported.Those who do not sign up by Thursday will be given three months' severance pay, Mr Musk said.The BBC has contacted Twitter for comment.In his email to staff, also seen by The Guardian, Mr Musk said that Twitter "will need to be extremely hardcore" in order to succeed."This will mean working long hours at high intensity. Only exceptional performance will constitute a passing grade," he said.Workers were told that they needed to click on a link by 17:00 EST on Thursday, if they want to be "part of the new Twitter".He added: "Whatever decision you make, thank you for your efforts to make Twitter successful."The world's richest man has already announced half of Twitter's staff are being let go, after he bought the company in a $44bn (????38.7bn) deal.Mr Musk said he had "no choice" over the cuts as the company was losing $4m (????3.51m) a day. He has blamed "activist groups pressuring advertisers" for a "massive drop in revenue".A host of top Twitter executives have also stepped down following his purchase of the firm.Last week, the entrepreneur told Twitter staff that remote working would end and "difficult times" lay ahead, according to reports.In an email to staff, the owner of the social media firm said workers would be expected in the office for at least 40 hours a week, Bloomberg reported.Mr Musk added that there was "no way to sugar coat the message" that the slowing global economy was going to hit Twitter's advertising revenues.But tech investor Sarah Kunst said the real reason Twitter is facing difficulties is because Mr Musk's takeover has saddled the company with debt. His behaviour since the takeover has also led some advertisers to pause their spending, she said. "He's now trying to inflict that pain and uncertainty on the employees," she said.She added that there was a question mark over how enforceable Mr Musk's email about hours to staff really was."Can you just send an email to staff who already work for you, and just unilaterally change their working contract? That remains to be seen."Mr Musk himself has been sleeping at Twitter in recent weeks, even while leading electric carmaker Tesla and rocket company SpaceX.He described his work habits in a US court on Wednesday, where he appeared to defend the eye-popping $56bn pay package he received from electric carmaker Tesla in 2018."I pretty much work all the time, with rare exceptions." he said. In response to questioning, he later added that the "fundamental organisational restructuring" at Twitter would be complete by the end of this week.Dan Ives, a senior equity analyst at Wedbush Securities, warned it was further evidence that Twitter's culture had "dramatically changed" with Mr Musk at the helm."Elon Musk is not going to be doing candlelight dinners and playing ping pong in Twitter's cafeteria and this is a shock to the system," he said."But he also needs to play nice in the sandbox because if key Twitter engineers and developers leave, this will be a major void in the Twitter ecosystem," he warned. "There's a careful balance ahead for him, in this tightrope act."Do you work long hours either willingly or because you feel you have no choice? Please share your story by emailing: haveyoursay@bbc.co.uk. Please include a contact number if you are willing to speak to a BBC journalist. You can also get in touch in the following ways:If you are reading this page and can't see the form you will need to visit the mobile version of the BBC website to submit your question or comment or you can email us at HaveYourSay@bbc.co.uk. Please include your name, age and location with any submission.

First-3 Sentence Result:
elon musk has told twitter staff that they must commit to working long hours at high intensity or else leave the company, according to reports. in an email to staff, the social media firm s new owner said workers should agree to the pledge if they wanted to stay, the washington post reported.
those who do not sign up by thursday will be given three months severance pay, mr musk said.

GPT-2 Result:
Elon Musk says "we need to be extremely hardcore" Musk's Twitter account has been hit by layoffs. Musk has said he will "force" employees to sign up at will. He has also said he will "force" them to leave Twitter.

T-5 Result:
Elon Musk says Twitter will need to be extremely hardcore in its new owner . Emails sent to staff say they have to work "long hours at high intensity" Staff must sign up by Thursday unless they want to stay . Those who don't sign up will be given three months' severance pay .

Bart Result:
Elon Musk tells Twitter staff that they must work "long hours at high intensity" Firms say some staff will not be paid; others may leave . The world's richest man has bought Twitter in a $44 billion deal . Twitter has said half of its staff are being let go .

Figure 9. Summarization Outcome Example

The team also tried GPT-3 playground to summarize the news, and the result is attached in Figure 10 below. Compared to the state-of-the-art GPT-3, our models performed very well.

Elon Musk has reportedly told Twitter staff that they must commit to working "long hours at high intensity" or else leave the company. In an email to staff, the social media firm's new owner said that workers who do not sign up by Thursday will be given three months' severance pay. Following his purchase of the firm, Mr Musk has announced half of Twitter's staff are being let go and has told staff they will be expected in the office for at least 40 hours a week

Figure 10. GPT-3 Summarization Results

Another interesting finding that the team has noticed is that extractive summarization (First three sentences approach) works well on news articles about a big event while it works worse on news articles about an individual's story. The potential reason might be that when writing a big event, the author tends to put the topic sentences at the beginning of the news article. However, when writing an individual's story, the author tends to start with some words by the individual which could be irrelevant to the topics of the news article.

9. Discussion and Learnings

In overall, the Seq2Seq transformer models (T5 & BART) work well and have a good balance in terms of accuracy of content, readability, and conciseness. Seq2Seq transformers seem to have better performance than GPT-2 in terms of summarization tasks. Seq2Seq transformers are smarter and more efficient in extracting ideas from the given texts.

In addition, the team has experimented with multiple decoding methods of transformer models. Greedy search makes the model easier to replicate the same sentence since it always chooses the next word with the highest probability. Beam search works better than Greedy search, the number of beams = 5 or 10 has reasonable performance. Besides, the team tried sampling and finetuning on different temperature settings. The team has found that the model seems easier to reproduce the existing sentences from the articles when the temperature is low, and the model is more likely to generate new sentences but makes less sense when the temperature is higher. Figure 11 below shows an example of generated summaries using different temperature values.

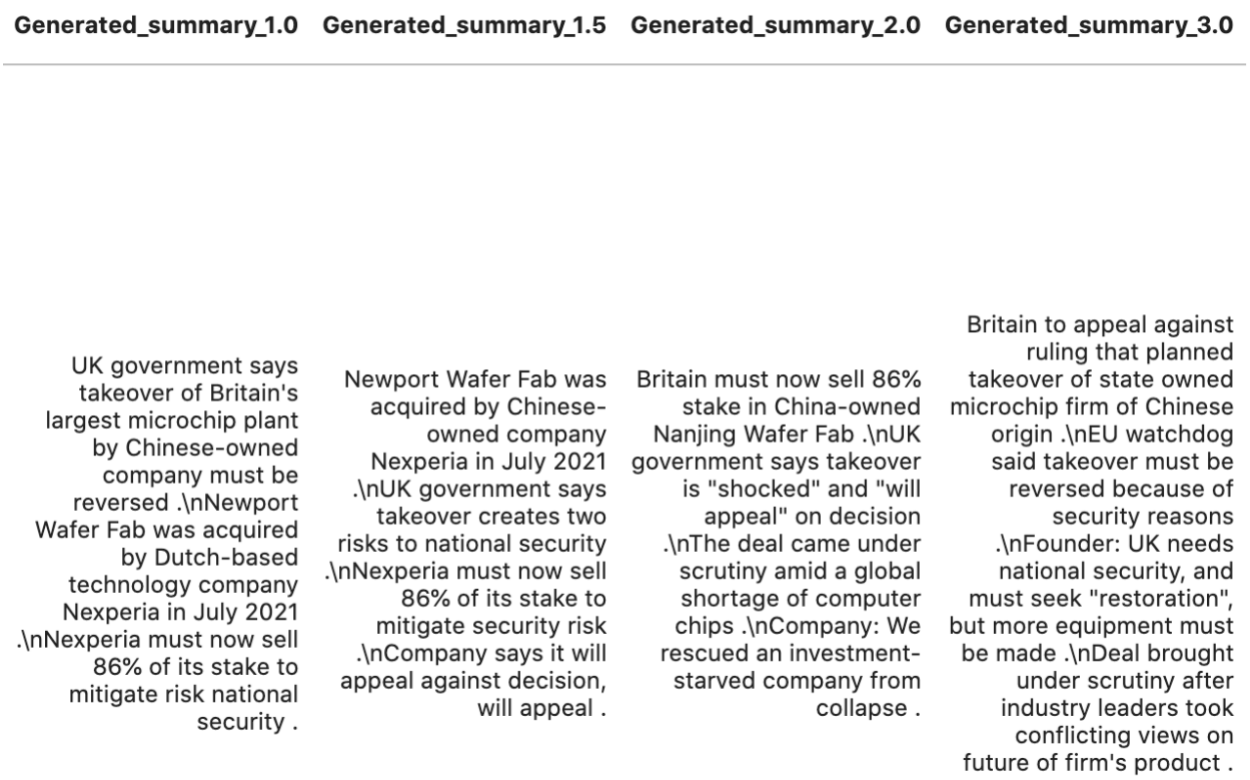

Figure 11. Generated Summary Using Different Temperature Values

Something that the team could improve is that the extractive and abstractive methods can be combined, and both of their advantages can be leveraged. The first three sentences are content-accurate in most cases, and the team can use them in abstractive methods as another input, hence the model would pay more attention to the first three sentences, and thus produce more precise results.

10. Individual Contributions

The team members have been actively involved in the project. Both team members have actively participated in Zoom meetings to communicate the progress as well as the difficulties of the project. The following section presents the specific work that each team member has been working on.

Yichen has worked on the following:

- Web-scraping recent news data from BBC
- Building the extractive baseline model
- Building and training the fine-tuned T5-base model
- Building and training the fine-tuned BART-base model
- Giving manual scoring to **half** of the summarization results on the recent news data
- Building the Gradio implementation of the user-facing side of the project

Zijie has worked on the following:

- Data preprocessing
- Implementing GRU encoder-decoder architecture
- Building the abstractive baseline model (GPT-2)
- Setting up manual scoring criteria and manual scoring table
- Giving manual scoring to **half** of the summarization results on the recent news data
- Performing qualitative results analysis

Reference

[1] Y. Zhang, J. Meng, and M. Pratama. Extractive document summarization based on convolutional neural networks. IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society.

[2] L. Liu, Y. Lu, M. Yang, Q. Qu, J. Zhu, and H. Li. Generative Adversarial Network for Abstractive Text Summarization. arXiv, 2017. doi: arXiv:1711.09357v1.

[3] HuggingFace, T5 Documentation, available at https://huggingface.co/docs/transformers/model_doc/t5, accessed December 11, 2022

[4] M. Lewis et al., BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv, 2019. doi: 10.48550/ARXIV.1910.13461.

[5] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, BERTScore: Evaluating Text Generation with BERT. arXiv, 2019. doi: 10.48550/arXiv.1904.09675.