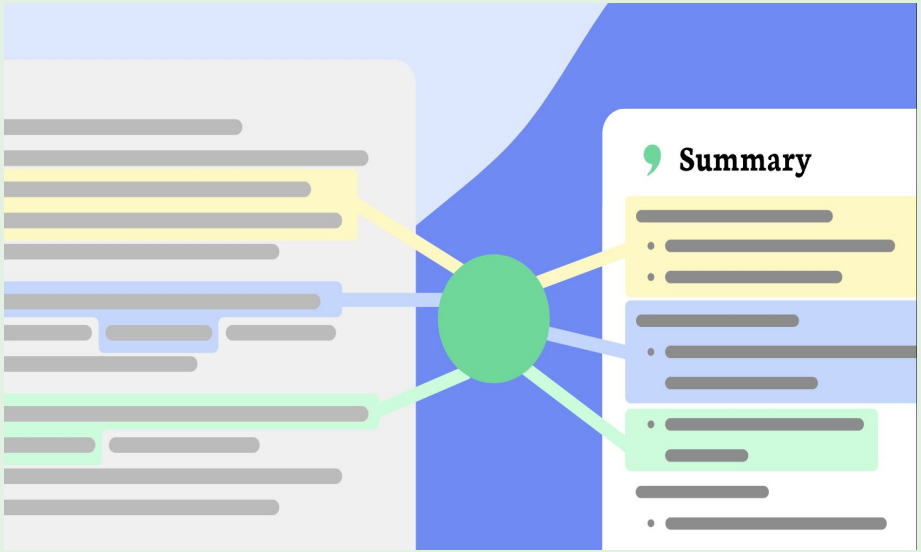
Text Summarization For Multiple News Using mt-5

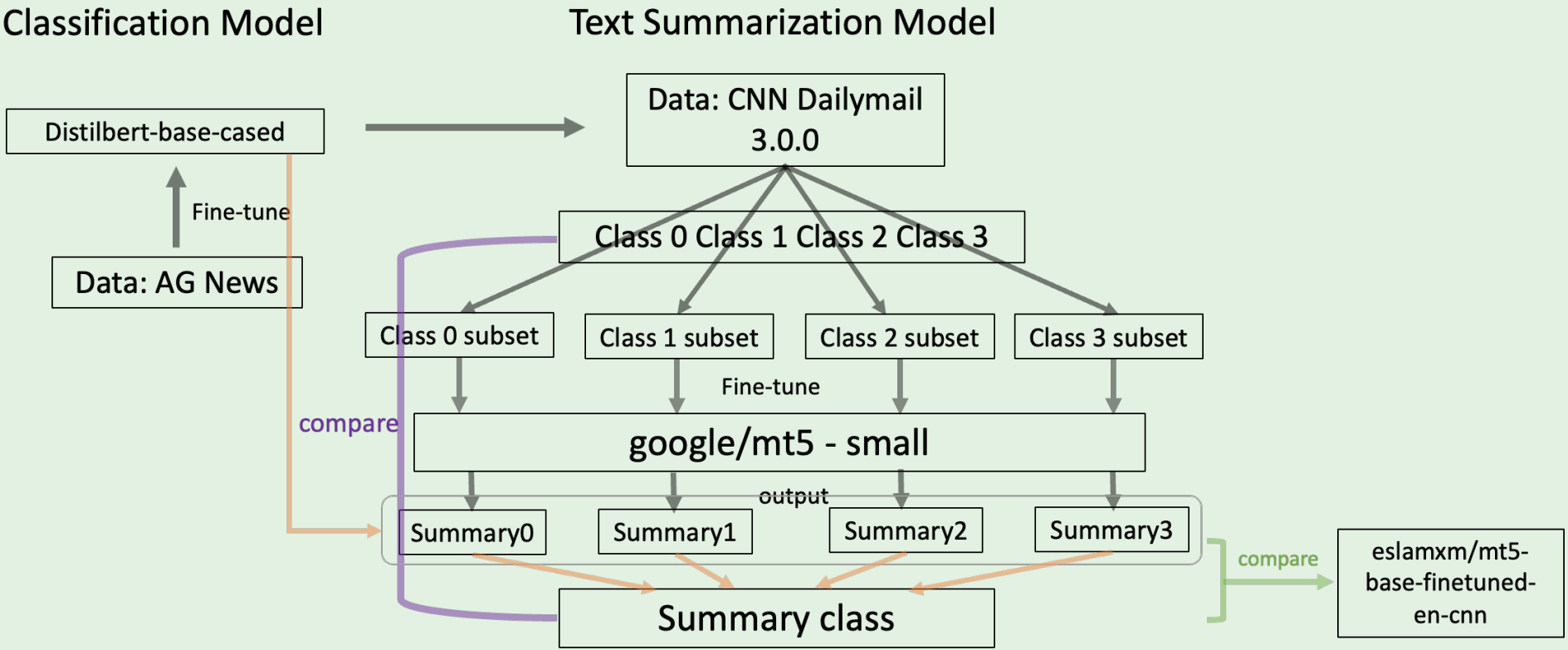
Yihui Liu, Sirui Wang, Rae Zhang

Definition

- Text summarization is turning larger documents into shorter and precise paragraphs or sentences
- The meaning of the paragraph stays the same
- Reduce time to understand large papers without skipping vital information



Workflow


Classification Model
Text Summarization Model
Distilbert-base-cased
Fine-tune
Data: AG News
Data: CNN Dailymail 3.0.0
Class 0 Class 1 Class 2 Class 3
Class 0 subset
Class 1 subset
Class 2 subset
Class 3 subset
Fine-tune
compare
google/mt5 - small
output
Summary0
Summary1
Summary2
Summary3
Summary class
compare
eslamxm/mt5-base-finetuned-en-cnn

Classification Model - Dataset Information

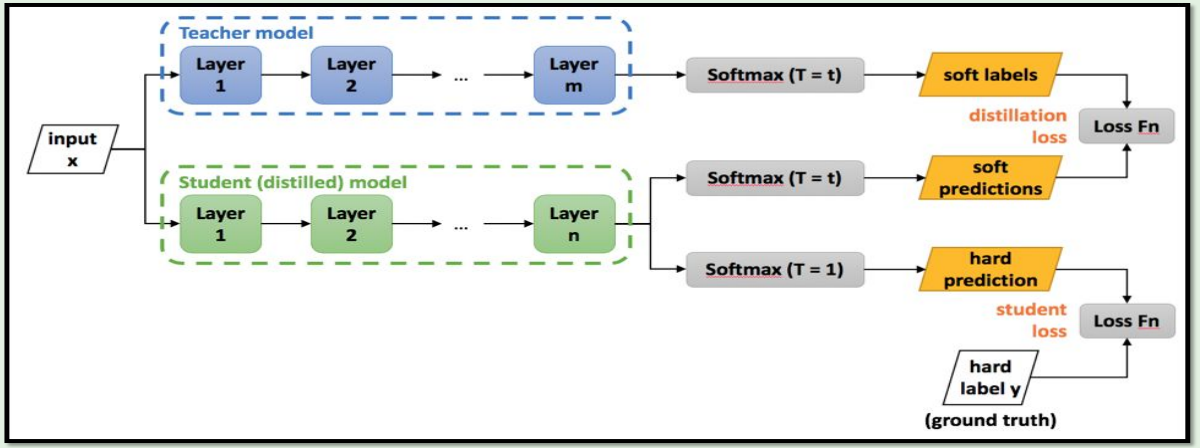
- AG is a collection of more than 1 million news articles. News articles have been gathered from more than 2000 news sources by ComeToMyHead in more than 1 year of activity
- Example:


```
{
    "label": 3,
    "text": "New iPad released Just like every other September, this one is no different. Apple is planning to release a bigger, heavier, fatter iPad that..."
}
```


- text: a string feature
- label: a classification label, with possible values including World (0), Sports (1), Business (2), Sci/Tech (3)

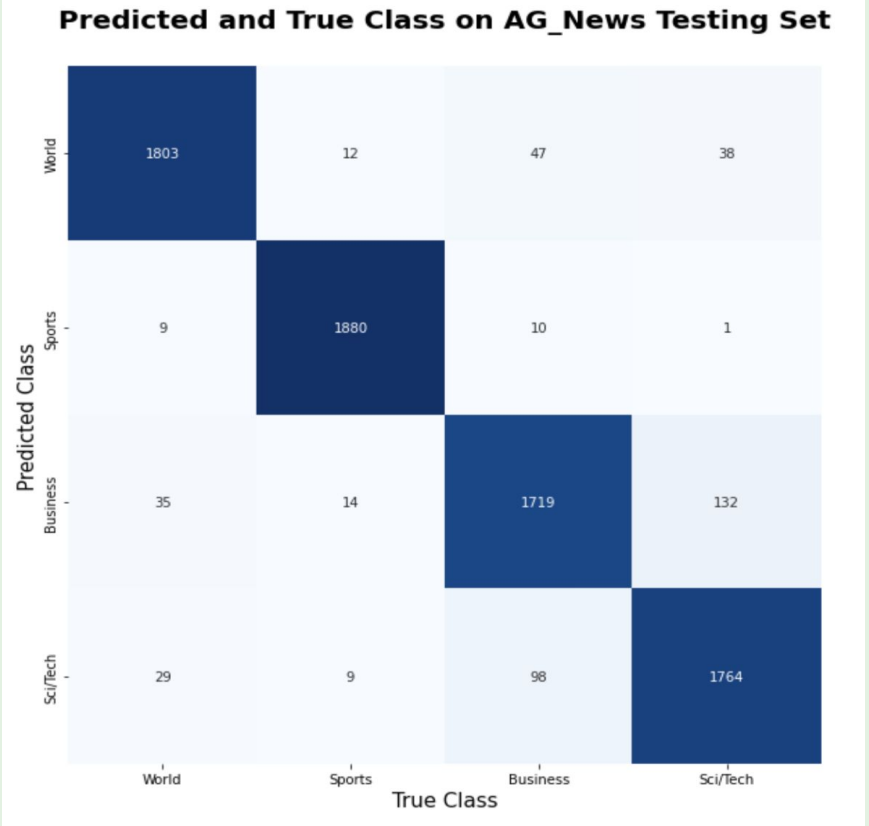
Classification Model - Model Information

- DistilBERT-base-cased is a transformer model, smaller and faster than BERT, which was pretrained on the same corpus in a self-supervised fashion, using the BERT base model as a teacher
- Pre-trained with three objectives: Distillation loss, Masked language modeling (MLM) and Cosine embedding loss



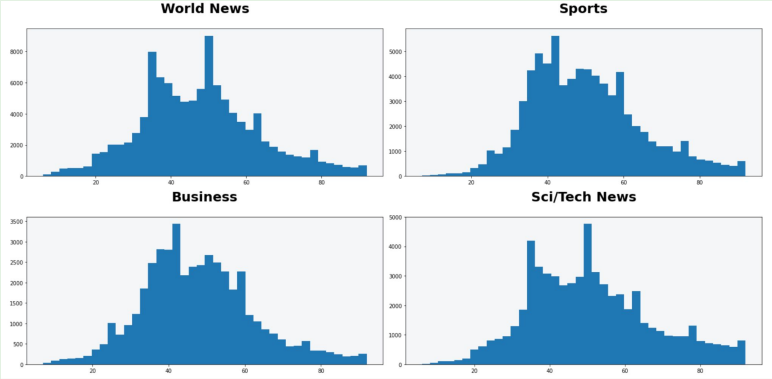
Classification Model - Results

	Loss	Accuracy
Train	0.125	95.81%
Test	0.170	94.29%

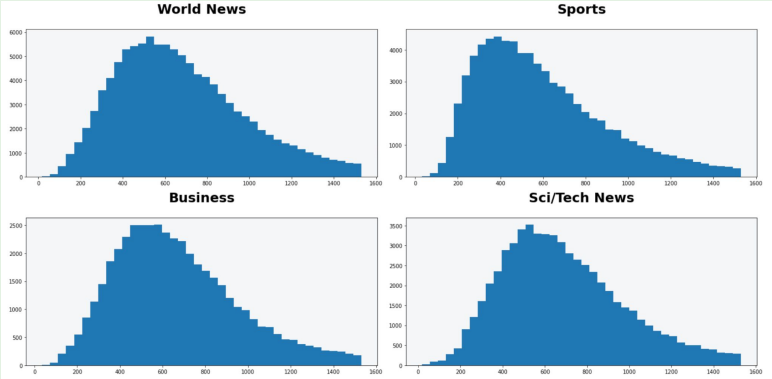


Logistics

- AG News and CNN Dailymail datasets are all news datasets → Use the classification model trained before on CNN dataset
- Trained classification model → split the highlights of CNN dataset into 4 classes
- Different kinds of news have different distribution for lengths of articles and highlights → different summary → different summarization model for each class



highlights length distribution



article length distribution

(a) Extractive Summarization

Source Text: Peter and Elizabeth took a taxi to attend the night party in the city.

While in the party, Elizabeth collapsed and was rushed to the hospital.

Summary: Peter and Elizabeth attend party city. Elizabeth rushed hospital.

(b) Abstractive Summarization

Source Text: Peter and Elizabeth took a taxi to attend the night party in the city.

While in the party, Elizabeth collapsed and was rushed to the hospital.

Summary: Elizabeth was hospitalized after attending a party with Peter.

Text Summarization Model - Dataset Information

- The CNN / DailyMail Dataset is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail.
- Example:

{'id': '0054d6d30dbcad772e20b22771153a2a9cbeaf62',

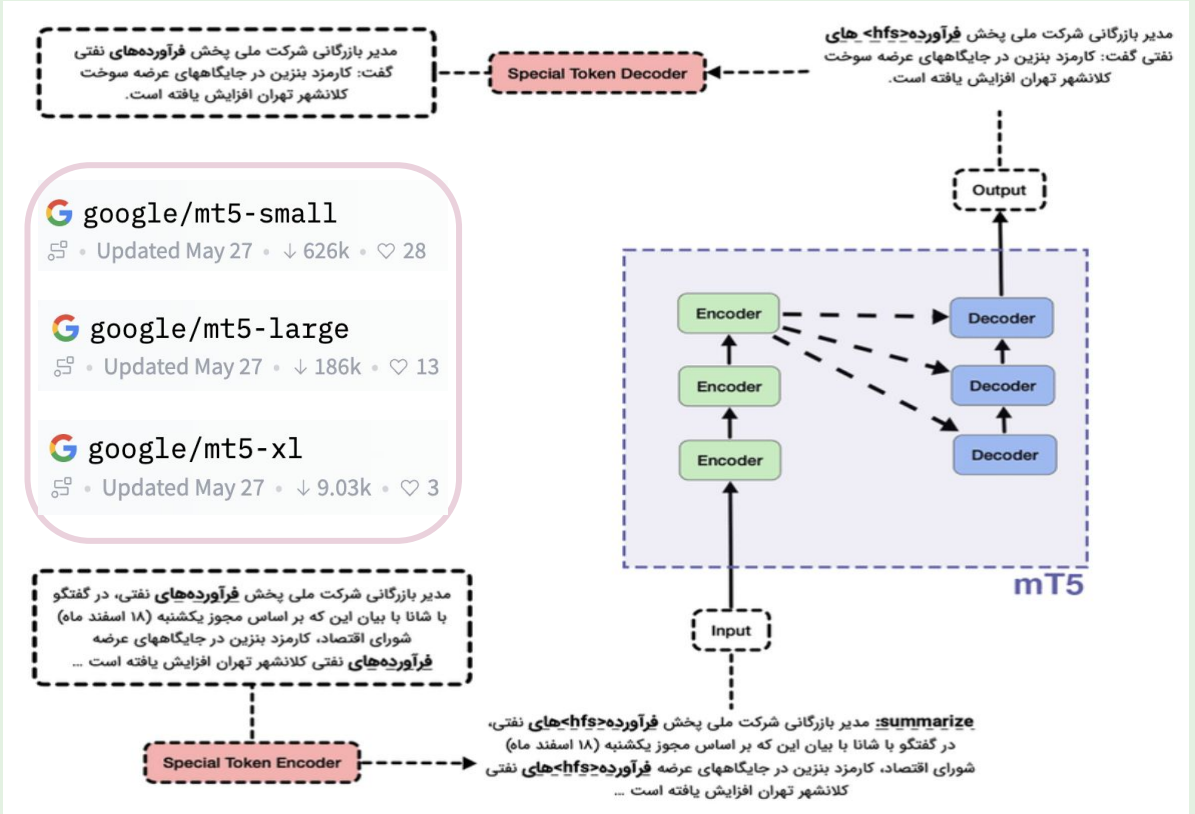
'article': '(CNN) -- An American woman died aboard a cruise ship that docked at Rio de Janeiro on Tuesday, the same ship on which 86 passengers previously fell ill, according to the state-run Brazilian news agency, Agencia Brasil. The American tourist died aboard the MS Veendam, owned by cruise operator Holland America. Federal Police told Agencia Brasil that forensic doctors were investigating her death. The ship's doctors told police that the woman was elderly and suffered from diabetes and hypertension, according the agency. The other passengers came down with diarrhea prior to her death during an earlier part of the trip, the ship's doctors said. The Veendam left New York 36 days ago for a South America tour.'

'highlights': 'The elderly woman suffered from diabetes and hypertension, ship's doctors say .\nPreviously, 86 passengers had fallen ill on the ship, Agencia Brasil says .'}

- id: a string containing the heximal formatted SHA1 hash of the url where the story was retrieved from
- article: a string containing the body of the news article
- highlights: a string containing the highlight of the article as written by the article author

Text Summarization Model - google/mT5-small

- mT5 - multilingual variant of “Text-to-Text Transfer” (T5) that was pre-trained on a new Common Crawl-based dataset covering 101 languages.



Text Summarization Model Results - Rouge Score

ROUGE refers … system summary is recovering o… geLsum)

F1 score

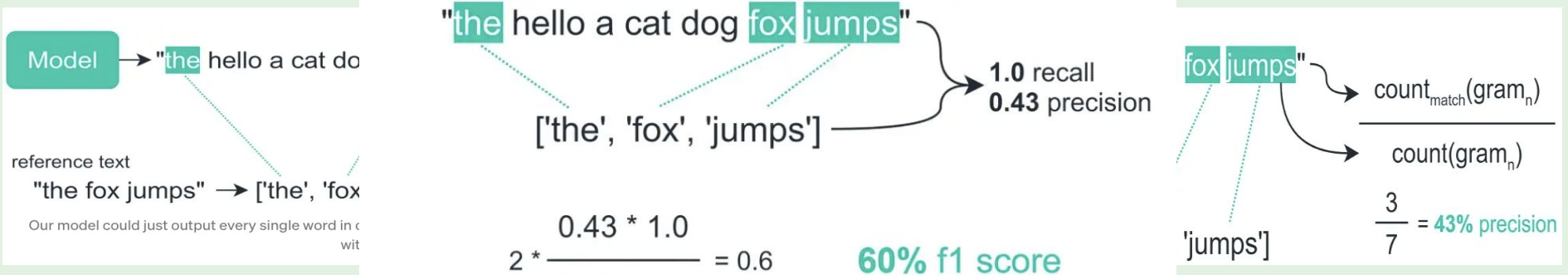
$$2 * \frac{precision * recall}{precision + recall}$$

Recall

number of n-grams four… … d in model and reference

number of n-gra… …rams in **model**

Let's apply that again to our previous example:



$$2 * \frac{0.43 * 1.0}{0.43 + 1.0} = 0.6$$

60% f1 score

Text Summarization Model Results - Rouge Score

	rouge1	rouge2	rougeL	rougeLsum
Class 'World' Model	26.99%	11.88%	21.42%	25.42%
Class 'Sports' Model	29.28%	13.57%	22.87%	27.74%
Class 'Business' Model	26.56%	12.03%	21.40%	24.93%
Class 'Sci/Tech' Model	26.23%	11.97%	20.84%	24.51%
Average	27.27%	12.36%	21.63%	25.65%

Text Summarization Model Results - Demo

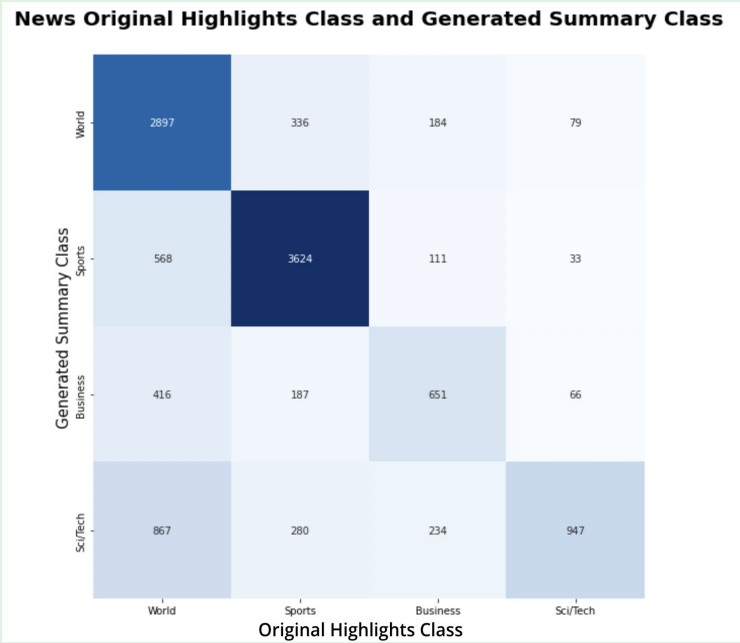
Article:
(CNN)A 32-year-old Massachusetts man is facing murder charges, authorities said Wednesday, four days after another man's remains were found in a duffel bag. The Middlesex District Attorney's Office said that Carlos Colina, 32, will be arraigned the morning of April 14 for murder in connection with the remains discovered Saturday in Cambridge. Earlier this week, Colina was arraigned on charges of assault and battery causing serious bodily injury and improper disposal of a body. A Middlesex County judge then revoked bail for Colina in another case he's involved in, for alleged assault and battery. The victim in that case is different from the one whose remains were found in recent days. Police were notified Saturday morning about a suspicious item along a walkway in Cambridge. Officers arrived at the scene, opened a duffel bag and found human remains. After that discovery, police say, a surveillance video led them to an apartment building, where more body parts were discovered in a common area. That location is near the Cambridge Police Department headquarters. The remains at both locations belonged to the same victim, identified Monday as Jonathan Camilien, 26. Camilien and Colina knew each other, according to authorities. "This was a gruesome discovery," District Attorney Marian Ryan said. CNN's Kevin Conlon contributed to this report.

Highlights:
Prosecutor: Carlos Colina, 32, will be arraigned on the murder charge next week . He's already been arraigned for alleged assault and battery, improper disposal of a body . Body parts were found in a duffel bag and a common area of an apartment building .

Model’s Summary:
Carlos Colina, 32, will be arraigned April 14 for murder . The remains were found Saturday in Cambridge .

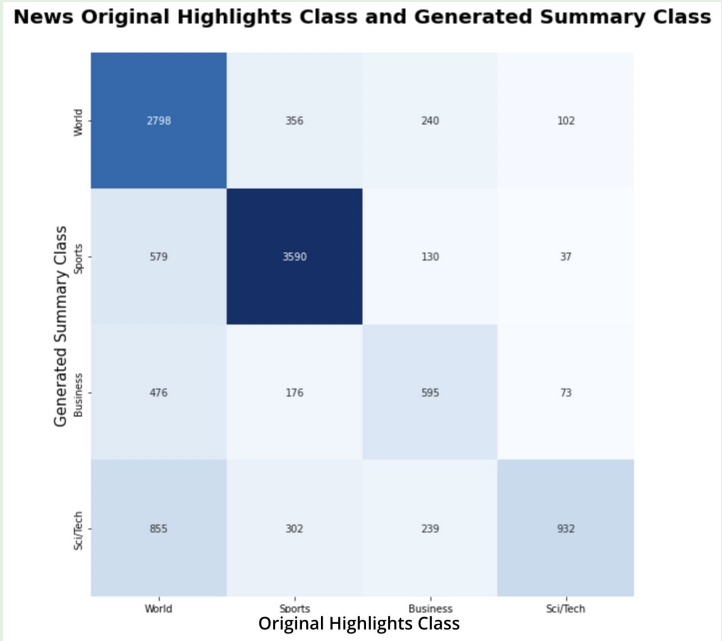
Model performance comparison – Confusion Matrix

Our fine-tune model



Accuracy – 70.72%

Compared model - mt5-base-finetuned-en-cnn



Accuracy – 68.95%

Reference: https://huggingface.co/eslamxm/mt5-base-finetuned-en-cnn

Model performance comparison – Rouge Score

Our fine-tune model
Rouge1: 27.26%
Rouge2: 12.36%
RougeL: 21.63%

Compared model - mt5-base-finetuned-en-cnn
Rouge1: 22.84%
Rouge2: 10.11%
RougeL: 21.8%

Reference: https://huggingface.co/eslamxm/mt5-base-finetuned-en-cnn

Result

- Text summary outputs for each class

	World	Sports	Business	Sci/Tech
Highlights	A French rescue team finds Rishi Khanal more than three days after the quake . A 4-month-old baby is reported to have been rescued after 22 hours in rubble .	Chelsea are six points clear with a game in hand in the Premier League . They have led the title race since a brilliant start to the season in August . Cesar Azpilicueta feels that consistency puts them in pole position to win .	Some of Jesus' most important financial backers were women, historians say. Joseph of Arimathea and Nicodemus, both men of stature and wealth, chipped in to help fund Jesus' ministry.	The "Star Wars" digital collection is set for release this week . Special features include behind-the-scenes stories on the unique alien sounds from the movie .
Summary from our model	Rishi Khanal was saved after a French search and rescue team found him under the rubble.	Cesar Azpilicueta is counting down the games until Chelsea win the Barclays Premier League title.	The New Testament says money was a concern, but not just as an impediment to salvation.	"Star Wars" fans will get more than they bargained for when the saga comes to digital HD on Friday.
Summary from compared model	Rishi Khanal was rescued after a French search and rescue team found him . He was stuck under rubble.	Chelsea sit top of the table ahead of a game in hand at Stamford Bridge .	Jesus told his Twelve Apostles to leave their day jobs and follow him on an itinerant mission .	"Star Wars" films will include special features . One focus of the features will be the sound effects of the movies .

Limitations and Expectations

- Limited computation time, computer units, and hardware condition
- Fine-tune on mt5-base or mt5-large
- More News data
- Apply more multilingual features in mT5 when doing fine-tuning

Thank you !

Questions?