Text Summarisation Project

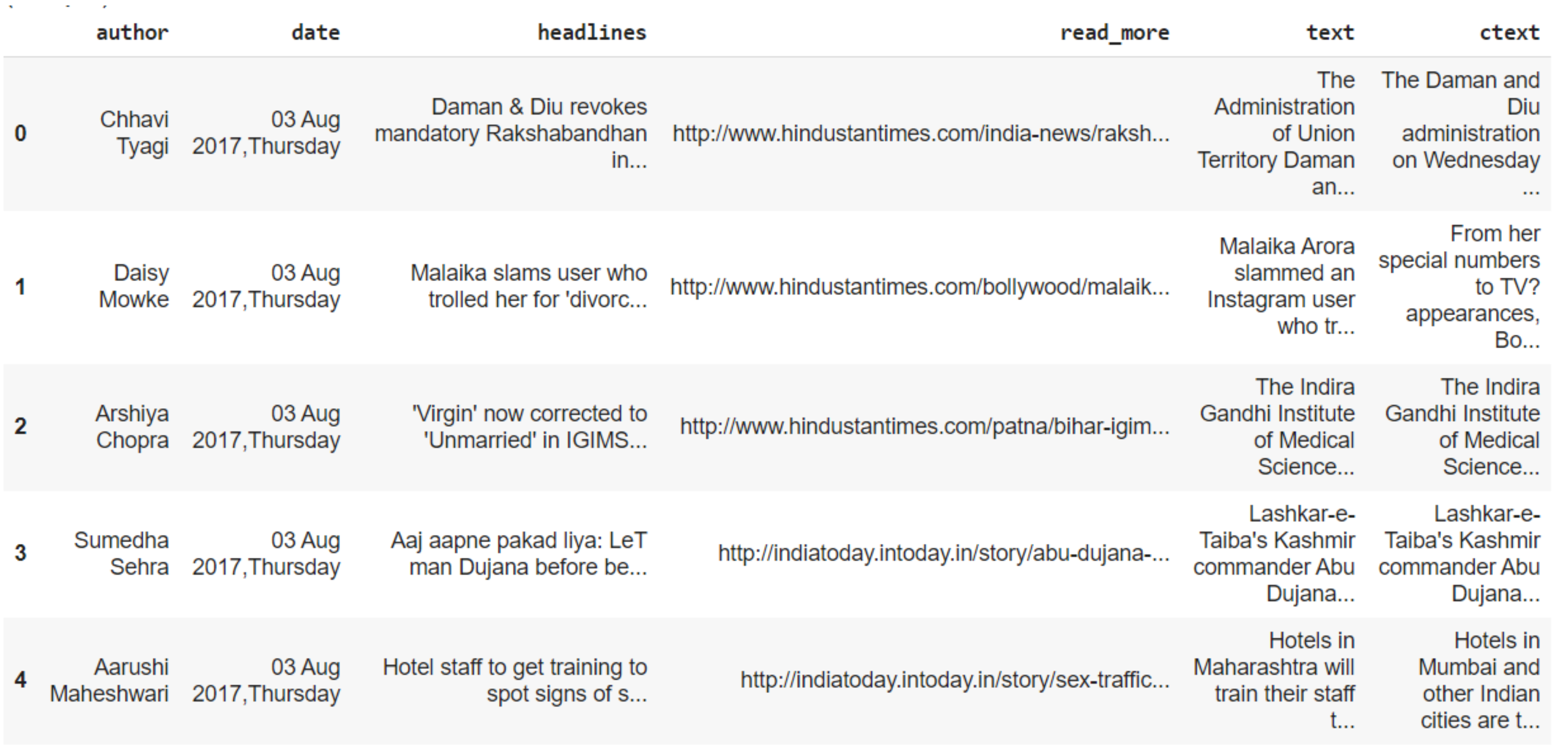
Dataset

For the abstractive summarisation task, we are using a news summary dataset taken from Kaggle.

Dataset Description

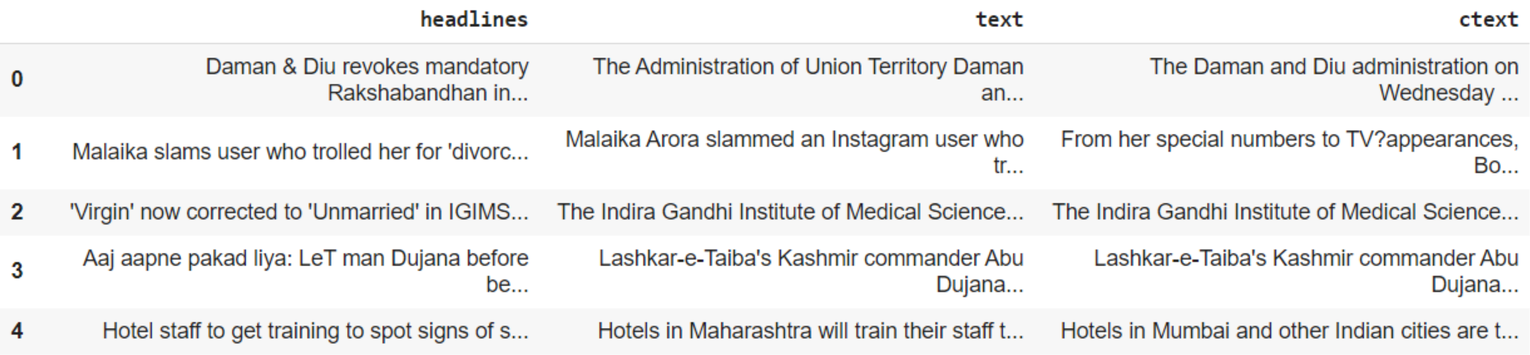
The dataset consists of 4514 examples.
Columns: Author_name, Headlines, Url of Article, Short text, Complete Article
This dataset was formed by summarizing news from Inshorts and scrapping the news articles from Hindu, Indian times and Guardian. The duration for which the news has been taken is february 2017 to august 2017.

Dataset Visualization:



	author	date	headlines	read_more	text	ctext
0	Chhavi Tyagi	03 Aug 2017,Thursday	Daman & Diu revokes mandatory Rakshabandhan in...	http://www.hindustantimes.com/india-news/raksh...	The Administration of Union Territory Daman an...	The Daman and Diu administration on Wednesday ...
1	Daisy Mowke	03 Aug 2017,Thursday	Malaika slams user who trolled her for 'divorc...	http://www.hindustantimes.com/bollywood/malaik...	Malaika Arora slammed an Instagram user who tr...	From her special numbers to TV? appearances, Bo...
2	Arshiya Chopra	03 Aug 2017,Thursday	'Virgin' now corrected to 'Unmarried' in IGIMS...	http://www.hindustantimes.com/patna/bihar-igim...	The Indira Gandhi Institute of Medical Science...	The Indira Gandhi Institute of Medical Science...
3	Sumedha Sehra	03 Aug 2017,Thursday	Aaj aapne pakad liya: LeT man Dujana before be...	http://indiatoday.intoday.in/story/abu-dujana-...	Lashkar-e-Taiba's Kashmir commander Abu Dujana...	Lashkar-e-Taiba's Kashmir commander Abu Dujana...
4	Aarushi Maheshwari	03 Aug 2017,Thursday	Hotel staff to get training to spot signs of s...	http://indiatoday.intoday.in/story/sex-traffic...	Hotels in Maharashtra will train their staff t...	Hotels in Mumbai and other Indian cities are t...

But among all these columns only `['headlines', 'text', 'ctext']` are useful for summarisation task so taking only those columns.



	headlines	text	ctext
0	Daman & Diu revokes mandatory Rakshabandhan in...	The Administration of Union Territory Daman an...	The Daman and Diu administration on Wednesday ...
1	Malaika slams user who trolled her for 'divorc...	Malaika Arora slammed an Instagram user who tr...	From her special numbers to TV?appearances, Bo...
2	'Virgin' now corrected to 'Unmarried' in IGIMS...	The Indira Gandhi Institute of Medical Science...	The Indira Gandhi Institute of Medical Science...
3	Aaj aapne pakad liya: LeT man Dujana before be...	Lashkar-e-Taiba's Kashmir commander Abu Dujana...	Lashkar-e-Taiba's Kashmir commander Abu Dujana...
4	Hotel staff to get training to spot signs of s...	Hotels in Maharashtra will train their staff t...	Hotels in Mumbai and other Indian cities are t...

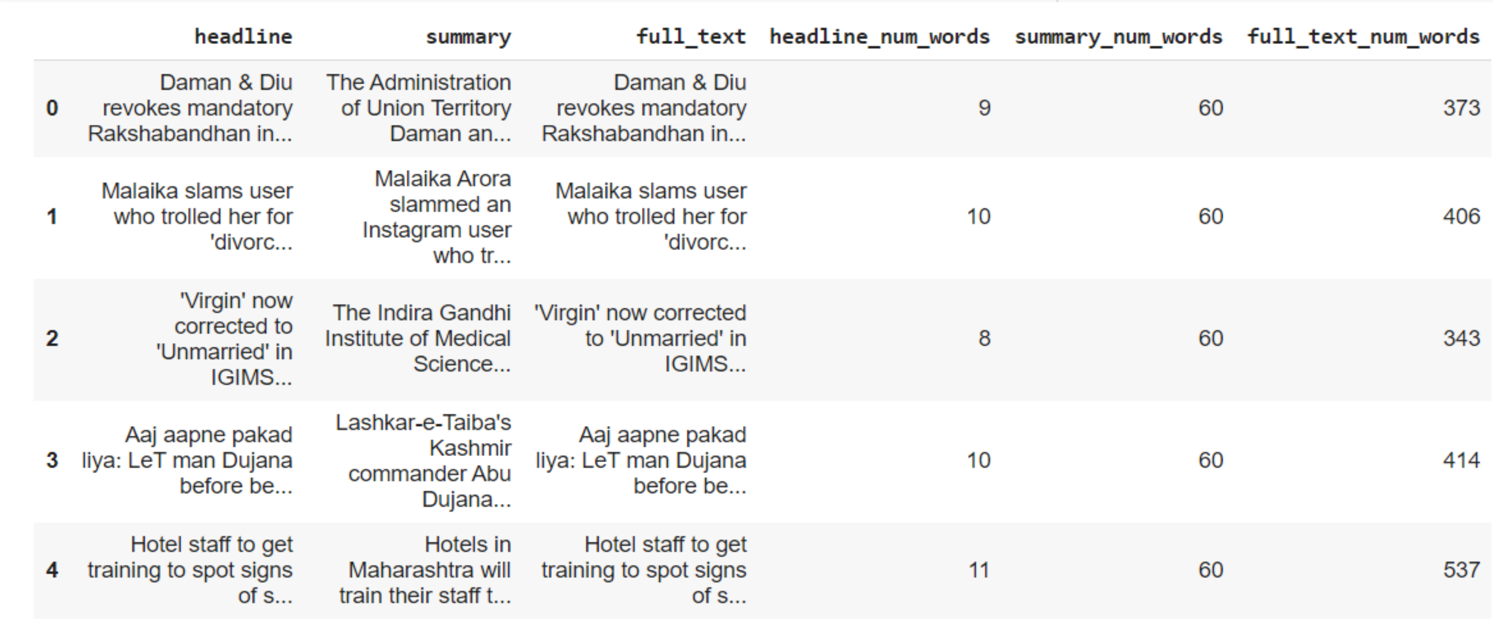
Dataset Pre-processing:

After observing the dataset, we found that out of 4514 news only 4396 news have a summary.

```
#   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   headlines  4514 non-null   object
 1   text       4514 non-null   object
 2   ctext      4396 non-null   object
```

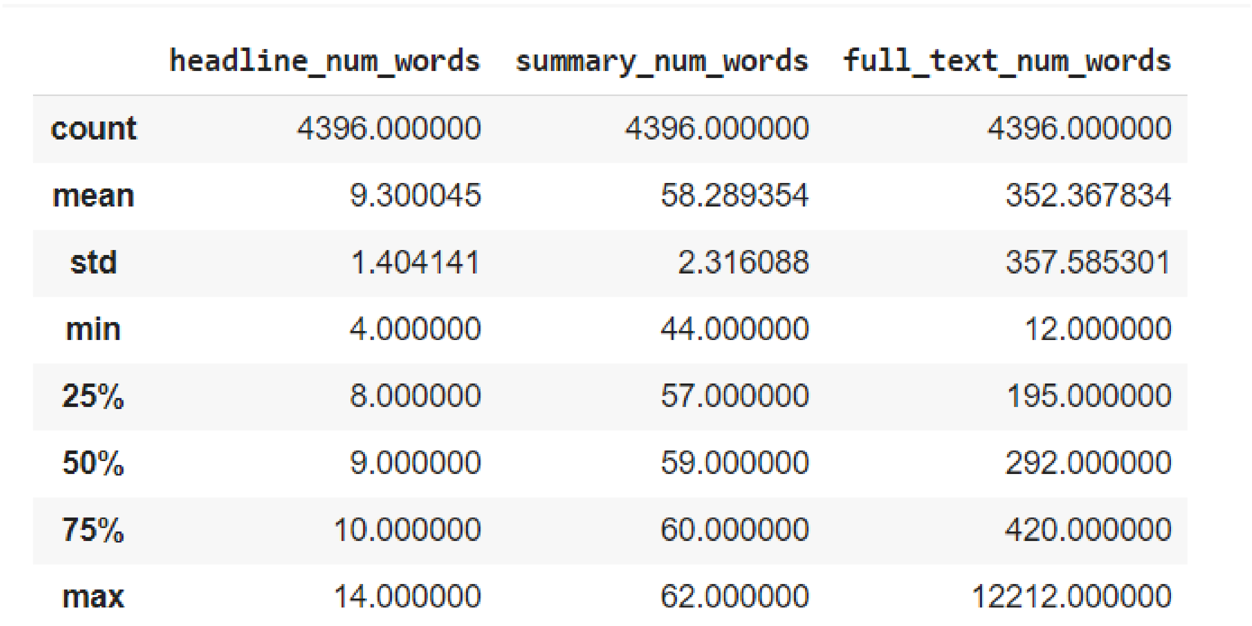
So we dropped the news for which there is no summary and hence now we have a total 4396 news with summary.

Length Analysis of news:



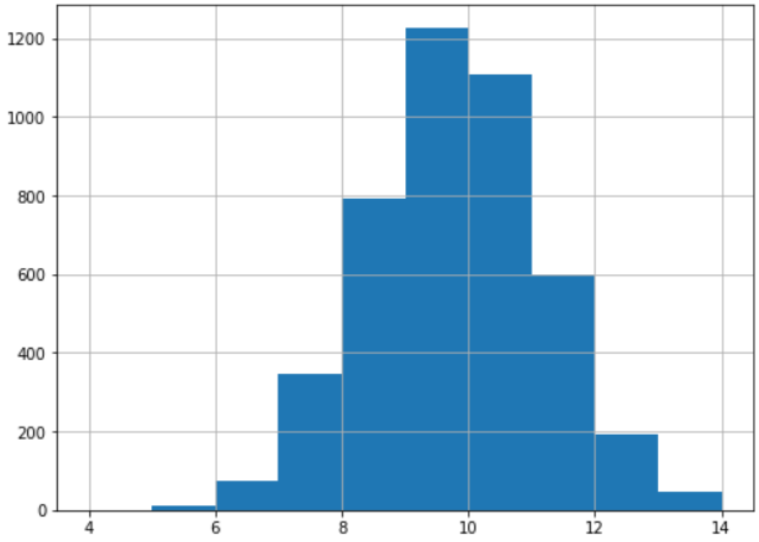
	headline	summary	full_text	headline_num_words	summary_num_words	full_text_num_words
0	Daman & Diu revokes mandatory Rakshabandhan in...	The Administration of Union Territory Daman an...	Daman & Diu revokes mandatory Rakshabandhan in...	9	60	373
1	Malaika slams user who trolled her for 'divorc...	Malaika Arora slammed an Instagram user who tr...	Malaika slams user who trolled her for 'divorc...	10	60	406
2	'Virgin' now corrected to 'Unmarried' in IGIMS...	The Indira Gandhi Institute of Medical Science...	'Virgin' now corrected to 'Unmarried' in IGIMS...	8	60	343
3	Aaj aapne pakad liya: LeT man Dujana before be...	Lashkar-e-Taiba's Kashmir commander Abu Dujana...	Aaj aapne pakad liya: LeT man Dujana before be...	10	60	414
4	Hotel staff to get training to spot signs of s...	Hotels in Maharashtra will train their staff t...	Hotel staff to get training to spot signs of s...	11	60	537

Now by using describe() function we find basic statistical details like count, mean, std etc. of the news data frame.

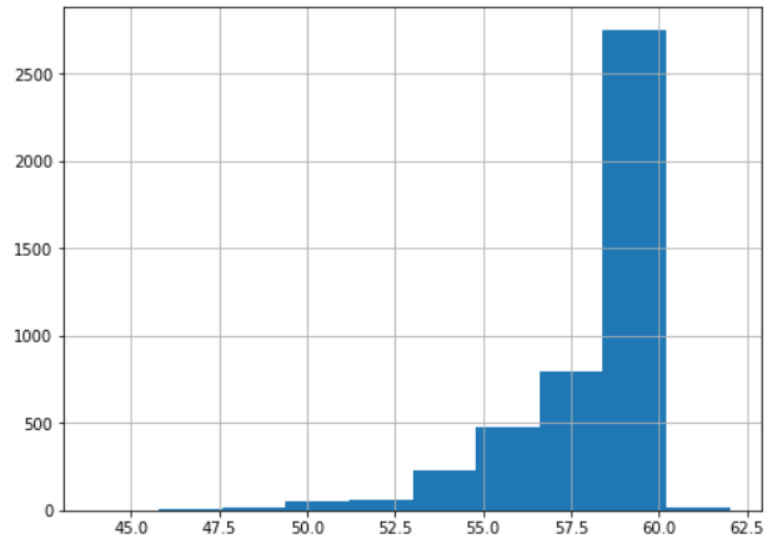


	headline_num_words	summary_num_words	full_text_num_words
count	4396.000000	4396.000000	4396.000000
mean	9.300045	58.289354	352.367834
std	1.404141	2.316088	357.585301
min	4.000000	44.000000	12.000000
25%	8.000000	57.000000	195.000000
50%	9.000000	59.000000	292.000000
75%	10.000000	60.000000	420.000000
max	14.000000	62.000000	12212.000000

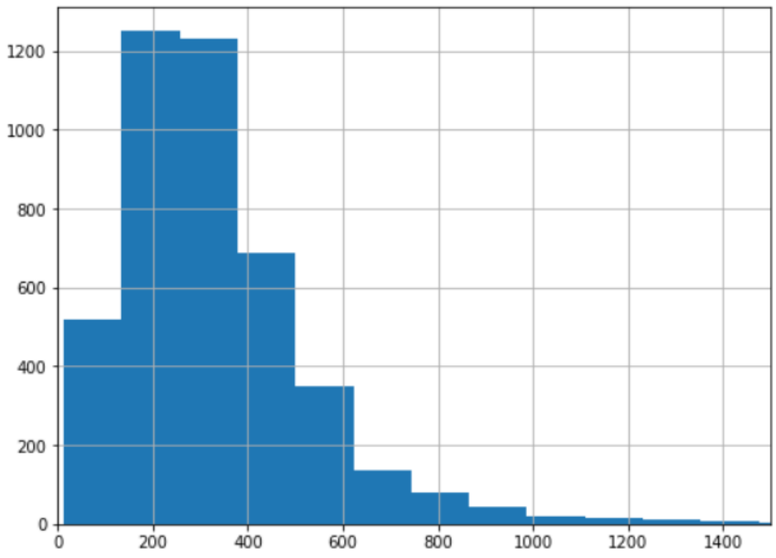
Plotting Number of words in headline:



Plotting Number of words in summary:



Plotting Number of words in full text:



So we can see from the above histograms that for around 1200 news the length of headlines is 10-12 words. Also there are more than 2500 news articles for which the length of the summary is 60 words. And for full text for maximum news the number of words is between 200-600.

Model

We use the idea from Text Summarization with Pretrained Encoders paper to start with a pretrained transformers model and finetune it on our summarization dataset.

We use T5 as our base language model. T5 in many ways is one of its kind transformers architecture that not only gives state of the art results in many NLP tasks, but also has a very radical approach to NLP tasks.

- **Text-2-Text** - According to the graphic taken from the T5 paper. All NLP tasks are converted to a text-to-text problem. Tasks such as translation, classification, summarization and question answering, all of them are treated as a text-to-text conversion problem, rather than seen as separate unique problem statements.
- **Unified approach for NLP Deep Learning** - Since the task is reflected purely in the text input and output, you can use the same model, objective, training procedure, and decoding process to ANY task. Above framework can be used for any task - show Q&A, summarization, etc.

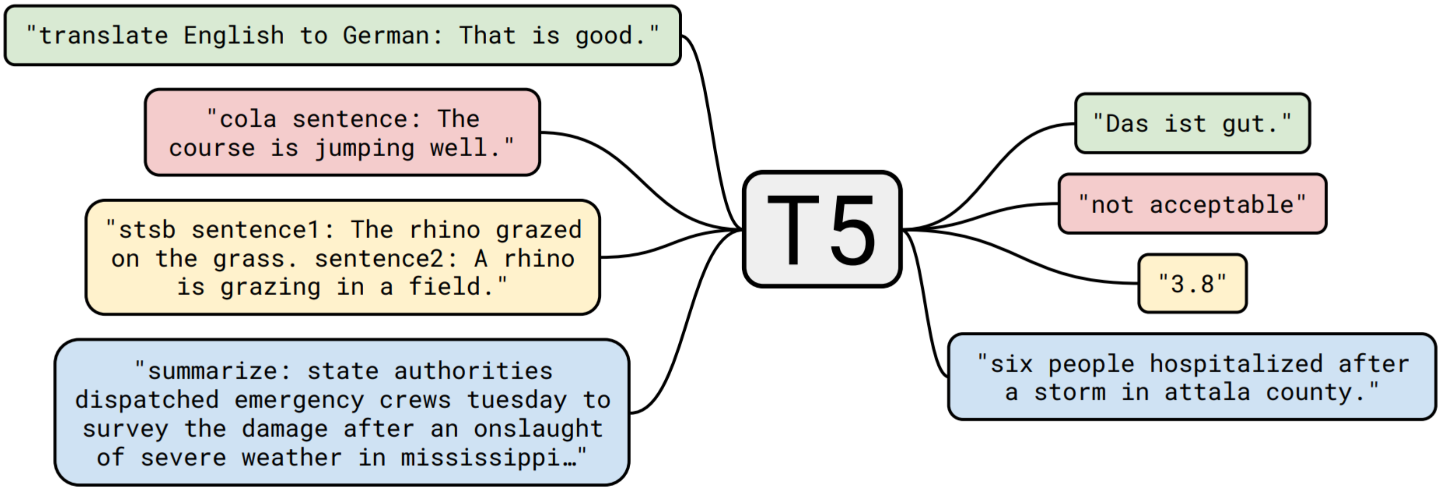

"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

Image From https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html

Preparing data for model:

Now Before feeding those texts to our model, we need to preprocess them. Transformer model is tied with a tokenizer. For each model it will have different tokenizer values. This is done by a Transformers `Tokenizer` which will tokenize the inputs by converting the tokens to their corresponding IDs in the pretrained vocabulary and put it in a format the model expects, as well as generate the other inputs that the model requires.

To do all of this, we instantiate our tokenizer with the AutoTokenizer.from_pretrained method, which will ensure:

- we get a tokenizer that corresponds to the model architecture we want to use,

- we download the vocabulary used when pretraining this specific checkpoint.

For example,

```
# pass list of sentences
tokenized_sample = tokenizer(["Hello, this one sentence!", "This
is another sentence."])
print(tokenized_sample)
```

Then we will get output something like,

```
{'input_ids': [[8774, 6, 48, 80, 7142, 55, 1], [100, 19, 430,
7142, 5, 1]], 'attention_mask': [[1, 1, 1, 1, 1, 1, 1], [1, 1, 1,
1, 1, 1]]}
```

Now in order to prepare the targets for our model, we need to tokenize them inside the `as_target_tokenizer` context manager. This will make sure the tokenizer uses the special tokens corresponding to the targets.

Model Specific Details

If our base model is t5 based, we have to prefix the inputs with `”summarize:”` . The prefix tells the model which task to perform (as the model can also do translation tasks). So this acts like a signal to update weights in a specific manner while training.

Our data has full_text and summary. Now we have defined a `preprocess_function` which will add the prefix tag `”summarize:”` to each document in the data and then we want to tokenize the data so we send the data to `Tokenizer.`

Now once the input which needs to be given to the model is obtained we give a target that is summary to the model.
For the summarisation task we have used a train-test split of 80:20 and

```
max_input_length = 600
max_target_length = 100
```

We have created a dataset object for train/test: which makes it easier to access a batch of data. In the output of `preprocess_function` we have ID's for the input, attention mask for input and ID's output or summary. So we apply `preprocess_function` to all elements of the dataset train/test object by using `map` method. This method will apply the given function on all the elements of all the splits in the dataset, so our training, validation and testing data will be preprocessed in one single command.

Fine-tuning the model:

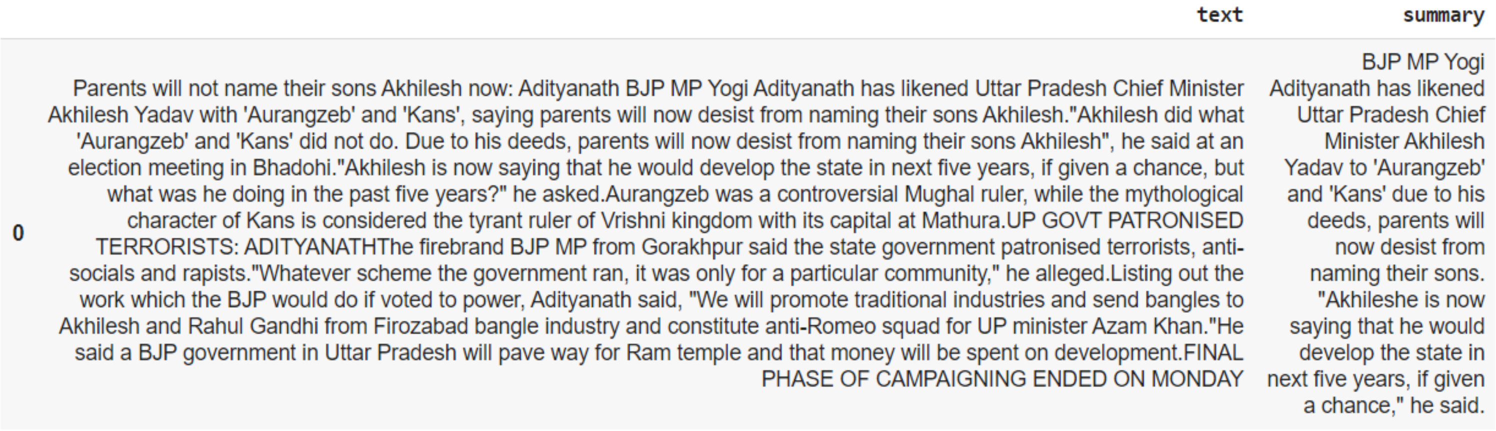
Once the data is ready then we download the pretrained model using `’from_pretrained’` method and then we perform fine-tuning on it. As our task is similar to sequence-to-sequence we have used `’AutoModelForSeq2SeqLM’` from auto classes provided by huggingface.

When we check the summary for news by loading the previously trained `’AutoModelForSeq2SeqLM’` model then observe some results.

Some sample outputs before training (Pre-trained Model) and after training the model:

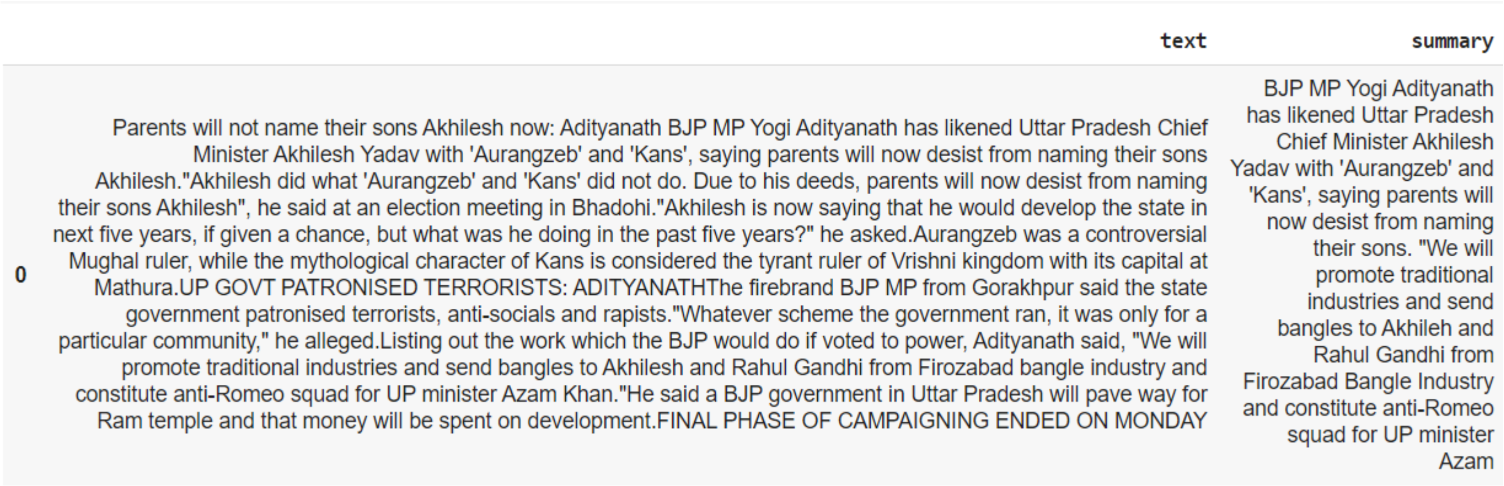
- **Example from train set:**

Summarisation before training:



	text	summary
0	Parents will not name their sons Akhilesh now: Adityanath BJP MP Yogi Adityanath has likened Uttar Pradesh Chief Minister Akhilesh Yadav with 'Aurangzeb' and 'Kans', saying parents will now desist from naming their sons Akhilesh."Akhilesh did what 'Aurangzeb' and 'Kans' did not do. Due to his deeds, parents will now desist from naming their sons Akhilesh", he said at an election meeting in Bhadohi."Akhilesh is now saying that he would develop the state in next five years, if given a chance, but what was he doing in the past five years?" he asked.Aurangzeb was a controversial Mughal ruler, while the mythological character of Kans is considered the tyrant ruler of Vrishni kingdom with its capital at Mathura.UP GOVT PATRONISED TERRORISTS: ADITYANATHThe firebrand BJP MP from Gorakhpur said the state government patronised terrorists, anti-socials and rapists."Whatever scheme the government ran, it was only for a particular community," he alleged.Listing out the work which the BJP would do if voted to power, Adityanath said, "We will promote traditional industries and send bangles to Akhilesh and Rahul Gandhi from Firozabad bangle industry and constitute anti-Romeo squad for UP minister Azam Khan."He said a BJP government in Uttar Pradesh will pave way for Ram temple and that money will be spent on development.FINAL PHASE OF CAMPAIGNING ENDED ON MONDAY	BJP MP Yogi Adityanath has likened Uttar Pradesh Chief Minister Akhilesh Yadav to 'Aurangzeb' and 'Kans' due to his deeds, parents will now desist from naming their sons. "Akhileshe is now saying that he would develop the state in next five years, if given a chance," he said.

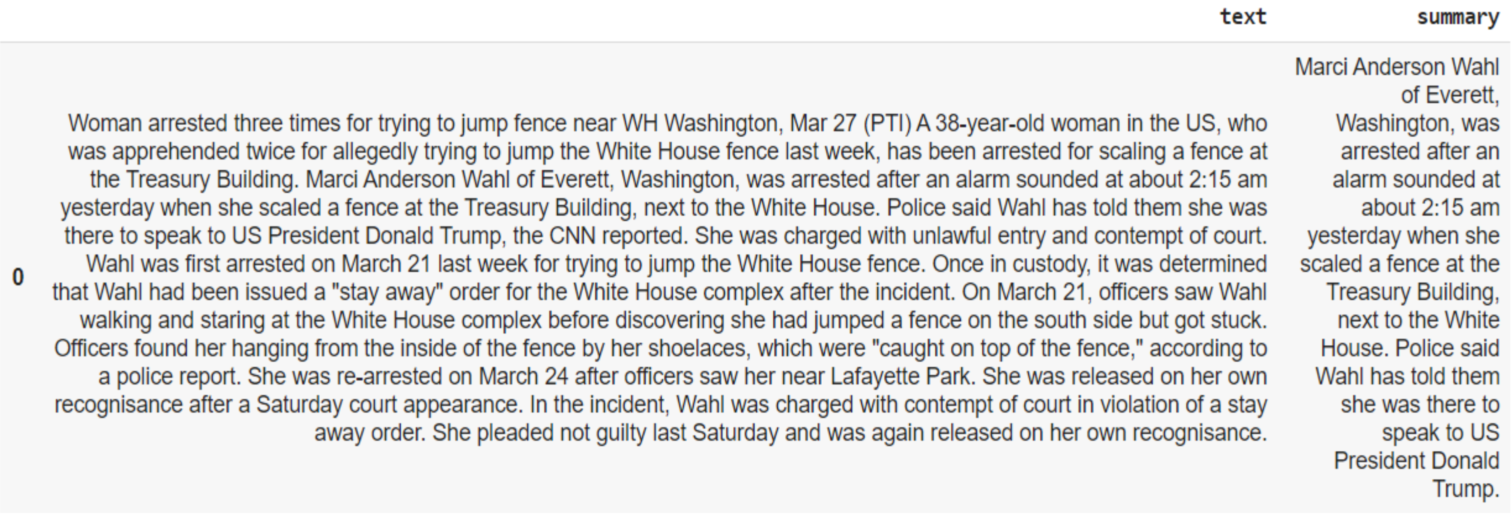
Summarisation after training:



	text	summary
0	Parents will not name their sons Akhilesh now: Adityanath BJP MP Yogi Adityanath has likened Uttar Pradesh Chief Minister Akhilesh Yadav with 'Aurangzeb' and 'Kans', saying parents will now desist from naming their sons Akhilesh."Akhilesh did what 'Aurangzeb' and 'Kans' did not do. Due to his deeds, parents will now desist from naming their sons Akhilesh", he said at an election meeting in Bhadohi."Akhilesh is now saying that he would develop the state in next five years, if given a chance, but what was he doing in the past five years?" he asked.Aurangzeb was a controversial Mughal ruler, while the mythological character of Kans is considered the tyrant ruler of Vrishni kingdom with its capital at Mathura.UP GOVT PATRONISED TERRORISTS: ADITYANATHThe firebrand BJP MP from Gorakhpur said the state government patronised terrorists, anti-socials and rapists."Whatever scheme the government ran, it was only for a particular community," he alleged.Listing out the work which the BJP would do if voted to power, Adityanath said, "We will promote traditional industries and send bangles to Akhilesh and Rahul Gandhi from Firozabad bangle industry and constitute anti-Romeo squad for UP minister Azam Khan."He said a BJP government in Uttar Pradesh will pave way for Ram temple and that money will be spent on development.FINAL PHASE OF CAMPAIGNING ENDED ON MONDAY	BJP MP Yogi Adityanath has likened Uttar Pradesh Chief Minister Akhilesh Yadav with 'Aurangzeb' and 'Kans', saying parents will now desist from naming their sons. "We will promote traditional industries and send bangles to Akhileh and Rahul Gandhi from Firozabad Bangle Industry and constitute anti-Romeo squad for UP minister Azam

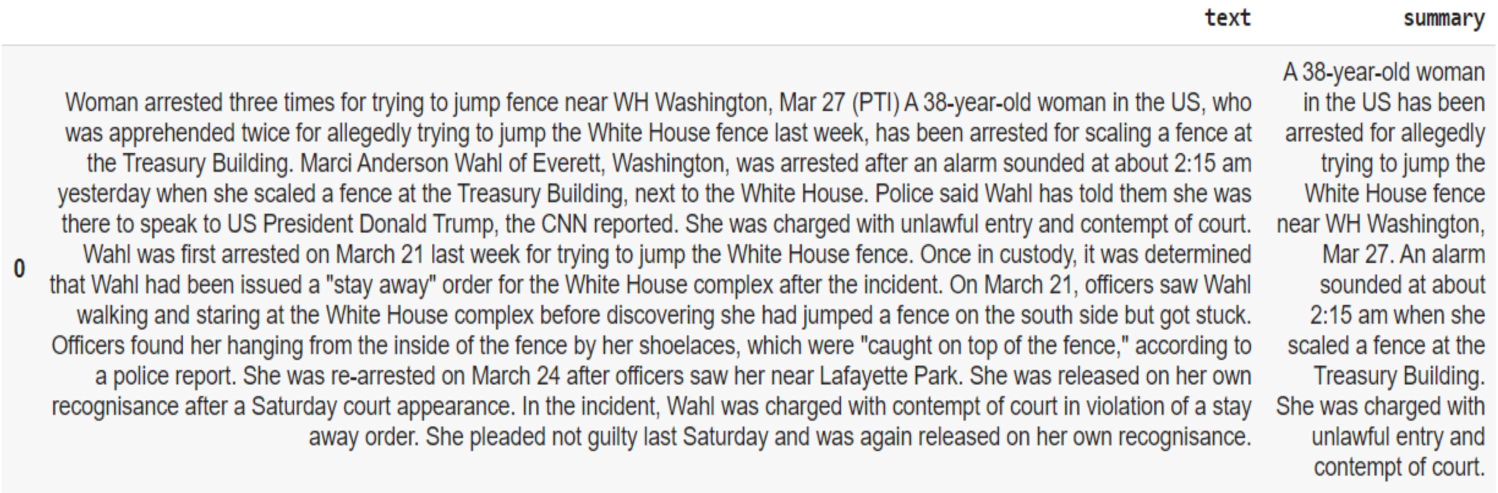
- **Example from test set:**

Summarisation before training:



	text	summary
0	Woman arrested three times for trying to jump fence near WH Washington, Mar 27 (PTI) A 38-year-old woman in the US, who was apprehended twice for allegedly trying to jump the White House fence last week, has been arrested for scaling a fence at the Treasury Building. Marci Anderson Wahl of Everett, Washington, was arrested after an alarm sounded at about 2:15 am yesterday when she scaled a fence at the Treasury Building, next to the White House. Police said Wahl has told them she was there to speak to US President Donald Trump, the CNN reported. She was charged with unlawful entry and contempt of court. Wahl was first arrested on March 21 last week for trying to jump the White House fence. Once in custody, it was determined that Wahl had been issued a "stay away" order for the White House complex after the incident. On March 21, officers saw Wahl walking and staring at the White House complex before discovering she had jumped a fence on the south side but got stuck. Officers found her hanging from the inside of the fence by her shoelaces, which were "caught on top of the fence," according to a police report. She was re-arrested on March 24 after officers saw her near Lafayette Park. She was released on her own recognisance after a Saturday court appearance. In the incident, Wahl was charged with contempt of court in violation of a stay away order. She pleaded not guilty last Saturday and was again released on her own recognisance.	Marci Anderson Wahl of Everett, Washington, was arrested after an alarm sounded at about 2:15 am yesterday when she scaled a fence at the Treasury Building, next to the White House. Police said Wahl has told them she was there to speak to US President Donald Trump.

Summarisation results after training:



	text	summary
0	Woman arrested three times for trying to jump fence near WH Washington, Mar 27 (PTI) A 38-year-old woman in the US, who was apprehended twice for allegedly trying to jump the White House fence last week, has been arrested for scaling a fence at the Treasury Building. Marci Anderson Wahl of Everett, Washington, was arrested after an alarm sounded at about 2:15 am yesterday when she scaled a fence at the Treasury Building, next to the White House. Police said Wahl has told them she was there to speak to US President Donald Trump, the CNN reported. She was charged with unlawful entry and contempt of court. Wahl was first arrested on March 21 last week for trying to jump the White House fence. Once in custody, it was determined that Wahl had been issued a "stay away" order for the White House complex after the incident. On March 21, officers saw Wahl walking and staring at the White House complex before discovering she had jumped a fence on the south side but got stuck. Officers found her hanging from the inside of the fence by her shoelaces, which were "caught on top of the fence," according to a police report. She was re-arrested on March 24 after officers saw her near Lafayette Park. She was released on her own recognisance after a Saturday court appearance. In the incident, Wahl was charged with contempt of court in violation of a stay away order. She pleaded not guilty last Saturday and was again released on her own recognisance.	A 38-year-old woman in the US has been arrested for allegedly trying to jump the White House fence near WH Washington, Mar 27. An alarm sounded at about 2:15 am when she scaled a fence at the Treasury Building. She was charged with unlawful entry and contempt of court.

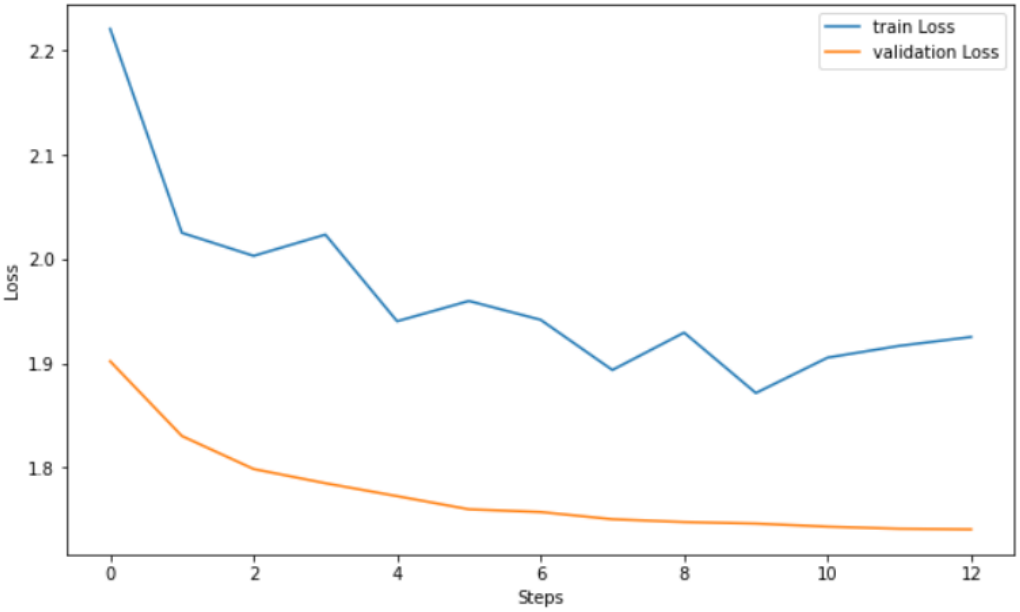
So from above tables we can see that even though results on pre-trained model are good there are still some minute details which are captured after training the model.

If we consider an example from a test set before training then we can see that it captures details like the name of the woman, timings at which the alarm sounded, Treasury building, name of US president, etc. While in the same summary obtained after training, details like age of woman, and the place where the event took place which is WH Washington among other details is captured.

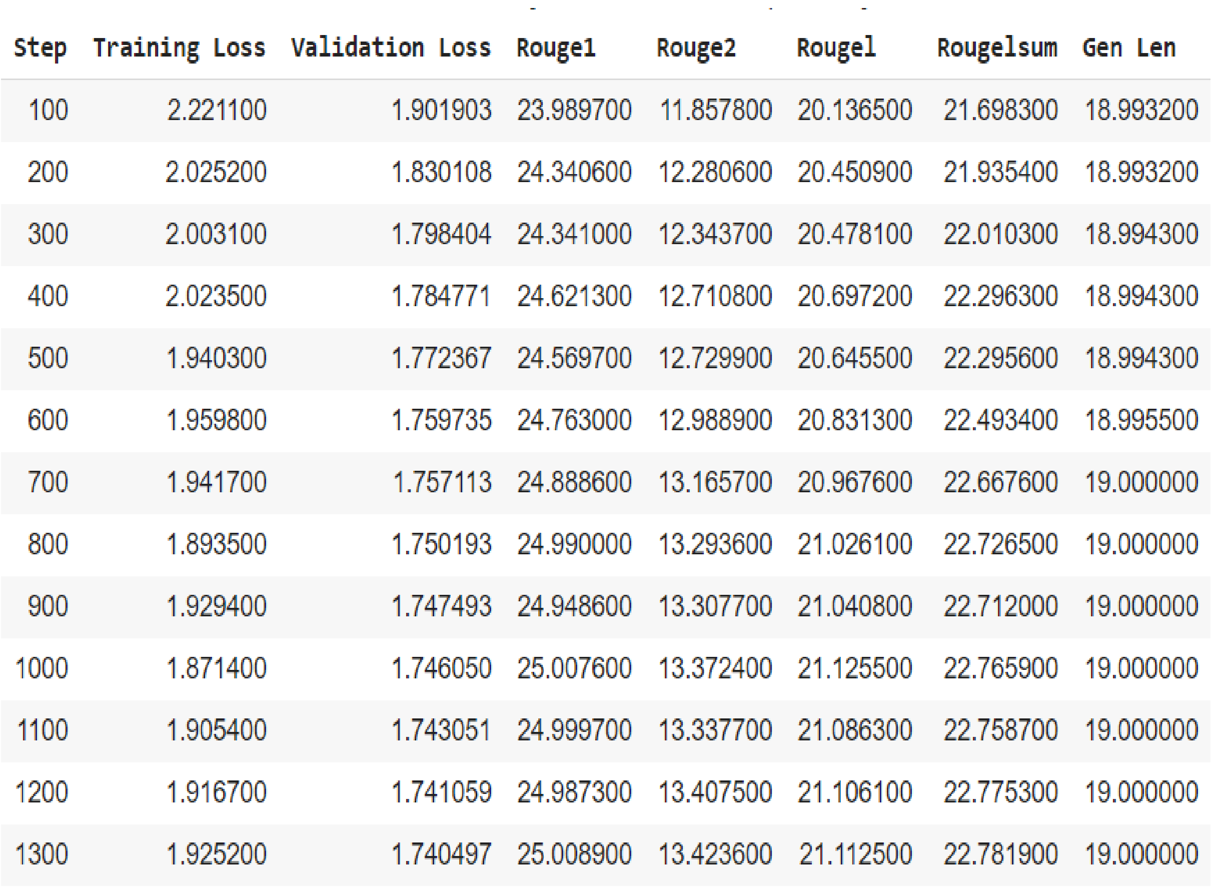
Model details:

- Size of trained model: 231MB
- Number of parameters of model:

Plotting steps vs loss for train and validation set:



Model Evaluation using Rouge scores:



Step	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
100	2.221100	1.901903	23.989700	11.857800	20.136500	21.698300	18.993200
200	2.025200	1.830108	24.340600	12.280600	20.450900	21.935400	18.993200
300	2.003100	1.798404	24.341000	12.343700	20.478100	22.010300	18.994300
400	2.023500	1.784771	24.621300	12.710800	20.697200	22.296300	18.994300
500	1.940300	1.772367	24.569700	12.729900	20.645500	22.295600	18.994300
600	1.959800	1.759735	24.763000	12.988900	20.831300	22.493400	18.995500
700	1.941700	1.757113	24.888600	13.165700	20.967600	22.667600	19.000000
800	1.893500	1.750193	24.990000	13.293600	21.026100	22.726500	19.000000
900	1.929400	1.747493	24.948600	13.307700	21.040800	22.712000	19.000000
1000	1.871400	1.746050	25.007600	13.372400	21.125500	22.765900	19.000000
1100	1.905400	1.743051	24.999700	13.337700	21.086300	22.758700	19.000000
1200	1.916700	1.741059	24.987300	13.407500	21.106100	22.775300	19.000000
1300	1.925200	1.740497	25.008900	13.423600	21.112500	22.781900	19.000000

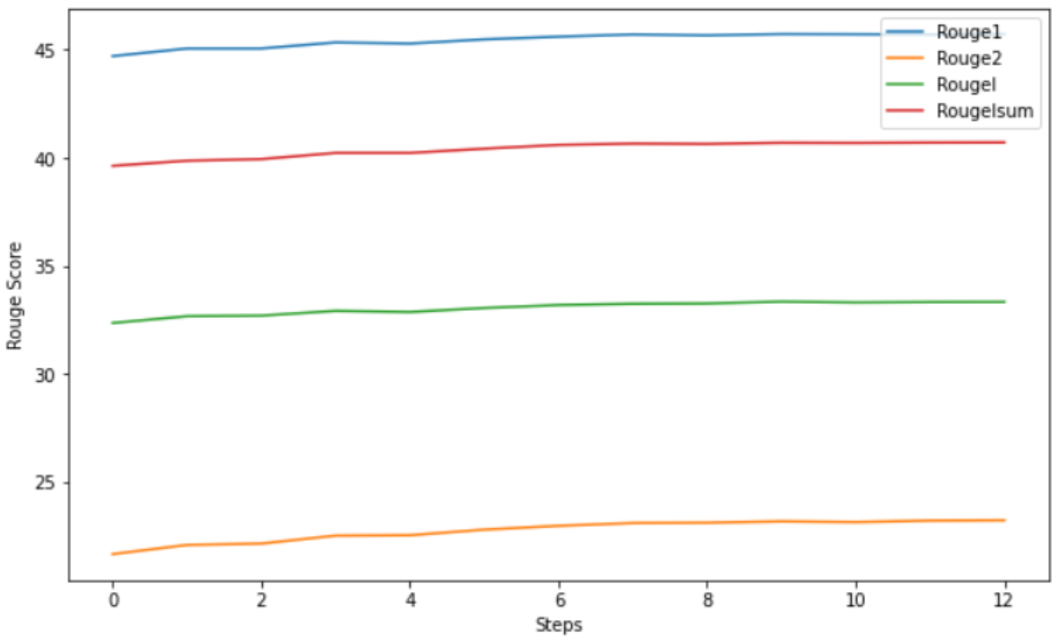
Rouge metric is most widely used for automatic evaluation of machine translation and text summarization.

Rouge-n: This is a recall based metric and it will calculate the score based on the number of matching n-grams between the text generated by the model and the reference text. Now when we consider single word which is overlapping then it will be a 1-gram model and hence the score will be Rouge-1, for bi-gram it will check for two overlapping words between target and input and hence it will be a Rouge-2 metric and so on. We can write a simple formula for Rouge-n as follows:

$$Rouge - n = \frac{Common\ n{-}grams\ obtained\ from\ model\ and\ reference\ summary}{number\ of\ n{-}grams\ obtained\ from\ reference\ summary}$$

Rouge-l: This measure is built on a concept of longest matching subsequence (LCS) between two summary sentences. This measure tells how similar both summaries are by considering the matches found in a sequence.

Plotting steps vs Rouge1, Rouge2, Rouge-L and Rouge-sum:



Types of evaluation:

The evaluation kind be done based on the various parameters and they include

- Fluency
- Concise
- Grammatical Correctness
- Readable
- Meaningful

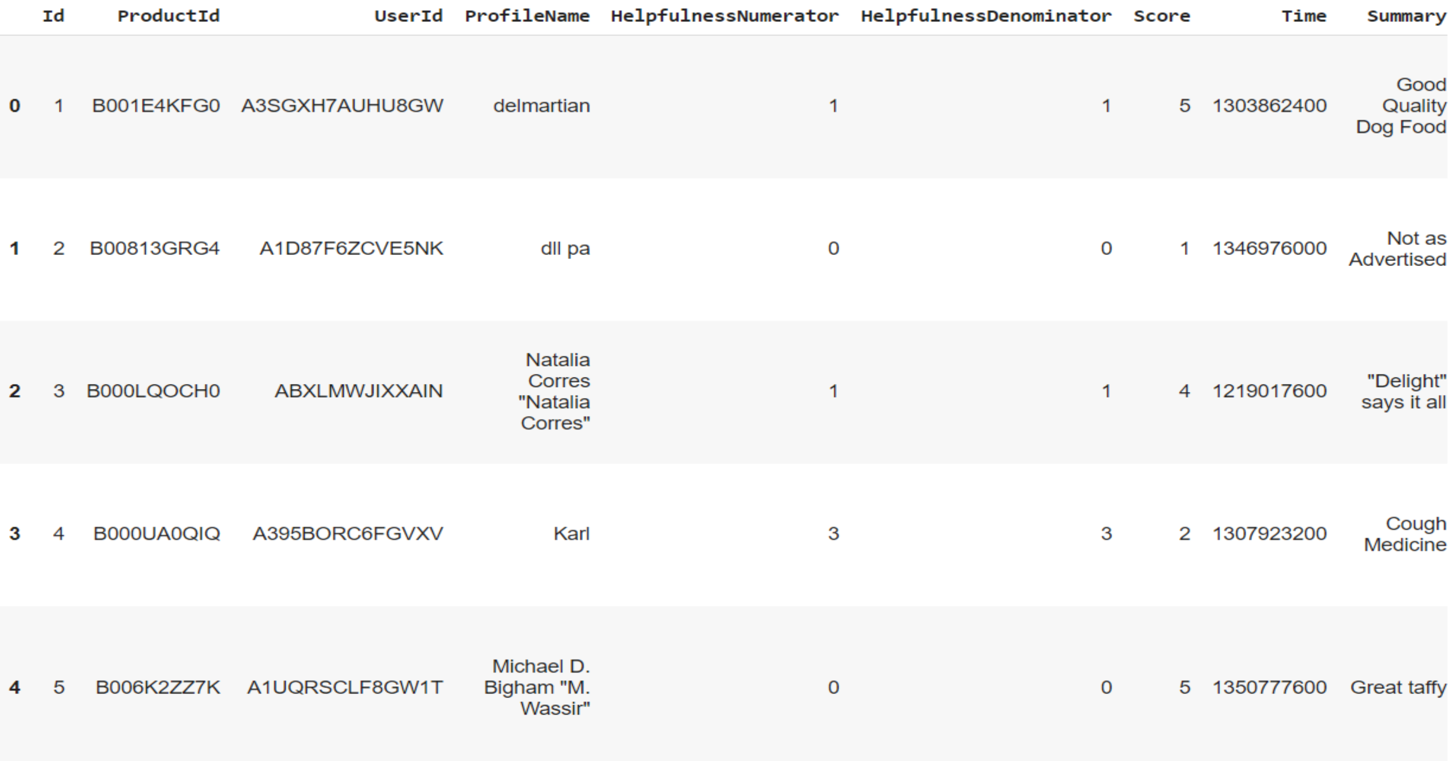
BONUS Question:

Amazon Food Dataset

Dataset Description:

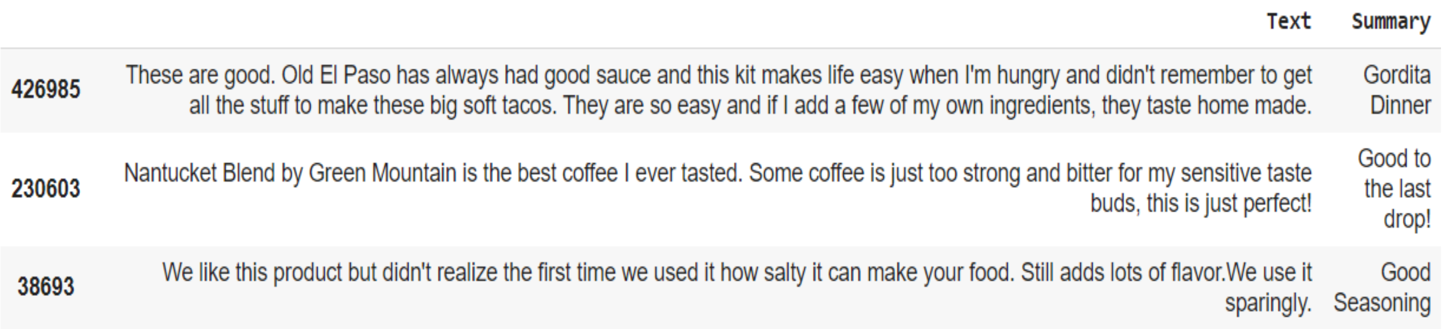
This dataset consists of reviews of fine foods from amazon. This dataset has around 500,000 reviews up to October 2012. The reviews include product and user information, ratings, and a plain text review. The reviews are from various Amazon categories.

Dataset visualization:



	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food
1	2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	1346976000	Not as Advertised
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy

Now as only text and summary columns are required so we take out text and summary columns

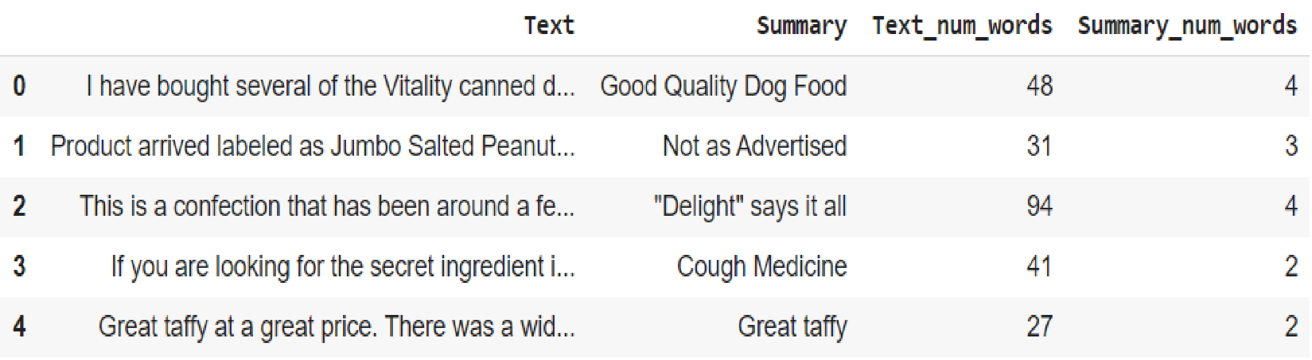


	Text	Summary
426985	These are good. Old El Paso has always had good sauce and this kit makes life easy when I'm hungry and didn't remember to get all the stuff to make these big soft tacos. They are so easy and if I add a few of my own ingredients, they taste home made.	Gordita Dinner
230603	Nantucket Blend by Green Mountain is the best coffee I ever tasted. Some coffee is just too strong and bitter for my sensitive taste buds, this is just perfect!	Good to the last drop!
38693	We like this product but didn't realize the first time we used it how salty it can make your food. Still adds lots of flavor.We use it sparingly.	Good Seasoning

Data pre-processing:

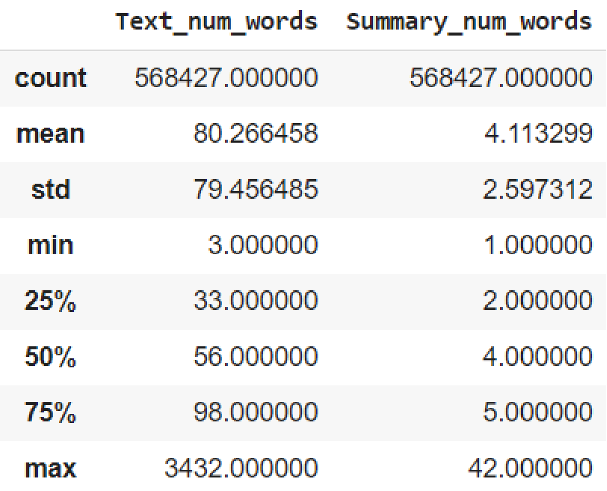
Out of 568454 reviews only 568427 reviews have summaries so we drop the reviews where there is no summary available.

Getting length of each reviews:



	Text	Summary	Text_num_words	Summary_num_words
0	I have bought several of the Vitality canned d...	Good Quality Dog Food	48	4
1	Product arrived labeled as Jumbo Salted Peanut...	Not as Advertised	31	3
2	This is a confection that has been around a fe...	"Delight" says it all	94	4
3	If you are looking for the secret ingredient i...	Cough Medicine	41	2
4	Great taffy at a great price. There was a wid...	Great taffy	27	2

Dataset statistics:



	Text_num_words	Summary_num_words
count	568427.000000	568427.000000
mean	80.266458	4.113299
std	79.456485	2.597312
min	3.000000	1.000000
25%	33.000000	2.000000
50%	56.000000	4.000000
75%	98.000000	5.000000
max	3432.000000	42.000000

Here we can see that the news dataset and the amazon food reviews dataset are drastically different in terms of length of reviews. Here the text is around 80 words and the summary is on average 4 words whereas in the news dataset summary itself was around 80 words.

5-fold validation:

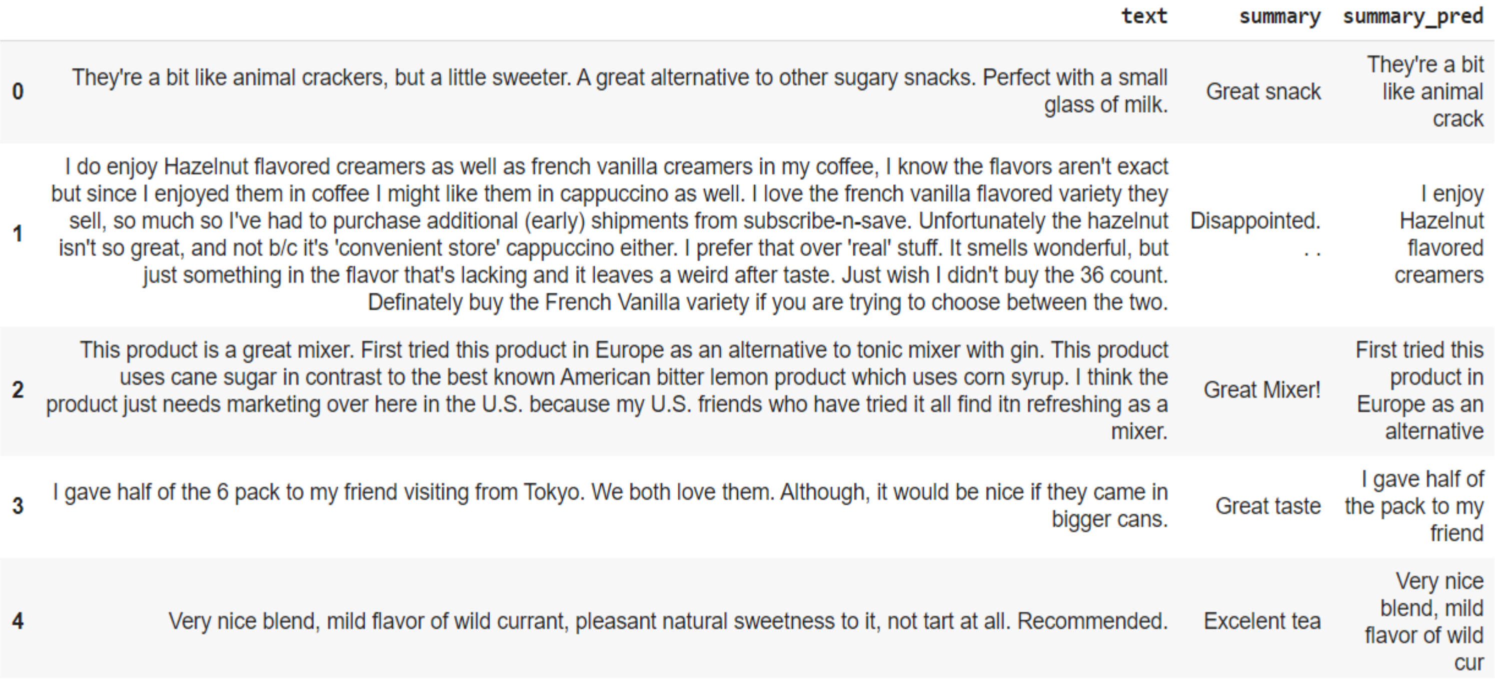
Here, we make 5 folds out of 20 samples, that is, we take a sample of 20 reviews from the data using a for loop and make 5 dataframes each having 20 samples. These data frames are stored in a list of dataframe named `df_folds`.

Now we get into each fold, take out its text, summary and predicted summary into fold_data list. Here we are using a batch of size 8. So as each fold has 20 samples, and we are taking them in a batch of 8 so there will be 3 batches for each fold. First batch will have 8 samples, the second batch will have 8 samples and the third batch will have 4 samples. This is because we are processing each fold, 8 samples at a time.

```
fold: 0
      batch: 0
      batch: 1
      batch: 2
fold: 1
      batch: 0
      batch: 1
      batch: 2
fold: 2
      batch: 0
      batch: 1
      batch: 2
fold: 3
      batch: 0
      batch: 1
      batch: 2
fold: 4
      batch: 0
      batch: 1
      batch: 2
```

So for each fold index 0 will have text, index 1 will have summary and index 2 will have predicted summary.

Showing 5 random samples from random fold :



	text	summary	summary_pred
0	They're a bit like animal crackers, but a little sweeter. A great alternative to other sugary snacks. Perfect with a small glass of milk.	Great snack	They're a bit like animal crack
1	I do enjoy Hazelnut flavored creamers as well as french vanilla creamers in my coffee, I know the flavors aren't exact but since I enjoyed them in coffee I might like them in cappuccino as well. I love the french vanilla flavored variety they sell, so much so I've had to purchase additional (early) shipments from subscribe-n-save. Unfortunately the hazelnut isn't so great, and not b/c it's 'convenient store' cappuccino either. I prefer that over 'real' stuff. It smells wonderful, but just something in the flavor that's lacking and it leaves a weird after taste. Just wish I didn't buy the 36 count. Definately buy the French Vanilla variety if you are trying to choose between the two.	Disappointed. . .	I enjoy Hazelnut flavored creamers
2	This product is a great mixer. First tried this product in Europe as an alternative to tonic mixer with gin. This product uses cane sugar in contrast to the best known American bitter lemon product which uses corn syrup. I think the product just needs marketing over here in the U.S. because my U.S. friends who have tried it all find itn refreshing as a mixer.	Great Mixer!	First tried this product in Europe as an alternative
3	I gave half of the 6 pack to my friend visiting from Tokyo. We both love them. Although, it would be nice if they came in bigger cans.	Great taste	I gave half of the pack to my friend
4	Very nice blend, mild flavor of wild currant, pleasant natural sweetness to it, not tart at all. Recommended.	Excelent tea	Very nice blend, mild flavor of wild cur

Results:

```
fold: 0
      rouge1 0.09307692307692308
      rouge2 0.055704099821746886
      rougeL 0.09077935222672065
      rougeLsum 0.0915587044534413

fold: 1
      rouge1 0.061743742368742374
      rouge2 0.0
      rougeL 0.05410846098346098
      rougeLsum 0.054542957042957044

fold: 2
      rouge1 0.11257066462948816
      rouge2 0.05
      rougeL 0.11210349291231644
      rougeLsum 0.11265682030387912

fold: 3
      rouge1 0.05193910256410257
      rouge2 0.0
      rougeL 0.045576923076923084
      rougeLsum 0.04536858974358974

fold: 4
      rouge1 0.10483544233544234
      rouge2 0.01818181818181818
      rougeL 0.10218253968253968
      rougeLsum 0.10328074703074705
```

Above scores are not as good as scores we got on the news dataset. There are multiple reasons for this.

- Even though both are summarization tasks, there is an important difference - news data is generating summary from full text while food data is more like generating headlines from given reviews. This causes a big change in how the model was trained.
    - We can avoid this length issue by training the original model of news summary to headline generation as news summary are of similar length to reviews.
- Another reason is the domain of the data itself, original data is generic in nature while food data is quite specific.

- For all folds, Rouge2 score is very low - this makes sense as Rouge2 measures bigram overlap and since output headlines are of few words (4 on average) only - there aren't many bigrams to match.

To Do

1. NER visulization
1. Rouge Scores explanation--- done
2. FineTuning Explanation
3. Add number of parameters for project
4. General Information about summarization systems
    a. Types
    b. Different types of approaches
    c. Prominent Datasets
5. Comparison of project with main paper models
6. Number of parameters and size of model -- project
7. RASA theory - overall structure
8. Maybe Some Information on Attention -- I think jyada ho jayega

RASA Links

An Open-Source Chatbot Made With Rasa

Watch this, will give a very good understanding of what we can do

Also report me help karega coz it explains the whole process quite well

How to Create Conversational AI Chatbot using RASA (Python) by Cisco Data Scientist

- response, utter, actions @ 50 min
  utter - static - hardcoded
  actions - find entities

  One class per actions @ 57 min

  https://in.springboard.com/blog/chatbot-using-rasa/
  https://in.springboard.com/blog/ai-chatbot-using-rasa/
  https://in.springboard.com/blog/best-ai-chatbot-using-rasa/

NLP Project: Automated Question Answering from FAQs Using Word-Embeddings

- Take questions and answers.
    - Augment this by adding chit-chat questions and answers
- For a given question, find it's similarity with given question and return the answer.

Building a chatbot with Rasa NLU and Rasa Core

-- this is old, baad me dekhte if we need it

https://rasa.com/docs/rasa-x/
https://rasa.com/docs/rasa/
- https://rasa.com/docs/rasa/generating-nlu-data
  - https://github.com/RasaHQ/**NLU-training-data**

- https://rasa.com/docs/rasa/writing-stories
- https://rasa.com/docs/rasa/chitchat-faqs
- https://rasa.com/docs/rasa/fallback-handoff

- https://rasa.com/docs/rasa/training-data-format
- https://rasa.com/docs/rasa/tuning-your-model

Data

https://github.com/benoitalvarez/**Covid-19-QBox-ChatbotModel**
- https://medium.com/qbox-nlp-performance-tooling/new-model-for-rapid-deployment-of-covid-19-trained-chatbots-8806fb94a8ff

https://github.com/Archish27/nora-covid-19-bot
- https://github.com/Archish27/nora-covid-19-bot/blob/master/actions/actions.py

https://github.com/narendraprasath/COVID-19-Tracker-Bot/
- https://github.com/narendraprasath/COVID-19-Tracker-Bot/blob/master/COVID-19-Tracker-Chatbot_Code/actions.py

Assignment2:
https://docs.google.com/document/d/15a1Jma2AJK9EEc5QAsQi7f6G8tgj76_5VTJyb2e8Sdg/edit