Qualitative Evaluation of abstractive text summarization using Google's Flan-T5 and Pegasus models on the Kaggle News articles Dataset and Hindi news articles

UNIVERSITY OF HOUSTON

Name: LEELA PRASAD BODDU

UHID: 2214707

Table of contents:

1) Novelty:

Using reference paper [5], LLMs trained models/ libraries mentioned in them, and by leveraging these neural networks and NLP techniques, this project aims to automatically generate concise summaries from lengthy texts, empowering individuals, and companies to quickly extract key information and improve information processing efficiency. The novelty of this project is to guide users on which model to choose based on the size, and training time of these models along with comparing the best model from my project reference paper [5] against the latest model Flan-T5 which was published on Dec 2022 in Kaggle article’s dataset (A different dataset from the dataset used in reference paper [5]). Another novel improvement is the project expands its scope on Hindi dataset to summarize to train and evaluate Flan-T5 model.

2) Introduction:

This project is a comparative analysis of the flan-t5 model and the PEGASUS algorithm, employing a Kaggle news article dataset as the testing ground. Its real-world applications are guiding users to select models based on their sizes, time complexity also expanding the scope of the summarization models. The primary aim is to assess and contrast their summarization prowess, with a particular focus on the quality of generated summaries. The novelty lies in the comprehensive investigation of the flan-T5 model's performance (The latest state-of-the-art model with no research papers on this model), as well as its potential to outperform the established PEGASUS model in terms of summarization quality. By leveraging Rouge-1, Rouge-2, and Rouge-L scores as performance metrics on the generated summaries against the actual summaries, a qualitative analysis of the models is performed.

3) Text summarization:

Text summarization is the process of condensing text into concise summaries that contain the key information about the text. Automated text summarization has become critical for enhanced productivity and ease of comprehension. This transformation saves time and empowers individuals to quickly comprehend the core message without delving into the entire document.

Text summarization techniques can be broadly categorized into two main types based on the methods used: extractive summarization and abstractive summarization. These techniques employ different strategies to condense text while retaining its essential information.

Extractive Text Summarization:

Extractive summarization involves selecting and extracting existing sentences or phrases from the original text to create a summary. These techniques retain the original wording and context, making them more interpretable and preserving the document's integrity. Extractive summarization, mostly use statistical techniques that perform sentence scoring technique which obtains the text's keyword. It is done by analyzing and filtering the words which are used most frequently in the text. The sentences with a high frequency of these words are used for generating a summary of the original text by using the sentences with high scores in decreasing order. They use computational resources and require less time to perform the task.

Abstractive Text Summarization:

Abstractive text summarization uses NLP to perform summarization. Summaries generated by abstractive summarization might not be composed of original sentences or words and might have been replaced by morphed sentences and new words. Summaries generated by abstractive summarization are more comparable to human-generated summaries. But these models consume a lot of time and computational resources to generate summaries.

This project uses two abstractive text summarization models and compares the quality of their summaries using rouge as performance metrics.

4) Google's Flan-T5 model:

Flan-T5 is a massive language model developed by Google in 2021 building on their previous T5 (Text-to-Text Transfer Transformer) architecture. Flan-T5 is a state-of-the-art natural language processing model trained on a 750 GB dataset, which performs a multitude of tasks (1.8k unique tasks), where it uses text input to generate text output. Five different sizes/ versions of this model are available: [1]

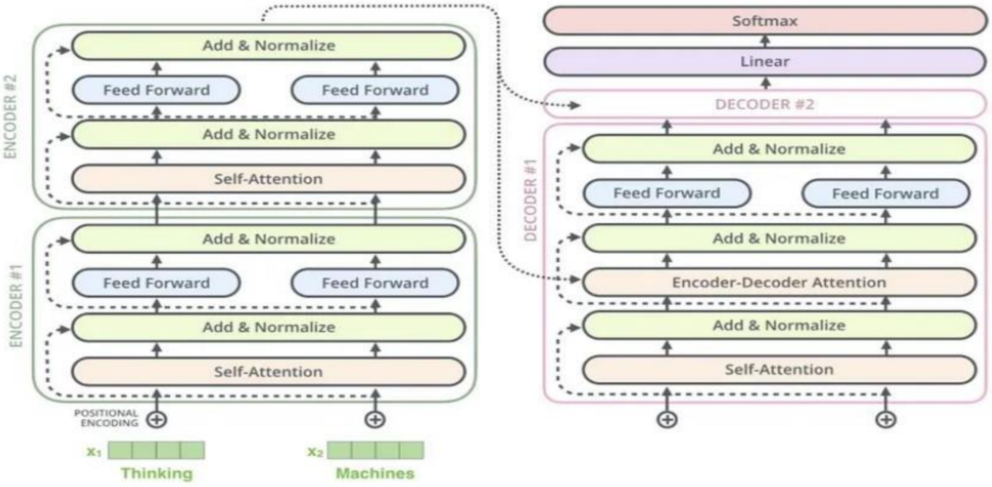
1) Flan-small: 77 million weights
2) Flan-base: 250 million weights
3) Flan-large: 780 million weights
4) Flan-XL: 3 billion weights
5) Flan-XLL: 11 billion weights

Architecture:

It is based on the T5 model architecture, which stands for Text-to-Text Transfer Transformer. The Flan-T5 Base model specifically has a size of 250 million weights. It follows the encoder-decoder structure, where the encoder processes the input text while also masking a small part of the text and the decoder generates the output text by first decoding the masked input and then using it along with the key sentences from the text input to generate an output that performs the specific task given in the input. The encoder maps an input text sequence into a continuous vector representation. It contains multiple transformer blocks, each with a multi-headed self-attention layer followed by a feedforward network.

The decoder takes this vector and generates an output sequence, one token at a time. The decoder also uses multi-headed self-attention between output tokens as well as attention over the input representation.

This architecture allows the model to take text as input and generate text as output. Pic [2]



Tokenization: The Flan-T5 base model utilizes tokenization to process input text. Tokenization is the process of breaking down the input text into smaller units called tokens. The input text is divided into individual words, punctuation marks, and other meaningful units. Each of these units is assigned a unique token. The tokenized input is then used as input to the Flan-T5-base model for further processing. Reference-1

In this project, the Flan-T5-base model which is almost 1GB in size, is used and trained on the Kaggle dataset to perform a specific task -- "Generate concise summaries from text". The process is discussed in the methodology section.

5) Pegasus Model:

(Pre-training with Extracted Gap-sentences for Abstractive Summarization) or PEGASUS is an abstractive summarization algorithm developed and published by Google in 2019. It is trained on more than 5TB datasets. PEGASUS is a sequence-to-sequence model designed for abstractive text summarization. It uses a pre-training objective called gap-sentences generation, where important sentences are removed from an input document and generated as one output sequence from the remaining sentences. The model is trained on a large variety of datasets and has demonstrated highly concise and coherent summaries. It is 2.3 GB in size and doesn't need training to generate summaries.

The PEGASUS model is a sequence-to-sequence model with a Transformer encoder-decoder architecture, but its novelty lies in its pre-training processes like Gap sentence Generation (GSG) and Masked Language Model (MLM).

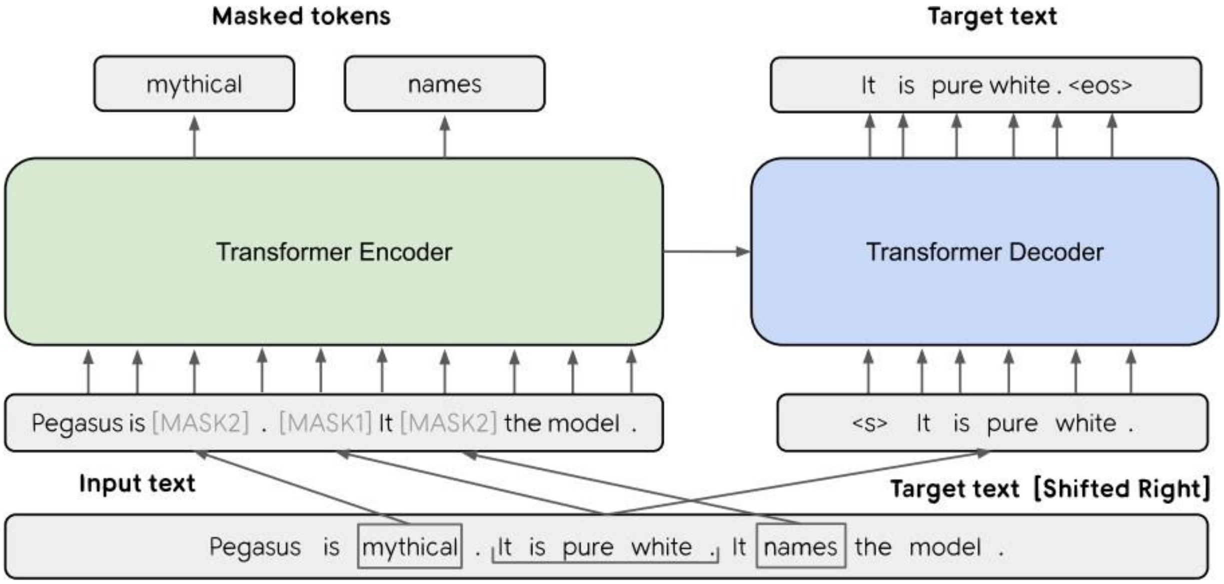
Gap sentence Generation (GSG):

In the GSG objective, important sentences are removed or masked from an input text, and the model generates these sentences as one output sequence from the remaining sentences.

Masked Language Model (MLM):

A certain percentage of tokens in the input text are replaced with a mask token and the encode is trained to predict the original token based on this masked token. Generally, 15% of the input text is masked for the encoder to predict.

Overall, PEGASUS offers a powerful approach to abstractive text summarization, leveraging pre-training with extracted gap sentences and achieving impressive results across diverse domains. [3]



Masked tokens
mythical
names
Target text
It is pure white . <eos>
Transformer Encoder
Transformer Decoder
Pegasus is [MASK2] . [MASK1] It [MASK2] the model .
<s> It is pure white .
Input text
Target text [Shifted Right]
Pegasus is mythical . It is pure white . It names the model .

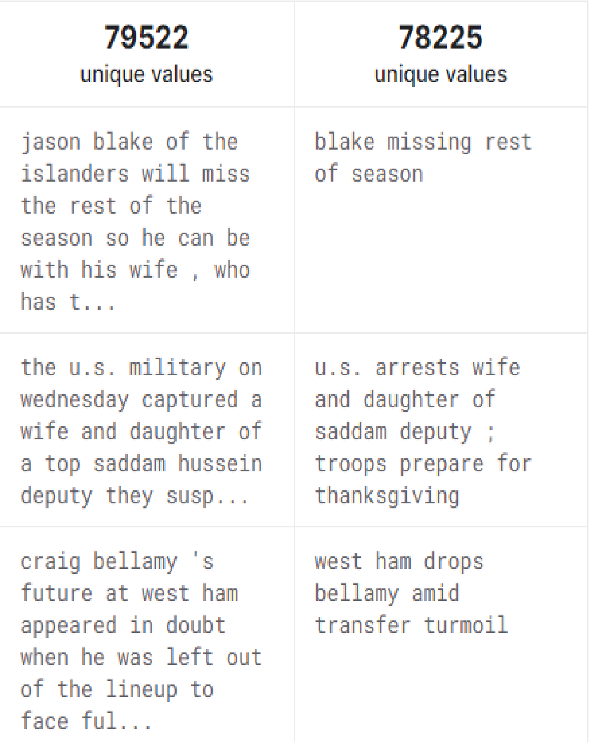
From the fig, we can see Mask 1 being GSG and Mask 2 being MLM used on the encoder-decoder architecture. The encoder predicts the mask 2 tokens and using the input text along with the predicted mask 2 tokens the decoder is trained to predict the target sentence in an abstractive method.

6) Dataset:

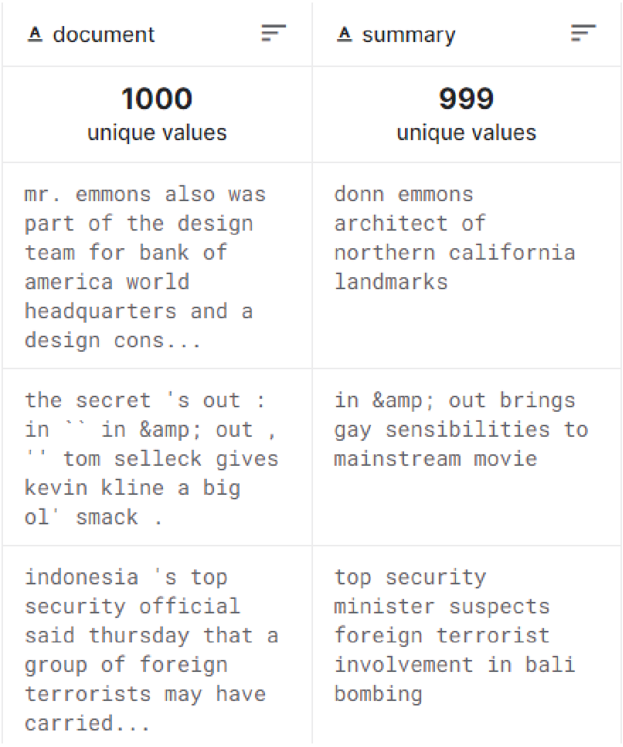
Using Kaggle's Text-summarization Dataset (source: BBC English summary datasets) , although it contains 5 datasets, this project utilizes train_ds.csv and val.csv which contain almost 80,000 articles (The Left table below) and their summaries and 999 articles (The right table below) and their summaries respectively.

From the train data, we use only 20000 rows after processing the articles for minimum 25 words and summaries with less than 18 words (these ranges were decided based on the histogram) and random sampling to optimize time and computational resources. So, the train to validation split is 5% (since only 1000 articles are used to evaluate).

The data preprocessing and tokenization (The longest article and summary length are 173 tokens and 50 tokens respectively) on the dataset are explained in the methodology.



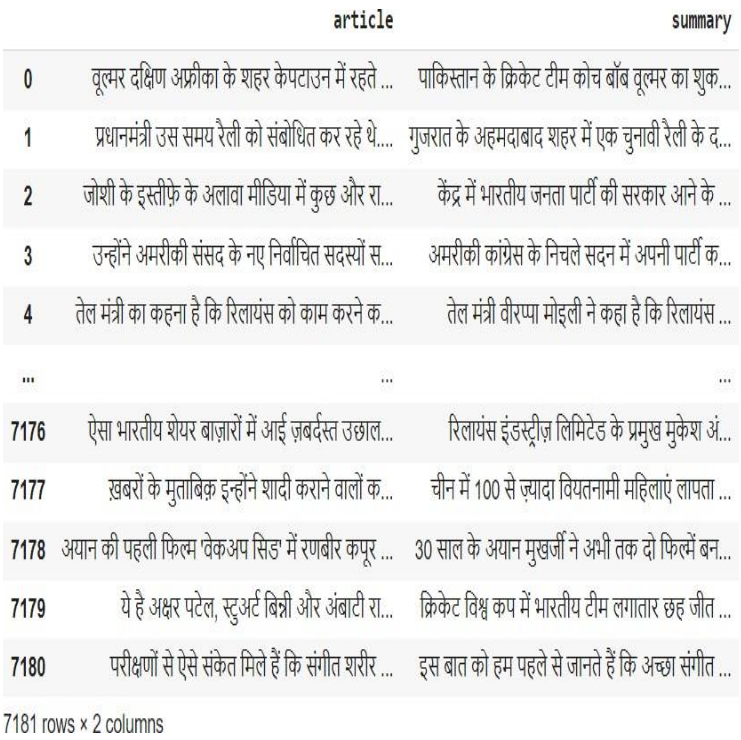
79522 unique values	78225 unique values
jason blake of the islanders will miss the rest of the season so he can be with his wife , who has t...	blake missing rest of season
the u.s. military on wednesday captured a wife and daughter of a top saddam hussein deputy they susp...	u.s. arrests wife and daughter of saddam deputy ; troops prepare for thanksgiving
craig bellamy 's future at west ham appeared in doubt when he was left out of the lineup to face ful...	west ham drops bellamy amid transfer turmoil



document	summary
1000 unique values	999 unique values
mr. emmons also was part of the design team for bank of america world headquarters and a design cons...	donn emmons architect of northern california landmarks
the secret 's out : in `` in &amp; out , '' tom selleck gives kevin kline a big ol' smack .	in &amp; out brings gay sensibilities to mainstream movie
indonesia 's top security official said thursday that a group of foreign terrorists may have carried...	top security minister suspects foreign terrorist involvement in bali bombing

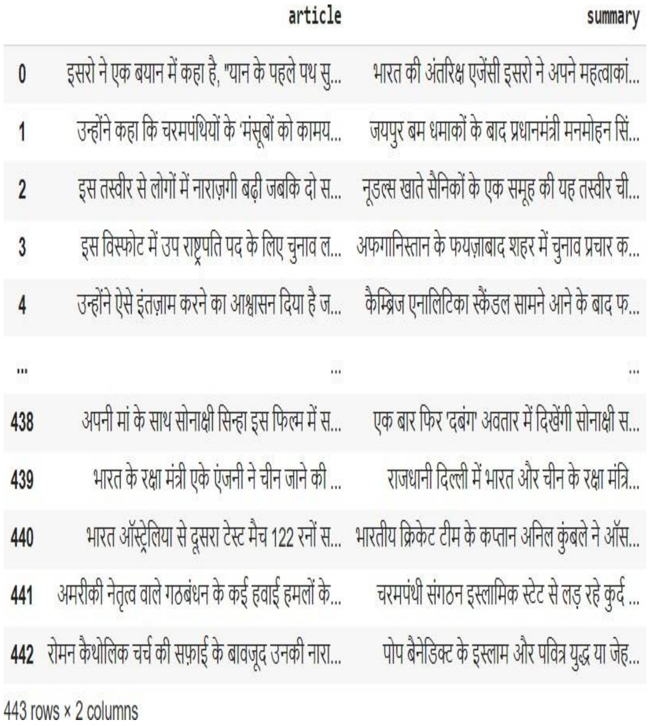
The datasets contain two columns Article and its respective human generated summaries. [4]

The new Hindi dataset is also from BBC Hindi articles it contains about 17500 articles and their respective summaries. The dataset is processed for articles with minimum of 50 words and less than 400 words, summaries with more than 19 words (these ranges are selected based on the histograms of the word count of the articles and summaries.), this processing left 7181 articles to feed into the model and 443 articles to evaluate the model. (these minimum word filters were placed as the tokenizer limit is reached and to optimize the computational resources). The longest article and summary length are 863 and 101 respectively.)



	article	summary
0	वूल्मर दक्षिण अफ्रीका के शहर केपटाउन में रहते ...	पाकिस्तान के क्रिकेट टीम कोच बॉब वूल्मर का शुक...
1	प्रधानमंत्री उस समय रैली को संबोधित कर रहे थे....	गुजरात के अहमदाबाद शहर में एक चुनावी रैली के द...
2	जोशी के इस्तीफ़े के अलावा मीडिया में कुछ और रा...	केंद्र में भारतीय जनता पार्टी की सरकार आने के ...
3	उन्होंने अमरीकी संसद के नए निर्वाचित सदस्यों स...	अमरीकी कांग्रेस के निचले सदन में अपनी पार्टी क...
4	तेल मंत्री का कहना है कि रिलायंस को काम करने क...	तेल मंत्री वीरप्पा मोइली ने कहा है कि रिलायंस ...
...	...	...
7176	ऐसा भारतीय शेयर बाज़ारों में आई ज़बर्दस्त उछाल...	रिलायंस इंडस्ट्रीज़ लिमिटेड के प्रमुख मुकेश अं...
7177	ख़बरों के मुताबिक़ इन्होंने शादी कराने वालों क...	चीन में 100 से ज़्यादा वियतनामी महिलाएं लापता ...
7178	अयान की पहली फिल्म 'वेकअप सिड' में रणबीर कपूर ...	30 साल के अयान मुखर्जी ने अभी तक दो फिल्में बन...
7179	ये है अक्षर पटेल, स्टुअर्ट बिन्नी और अंबाटी रा...	क्रिकेट विश्व कप में भारतीय टीम लगातार छह जीत ...
7180	परीक्षणों से ऐसे संकेत मिले हैं कि संगीत शरीर ...	इस बात को हम पहले से जानते हैं कि अच्छा संगीत ...

7181 rows × 2 columns



	article	summary
0	इसरो ने एक बयान में कहा है, "यान के पहले पथ सु...	भारत की अंतरिक्ष एजेंसी इसरो ने अपने महत्वाकां...
1	उन्होंने कहा कि चरमपंथियों के 'मंसूबों को कामय...	जयपुर बम धमाकों के बाद प्रधानमंत्री मनमोहन सिं...
2	इस तस्वीर से लोगों में नाराज़गी बढ़ी जबकि दो स...	नूडल्स खाते सैनिकों के एक समूह की यह तस्वीर ची...
3	इस विस्फोट में उप राष्ट्रपति पद के लिए चुनाव ल...	अफ़गानिस्तान के फयज़ाबाद शहर में चुनाव प्रचार क...
4	उन्होंने ऐसे इंतज़ाम करने का आश्वासन दिया है ज...	कैम्ब्रिज एनालिटिका स्कैंडल सामने आने के बाद फ...
...	...	...
438	अपनी मां के साथ सोनाक्षी सिन्हा इस फिल्म में स...	एक बार फिर 'दबंग' अवतार में दिखेंगी सोनाक्षी स...
439	भारत के रक्षा मंत्री एके एंजनी ने चीन जाने की ...	राजधानी दिल्ली में भारत और चीन के रक्षा मंत्रि...
440	भारत ऑस्ट्रेलिया से दूसरा टेस्ट मैच 122 रनों स...	भारतीय क्रिकेट टीम के कप्तान अनिल कुंबले ने ऑस...
441	अमरीकी नेतृत्व वाले गठबंधन के कई हवाई हमलों के...	चरमपंथी संगठन इस्लामिक स्टेट से लड़ रहे कुर्द ...
442	रोमन कैथोलिक चर्च की सफ़ाई के बावजूद उनकी नारा...	पोप बैनेडिक्ट के इस्लाम और पवित्र युद्ध या जेह...

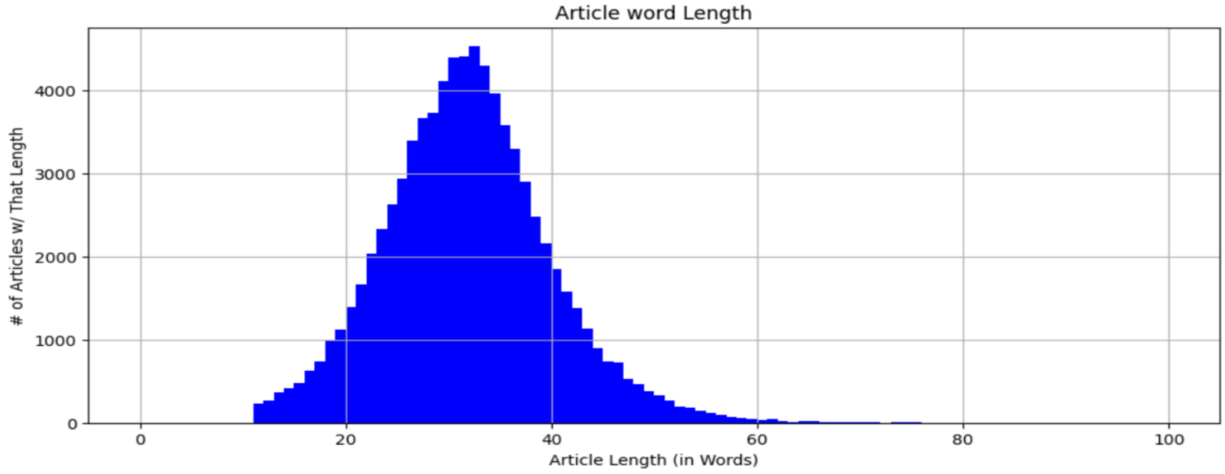
443 rows × 2 columns

The train and evaluation split for this evaluation is 5% after the preprocessing.

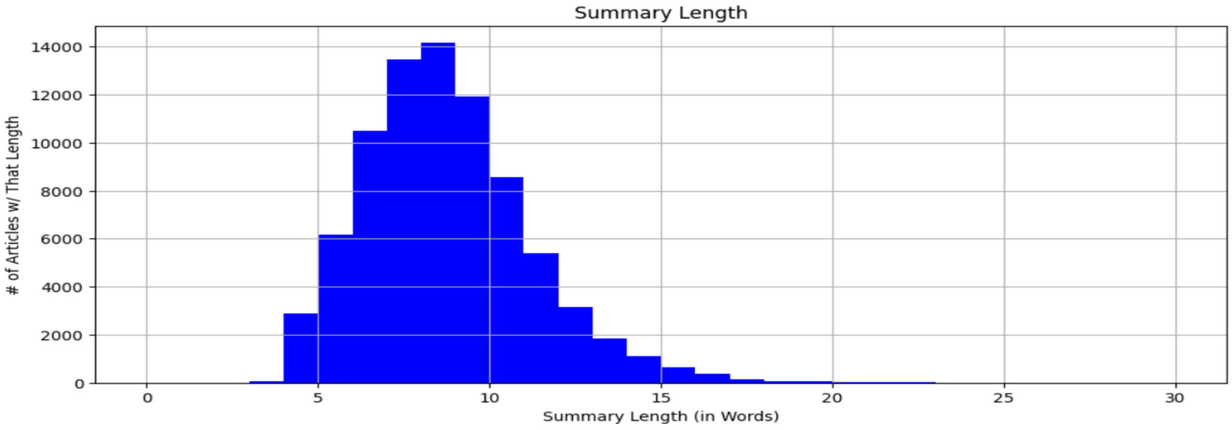
7) Methodology:

This project uses the Kaggle news article dataset—an collection of diverse news articles that serves as the substrate for evaluation. The flan-t5 model and the PEGASUS algorithm will be subjected to the dataset, generating summaries that encapsulate the articles' essence. By employing Rouge-1, Rouge-2, and Rouge-L scores—a recognized standard in evaluating summarization qualities.

```
The longest string (in words) in this dataset is 86 words.
```



```
The longest string (in words) in this dataset is 28 words.
```



The figures above show us a histogram, the word count of each article, and summaries (train and validation data).

The Flan-T5 model is trained on the train.csv data, where first, the articles are tokenized. The longest article and summary length are 173 tokens and 50 tokens respectively.

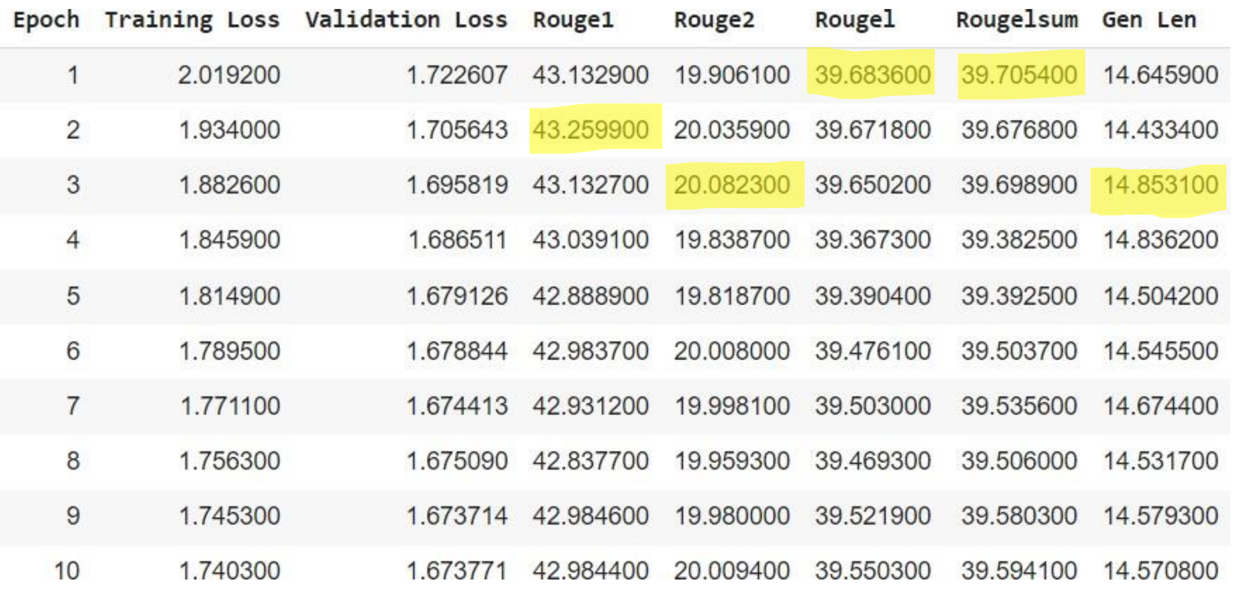
Loss Function: Both models use cross entropy as the loss function to minimize. Cross-entropy loss measures the dissimilarity between predicted and actual summary. For each token in the predicted sequence, the cross-entropy loss calculates the difference between the predicted probability distribution and the actual target distribution using negative log-likelihood, it penalizes deviations from the ground truth.

Rouge metrics: ROUGE measures the overlap between the n-gram units (words or phrases) in the generated summary and the reference summary.

- **Rouge-1:** It focuses on the overlapping unigrams among the generated and reference summaries.
  Rouge-1= no. of overlapping unigrams/ no. of unigrams in reference summary
- **Rouge-2:** It focuses on the overlapping bigrams among the generated and reference summaries.
  Rouge-2= no. of overlapping bigrams/ no. of bigram in reference summary
- **Rouge-L:** It focuses on the overlapping unigrams among the generated and reference summaries.
  Rouge-L= no. of overlapping Longest common substring/ no. of words in reference summary.

8) Results and Discussion:

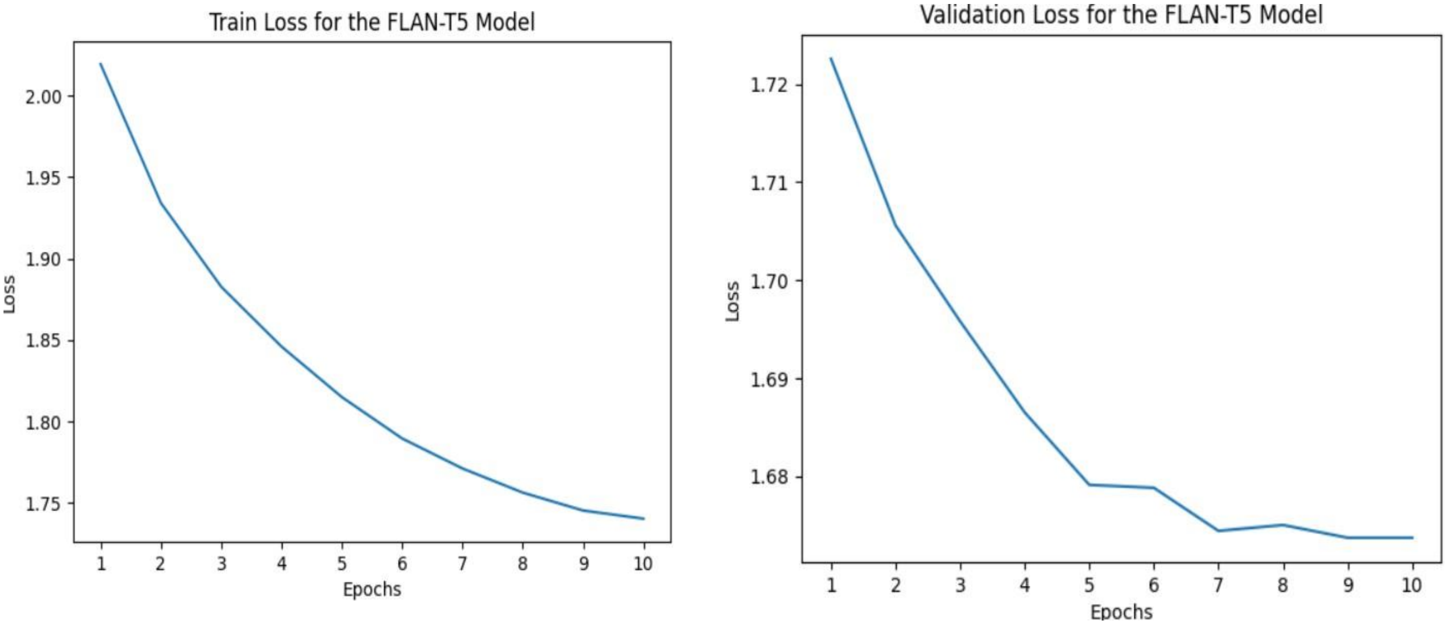
The Flan-T5-base model is trained on the train data. The model is iterated over this data over 10 epochs, (from the table below) it is observed the performance of the model on the rouge metrics dipped after the 3rd epoch and the loss reduction on the validation set was minimal after 4th epoch.



Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
1	2.019200	1.722607	43.132900	19.906100	39.683600	39.705400	14.645900
2	1.934000	1.705643	43.259900	20.035900	39.671800	39.676800	14.433400
3	1.882600	1.695819	43.132700	20.082300	39.650200	39.698900	14.853100
4	1.845900	1.686511	43.039100	19.838700	39.367300	39.382500	14.836200
5	1.814900	1.679126	42.888900	19.818700	39.390400	39.392500	14.504200
6	1.789500	1.678844	42.983700	20.008000	39.476100	39.503700	14.545500
7	1.771100	1.674413	42.931200	19.998100	39.503000	39.535600	14.674400
8	1.756300	1.675090	42.837700	19.959300	39.469300	39.506000	14.531700
9	1.745300	1.673714	42.984600	19.980000	39.521900	39.580300	14.579300
10	1.740300	1.673771	42.984400	20.009400	39.550300	39.594100	14.570800

The Cross-Entropy Loss is used to calculate the Loss of the model through each epoch discussed in methodology. The model reduces the train loss through each epoch although the train loss and validation loss are reducing after each epoch, the loss reduction is minimal in the order of 0.005 and the performance is not improving significantly even after the first epoch (The epoch rouge scores even after perfectly optimizing the loss would not improve significantly as the model is close to overfitting).

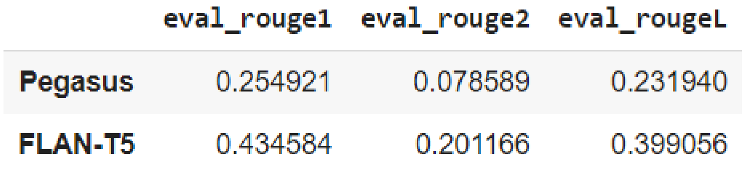
The batch size for training is 25, learning rate is 0.00002, and weight decay is 0.01. These parameters were chosen to optimize training time and computational resources.



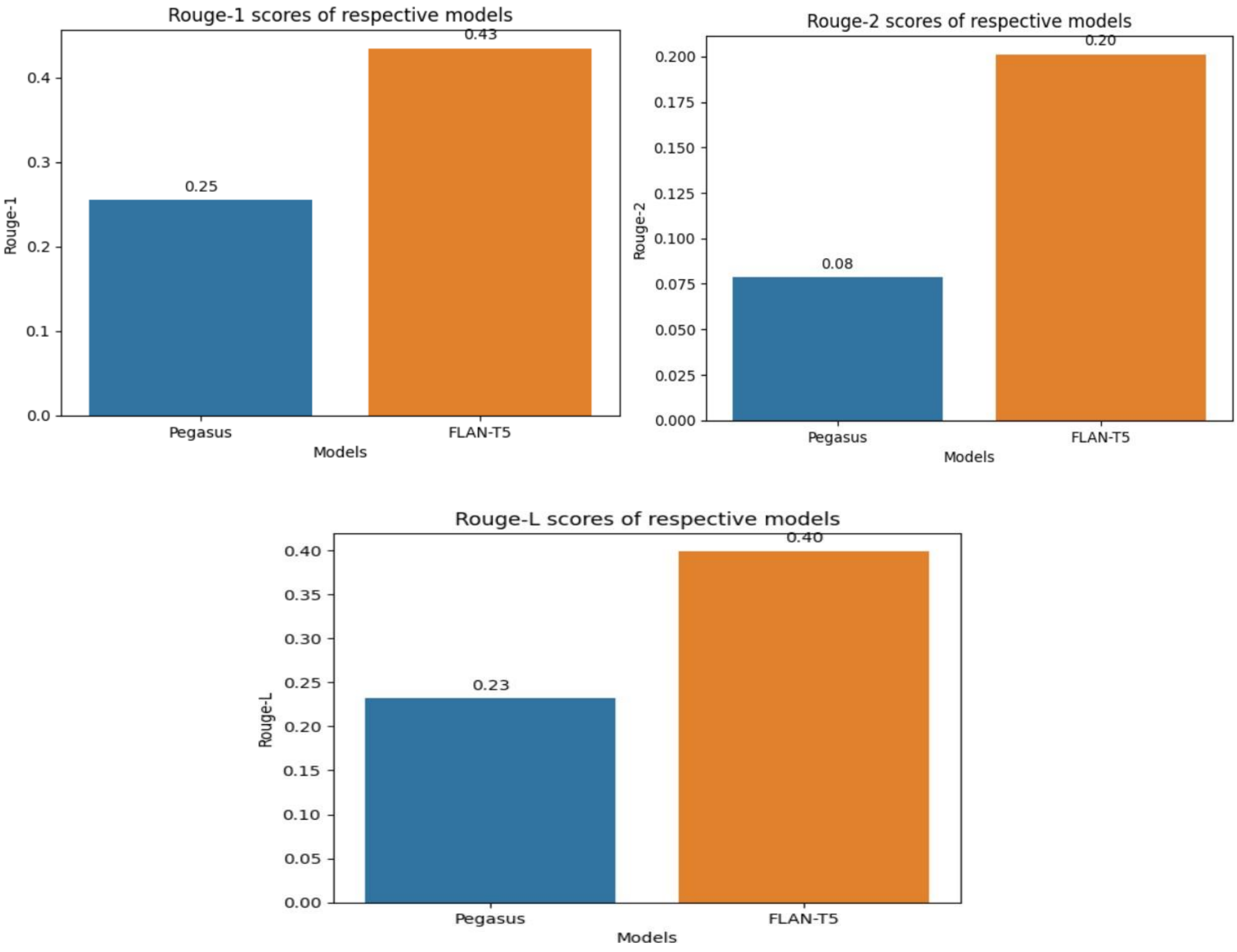
From the table above we can generate these figures of loss plots for training and validation data.

From the Validation Loss plots we can see that the loss reduced steadily till epoch 5 and fluctuated after that and the loss reduction was minimal (order of 0.0001). Another insight is that although the loss decreased over the epochs, the rouge score was slightly better for epoch 2.

The Final results of the models on the evaluation dataset are mentioned in the table below.



	eval_rouge1	eval_rouge2	eval_rougeL
Pegasus	0.254921	0.078589	0.231940
FLAN-T5	0.434584	0.201166	0.399056

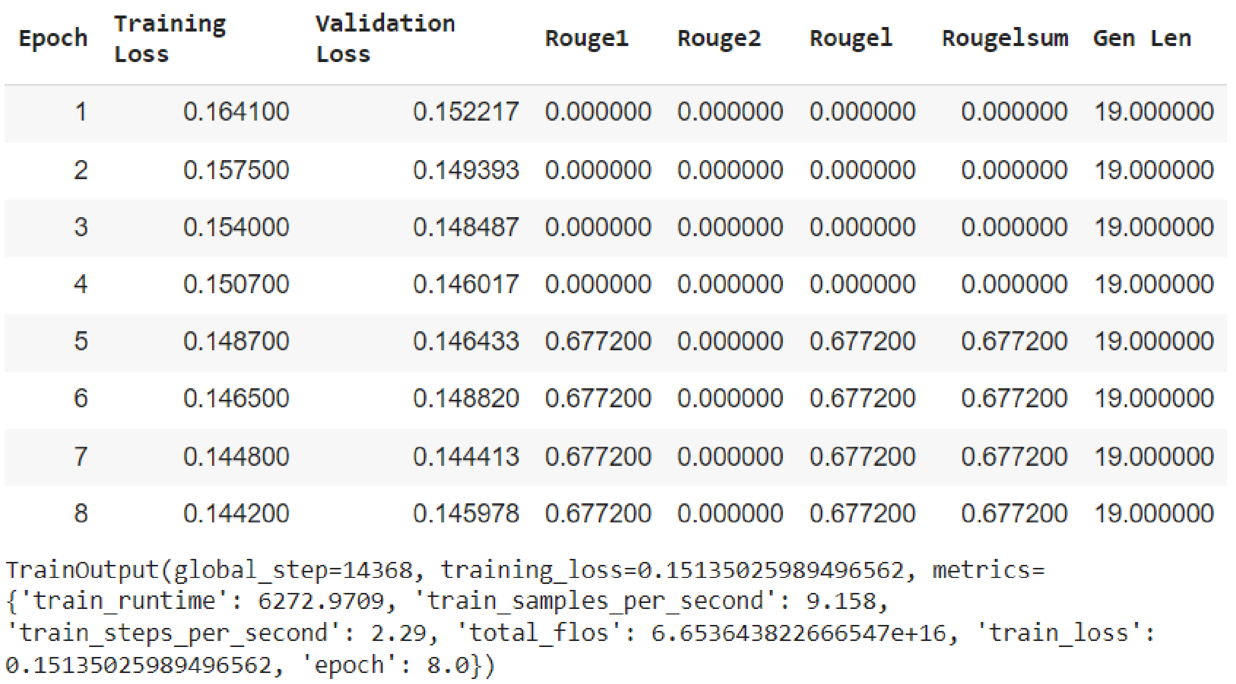


The above figures are the rouge-1, rouge-2, and rouge-L scores of the respective models.

The rouge-1, rouge-2, and rouge-L scores of Pegasus and Flan-T5 models are 0.25, 0.43; 0.07, 0.2; and 0.23, 0.4 respectively (rouge scores above 0.4 are considered good). The Flan-T5 model clearly performed better on the dataset on all the rouge metrics. The figures provide empirical evidence that the performance of the Flan-T5 model has outperformed Pegasus in every Rouge metric.

Results from Hindi dataset used to train Flan-T5 model:

The Flan-T5 base model was trained on a Hindi training dataset of about 7200 articles after preprocessing to meet the model tokenizer limits and optimize resources. The training batch size is 4, epochs are 8, and learning rate is 0.00002.



Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
1	0.164100	0.152217	0.000000	0.000000	0.000000	0.000000	19.000000
2	0.157500	0.149393	0.000000	0.000000	0.000000	0.000000	19.000000
3	0.154000	0.148487	0.000000	0.000000	0.000000	0.000000	19.000000
4	0.150700	0.146017	0.000000	0.000000	0.000000	0.000000	19.000000
5	0.148700	0.146433	0.677200	0.000000	0.677200	0.677200	19.000000
6	0.146500	0.148820	0.677200	0.000000	0.677200	0.677200	19.000000
7	0.144800	0.144413	0.677200	0.000000	0.677200	0.677200	19.000000
8	0.144200	0.145978	0.677200	0.000000	0.677200	0.677200	19.000000

```
TrainOutput(global_step=14368, training_loss=0.15135025989496562, metrics=
{'train_runtime': 6272.9709, 'train_samples_per_second': 9.158,
'train_steps_per_second': 2.29, 'total_flos': 6.653643822666547e+16, 'train_loss':
0.15135025989496562, 'epoch': 8.0})
```

The results of the model on this dataset are 0.0067 rouge-1 score and all the other metrics are zero, This is interesting as the model is performing very poorly on the dataset although it is trained on Hindi language. We can see the model validation loss decreased and started to increase again at 7th epoch.

9) Conclusion and future scope:

In conclusion, this project provides insight into the need for automatic text summarization. The models used in the project are discussed in detail, using rouge as evaluation metrics we conclude that the Flan-T5 model has outperformed the Pegasus model on English BBC dataset. Using a new focused dataset this project was able to evaluate the models on generalized topics for summarization. Flan-T5 being the latest state-of-the-art model, assessing its summarization prowess is novel and valuable.

The performance of Flan-T5-base model on Hindi dataset is very poor, possible reasons for this performance might be the relatively small dataset used in training, possible improper tokenization, this raises a question whether the model is learning anything as the loss in very small even in the first epoch, which saw a steady decline till 4th epoch and reached a minimal validation loss for 7th epoch.

Future development:

The possible reasons for this outcome might be contributed to the training of the Flan-T5 model over the direct evaluation of the Pegasus model. Also, Pegasus parameters can be tuned to improve the performance of the model. The training approach for Flan-T5 vs the pre-trained Pegasus model can impact relative performance.

Another possible improvements in the Hindi summarization model are using a larger dataset, large number of epochs to know the ground truth, better tokenization methods possible change in evaluation metrics other than rouge scores (BLEU scores).

Overall, even with the suggested improvements, the Pegasus model might not outperform the Flan-T5 model. Although Flan-T5 model is a latest state of the art model trained on generalized on multiple domains, it struggled on the Hindi BBC dataset which can be the researched as future scope of the project.

10) References:

1. Flan-T5 models: https://arxiv.org/pdf/2210.11416v5.pdf
2. T5- Transformer architecture: medium.com/analytics-Vidhya
3. Pegasus model: https://arxiv.org/pdf/1912.08777v2.pdf
4. Kaggle dataset: Kaggle Text summarization data
5. Qualitative analysis reference: Research paper