Group 4
Summarization Project

Alexis Kaldany
Ricardo Diaz
Sagar Tripathi

Scope

Analyzing the relationship between the summary/text ratio and the effectiveness of the model. Our hypothesis is that there is no relationship between the summary/text ratio and the effectiveness of the model.

HUGGING FACE

text_ratio =

"summary_words_length" / "text_words_length"

Dataset

These datasets contain face-to-face spoken dialogues that cover a wide range of daily-life topics, including schooling, work, medication, shopping, leisure, travel. Most conversations take place between friends, colleagues, and between service providers and customers.

text: text of dialogue.

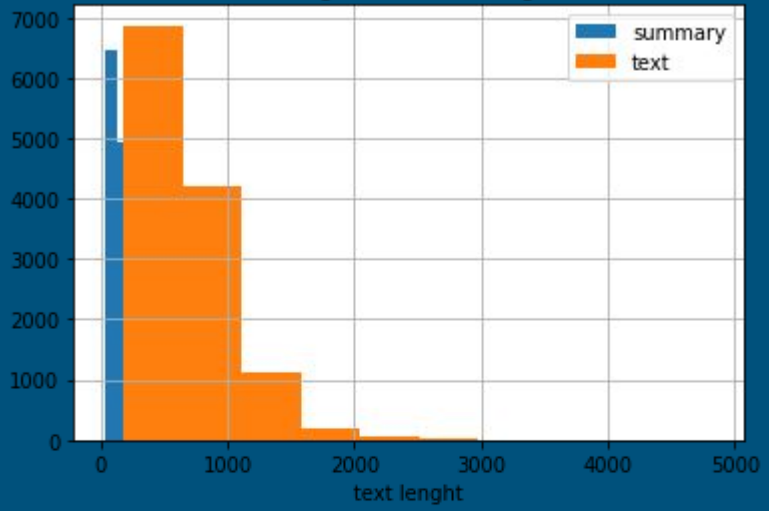
summary: human written summary of the dialogue.

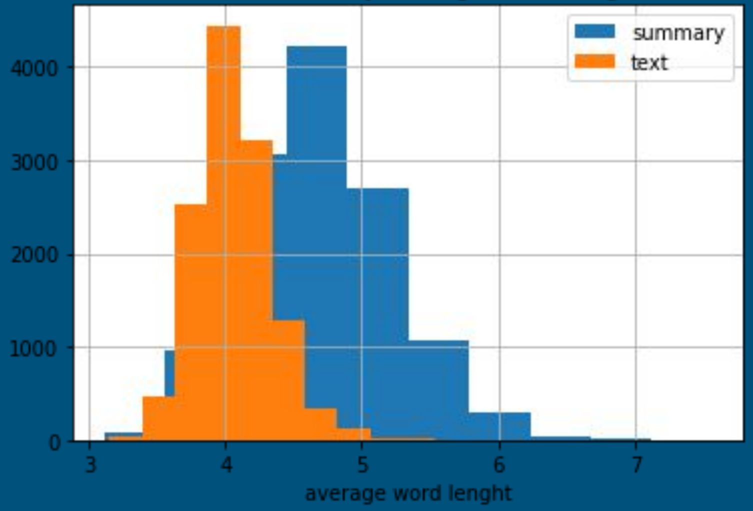
topic: human written topic/one liner of the dialogue.

id: unique file id of an example.

knkarthick/dialogsum · Datasets at Hugging Face

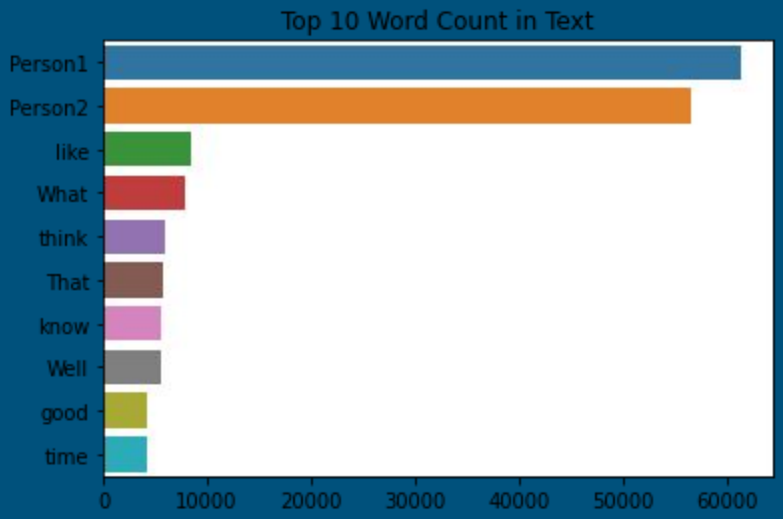
EDA


Histogram of Text Lenght


Text vs Summary Average Word Lenght

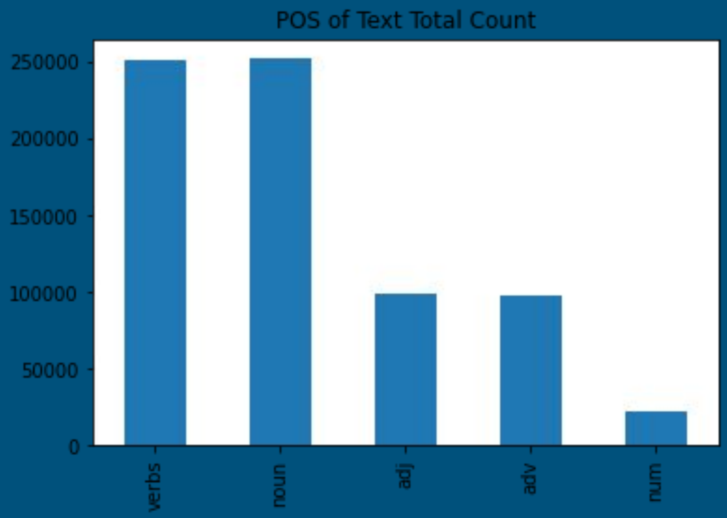
Dialogue between two people

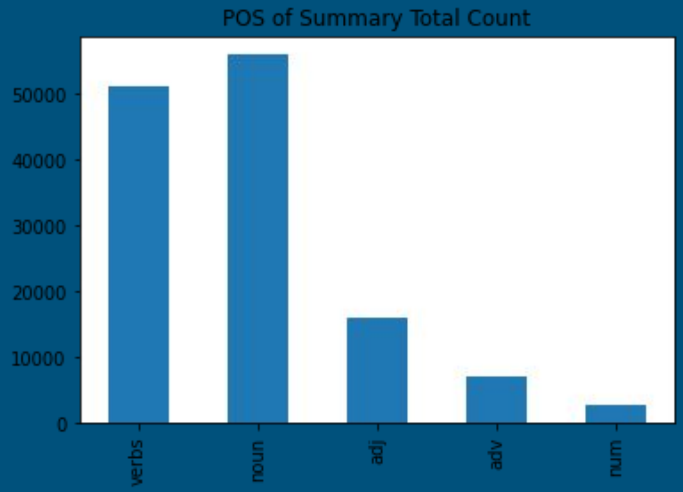
After getting the top word count from the dataset we noticed the data is a dialogue is between two people.



Difference between Part of Speech

In the summarize text the POS with more count are nouns, comparing to the normal text that has more verbs



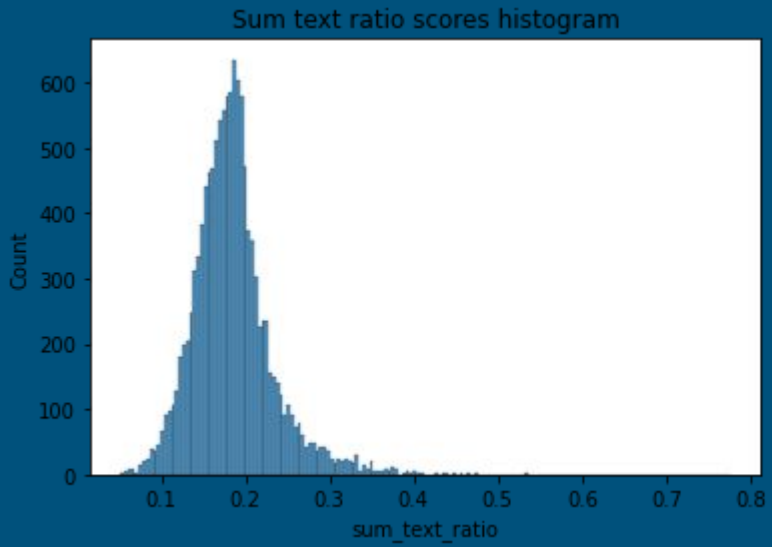


Metric Scores Distribution

The metric created text_ratio, which compares the difference between numbers of words in summarize and the text, has a normal distribution.

text_ratio =

"summary_words_length" / "text_words_length"

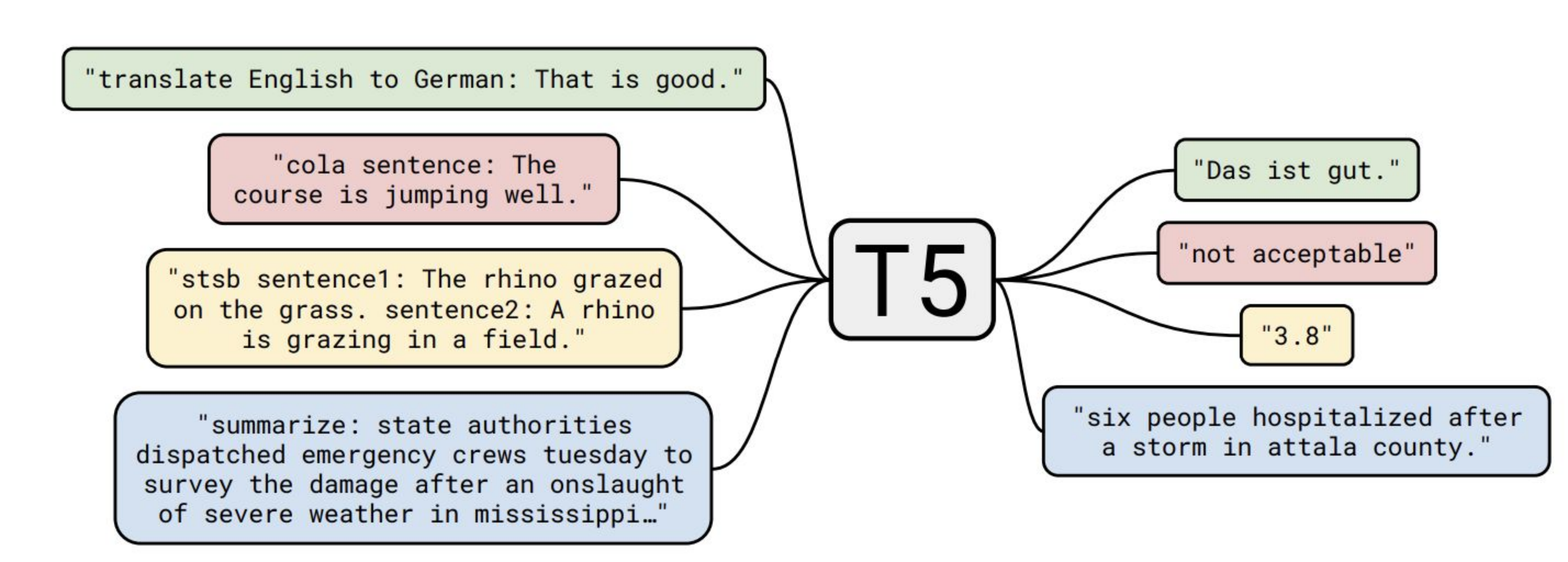


Summarization Plan

1. Identify model for summarization task
2. Identify metrics to calculate effectiveness of model
3. Train the model till metrics plateau.
4. Run model on evaluation set and save predicted summaries
5. Compare predicted summaries to correct summaries to generate metrics which will be used for hypothesis testing

Model Choice for Summarization Task
T5 = **Text-to-Text Transfer Transformer**

- Tried a variety of models using AutoModel, decided on T5-small
  - "facebook/blenderbot_small-90M",”bert-base-uncased”, “t5”, “t5-small”
- 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.
- T5 works well on a variety of tasks out-of-the-box by prepending a different prefix to the input corresponding to each task
  - for translation: “translate English to German: …”
  - for summarization: “summarize: ….”
- T5-small pre-trained on “Colossal Clean Crawled Corpus (C4)”
- T5 is primarily a sequence to sequence model



"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."

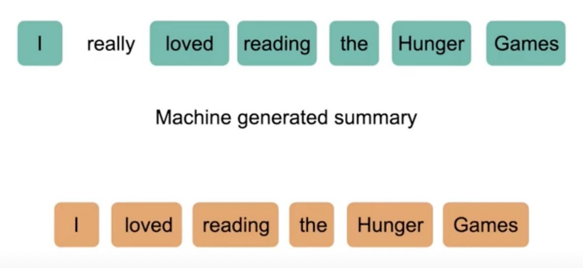
Cola sentence= is sentence grammatically correct
Stsb = sentence similarity

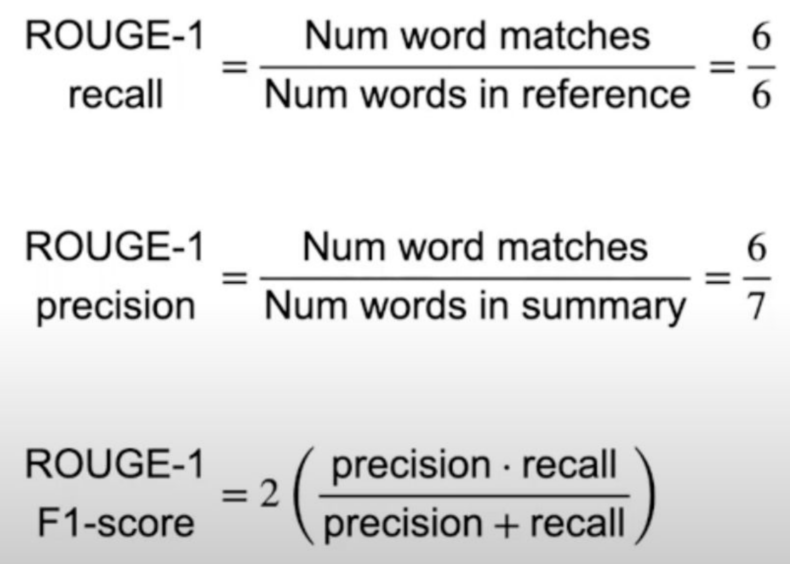
Parameters and Training

- Parameters:
  - Learning Rate = 2e-5
  - Optimizer = AdamW algorithm
- Training
  - Number of Epochs = 5
  - Each epoch takes ~2 hours
- Evaluation
  - ROUGE-1
  - ROUGE-2
  - ROUGE-L

ROUGE

- ROUGE = **R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation
- Compares an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation
- ROUGE-1 refers to the overlap of unigram (each word) between the system and reference summaries
- ROUGE-2 refers to the overlap of bigrams between the system and reference summaries
- ROUGE-L: Longest Common Subsequence (LCS)[3] based statistics. Longest common subsequence problem takes into account sentence level structure similarity naturally and identifies longest co-occurring in sequence n-grams automatically


I really loved reading the Hunger Games
Machine generated summary
I loved reading the Hunger Games



$$\text{ROUGE-1 recall} = \frac{\text{Num word matches}}{\text{Num words in reference}} = \frac{6}{6}$$

$$\text{ROUGE-1 precision} = \frac{\text{Num word matches}}{\text{Num words in summary}} = \frac{6}{7}$$

$$\text{ROUGE-1 F1-score} = 2\left(\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}\right)$$

Examples of Output Improving

- Target = “#Person2# has trouble breathing. The doctor asks #Person2# about it and will send #Person2# to a pulmonary specialist.” (rouge-1= 1.0)

- Epoch 1= “#Person2# has **been** breathing **lately** # doctor **wills #Person1**# about #. ask send #Person1# to a pulmonary specialist.” (rouge-1=0.5882)

- Epoch 5= “#Person2# has **been** breathing **lately** # doctor tells #Person2# about #. tell send #Person2# to a pulmonary specialist.” (rouge-1 =0.7059)

Effect of Feature on Target

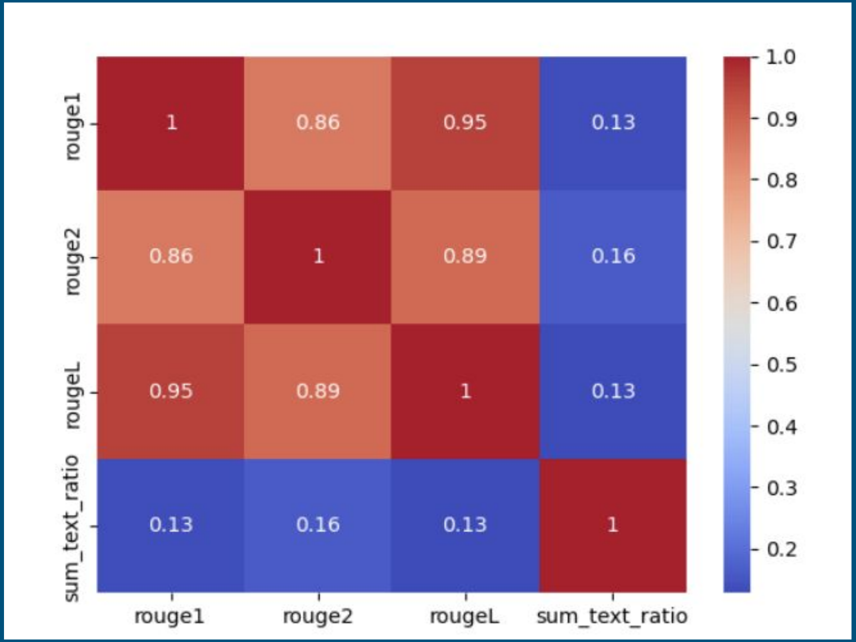
- Feature:
  - Summary to text ratio referred as sum_text_ratio

- Targets:
  - Rouge1
  - Rouge2
  - RougeL

- Models Applied for evaluation:
  - Linear Regression

Relationship between Feature and Target

Target variables have a strong positive correlation with one another, but there is no clear relation between the target variables and feature variable

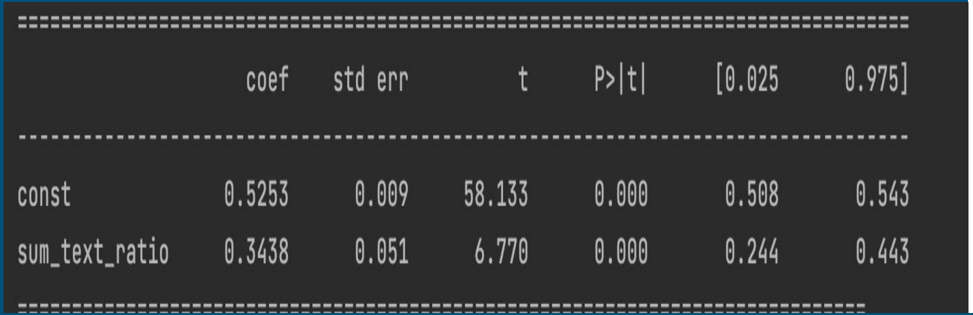


Observation of Model evaluation

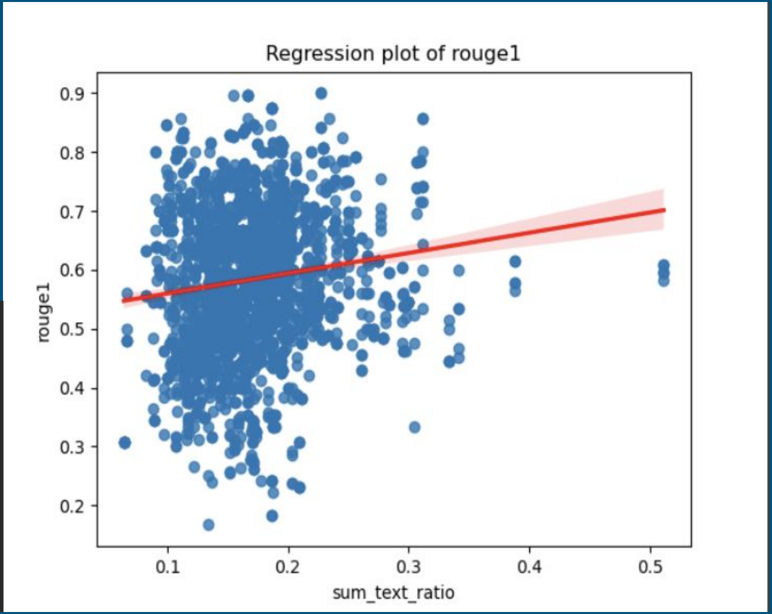
Linear Regression

Feature: Summary to text ratio
Target: Rouge1

R^2 : 0.018 & Adjusted R^2: 0.018



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.5253      0.009     58.133      0.000       0.508       0.543
sum_text_ratio 0.3438      0.051      6.770      0.000       0.244       0.443
==============================================================================
```



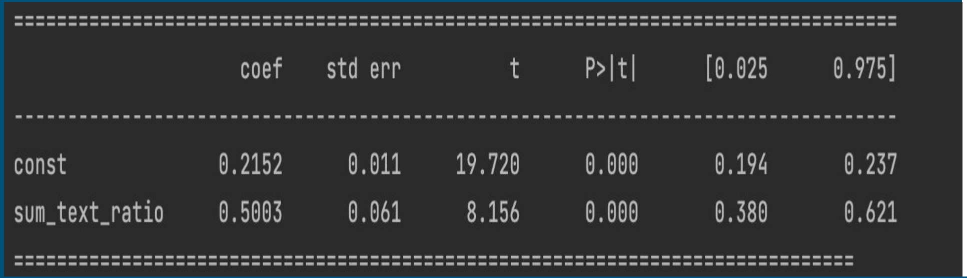
Observation of Model evaluation

Linear Regression

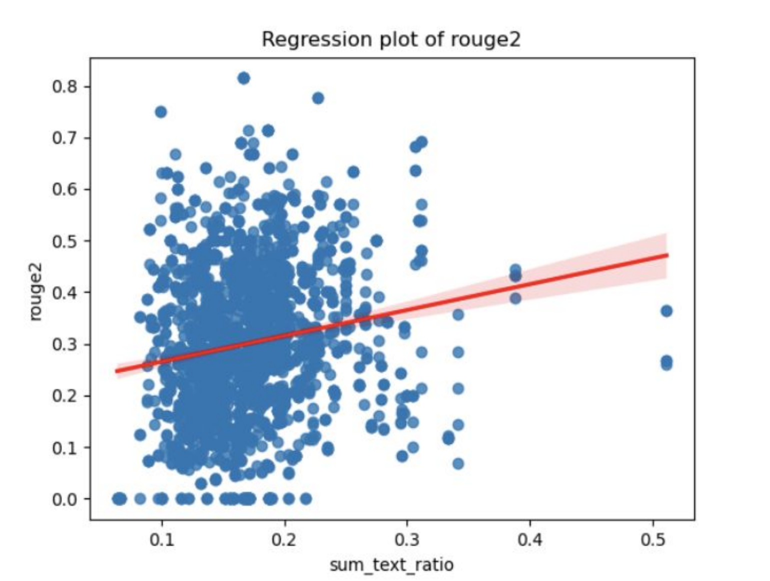
Feature: Summary to text ratio

Target: Rouge2

R^2 : 0.026 & Adjusted R^2: 0.026



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.2152      0.011     19.720      0.000       0.194       0.237
sum_text_ratio 0.5003      0.061      8.156      0.000       0.380       0.621
==============================================================================
```



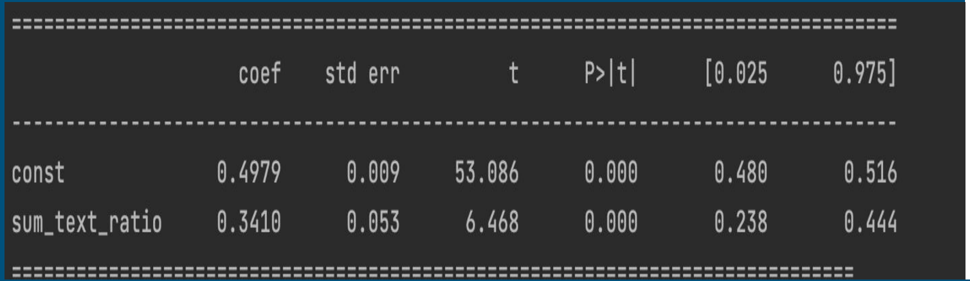
Observation of Model evaluation

Linear Regression

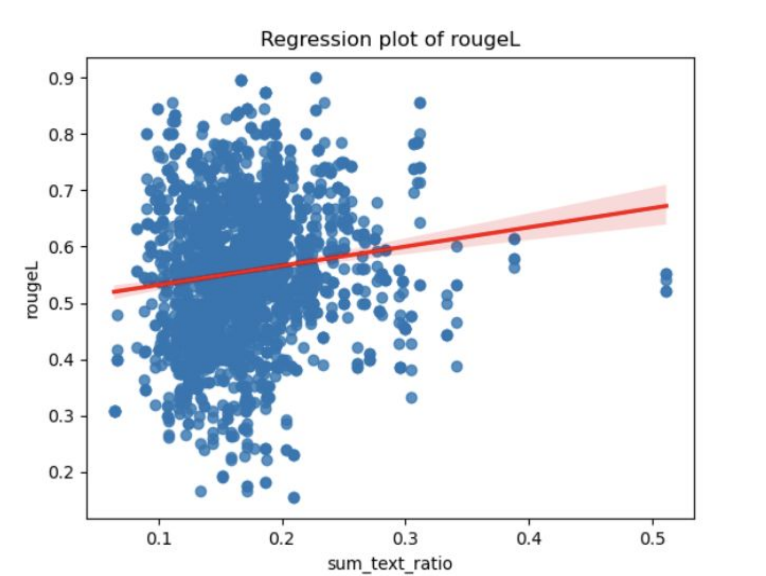
Feature: Summary to text ratio

Target: RougeL

R^2 : 0.016 & Adjusted R^2: 0.016



```
==============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const           0.4979      0.009     53.086      0.000       0.480       0.516
sum_text_ratio  0.3410      0.053      6.468      0.000       0.238       0.444
==========================================================================
```



Conclusion/Inference from the Model:

Conclusion from Linear Regression:

We reject the null hypothesis and accept the alternative hypothesis as $p<0.05$. so, we can say that there is a relationship between Targets (Rouge1,Rouge2 and RougeL) and feature (Summary to text ratio).

As R^2 (Rouge1: 0.018, Rouge2: 0.026, and RougeL: 0.016) and Adjusted R^2 (Rouge1: 0.018, Rouge2: 0.026, and RougeL: 0.016) of the all target variables are near to zero but not zero. Hence, we can conclude that the response variables can be explained by the predictor variable so, there is some relationship between target variables and feature variable.