Group 4 Final Project

Group Members:

- Alexis Kaldany
- Ricardo Diaz Pantoja
- Sagar Tripathi

Introduction.

Analyzing the relationship between the summary/text ratio and the effectiveness of the model. To achieve this evaluation we used a variety of rogue metrics and created a text ratio metric which measures the comparison between summarize and text length.

Hypothesis Statement

There is no relationship between the summary/text ratio and the effectiveness of the model.

Dataset Description

DialogSum is a large-scale dialogue summarization dataset, consisting of 13,460 variables. This 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.

Data Fields

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.

Model

To accomplish the summarization task we used AutoModelForSeq2Seq to run quick tests on a variety of models, to get a sense of the speed of training as well as the effectiveness of fine-tuning the various models. Tests were run on "facebook/blenderbot_small-90M" (4), ”bert-base-uncased”(3), “t5” (2) and “t5-small” (2).

The blenderbot model is a chatbot model, which uses similar dialogue text as our data to generate a sequence output, but the blenderbot model is more trained to generate the next sentence in the conversation, rather than the summary of the preceding/input text. It never really improved in its metrics as it was trained, but it did train very fast, which was desirable.

The ”bert-base-uncased” model did show iterative improvement while training, but it began from a relatively low level compared to other models, and took a very long time to train. It was decided that large models were not desirable for our purposes and so “bert-base-uncased” was not used. It is likely that if we had trained it for many epochs it could have ended up with the highest metrics.

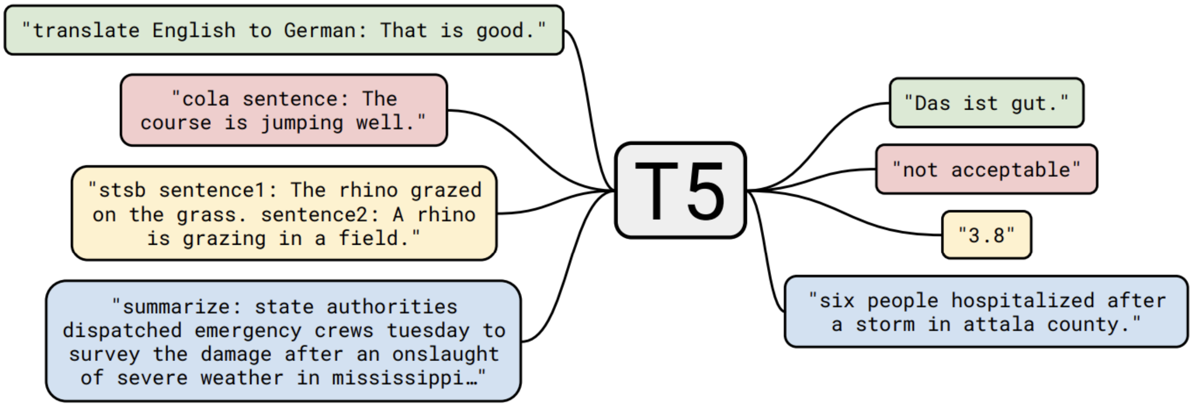

"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather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

Image of how T5 can be used for various purposes. Altering the text at the start of the input text changes the behavior of the model “t5” and “t5-small” are two version of the T5 model. 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. T5 uses relative scalar embeddings. Encoder input padding can be done on the left and on the right. “T5-small” was chosen as it trained significantly faster than the base model.

T5 is designed for sequence to sequence mapping, and so seemed appropriate, compared to the other models, for summarization.

Hyper-parameters

- max_lenght= 512
- epochs = 5
- batch_size = 8
- learning_rate=3e-5,
- weight_decay=0.01,
- adam_beta1=0.9,
- adam_beta2=0.98,
- adam_epsilon=1e-6,
- lr_scheduler_type="linear"

The approach to detect and prevent overfitting is observing the scores by epoch and checking the loss, depending on the scores we will decide which parameters are the best for the model.

Model Results

Table of Average Evaluation Results Per Epoch

epoch	loss	rouge1	rouge2	rougeL

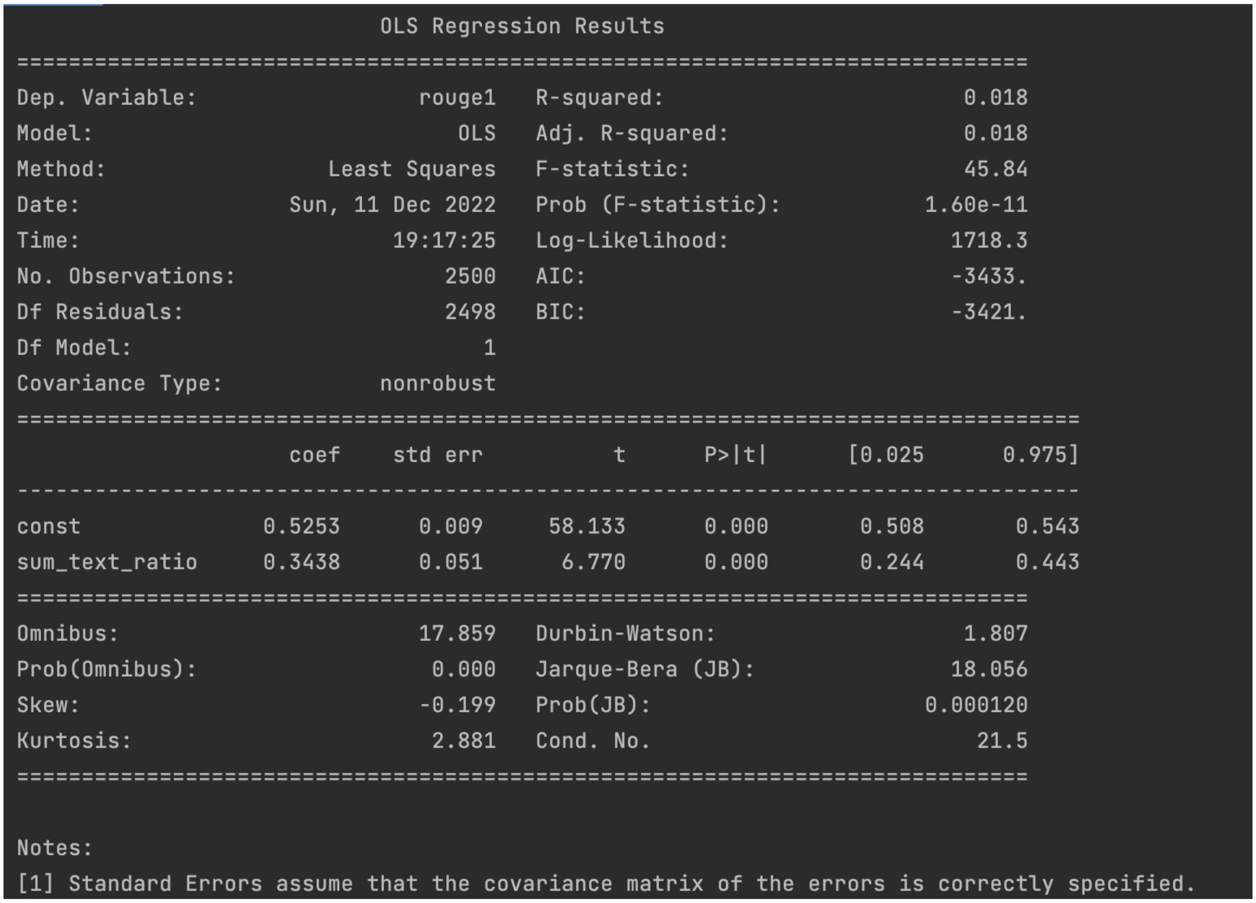
1	**1.3025**	**0.5638**	**0.2747**	**0.5328**
2	**1.2435**	**0.5811**	**0.2956**	**0.5524**
3	**1.2172**	**0.591**	**0.3087**	**0.5639**
4	**1.2061**	**0.5931**	**0.3118**	**0.5661**
5	**1.2053**	**0.5918**	**0.3138**	**0.5663**

As seen in the table above, the model essentially stopped improving after epoch 4, so I ceased training the model at that stage. The largest improvements are seen in rouge2 scores in the early epochs. All evaluation predictions are saved and individual rouge scores analyzed in the next section.

Results:

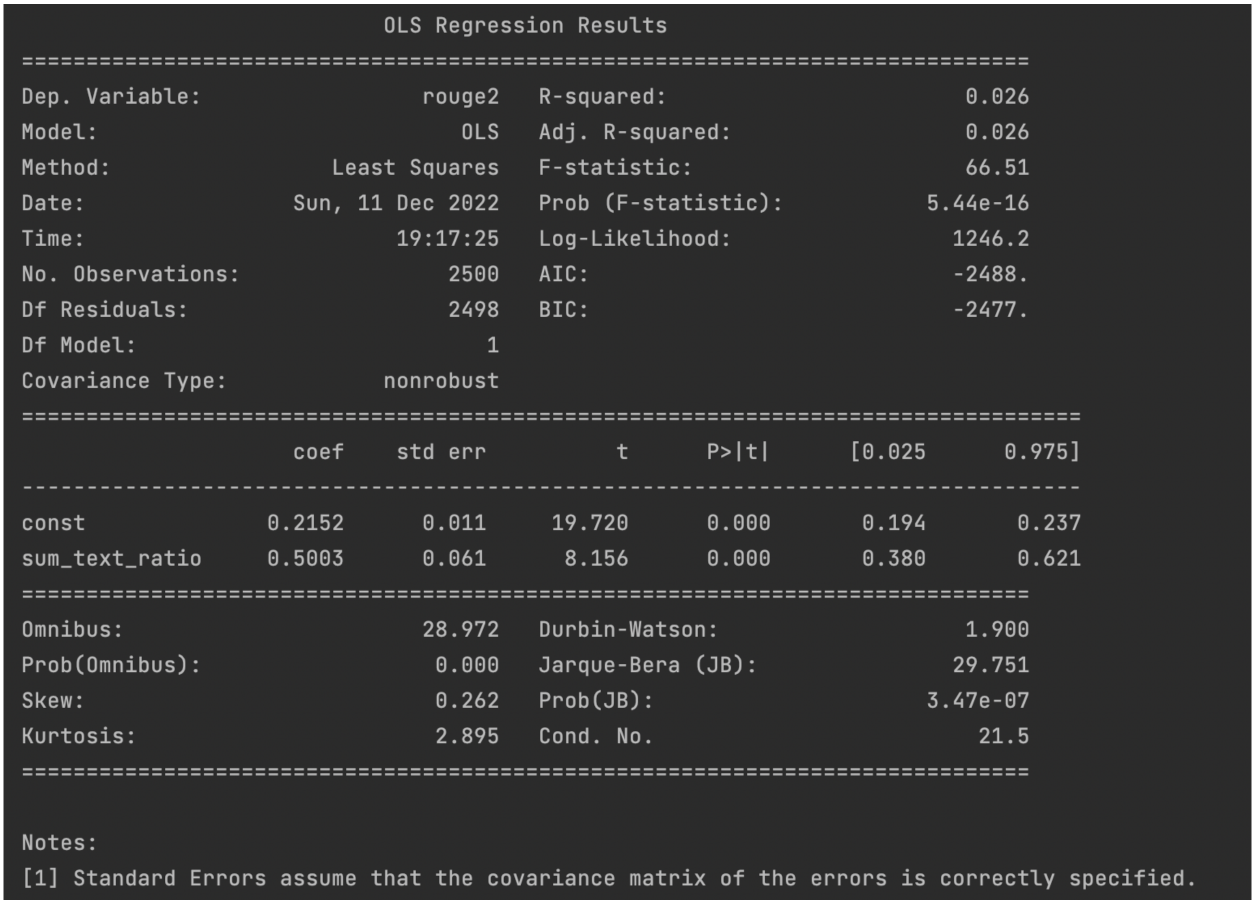
Regression models use independent variables to forecast the results of the dependent variables. Regression analysis takes significance into account to address the most challenging issues. How can we properly analyze statistical evidence for relationships between the observed variables while adjusting for the existence of additional factors is a key challenge in regression analysis? If you are not an expert statistician, regression analysis can be abused. It is a powerful instrument for explaining complicated phenomena and a very persuasive technique to show links between them. Additionally, there are numerous traps that can befall the execution and interpretation of linear regression analysis. In this project, the purpose of statistical evaluation of the Dialog Sum dataset is often to describe relationships between two variables or among several variables. For example, we would like to know whether the Summary text ratio has any influence on any of the scores interpreted by the models. The variables to be explained (Rouge1, Rouge2, and RougeL) are called the dependent variables, or, alternatively, the response variable; the variable that explains it (Summary text ratio) is called independent variables or predictor variables.

To ascertain the connection between the scores obtained from the models and the summary to text ratio, we used linear regression. First, we did linear regression on the relevant variables, using Rouge1 as the target (dependent variable) and summary to text ratio as the feature (independent variable). Secondly, we did linear regression using Rouge2 as the target (dependent variable) and summary to text ratio as the feature (independent variable) and finally, we did linear regression using RougeL as the target (dependent variable) and summary to text ratio as the feature (independent variable). The outputs of the regression analysis are mentioned below.



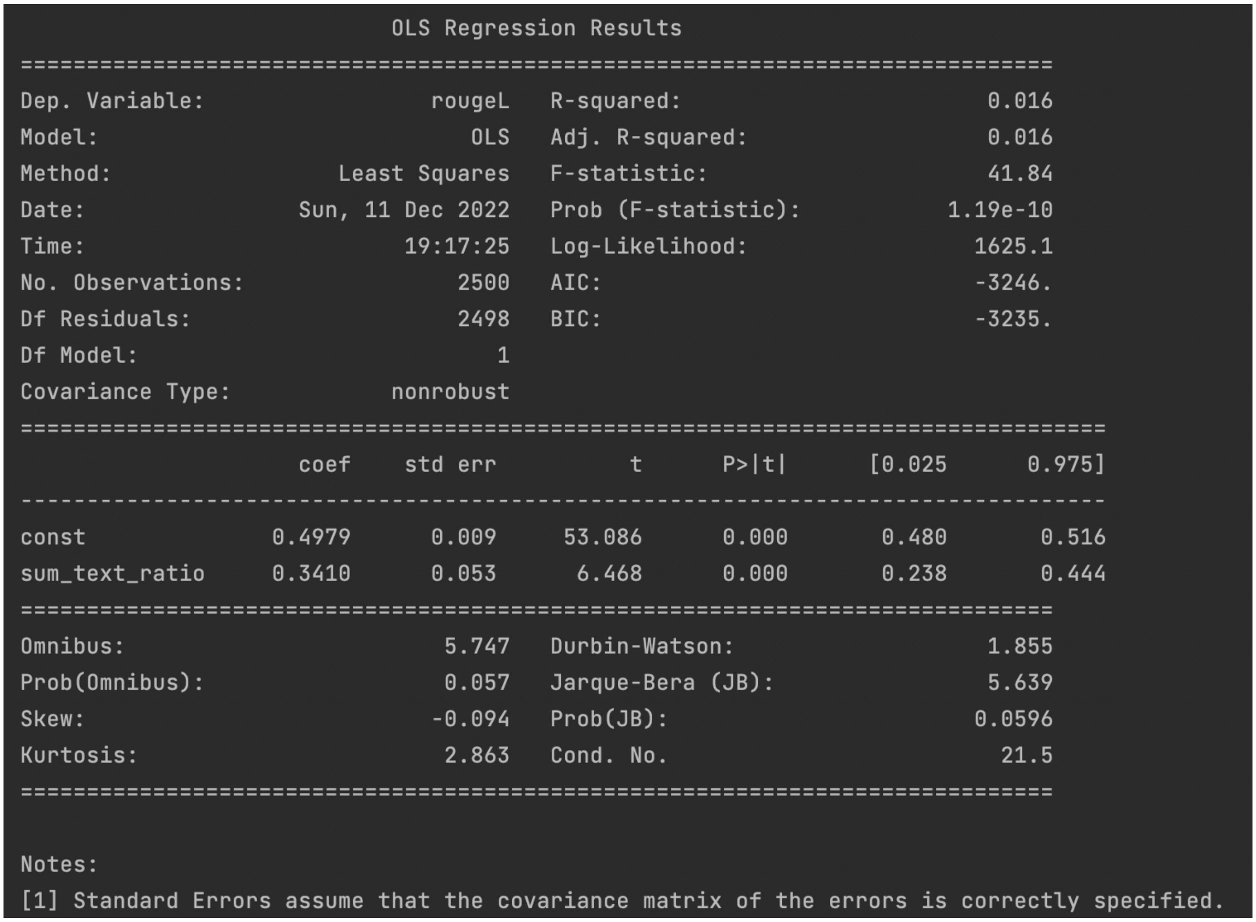
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 rouge1   R-squared:                       0.018
Model:                            OLS   Adj. R-squared:                  0.018
Method:                 Least Squares   F-statistic:                     45.84
Date:                Sun, 11 Dec 2022   Prob (F-statistic):           1.60e-11
Time:                        19:17:25   Log-Likelihood:                 1718.3
No. Observations:                2500   AIC:                            -3433.
Df Residuals:                    2498   BIC:                            -3421.
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     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
==============================================================================
Omnibus:                       17.859   Durbin-Watson:                   1.807
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               18.056
Skew:                          -0.199   Prob(JB):                     0.000120
Kurtosis:                       2.881   Cond. No.                         21.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Img 1a. Rouge1 and Summary to text ratio



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 rouge2   R-squared:                       0.026
Model:                            OLS   Adj. R-squared:                  0.026
Method:                 Least Squares   F-statistic:                     66.51
Date:                Sun, 11 Dec 2022   Prob (F-statistic):           5.44e-16
Time:                        19:17:25   Log-Likelihood:                 1246.2
No. Observations:                2500   AIC:                            -2488.
Df Residuals:                    2498   BIC:                            -2477.
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     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
==============================================================================
Omnibus:                       28.972   Durbin-Watson:                   1.900
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               29.751
Skew:                           0.262   Prob(JB):                     3.47e-07
Kurtosis:                       2.895   Cond. No.                         21.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Img 1b. Rouge2 and Summary to text ratio

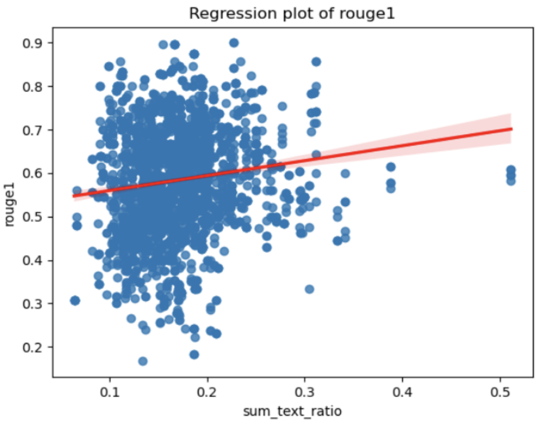


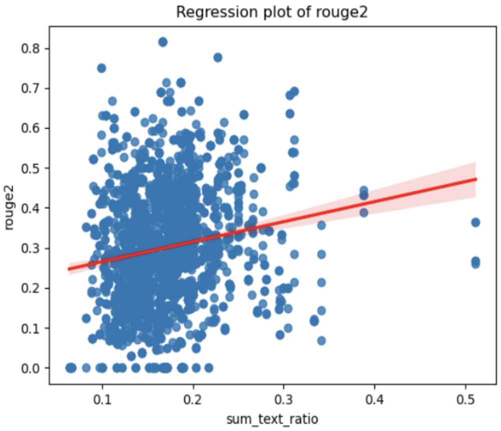
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 rougeL   R-squared:                       0.016
Model:                            OLS   Adj. R-squared:                  0.016
Method:                 Least Squares   F-statistic:                     41.84
Date:                Sun, 11 Dec 2022   Prob (F-statistic):           1.19e-10
Time:                        19:17:25   Log-Likelihood:                 1625.1
No. Observations:                2500   AIC:                            -3246.
Df Residuals:                    2498   BIC:                            -3235.
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     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
==============================================================================
Omnibus:                        5.747   Durbin-Watson:                   1.855
Prob(Omnibus):                  0.057   Jarque-Bera (JB):                5.639
Skew:                          -0.094   Prob(JB):                       0.0596
Kurtosis:                       2.863   Cond. No.                         21.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Img 1c. RougeL and Summary to text ratio

As per image Img1a, Img1b, Img1c R2 (Rouge1: 0.018, Rouge2: 0.026, and RougeL: 0.016) and Adjusted R2 (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 variables.

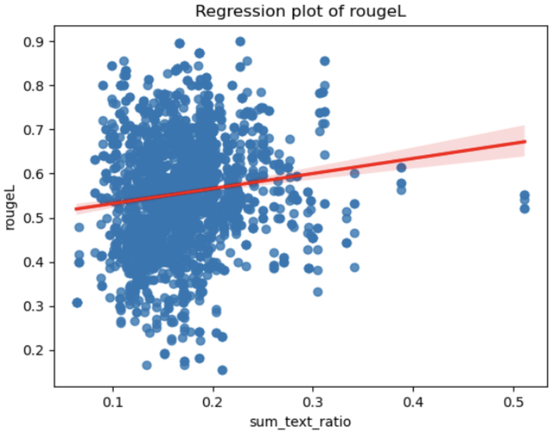
The Regression plots of the above mentioned three linear regression are show in below mentioned graphs 3a,3b,3c:





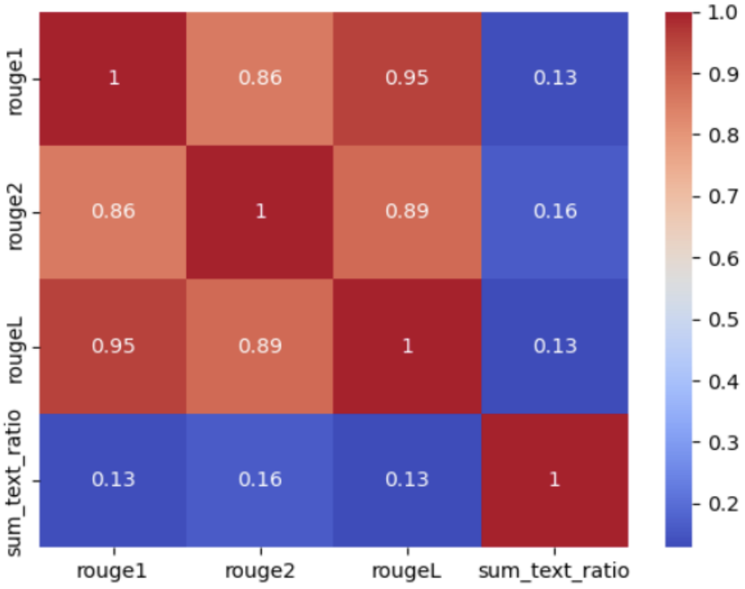
Img 3a. Regression plot of summary text ratio vs Rouge1

Img 3b. Regression plot of summary text ratio vs Rouge2



Img 3c. Regression plot of summary text ratio vs RougeL

We used the heatmap correlation between the target and feature variables to confirm the relationship hypothesis.



Img 3d correlation plot between target and feature

Conclusion:

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 R2 (Rouge1: 0.018, Rouge2: 0.026, and RougeL: 0.016) and Adjusted R2 (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 variables. Seeing that the models trained on the DialogSum dataset for fine tuning it to text summarization, we built an easy text summarization Machine Learning model from facebook/blenderbot_small-90M" , "bert-base-uncased", “t5” and “t5-small” to compare predicted summaries to correct summaries to generate metrics which will be used for hypothesis testing . The examples above illustrate that it works really well, which is really impressive!. In future we can apply different NLP models for better predictions of summary via text and can generate better score

Reference

1. Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A Real-Life Scenario Dialogue Summarization Dataset. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 5062–5074, Online. Association for Computational Linguistics.

2. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21(140), 1-67.

3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

4. Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., & Edunov, S. (2019). Facebook FAIR's WMT19 news translation task submission. arXiv preprint arXiv:1907.06616.