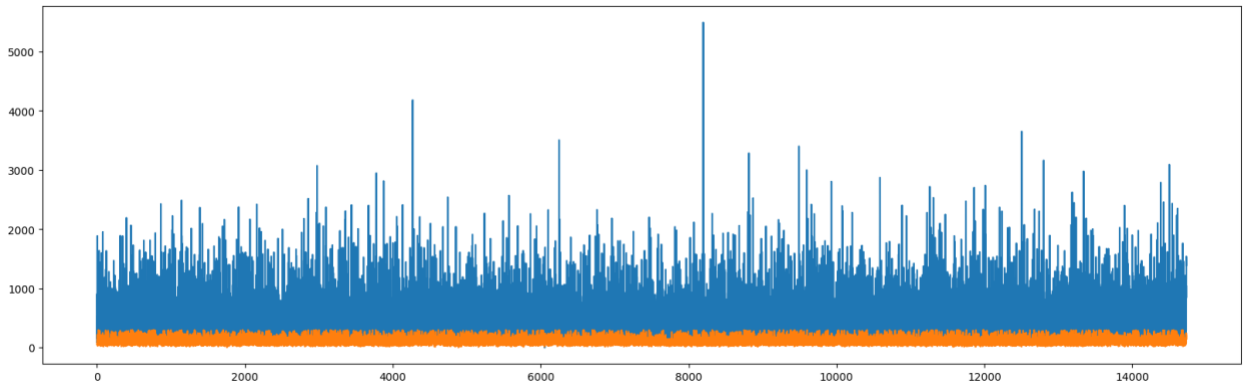
Assignment-3 Report

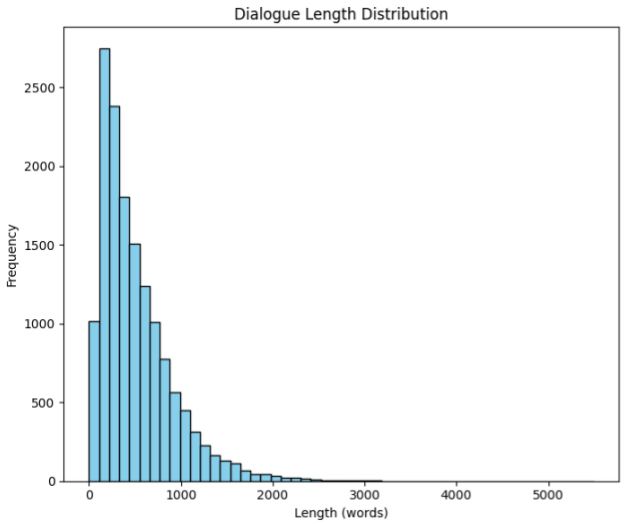
Harshkumar Modi

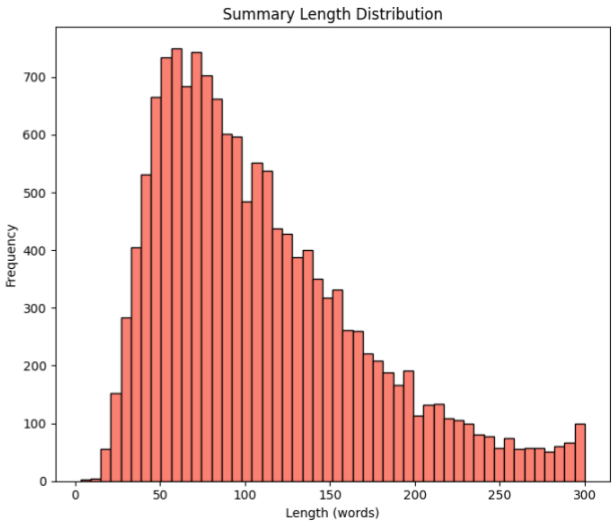
Aim: Use SAMsum dataset to compare different text summarization models and fine-tune a model to compare improvements

Dataset: Here we visualize the dataset we will be using to fine-tune the model. We can see the lengths of our input data vs the length of summaries. Upon rough estimation, we can see that the mean ratio of summaries to input data is 29%.

```
Mean ratio of summary to dialogue: 0.290316633016083
```







Results: For our comparison, I choose three text-summarization models from hugging face. My choices were based on size and popularity of models. One of the most downloaded models I found was the FalconAI's Text Summarization model. Another model that I chose was Google's t5-large model and Facebook's Bart model.

All these models have their own set achievements. When tested on the same sample input, I did observe that FalconAI's model didn't summarize any of the text. Faceboo's Bart model clipped the important snippets from the input to give only relevant parts of the conversation. However, Google's t5 model was able to generate summaries of the text which resembling to the expected output. The reason for such varied results could be because of the training dataset used in preparing these models. Conversation is something that is new to all these databases and hence their outputs vary. The models are trained for certain use cases and summarizing a conversation might not be one of those. If we prepare the data according to a paragraph instead of the conversation, we will use the semantics as the dataset will not capture who spoke which line in the conversation.

Further, to use the weights and determine how fine-tuning a model trained on a huge dataset can help get preferred outputs of a custom use-case, conversation summarizing in this case, I choose the Facebook's Bart model. The process of finetuning was carried out using the trainer packages from the transformer's library. The steps involved pre-processing the dataset and using the weights from the Seq2SeqLM model (Bart). Training it on AWS Sage Maker, we achieved a good training and validation loss.

To compare the performances of the model, we use the rouge evaluation parameter on our test dataset. We subset 250 random samples from our test set and used it to determine rouge scores of each model.

Rouge scores vary from 0 to 1, where 0 defines a bad score and 1 defines a perfect score. The results achieved are as follows:

	FAIModel	googleT5	bartFB	fineTunedModel
rouge1	0.293357	0.471516	0.300304	0.523027
rouge2	0.085484	0.240905	0.106826	0.286219
rougeL	0.222210	0.398962	0.226886	0.435865
rougeLsum	0.222767	0.399145	0.226734	0.436695

As observed, we were able to use Bart model which initially was performing poorly to beat the Google's t5 model which happen to be working better than any other pretrained model without any tuning.