CUSTOMER FEEDBACK SUMMARIZATION

TABLE OF CONTENTS

04

MODELING

Model architecture for customer feedback summarization

05

TRAINING

Training scoring and summarization model

06

INFERENCE

Get real time summarization and annotation score of customer reviews

DATA

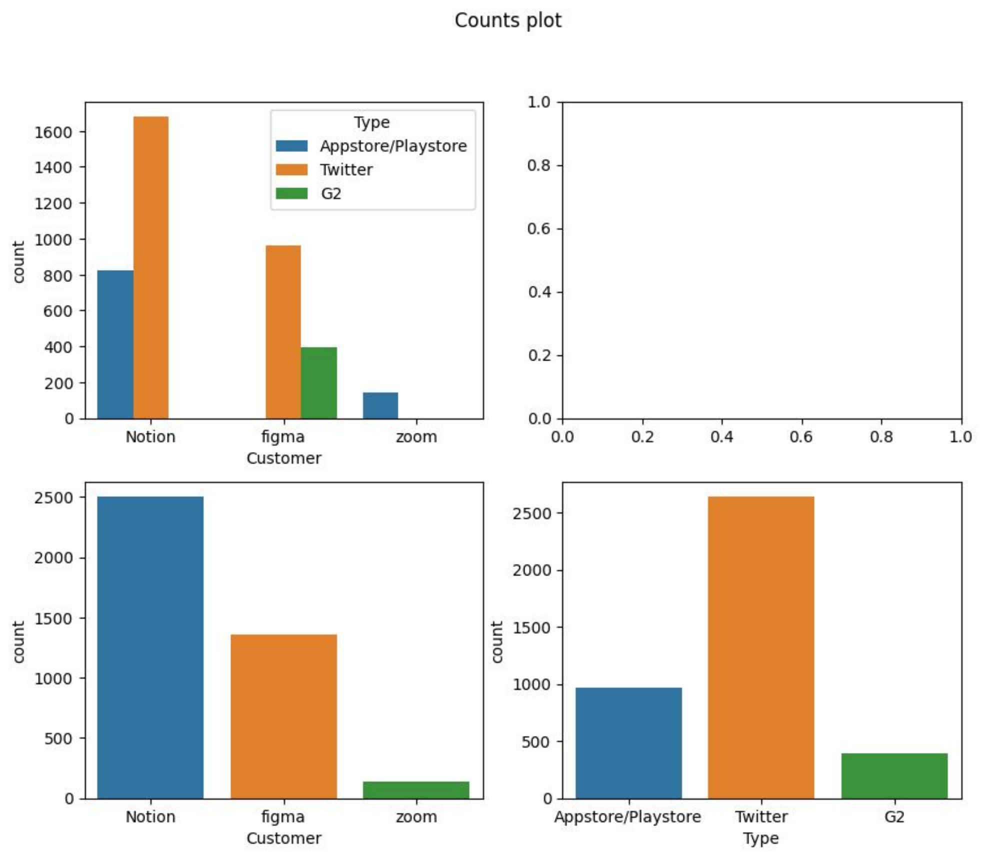
The data consists of following columns

1. Customer: Any one of the following customers:
   a. Zoom/figma/Notion
2. Type: Any one of the following platforms for user reviews:
   a. Playstore/G2/Twitter
3. Text: User review for that Customer
   a. Example: (User: Latest update killed the app.)
4. Summary: Summarized text for the User review
   a. Example: (User is disappointed with the latest update, which has killed the app.)

ANALYSIS

Observations from counts plot:

1. Most of the user reviews are for Notion
2. Zoom has the least number of user reviews
3. User reviews for Notion are not collected from G2
4. User reviews for Figma are not collected from Playstore
5. User reviews for Zoom are only collected from Playstore



Observations from counts plot before and after removing Nans/None:

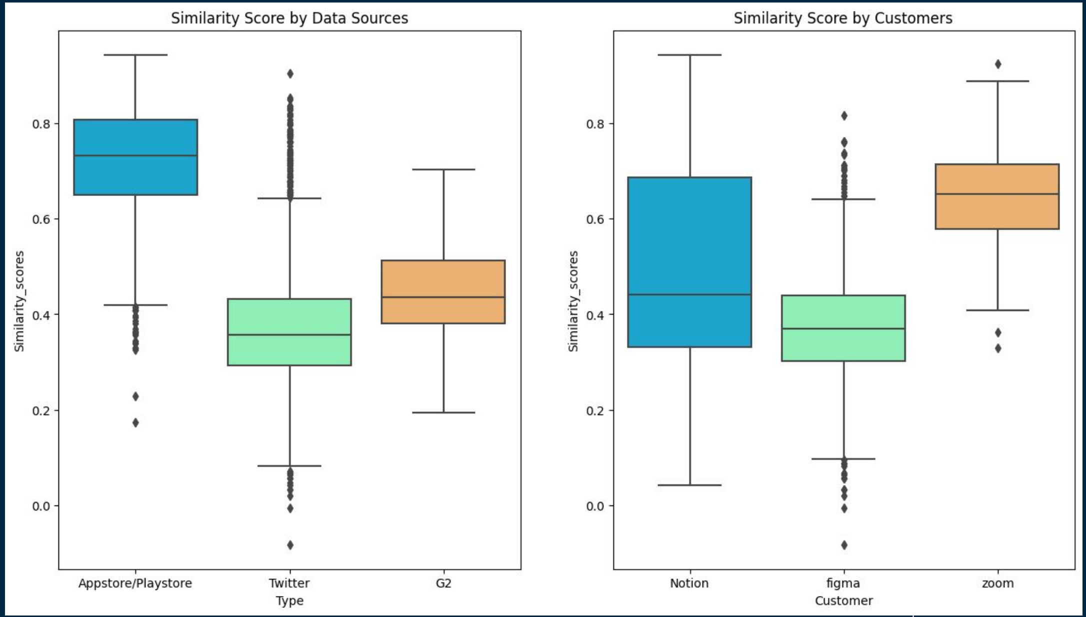
1. Zoom has the most number of rows with None or empty Summaries.


```python
zoom_playstore_df = df[(df["Customer"]=="zoom") & (df["Type"]=="Appstore/Playstore")]
zoom_playstore_df.head()
```

	Customer	Type	Text	Summary
4368	zoom	Appstore/Playstore	User: Best app ever	User praises Zoom as the best app ever.
4369	zoom	Appstore/Playstore	User: Good	None
4370	zoom	Appstore/Playstore	User: Nay Lin Wati Khit	None
4371	zoom	Appstore/Playstore	User: Excellent	None
4372	zoom	Appstore/Playstore	User: Ok 👍	None

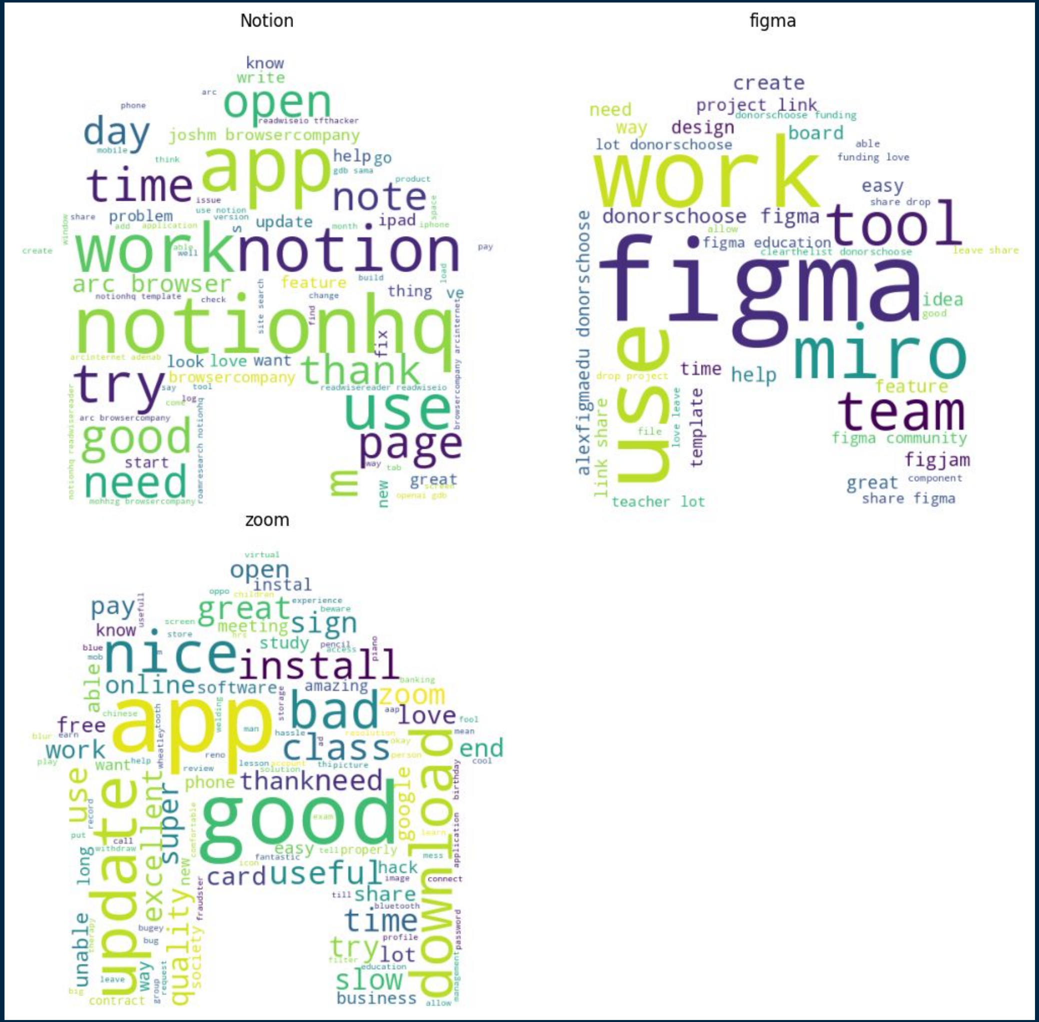
Using semantic similarity score between text and summary, observations are:

1. User reviews from twitter has not been summarized properly.
2. Data from playstore have decent quality of summarized text
3. User reviews for figma has bad quality of summarized text



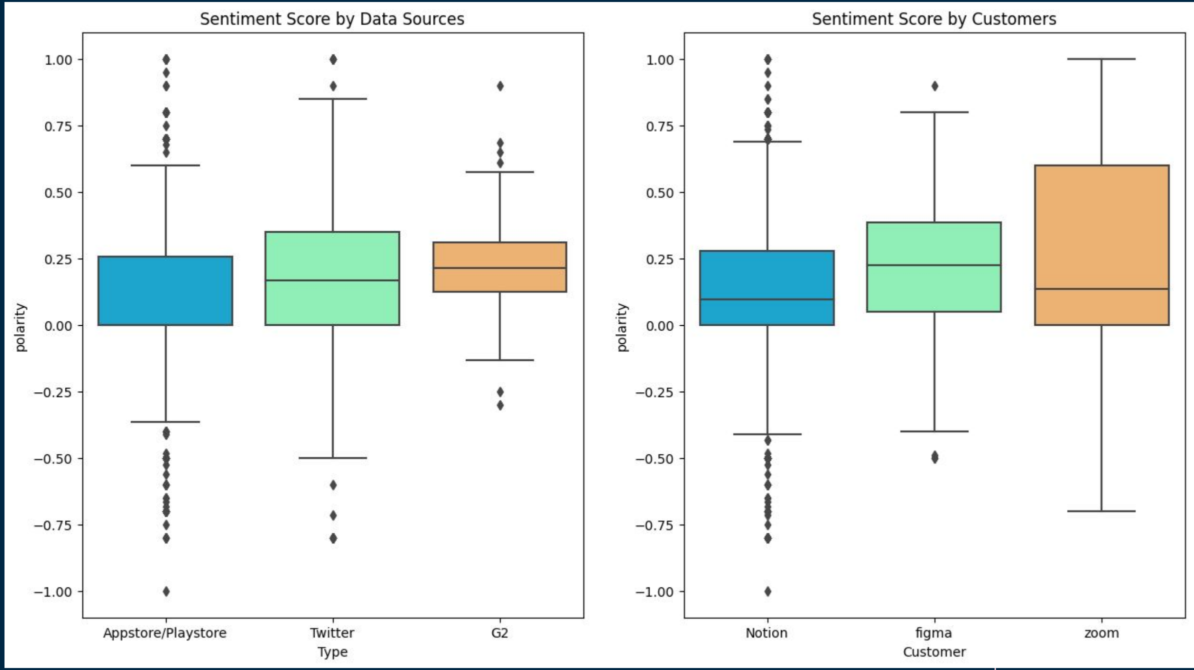
Using wordclouds, observations are:

1. Users use notion from browsers/ipads/iphone.
2. Users use Miro plugin for figma
3. Figma is easy to use and users use it to ideate, design or create features.
4. Many users have positive feedback for zoom



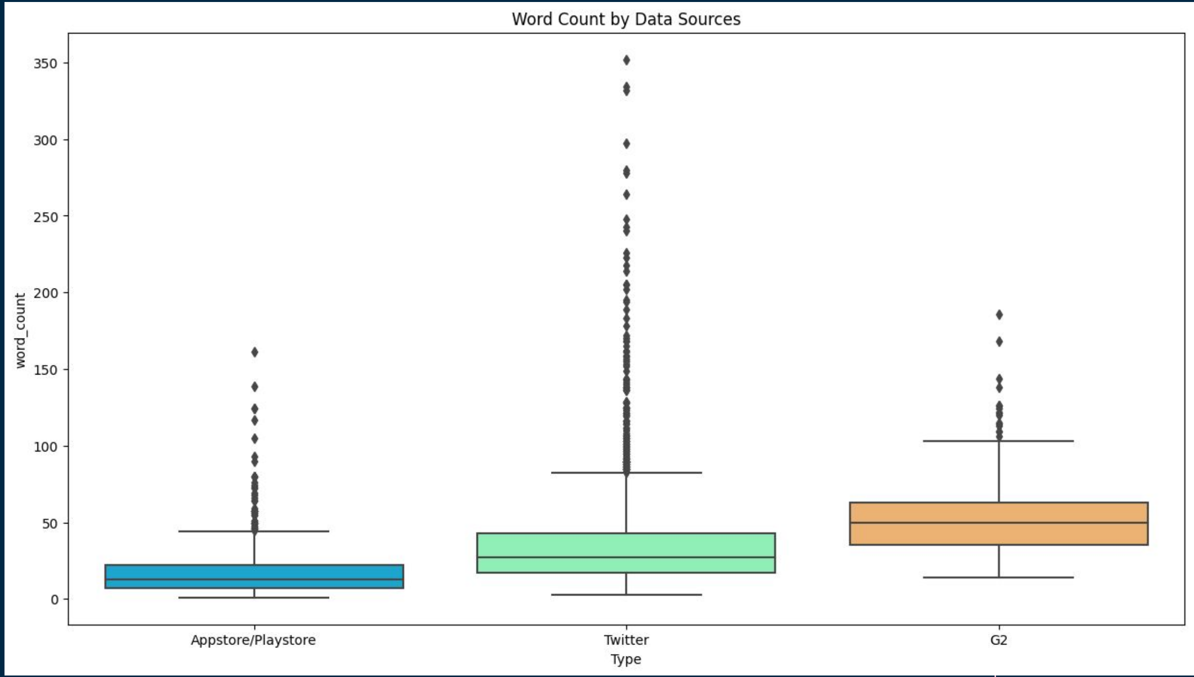
The same observations which has been made from wordcloud diagrams can be validated here as well such as.

1. Zoom generally has a positive user feedback as compared to other customers



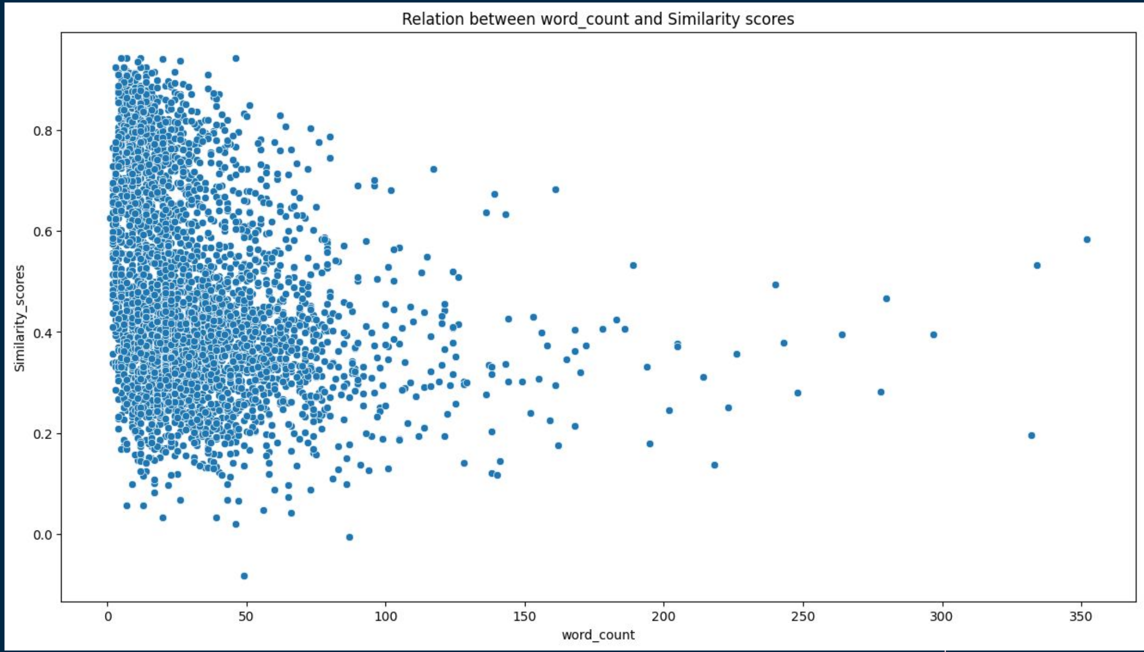
Wordcount distribution observations for platforms:

1. G2 generally has longer text for user reviews on any product followed by Twitter and then Playstore



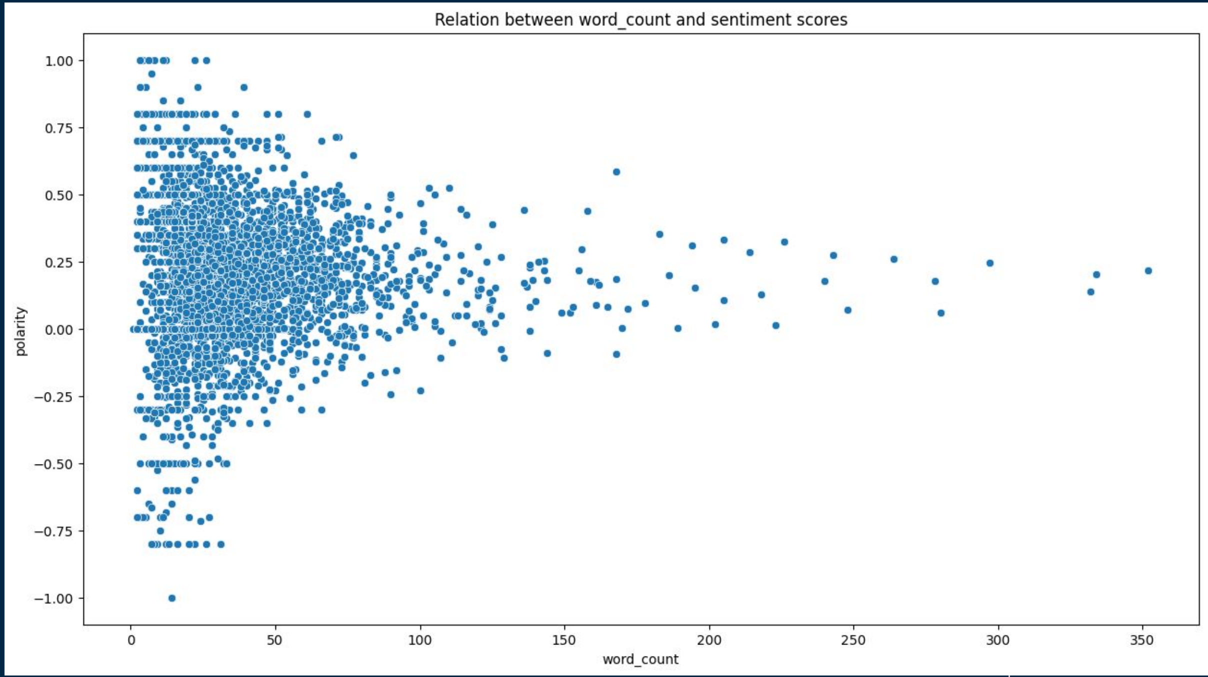
Relation between word count and semantic similarity scores:

1. There exists no relation which means that length of user review has no effect on how well the text has been summarized



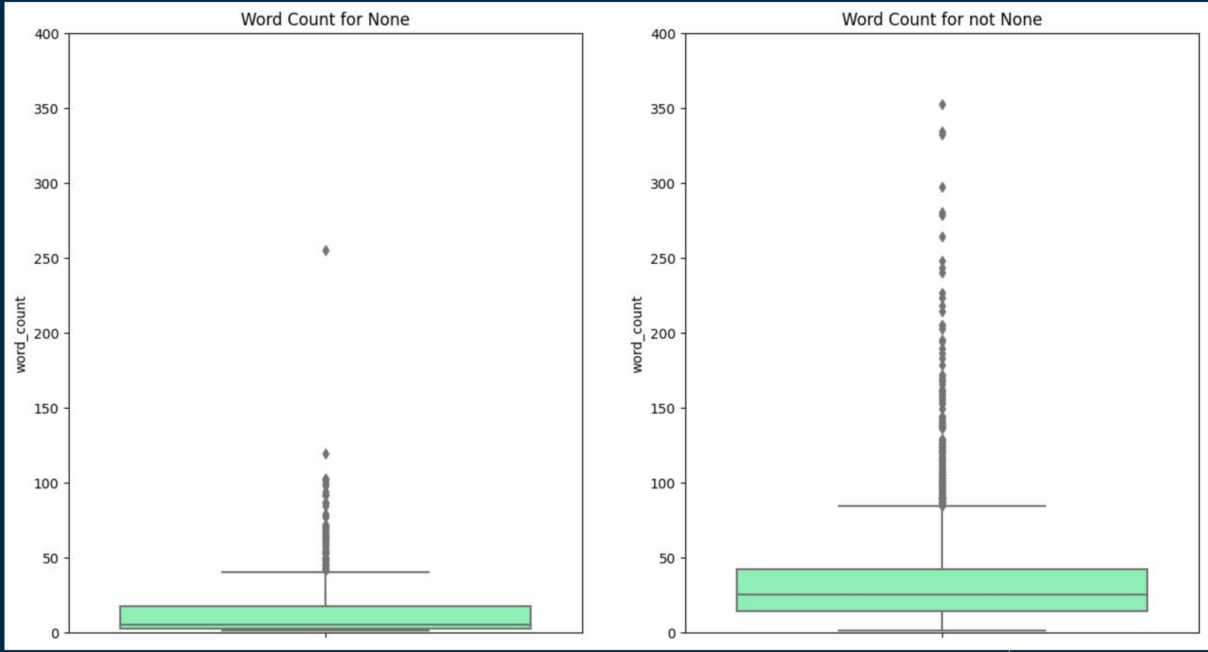
Relation between word count and sentiment scores:

1. There exists no relation which means that length of user review has no effect on user's sentiment about that review



Word count distribution for “None” and other “Summary”:

1. Shorter user reviews have None as "Summary" due to not sufficient content to summarize.



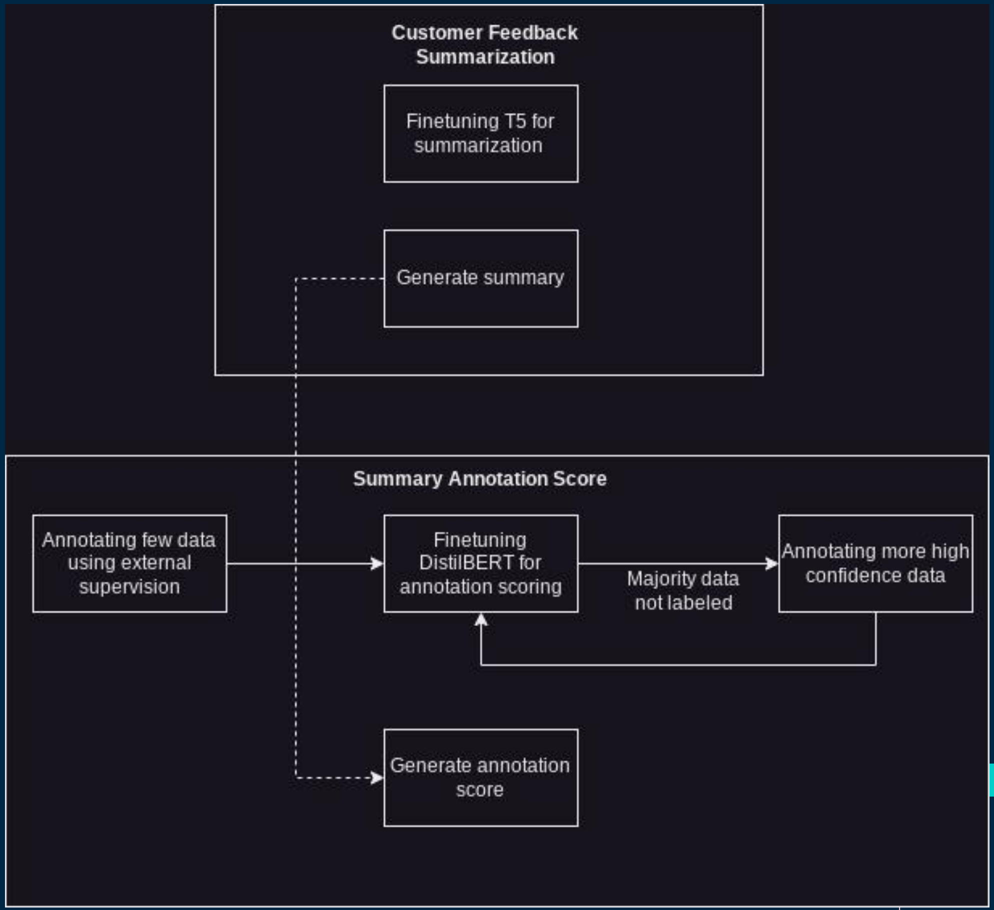
Low quality twitter summary observations:

1. Summary is generated from a Seq2Seq model which suffers from repetition problem.
    a. Ex: User: I'm not sure if this is a bug or not, but I'm not able to see the "Add to Home" button in the "Add to Home" section of the "Add to Home" tab .... (contd.)
2. Quality of Summary is not good as the generated summary is not even relevant to the text content.
    a. Text: User: To replicate or to do something different? <STRICT_LINK> User: Settled for something different 😌. with @figma. Animating this will be cool sha.
    b. Summary: Hiring a Data Scientist - Strategic Finance in the United States or Canada. #remotework #remotejobs #workfromhome #wfh #remoteworking #futureofwork
3. 3. Last message does not enough content to generate a good summary
    a. User: To replicate or to do something different? <STRICT_LINK> User: Settled for something different 😌. with @figma. Animating this will be cool sha. <STRICT_LINK>

MODELING

The architecture diagram is used to generate feedback summary and annotation score using:

1. Feedback Summarization: Finetuned T5 model on summarization task
2. Annotation Score: Finetuned DistilBERT in a semi supervised way to generate annotation scores



TRAINING: Summary Generation

Experiments:

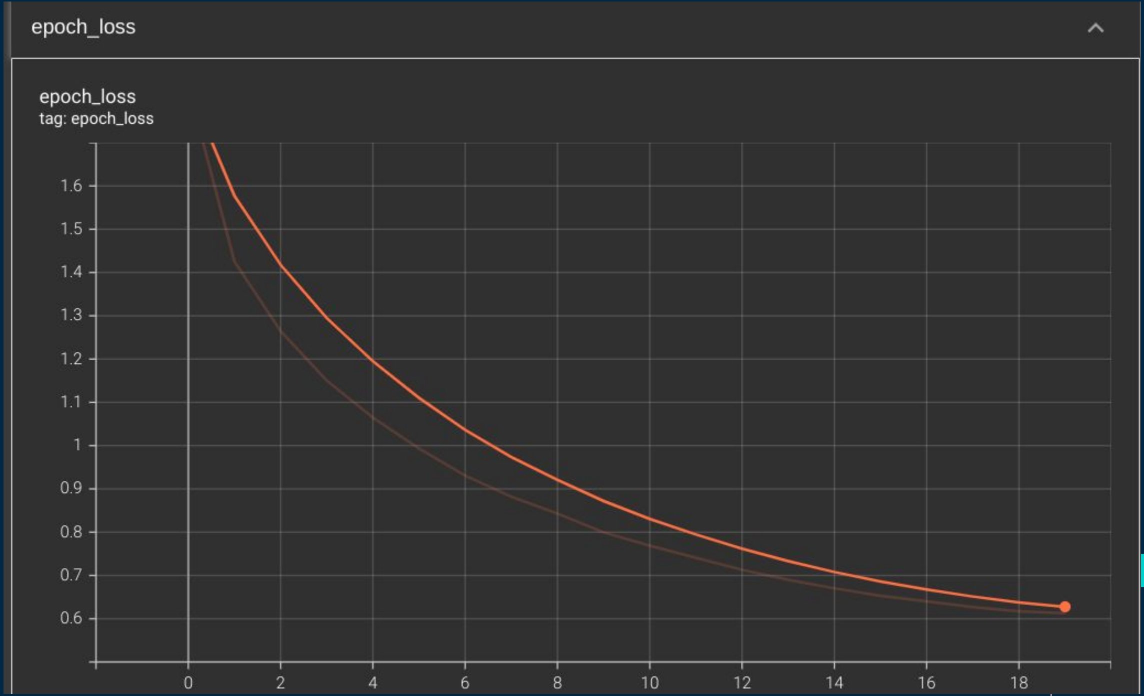
1. Raw_data: Use the complete data to train
2. Filtered_50: Use data with semantic similarity > 0.5
3. Filtered_60: Use data with semantic similarity > 0.6

RAW_DATA

Training

Metrics/Loss:

1. Loss: 0.61
2. Rouge1: 59.33
3. Rouge2: 33.97
4. RougeL: 55.82
5. RougeLsum: 56.28

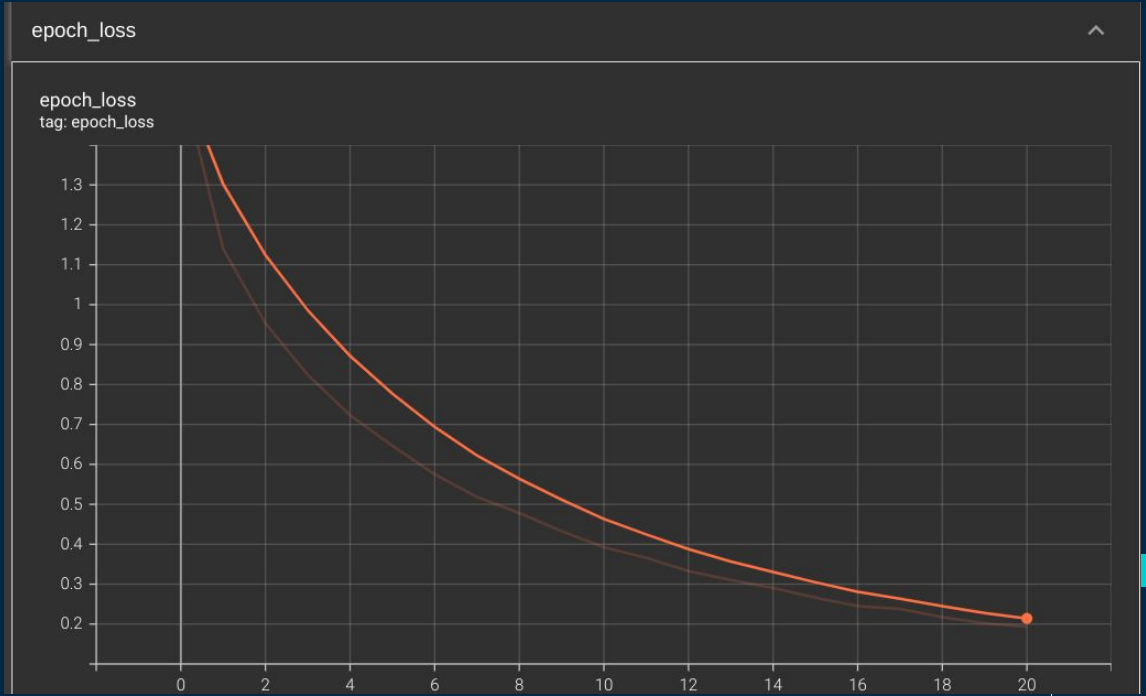


FILTERED_50

Metrics/Loss:

1. Loss: 0.432
2. Rouge1: 75.68
3. Rouge2: 27.57
4. RougeL: 72.25
5. RougeLsum: 73.08

Training

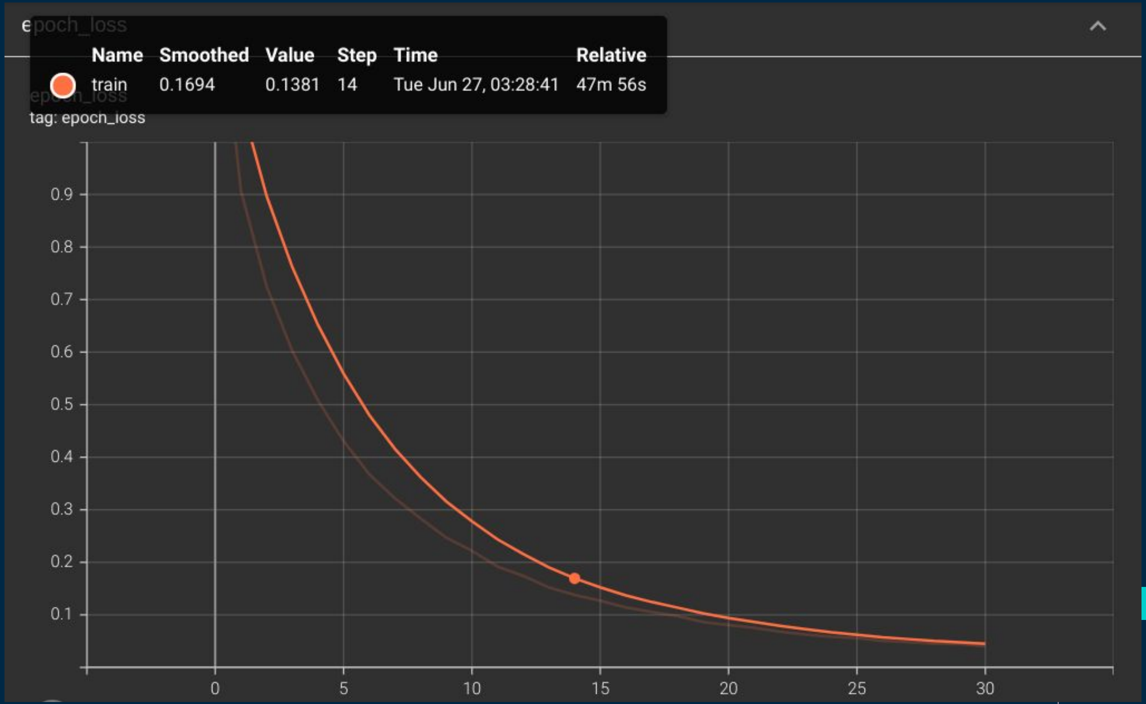


FILTERED_60

Training

Metrics/Loss:

1. Loss: 0.24
2. Rouge1: 73.55
3. Rouge2: 20.73
4. RougeL: 69.68
5. RougeLsum: 70.71



TRAINING: Annotation Scoring

Steps:

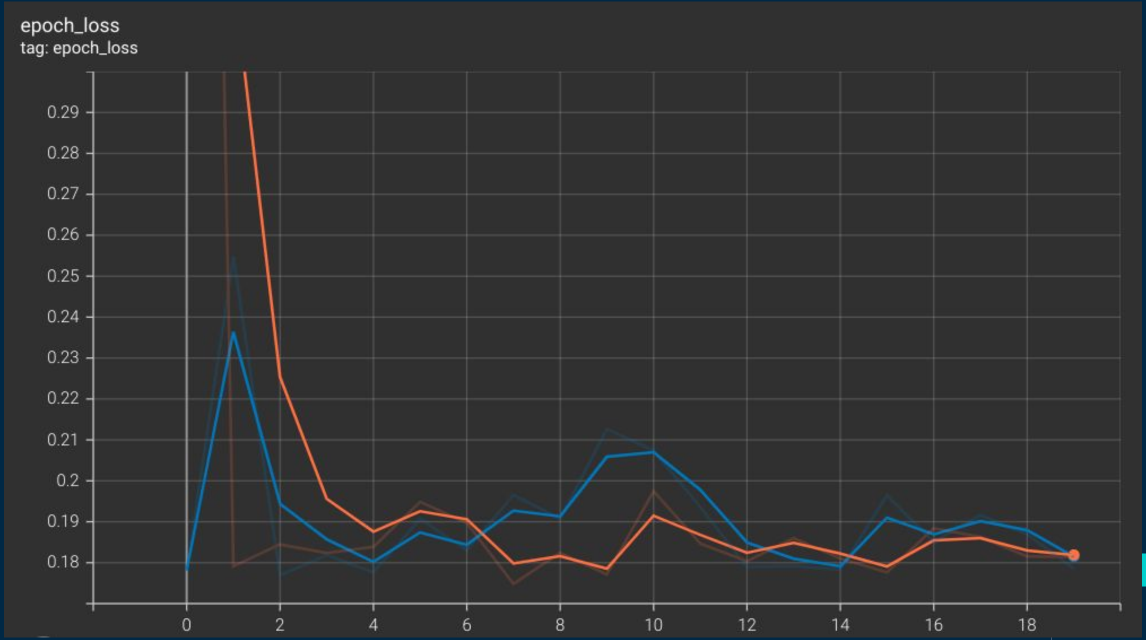
1. Stratify data based on semantic similarity score
2. Select 200 stratified data in the same proportion
3. Label the 200 data using external supervision
4. Finetune DistilBERT on the 200 labeled data
5. Use semi supervised learning iteratively to label most of the data

Finetune DistilBERT

Loss (MSE):

1. Train Loss: 0.18
2. Validation Loss: 0.17

Training



INFERENCE

An end to end training of both summarization and annotation scoring model is done and pushed to huggingface hub and an inference containerized application in huggingface spaces.
For demo use the following:

DEMO LINK