1. First, I unzipped and extracted all the text files using: 
2. Filtering text files: 
   1. Filtered for text files from June-Aug 2024, (output folder: ‘ ‘June-Aug-2024’)
   2. Filtered for 10 companies at random. (folder ‘filtered\_companies\_10’)
3. Cleaned the text files by removing things that are not the conversation: 

(output folder: ‘final\_cleaned\_transcripts’)

1. A) All the cleaning, tokenizing and sentiment analysis functions are in this: 
   1. Outputs are pickles.
2. A) I then put these pickles onto colab to do lemmenization and sentiment analysis (computer cant handle it lol) <https://drive.google.com/drive/u/0/folders/1v7ggUg5opYWMSgqkm2NXlQdV1rK8Gn3s>
   1. The outputs of these are also pickles—lem and vader.

After step 3), As a separate process, I tokenized every text file into sentences, and retained a column which said which text file it was from. Output was a csv with individual sentences 

I did step 4 and 5 again on this data. 4B and 5B.

Outputs are pickles for the sentence lem, tfidf, and vader and senti scores.

The final result is a csv which has the average sentiment for each of the 10 files. “avg\_file\_vader\_score.csv”

Summary: cleaned the data and did the sentiment analysis.

Please check this folder on GDrive. https://drive.google.com/drive/u/0/folders/1v7ggUg5opYWMSgqkm2NXlQdV1rK8Gn3s

Essentially, I cleaned the text files (partly locally on spyder-- then I moved to colab for lemmenization) and did sentiment analysis on the text files. The sentiments here were odd.

So, in the "sentences\_processing\_sentiment" folder, I tokenized every text file into individual sentences while retaining the file\_number as a column. I did a sentiment analysis on each sentence, and I average the sentiment score for each sentence that belong to the same file. This final result is the "avg\_file\_vader\_score.csv" in the nested folder.

I have saved .pk files for every stem (stopwords, lem, vader\_sentiment, analyze\_sentiment) for both the text file analysis and the sentence-wise analysis.

The Github has all the final code, .pk and txts (except for code used to extract the original zip after webscraping. The original webscraped is on the Whatsapp group—it is too large for github).