Sentiment Analysis on Starbucks Review Using OpenAI Model

Introduction:

In sentiment analysis, there is polarity analysis where the text is classified into positive, negative, and neutral. There is emotion detection where different emotions like anger, sadness, gratitude, love, fear, anger and many more emotions can be obtained from a review, all these can be done with the help of OpenAI models such as GPT-4 and RoBERTa. I have taken Starbucks dataset because Starbucks is found in almost every country. Since it has established itself as one of the top marketing brands in the world, it will have a huge number of customers. These customers use a wide variety of social media platforms like Facebook, Instagram, Reddit, Twitter where they give their opinions. These opinions may range from highly positive reviews to worst possible reviews. I have taken GPT-4 model for this analysis because it has much better architecture than other models, and it has better understanding of the textual data.

"Sentiment analysis is a process of extracting textual information from a given text. (Hovy, 2014)"

The different types of OpenAI models are BERT, RoBERTa and GPT models like GPT-3, GPT-3.5 and GPT4.

Related Work:

"The GPT models can detect emotions and generate results, but it lacks accuracy of fine – tuned models. So, in order to overcome this, they combined GPT models with RoBERTa model to get better accuracy. After conducting an experiment on goEmotion dataset they got macro-F1 score of 0.49 for RoBERTa, 0.17 for GPT-3.5 Turbo and 0.22 for GPT-4. But when they combined LangChain GPT-3.5 model with RoBERTa model they got best performance among other models with a score of 0.51 macro-F1 score (*Intertwining Two Artificial Minds: Chaining GPT and RoBERTA for Emotion Detection*, 2023b)".

Here Nandwani and Verma (2021b) talks about some instances where the machine faces difficulties in detecting sentiment analysis and emotional analysis. These days many young people use social media sites like Facebook, Instagram, Snapchat where they use sentences like 'My favorite party trick is not going". It is a sarcastic sentence which the machine will find hard to predict the emotion. Another problem is usage of words like 'Brb' which means be right back and 'Asap' which means as soon as possible which machine doesn't understand. Also, when there are multi – opinionated sentences like 'Antarctica is a nice place, but people hardly live there due to the climate and cold weather' in this there are 2 emotions, the first part of the sentence has emotion as admiration, but the second part of the sentence talks about disapproval. So, it is hard for machines to predict what emotion needs to be displayed. It’s also hard to detect polarity from comparative sentences like ‘Apple is better than android’ and ‘android is better than apple’ here the better words should mean that the polarity is positive, but this sentence means that it is negative, and they both oppose each other. Nandwani and Verma (2021b)

Azad (2024) compared human annotator with GPT-4 annotator. The GPT-4 annotator underperformed compared to the human annotator. It got a low accuracy score in the sentiment analysis. Although GPT-4 performs much faster and is ideal for FinTech markets where time is everything, and human annotator being slow but still having more knowledge about the domain in which they work, so the human annotator was considered more superior compared to the GPT-4 annotator. Azad (2024)

Data:

The “Starbucks review dataset” is obtained from Kaggle which is an open-source platform where anyone can download, read the data, and perform analysis. Keeping ethical issues into consideration I thought that this is the best site where a lot of datasets are available at free of cost and can be used for my analysis.

This is a secondary dataset which in turn is obtained from Consumer Affairs Website by the method of web scraping on customer review and rating columns. This dataset has 850 reviews out of which 812 reviews are unique and 38 values are No Review Text values. The review column contains symbols, emotions, abbreviations and acronyms, and numerical rating along with textual reviews of people’s opinions on Starbucks. The size of the dataset is 454kb.

Methodology:

The methodology I am using for this dataset are sentiment analysis using GPT-4 and Apps Script, RoBERTa using hugging face and for plotting graph I have used Seaborn and matplotlib.

GPT-4 and Apps Script:

The excel sheet containing the Starbucks dataset was fed into the google drive. From the google drive the file was opened in Apps Script environment through the Extensions tab. Now the code was written in the text editor environment. The function call was given as “function getReviewSentiments(Review)” which took Review column as input. Inside the function definition the excel sheet was read with the help of “OPENAI_key”. The model used for this analysis was “gpt-4-1106-preview”. The Review column was taken in the first column and the second column was empty and named Sentiment. With the help of OPENAI_key the Apps Script read all the data from Review column and updated the Sentiment column for each review with polarity score which was either ‘Positive’, ‘Negative’ or ‘Neutral’. This was done using the command “var prompt = `Classify the sentiment in this reviews: ${tweet}. Answer with a single word choosing from: Positive, Negative or Neutral`”.

Similarly, like polarity detection, the emotion detection was also done by giving the command as “var prompt = 'Classify the emotions in this reviews: ${tweet}. Answer with a single word choosing from: Joy, Happy, Sadness, Angry and many more emotions”. Here also the second column was empty and named as Emotions and the Apps Script ran through the Review columns and gave emotions for each Reviews.

Plotting graph using matplotlib and seaborn:

Both matplotlib and seaborn are python data visualization libraries. After the Polarity and emotion detection is done. I have used jupyter notebook to do the plotting of these analysis. First matplotlib, seaborn and pandas are imported and then the csv file is read into the jupyter notebook. I have used matplotlib for plotting polarity analysis using GPT-4 and emotion detection using GPT-4. I have used Seaborn to plot emotion detection using RoBERTa.

Analysis of the Result:

Polarity analysis using GPT-4:

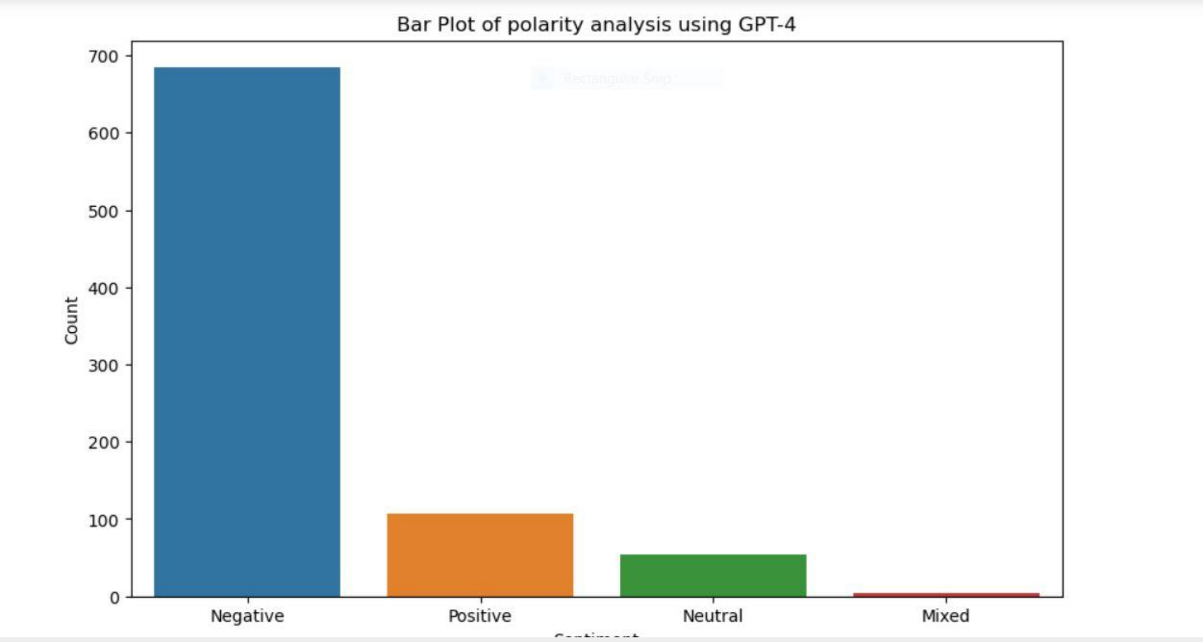

Fig.1 Polarity analysis using GPT-4

Here in the graph, I have taken the Sentiments in the X-axis and Count of reviews in the Y-axis and the title is given as Bar plot of polarity analysis using GPT-4. From the graph we can clearly see that there are lots of negative polarity scores. It has 660 reviews. The next one is positive polarity which has 110 reviews, and the neutral tweets has 38 tweets, and the mixed tweets are just 4. From this we can see that most of the customers are not happy with the service provided by Starbucks, and they have given bad reviews for the service received. Starbucks should make sure that the customers are happy by providing better customer service. They can make the customers happy by providing the orders on time and staff being friendly with the customers.

Emotion Detection using GPT-4

Here the X-axis is used for Count of reviews and Y-axis is used for different emotions. The Title is given as Bar plot of emotion detection using GPT-4.

From the graph we can clearly see that Disappointment is the most seen emotion, followed by Frustration, Dissatisfaction, Discontent, Disgust which are all constituting to Negative review which was also seen in the Polarity analysis. Appreciation, Mixed, Gratitude, Satisfaction, Positive, joy all come under the positive emotions. So, this graph matches with the polarity analysis as well as per the emotions obtained. But there are some emotions which are more than 1 like ‘Frustration, Disappointment, Displeasure, Dissatisfaction, Discontent’. For most of the reviews it is able to classify properly, but for very few reviews it is not able to give the exact emotion. This is also seen in other emotion classification as well where it has given as ‘The emotion in this text seems to be neutral’, ‘Based on the provided text, the emotion appears to be Disappointment’ and finally it has also given as ‘The emotion in this text could be classified as neutral’ So from this we can see that the GPT-4 model is not highly accurate. It can sometimes not understand what the review is conveying.

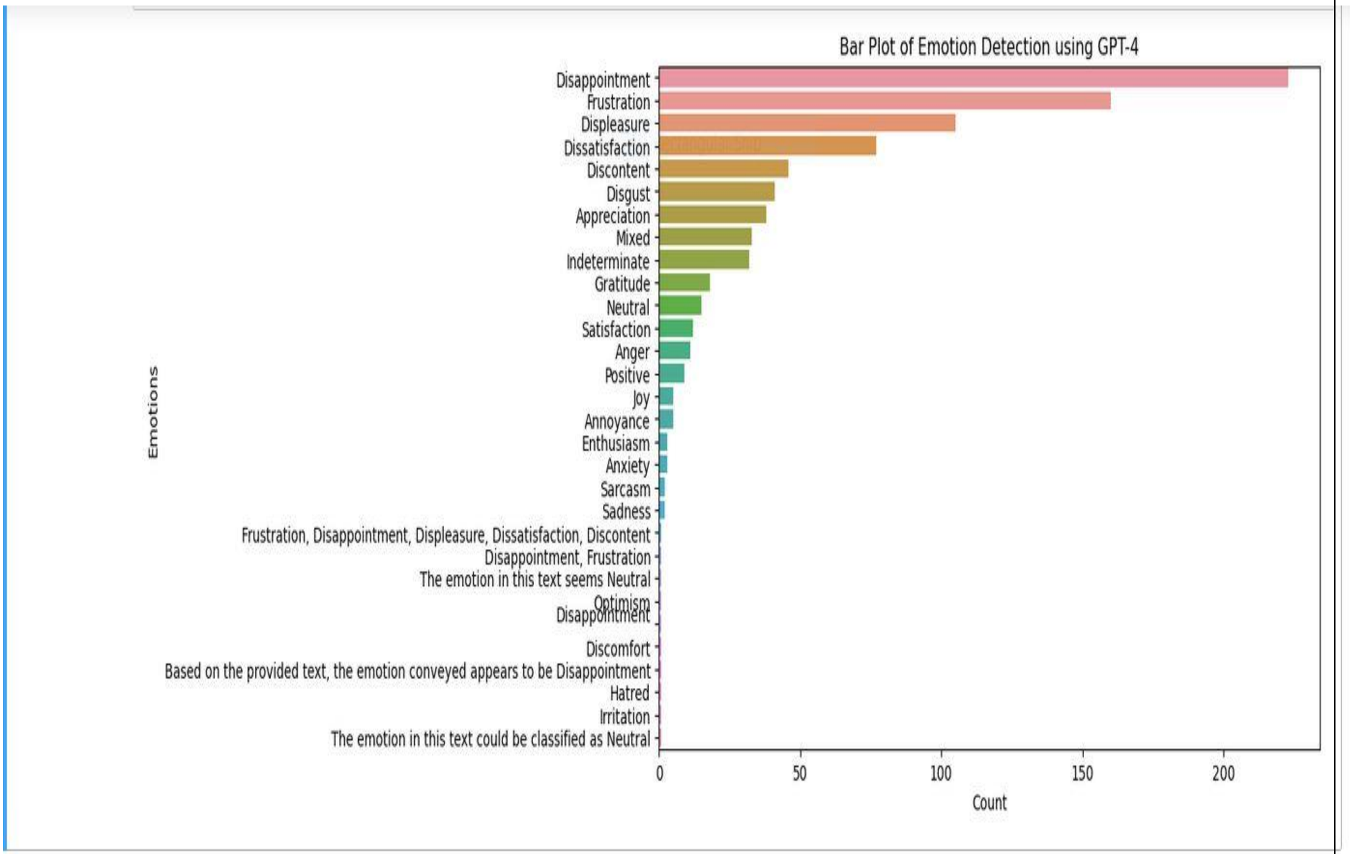

Fig.2 Emotion Detection using GPT-4

Emotion Detection using RoBERta model:

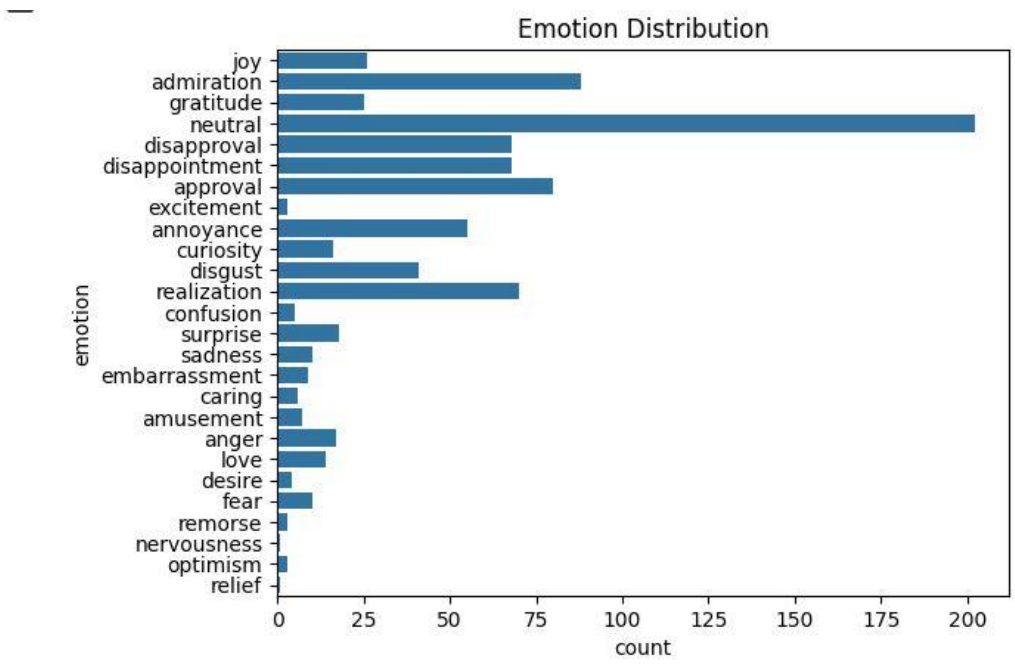

Fig.3 Emotion Detection using RoBERTa model:

This was done using hugging face where many emotions like sad, happy, angry and many more are pre-trained

The X-axis is used for count of reviews and Y-axis is used for different emotions. The title is given as emotion distribution. Here, we can see a different trend, surprisingly the neutral emotion distribution is the highest followed by admiration, approval which are positive emotions and followed by realization, disapproval and disappointment, and annoyance which are towards negative emotions except the realization. This model is not able to distinguish if 2 emotions are present in the same review. If there are 2 emotions, then it is treating the emotion as neutral.

Comparison between Emotion detection using GPT-4 and Emotion detection using RoBERTa:

Now if we compare the emotion between GPT-4 and RoBERTa models. The neutral values from the GPT-4 models do not give proper outputs.

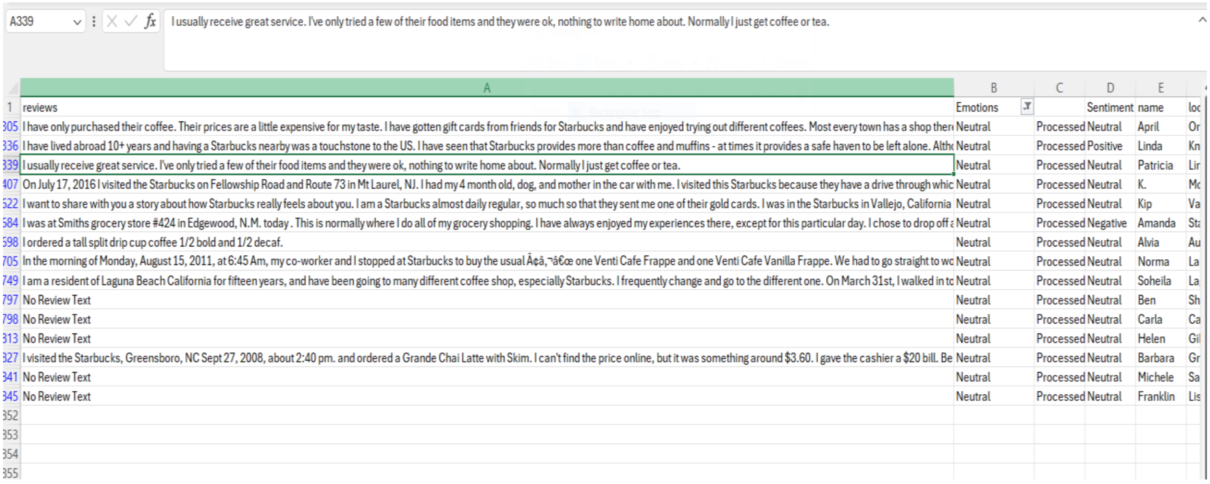

Fig.4 Neutral reviews emotion detection using GPT-4

We can see that most of the neutral emotion detection from GPT-4 model is not accurate. When I paste the same code in the RoBERTa model it gives an output which is much better. “I usually receive great service. I've only tried a few of their food items and they were ok, nothing to write home about. Normally I just get coffee or tea.” This was the review used for the comparison.

Fig.5 the same review but in the RoBERTa model

We can see here that it is showing the output as admiration, That is better output than the GPT-4 model.

Now let us take another example, the emotion which showed Frustration, Disappointment, Displeasure, Dissatisfaction, Discontent in the gpt model, if we put that as an input to the RoBERTa model it gives the output as neutral which is not true. The outputs which is obtained from the GPT-4 model were much better even though it gave 4 different emotions. “We had to correct them on our order 3 times. They never got it right then the manager came over to us and said we made her employee uncomfortable because we were trying to correct our order. The manager tried was racist against my stepmom (Chinese) taking over her but when I (**) would talk she would stop talking and listen to me.” This was the Review used for comparison.

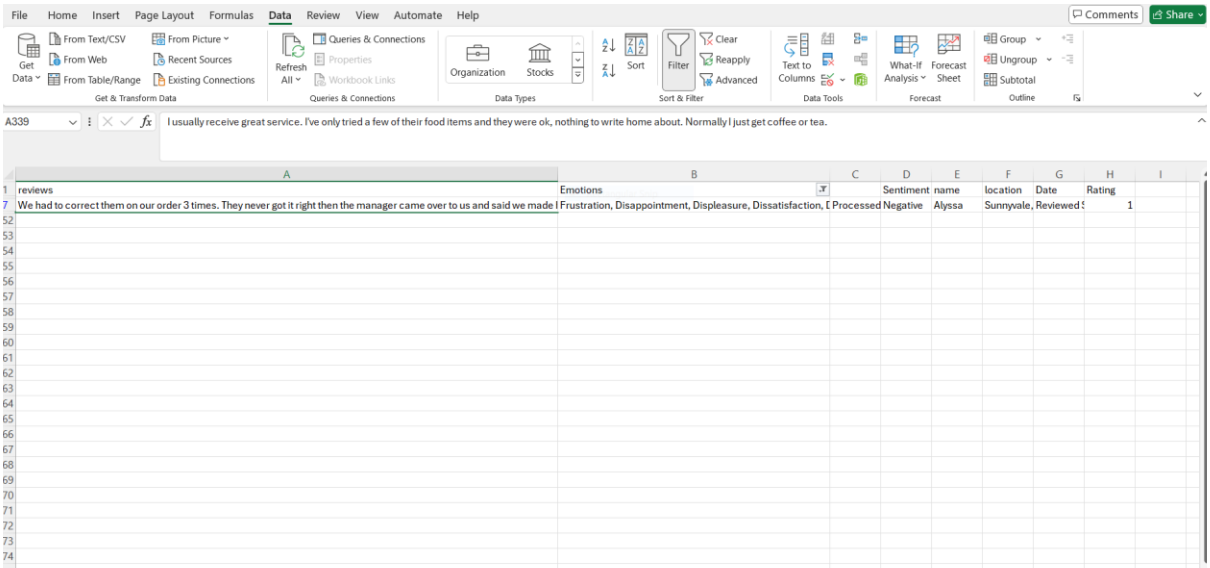

Fig.6 Frustration, Disappointment, Displeasure, Dissatisfaction, Discontent using GPT-4 model

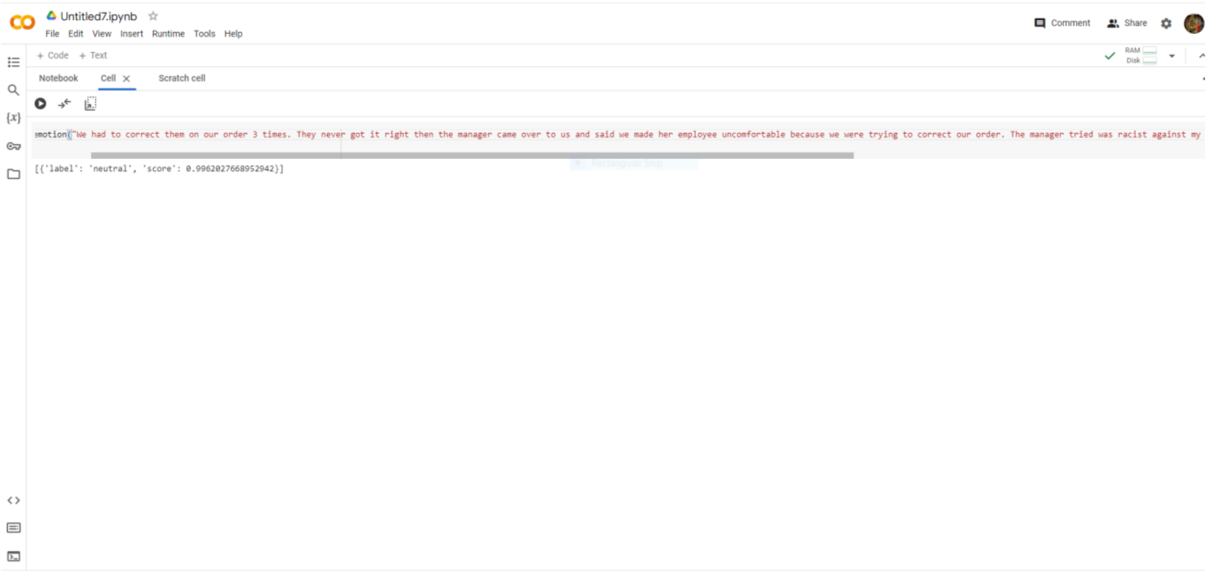

Fig.7 Frustration, Disappointment, Displeasure, Dissatisfaction, Discontent using RoBERTa model

But when both the models were compared on overall analysis. The GPT-4 model had very good accuracy compared to RoBERTa model even though for some of the emotions, it couldn't give a single emotion as output.

Conclusion:

There are advantages and disadvantages of both the models. The RoBERTa model could classify all the emotions, but the accuracy at which it classifies is very poor compared to the GPT-4 model. The GPT-4 model on the other hand was not able to give emotions for some of the reviews with one emotion. Sometimes it gave with 4-5 emotions and in some cases, it wasn't sure if the output it generated was correct. But when we compare both the models are overall. Then the GPT-4 model is very much better than the RoBERTa model. As it could classify multiple emotions and give high accuracy results for the majority of the results and couldn't give the emotion perfectly only in some cases. Also, the GPT-4 model requires an OPENAI_key to run the results, to obtain a OPENAI_key there has to be a subscription. While it does provide helpful information, we need to be cautious as it is only a predictive model. It can be exposed to being inaccurate in some cases. So, that is better to use it as support rather than main source of information. For future work, the GPT-4 model, which is giving an high accuracy result, can be improved so that the few instances where it couldn't give the output properly can be solved.

Acknowledgement.

I have used the code for GPT-4 and Apps Script which was provided by Professor Lorenzo and Elena.

References:

Intertwining two artificial minds: chaining GPT and RoBERTA for emotion detection. (2023c, December 4). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/10487718/authors#authors

Nandwani, P., & Verma, R. (2021b). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, *11*(1). https://doi.org/10.1007/s13278-021-00776-6

Azad, S. (2024b). The effectiveness of GPT-4 as financial news annotator versus human annotator in improving the accuracy and performance of sentiment analysis. In *Lecture notes in networks and systems* (pp. 105–119). https://doi.org/10.1007/978-981-99-8129-8_10

Hovy, E. (2014). What are Sentiment, Affect, and Emotion? Applying the Methodology of Michael Zock to Sentiment Analysis. In *Text, speech and language technology* (pp. 13–24). https://doi.org/10.1007/978-3-319-08043-7_2 https://www.kaggle.com/datasets/harshalhonde/starbucks-reviews-dataset

1littlecoder. (2022, January 8). *Text Emotion Detection in Python Hugging Face | Applied NLP Tutorial* [Video]. YouTube. https://www.youtube.com/watch?v=b9-0GpCqAQw

(for RoBERTa model)