Reviewing the Amazon Review Scale: A Sentiment Analysis Approach

Zhiqing Li, Zhourong Li, Pruthvi Reddy, Shreyans Kothari

Columbia University in the City of New York

Dec 11, 2021

Abstract

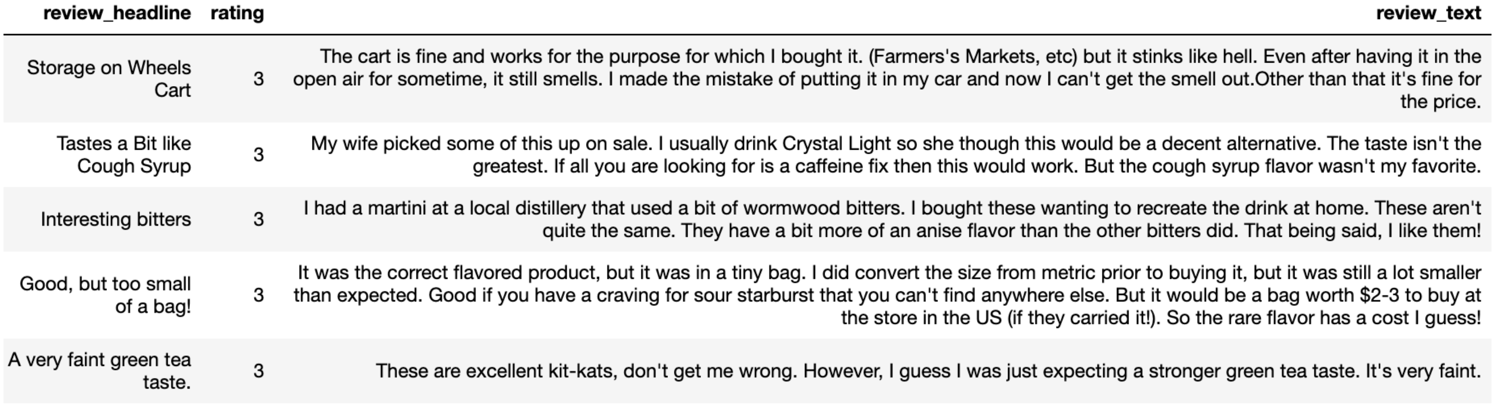
This paper seeks to explore the relationship between product reviews and their associated star ratings on Amazon. We hypothesize that the product rating system on the eCommerce website is a binary rating system where a 5-star rating is associated with positive reviews, and any stars less than 5 are associated with negative reviews. We test this theory by conducting sentiment analysis through three natural language processing models: BERT, GPT-3, and NLTK VADER. Prior to running the sentiment analyzer, we preprocess the data by removing special characters, removing stopwords, and stemming the text. However, in our final analysis, we only utilize the clean (without special characters) unstemmed text version of the product reviews, which includes the stopwords.

Introduction

The majority of the online shopping platforms design a 5-star rating system, allowing people to rate their shopping experience. Intuitively, if online shoppers are satisfied with a product they purchase online, they will give a five-star rating; by contrast, if they are slightly unsatisfied with it, they will tend to give a star rating lower than 5 and state their un-satisfactions in detail in the review sections. Theoretically, emotions and sentiments attached to the comments should be positively correlated with the star ratings. This is to say, the overall sentiments of a 4-star rating should be more positive than a

3-star rating, and a 3-star rating comment should be more positive than a 2-star rating comment and so on.

However, by peeking at the comments relative to the star rating, we find something interesting—nearly all the comments with lower than 5-star ratings seem to have strong negative sentiments that are inconsistent with their star ratings. For example, in the 3-star comments extracted from Amazon.com under the groceries category below, we can see that nearly the entire comments are criticizing the products or expressing disappointment; nevertheless, 3-star ratings are supposed to be associated with neutral feelings.



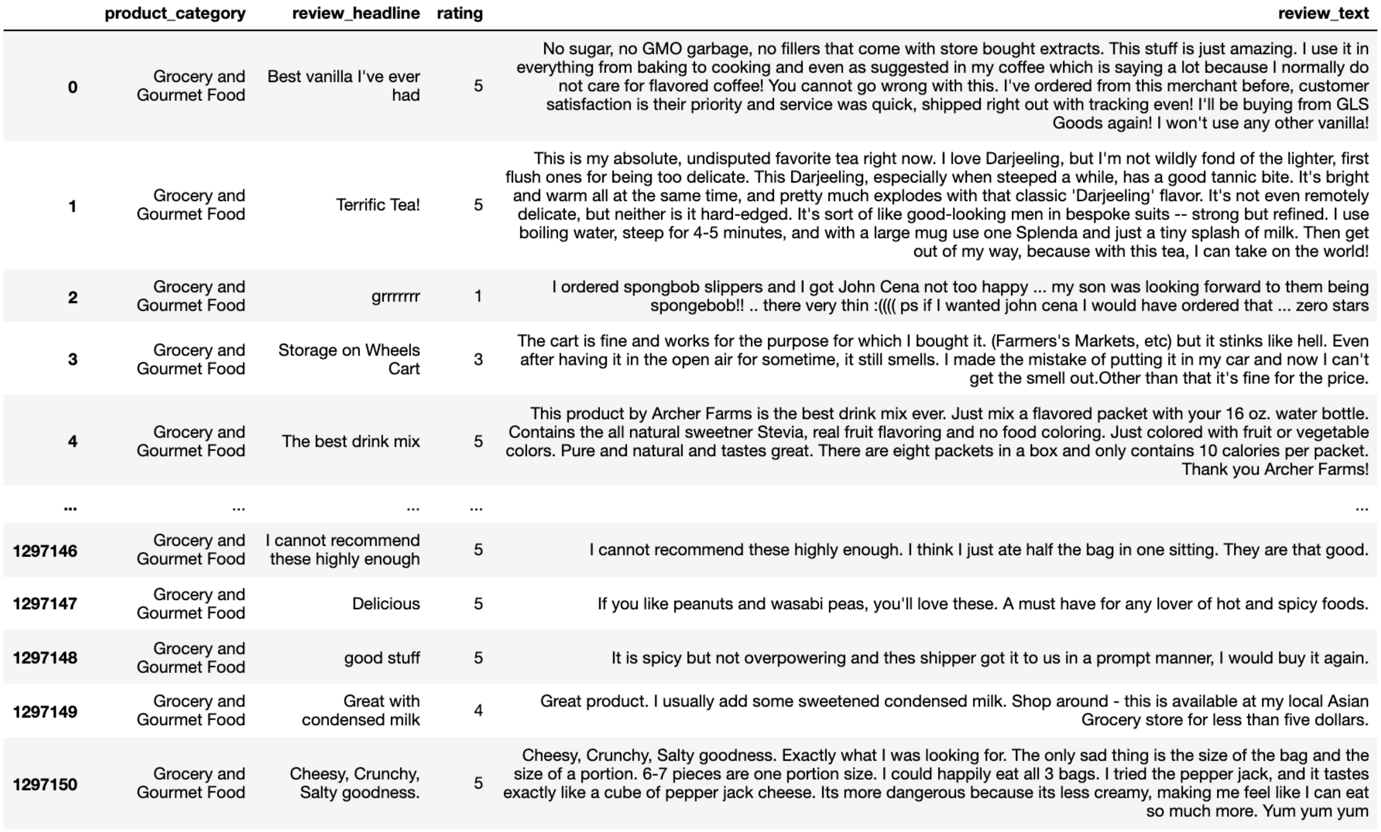
review_headline	rating	review_text
Storage on Wheels Cart	3	The cart is fine and works for the purpose for which I bought it. (Farmers's Markets, etc) but it stinks like hell. Even after having it in the open air for sometime, it still smells. I made the mistake of putting it in my car and now I can't get the smell out.Other than that it's fine for the price.
Tastes a Bit like Cough Syrup	3	My wife picked some of this up on sale. I usually drink Crystal Light so she though this would be a decent alternative. The taste isn't the greatest. If all you are looking for is a caffeine fix then this would work. But the cough syrup flavor wasn't my favorite.
Interesting bitters	3	I had a martini at a local distillery that used a bit of wormwood bitters. I bought these wanting to recreate the drink at home. These aren't quite the same. They have a bit more of an anise flavor than the other bitters did. That being said, I like them!
Good, but too small of a bag!	3	It was the correct flavored product, but it was in a tiny bag. I did convert the size from metric prior to buying it, but it was still a lot smaller than expected. Good if you have a craving for sour starburst that you can't find anywhere else. But it would be a bag worth $2-3 to buy at the store in the US (if they carried it!). So the rare flavor has a cost I guess!
A very faint green tea taste.	3	These are excellent kit-kats, don't get me wrong. However, I guess I was just expecting a stronger green tea taste. It's very faint.

Hence, we develop two hypotheses: (1) Despite having a 5-point scale for rating, the Amazon rating system is a binary rating system: 5-stars are associated with positive reviews, and stars less than 5 are associated with negative reviews; (2) the Amazon reviews with 3-4 stars would have a negative sentiment value attached to it.

This paper is organized as follows: (1) in the first section, we will give a brief introduction of our dataset; (2) in the second section, we will articulate our methodologies adopted to testify the hypotheses, including pre-processing techniques, sentiment analysis rationales, and three natural language processing models-BERT, NLTK VADER, and GPT-3; (3) in the last section, we will compare the results of these three mentioned models and deliver our final conclusions.

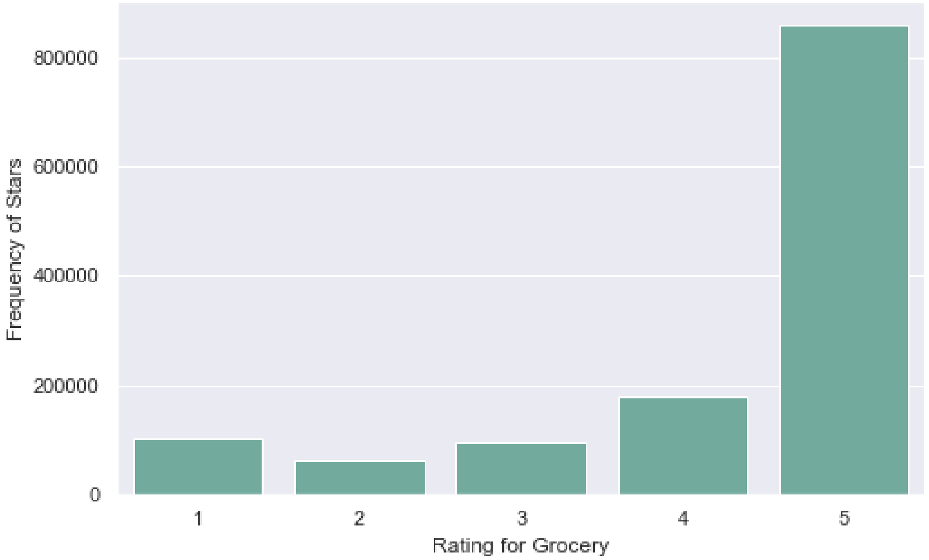
Data

The dataset we use is extracted from Amazon.com and organized by Julian McAuley at UCSD[1] under the *Grocery and Gourmet Food* section. The reviews span from May 1996 to July 2014 and include reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs). Our dataset has 1,297,150 review in total, and it is imbalanced in terms of rating distribution.



	product_category	review_headline	rating	review_text
0	Grocery and Gourmet Food	Best vanilla I've ever had	5	No sugar, no GMO garbage, no fillers that come with store bought extracts. This stuff is just amazing. I use it in everything from baking to cooking and even as suggested in my coffee which is saying a lot because I normally do not care for flavored coffee! You cannot go wrong with this. I've ordered from this merchant before, customer satisfaction is their priority and service was quick, shipped right out with tracking even! I'll be buying from GLS Goods again! I won't use any other vanilla!
1	Grocery and Gourmet Food	Terrific Tea!	5	This is my absolute, undisputed favorite tea right now. I love Darjeeling, but I'm not wildly fond of the lighter, first flush ones for being too delicate. This Darjeeling, especially when steeped a while, has a good tannic bite. It's bright and warm all at the same time, and pretty much explodes with that classic 'Darjeeling' flavor. It's not even remotely delicate, but neither is it hard-edged. It's sort of like good-looking men in bespoke suits -- strong but refined. I use boiling water, steep for 4-5 minutes, and with a large mug use one Splenda and just a tiny splash of milk. Then get out of my way, because with this tea, I can take on the world!
2	Grocery and Gourmet Food	grrrrrrr	1	I ordered spongbob slippers and I got John Cena not too happy ... my son was looking forward to them being spongebob!! .. there very thin :((((ps if I wanted john cena I would have ordered that ... zero stars
3	Grocery and Gourmet Food	Storage on Wheels Cart	3	The cart is fine and works for the purpose for which I bought it. (Farmers's Markets, etc) but it stinks like hell. Even after having it in the open air for sometime, it still smells. I made the mistake of putting it in my car and now I can't get the smell out.Other than that it's fine for the price.
4	Grocery and Gourmet Food	The best drink mix	5	This product by Archer Farms is the best drink mix ever. Just mix a flavored packet with your 16 oz. water bottle. Contains the all natural sweetner Stevia, real fruit flavoring and no food coloring. Just colored with fruit or vegetable colors. Pure and natural and tastes great. There are eight packets in a box and only contains 10 calories per packet. Thank you Archer Farms!
...	...	...	...	...
1297146	Grocery and Gourmet Food	I cannot recommend these highly enough	5	I cannot recommend these highly enough. I think I just ate half the bag in one sitting. They are that good.
1297147	Grocery and Gourmet Food	Delicious	5	If you like peanuts and wasabi peas, you'll love these. A must have for any lover of hot and spicy foods.
1297148	Grocery and Gourmet Food	good stuff	5	It is spicy but not overpowering and thes shipper got it to us in a prompt manner, I would buy it again.
1297149	Grocery and Gourmet Food	Great with condensed milk	4	Great product. I usually add some sweetened condensed milk. Shop around - this is available at my local Asian Grocery store for less than five dollars.
1297150	Grocery and Gourmet Food	Cheesy, Crunchy, Salty goodness.	5	Cheesy, Crunchy, Salty goodness. Exactly what I was looking for. The only sad thing is the size of the bag and the size of a portion. 6-7 pieces are one portion size. I could happily eat all 3 bags. I tried the pepper jack, and it tastes exactly like a cube of pepper jack cheese. Its more dangerous because its less creamy, making me feel like I can eat so much more. Yum yum yum

[1] The dataset can be accessed here: https://jmcauley.ucsd.edu/data/amazon/



Methodology

Preprocessing

Reviews have a peculiar nature to them, people express their views on products with interesting words and sometimes context too hence the pre-processing step was crucial. We need to keep as much context as possible and sometimes extract context too, for example from emojis and emoticons. The following steps of preprocessing were conducted

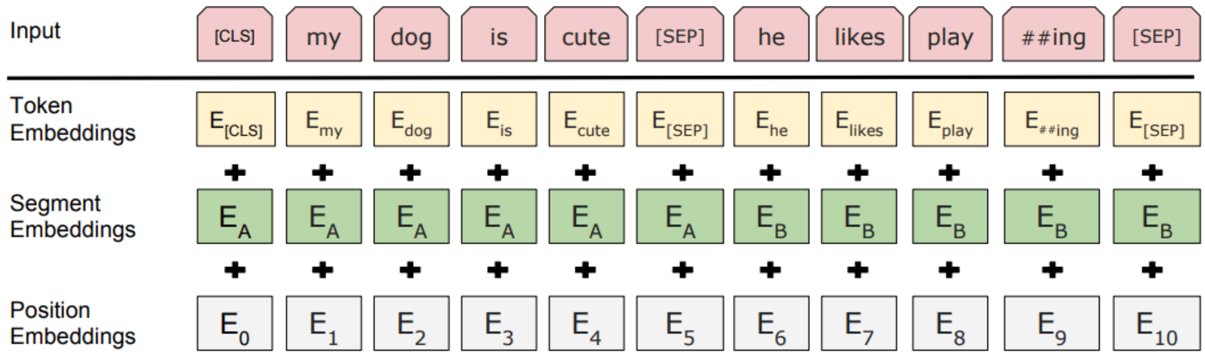
- **Removing URLs:** Finding and removing any URLs present in the review text.
- **Removing HTML tags:** Removing HTML tags that may exist.
- **Chat word conversions:** Converting chat abbreviations to text based on a dictionary
- **Emojis and Emoticons conversion to text:** These are clear indicators of sentiment

Later we removed stop words from the text and decided, based on trial and error, which text to use.

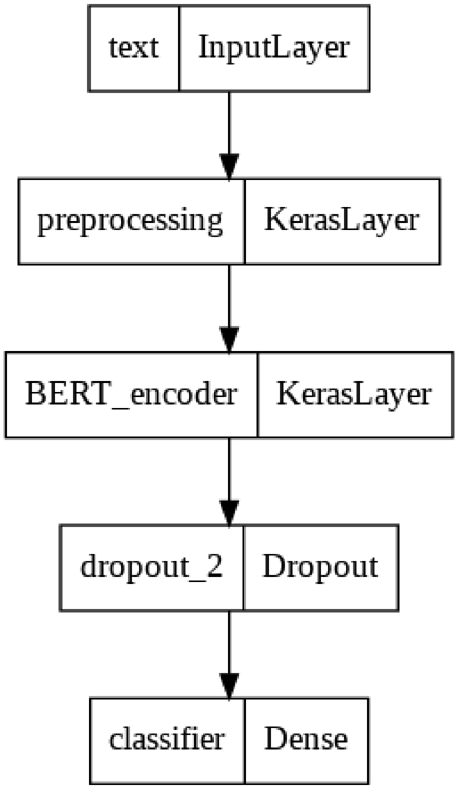
BERT

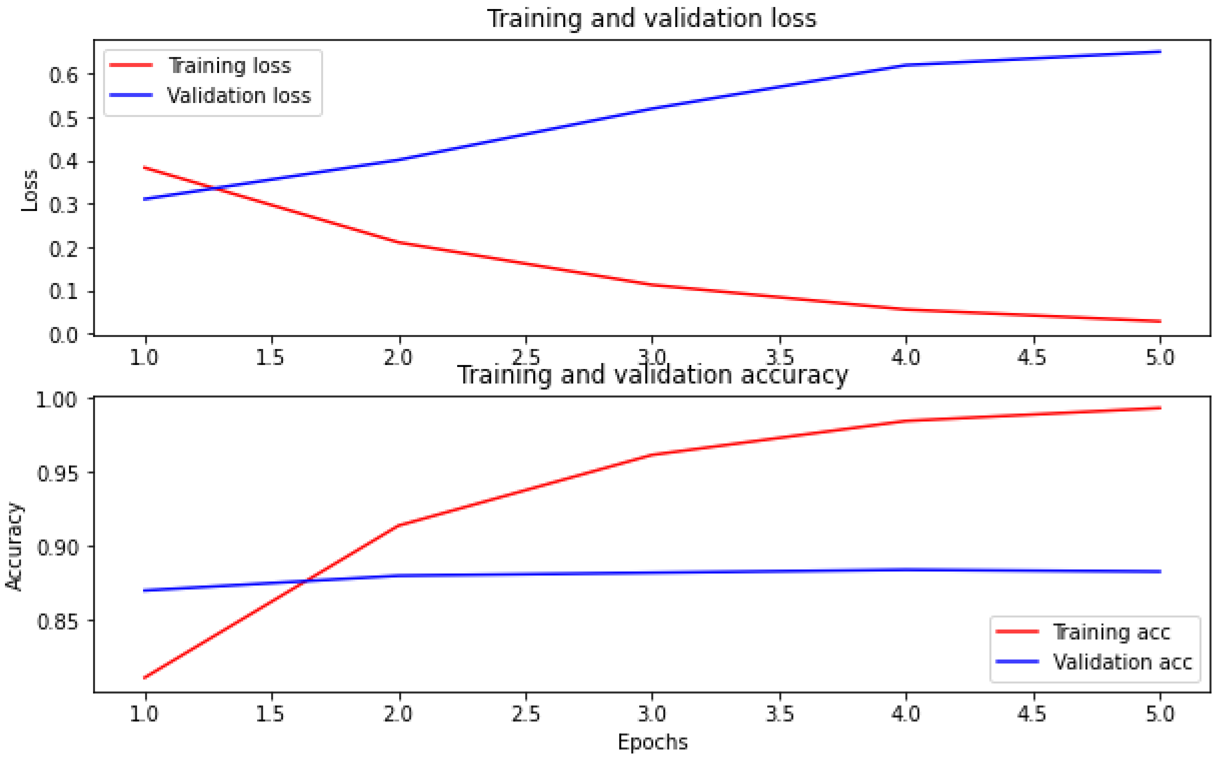
BERT i.e Bidirectional Encoder Representation from Transformers is based on a transformer Neural Network which was an ecolution over the LSTM Networks. Even bi-directional LSTMs were not as good at understanding context or developing semantic understanding since the way it worked was to concatenate the encodings of each direction. The separation of tasks by encoder and decoder in a Transformer architecture is what enables or allows it to understand language semantically. We pretrain BERT to understand language first and then we can fine tune it to a specific use-case like sentiment analysis, Q & A or text-summarization etc.

The pre-training is done with masking words known as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). BERT relies on a Transformer (the attention mechanism that learns contextual relationships between words in a text). A basic Transformer consists of an encoder to read the text input and a decoder to produce a prediction for the task. Since BERT's goal is to generate a language representation model, it only needs the encoder part. The input to the encoder for BERT is a sequence of tokens, which are first converted into vectors and then processed in the neural network. But before processing can start, BERT needs the input to be massaged and decorated with some extra metadata:

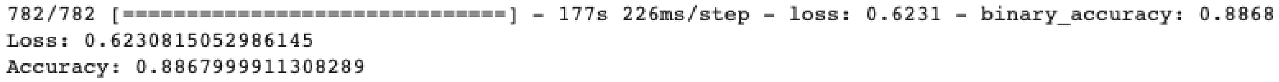


We took the Bert-Uncased pre-trained model and fine tuned it further with the IMDB Reviews Dataset which was classified hence something like a ground truth. The model architecture we ended up using was as follows.


text
InputLayer
preprocessing
KerasLayer
BERT_encoder
KerasLayer
dropout_2
Dropout
classifier
Dense


Training and validation loss
Training loss
Validation loss
Loss
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Training and validation accuracy
Accuracy
0.85
0.90
0.95
1.00
Training acc
Validation acc
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Epochs

The Training was done on 50k IMDB reviews with a 10% validation set. A training accuracy of 99% was achieved. The Model Performed well on the validation set with an accuracy of 88.6%.



```
782/782 [==============================] - 177s 226ms/step - loss: 0.6231 - binary_accuracy: 0.8868
Loss: 0.6230815052986145
Accuracy: 0.8867999911308289
```

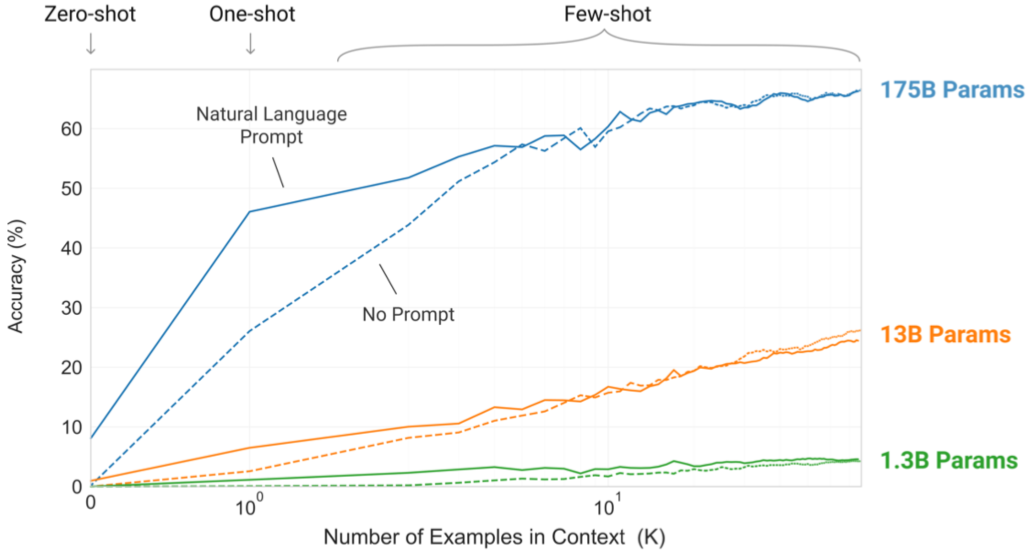
Later on, with this model we predicted 125k records out of the over 1.2 million records of review we had. This gave us a good sample size to visualize and explore and at the same time something we could process with resources that didn't cost too much.

GPT-3

The second sentiment analysis model we used is Generative Pretrained Transformer 3 (GTP-3). GPT-3 is a pre-trained deep learning model that was trained using large amount of datasets. The datasets being used include Common Crawl, Web Text2, Books and Wikipedia. GPT-3 is an autoregressive language model with 175 billion parameters, which is 10x more than any previous non-sparse language model.

Most existing NLP models are being trained with task-specific fine-tuning datasets of thousands or tens or thousands of examples, and might perform poorly when the given examples are limited. The developers team of GPT-3 want to demonstrate that scaling up the language models can greatly improve the task-agnostic, few shot performance, and sometimes even performs better than the state-of-art fine-tuning model that trained specifically towards the task[2]. For all tasks, GPT-3 was trained with no gradient updates or fine-tuning. With providing tasks and few-shots examples, GPT-3 can achieve high performance in many NLP tasks.

[2] https://arxiv.org/abs/2005.14165 pg1



The figure above shows that, with an increasing number of parameters and sizes, the performance of the model has highly increased in the zero-shot, one-shot, and few-shot tasks.

In our project, we used the API provided by Open AI that connects to the GPT-3 model online. We then provided the GPT-3 using API with few examples like follows,

Sentence: "I loved the new Batman movie!"

Sentiment: Positive

Sentence: "I hate it when my phone battery dies💢."

Sentiment: Negative

Sentence: "My day has been 👍"

Sentiment: Positive

After providing a few examples, we can then transfer reviews from our dataset to the model using the API, and let GPT-3 classify all our reviews. We would then clean up the feedback from the API, and record the sentiment and log probability returned by the model.

From the above example we can see GPT-3 can actually classify sentiment from the emojis as well, which is a task that cannot be easily attained using other NLP models. We ran the GPT-3 model over 10000 Amazon reviews, and collected the feedback provided by the model. The model would directly provide a response of “positive” or “negative” instead of a sentiment score, and with a log-probability, indicates the confidence of the prediction. In the result section we will further analyze sentiment classification results given by the GPT-3 model.

NLTK VADER

For the final sentiment analysis model, we used the Natural Language Toolkit Valence Aware Dictionary and sEntiment Reasoner (NLTK VADER). The NLTK VADER sentiment intensity analyzer is a pre-trained model that is commonly used for text analysis; the analyzer is sensitive to the intensity/strength of the emotion, as well as polarity: positive or negative. The model is pre-trained on text data from social media, including microblogs, and outperforms human raters on classifying sentiments for social media text, movie reviews, and product reviews into positive, neutral, or negative classes.[3] VADER sets itself apart from other sentient analysis models in that it is fine-tuned for sentiment expression on social media text, and yet it also scores high(er) accuracy metrics (F1) in all other contexts.[4]

We ran the VADER sentiment intensity analyzer on 125,000 Amazon product reviews and got back compound scores for each review. The compound score is essentially the combined normalized positive, neutral, and negative score. It ranges between -1 and 1. Commonly, 0.05 is used as a threshold for positive sentiments; a compound score above 0.05 is considered positive. Similarly, a compound score below -0.05 is considered negative. Compound scores lying between -0.05 and 0.05 are considered neutral. For the sake of our research, we used 0 as the threshold since -0.05 to 0.05 was a really small range and virtually no observations fell in it. All observations with a compound score greater than 0 are considered positive, and all under 0 are considered negative.

[3] https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109/8122 pg 1

[4] https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109/8122 pg 1

Comparing the models

Using 5 different randomly picked reviews, one from each star category, we compare the three models.

On the review associated with 5 stars, all three models perform extremely well. Both BERT and VADER return a high score associated with positive sentiments. In the case of GPT-3, it does not return a likelihood score, rather it returns classifications; in this case, it returns a positive classification. The reviewer used strong positive words such as “absolute, undisputed favorite”, “love”, “good-looking”, etc. All three models are able to pick up on the lexicons despite the fact that not all the words that were used in the review are objectively ‘positive’ by themselves. This depicts the models’ ability to pick up on the context.

Review (5 Stars)	Model	Score
This is my absolute, undisputed favorite tea right now. I love Darjeeling, but I'm not wildly fond of the lighter, first flush ones for being too delicate. This Darjeeling, especially when steeped a while, has a good tannic bite. It's bright and warm all at the same time, and pretty much explodes with that classic 'Darjeeling' flavor. It's not even remotely delicate, but neither is it hard-edged. It's sort of like good-looking men in bespoke suits -- strong but refined. I use boiling water, steep for 4-5 minutes, and with a large mug use one Splenda and just a tiny splash of milk. Then get out of my way, because with this tea, I can take on the world!	BERT range [0,1]	0.99
	GPT-3	Positive
	NLTK (compound) range [-1,1]	0.975

For the 4 star review, both BERT and NLTK VADER return a positive score (closer to 1). However, the score is less than the score for the product with a 5 star rating. GPT-3, on the other hand, assigns a negative sentiment to this review. The reviewer used positive words like “pretty-nice”, and these were juxtaposed against negative words like “overpriced.” Overall, just by reading the review, one might not be able to guess that this review is associated with a 4 star rating. From this comparison, it seems that the GPT-3 model is more sensitive to negative words than the other two models.

Review (4 Stars)	Model	Score
And they're pretty nice! One set looks like ice, and you get a set of four thrice. Kind of overpriced.	BERT range [0,1]	0.96
	GPT-3	Negative
	NLTK (compound) range [-1,1]	0.8313

In comparing the review for the product with a 3 star rating, we find that all three models return a negative score/classification. The review does not use a lot of negative terms outrightly, however, the three models are still able to pick up on the context and appropriately classify this review as negative.

Review (3 Stars)	Model	Score
The cart is fine and works for the purpose for which I bought it. (Farmers's Markets, etc) but it stinks like hell. Even after having it in the open air for sometime, it still smells. I made the mistake of putting it in my car and now I can't get the smell out.Other than that it's fine for the price.	BERT range [0,1]	0.019
	GPT-3	Negative
	NLTK (compound) range [-1,1]	-0.7992

In the case of 2 stars, the reviewer uses positive words like “best” and “better”, but these positive words are nullified either by use of negation, or through comparison against another product. All three models are able to pick up on this and return a low(er) score. Both BERT and GPT-3 classify this review as negative. NLTK VADER, on the other hand, assigns this product review a low score, but the score still lies on the positive side of the spectrum. This phenomenon possibly arises due to

VADER's inability to fully understand the context. Another explanation could be the model's inability to comprehend negation.

Review (2 Stars)	Model	Score
Not the best chai taste. Pretty grainy. Is OK, at best. Have had better tasting ones.	BERT range [0,1]	0.0039
	GPT-3	Negative
	NLTK (compound) range [-1,1]	0.3176

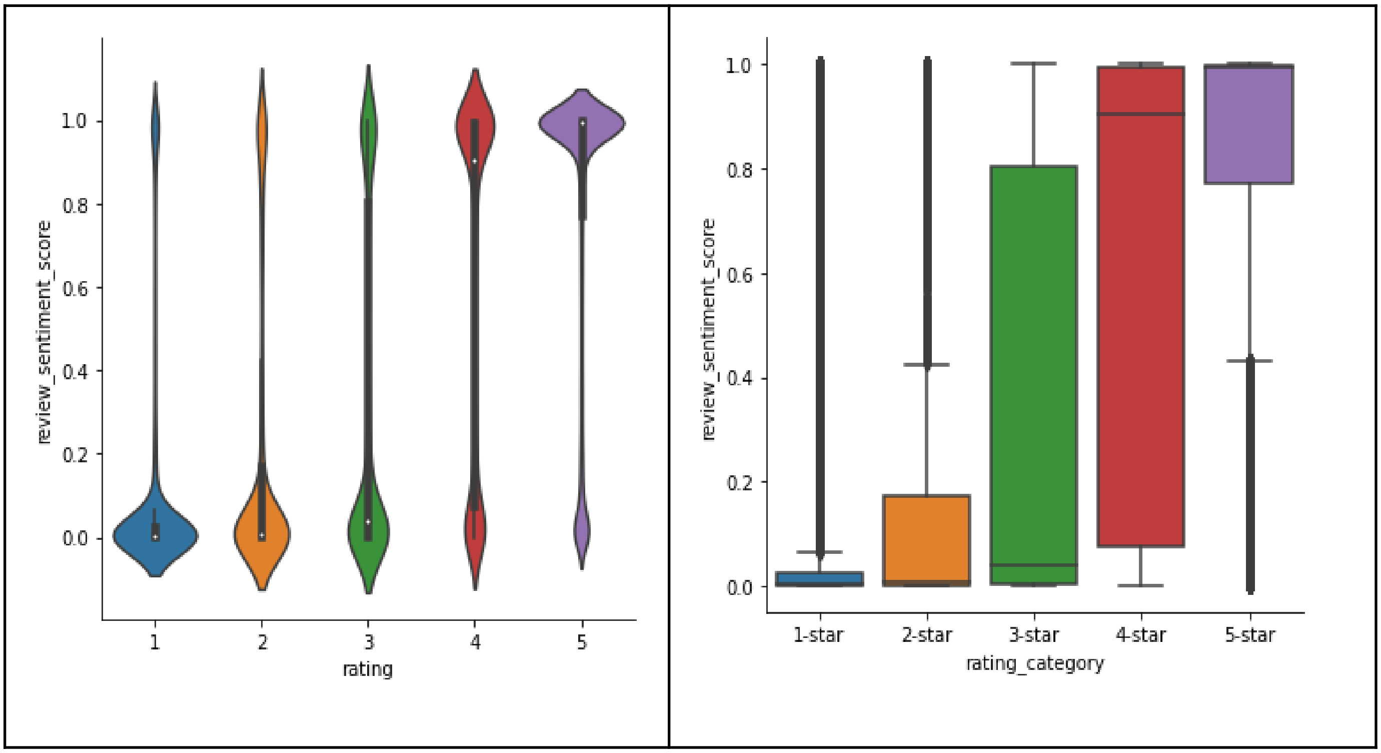
In the case of the 1 star rating, the review is assigned a negative sentiment by all three models. The reviewer goes into grave detail regarding their experience with the product, and except for "damaged" and "caution," they do not necessarily use words that would be assigned a negative score by themselves. In this case, all three models do a great job of analyzing and scoring based on the context of the review.

Review (2 Stars)	Model	Score
Not the best chai taste. Pretty grainy. Is OK, at best. Have had better tasting ones.	BERT range [0,1]	0.0039
	GPT-3	Negative
	NLTK (compound) range [-1,1]	0.3176

Overall, all three models perform quite well. GPT-3 is the only model we utilized that fails to reject our hypothesis that any ratings below 5 are associated with a negative review. BERT and NLTK VADER are slightly less sensitive to negative words/reviews than the GPT-3 model.

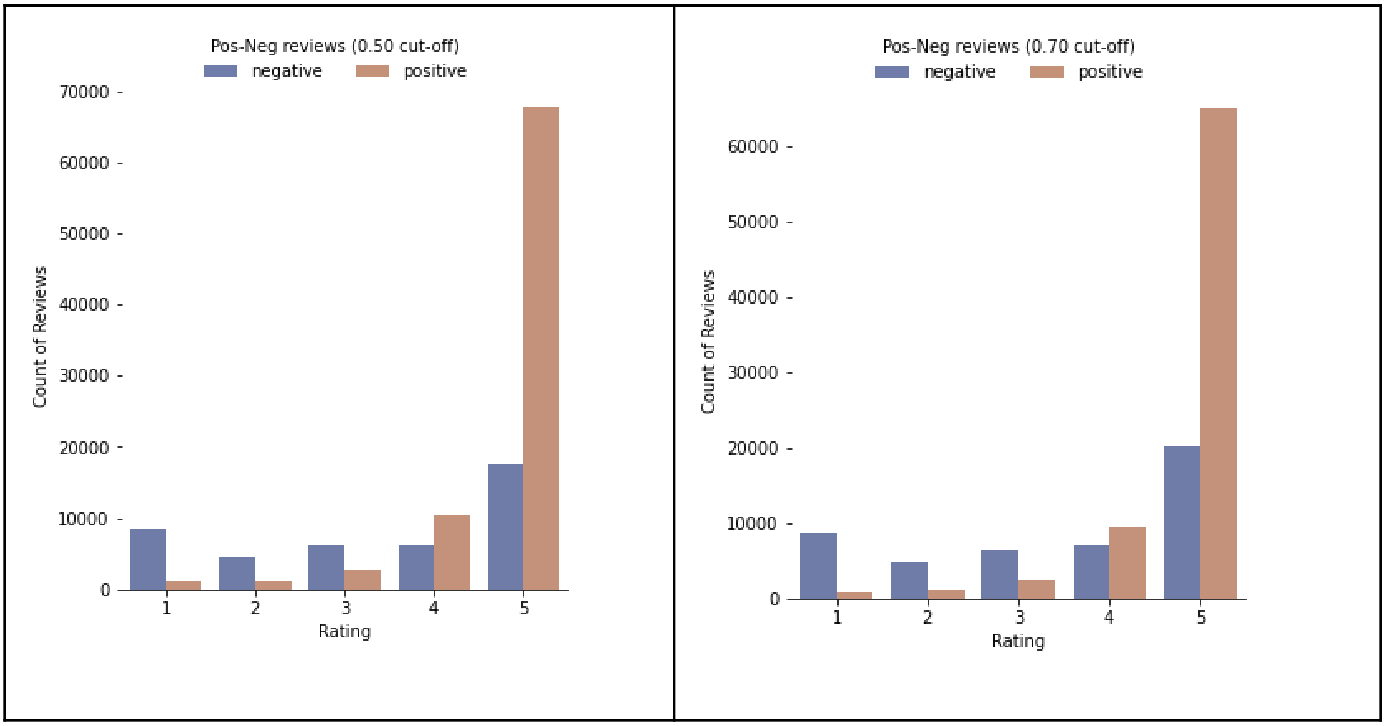
Results

BERT



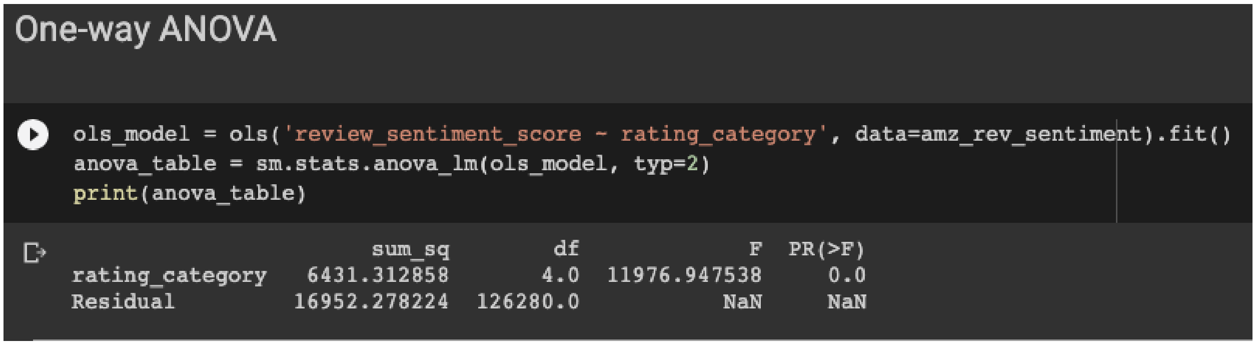
As observed in the above two graphs we can see that the 5-star reviews have 50%tile comfortable on the positive sentiment. The 4-star rated ones though, the 50%tile is spread out which is not what our hypothesis hoped to find. If our hypothesis was indeed true, the medians of both 3-star rating and the 4-star rating should be comfortable or at least just below the 0.5 mark on the sentiment score.

Even after classifying the positive or negative reviews based on a slightly more biased threshold of 0.7 instead of 0.5, the distribution didn't move much in favor of the hypothesis. The plot below shows the comparison. That said, either of them need to be confirmed with a more clearer hypothesis testing mechanism like a t-test or a ANOVA.


Pos-Neg reviews (0.50 cut-off)

Pos-Neg reviews (0.70 cut-off)

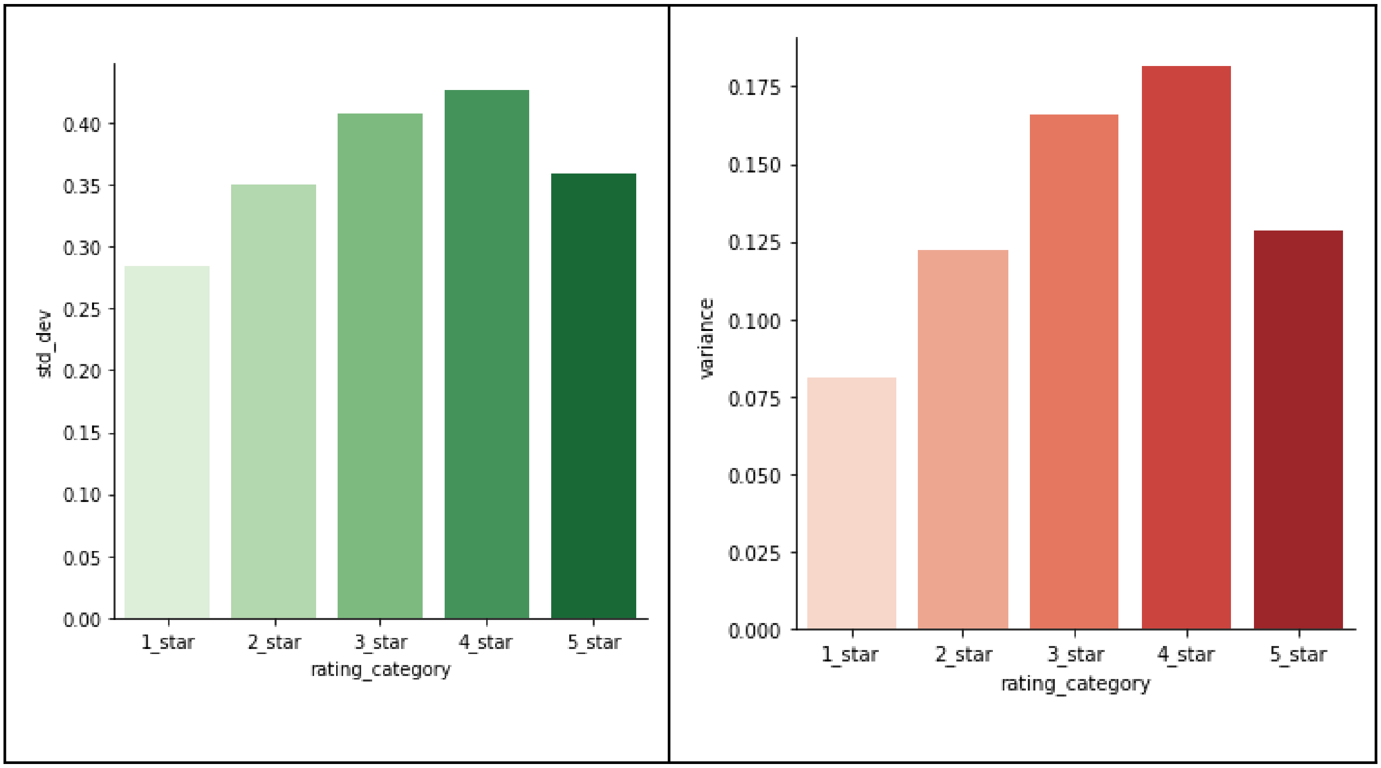
We performed a quick ANOVA test to establish and verify if these groups of rating were indeed separate or they have such high overlapping traits and behavior that the difference is not resulting in any distinction. Turns out, a simple one-way ANOVA confirms that the groups are indeed distinct and have their medians not overlapping at all with a p-value of 0.00, very affirmative!


One-way ANOVA

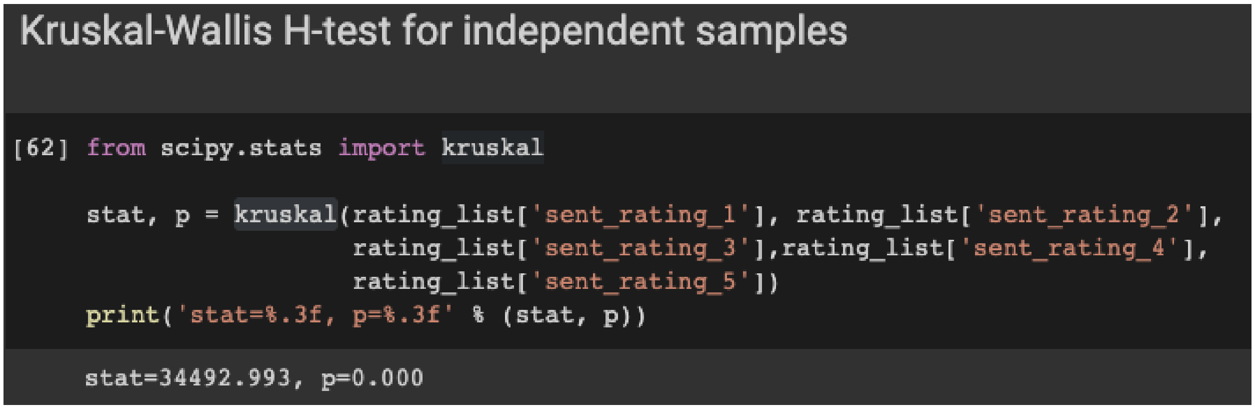
```
ols_model = ols('review_sentiment_score ~ rating_category', data=amz_rev_sentiment).fit()
anova_table = sm.stats.anova_lm(ols_model, typ=2)
print(anova_table)
```

```
                       sum_sq        df             F  PR(>F)
rating_category   6431.312858       4.0  11976.947538     0.0
Residual         16952.278224  126280.0           NaN     NaN
```

To reason for it we can look at how the variance and standard deviation are there in each of these groups of ratings. If one group or other is really indeed behaving in an exceptional way. Turns out, nothing significant can be observed here too.



Lastly, we compute the Kruskal-Wallis H-test for independent samples, which is just another way to assert if these groups are indeed distinct.The Kruskal-Wallis H-test tests the null hypothesis that the population medians of all of the groups are equal. It is a non-parametric version of ANOVA. The test works on 2 or more independent samples, which may have different sizes. Note that rejecting the null hypothesis does not indicate which of the groups differs. The independence is confirmed here too hence, disproving our hypothesis with a p-value of 0.00, very affirmative as well.



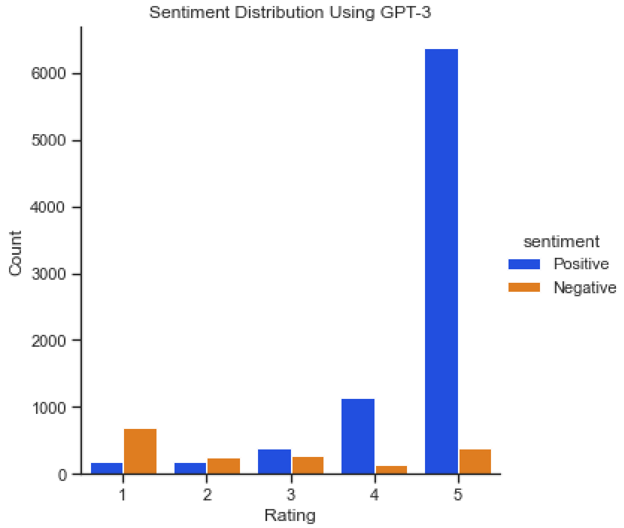
Kruskal-Wallis H-test for independent samples

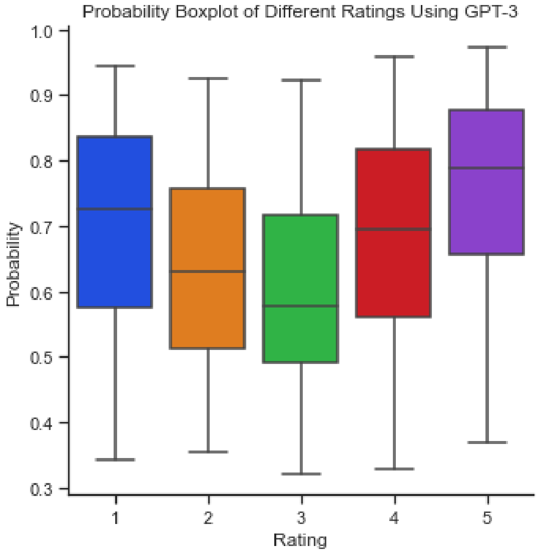
```
[62] from scipy.stats import kruskal

     stat, p = kruskal(rating_list['sent_rating_1'], rating_list['sent_rating_2'],
                       rating_list['sent_rating_3'],rating_list['sent_rating_4'],
                       rating_list['sent_rating_5'])
     print('stat=%.3f, p=%.3f' % (stat, p))
```

```
     stat=34492.993, p=0.000
```

GPT-3





GPT-3 Classification	1 Star	2 Star	3 Star	4 Star	5 Star
Positive Count	173	175	393	1149	6373
Negative Count	687	249	269	144	388
Total Count	860	424	662	1293	6761
Positive Proportion	0.2012	0.4127	0.5937	0.8886	0.9426
Negative Proportion	0.7988	0.5873	0.4063	0.1114	0.05739
Total Proportion	1	1	1	1	1

From the sentiment distributions plot on the left, we can see there are more positive reviews than the negative reviews in the ratings of 3 star, 4 star and 5 star. For the plot on the right, we can see the GPT-3 model has high confidences (probabilities) when classifying the one star review and five star review. GPT-3 has the lowest confidence when classifying the three star reviews.

Since we do not have the same number of comments among all the ratings. It would be more meaningful to compare the proportion of positive reviews and negative reviews among different ratings. From the table above, we can observe that reviews with 5 stars, 4 stars and 3 stars all have positive proportions that are higher than 50%, so these ratings have more positive reviews than negative ones. Only for reviews with one star and two stars, the negative proportions started to surpass the positive proportions.

GPT-3 Output	5 Stars Reviews	Non 5 Stars Reviews
Positive Count (Percentage)	6373 (0.9426)	1890 (0.5835)
Negative Count (Percentage)	388 (0.0574)	1349 (0.4165)
Total	6761	3239

From the result generated by GPT-3 above, we can safely conclude that 5 star reviews are associated with positive reviews since 94.26% of 5 stars reviews are classified as positive, and only 5.74% of the 5 star reviews are classified as negative. However, for the reviews that are under 5 stars, the percentage of positive reviews is also more than half, which is 58.35%. The reviews under 5 stars that are being classified as negative only have a percentage of 41.65%. Therefore, the reviews that under 5 stars seem not quite associated with negative sentiment.

However, we can conduct a two proportion Z-Test to see if the proportion of positive reviews in 5 star reviews are significantly different from the proportion of positive reviews in the below 5 star reviews. Our null hypothesis is

$$H_0 : P_1 = P_2$$

which means the two proportions are not significantly different. And our alternative hypothesis is

$$H_0 : P_1 \neq P_2$$

Which indicates these two proportions are significantly different. We can use the following test statistic

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where

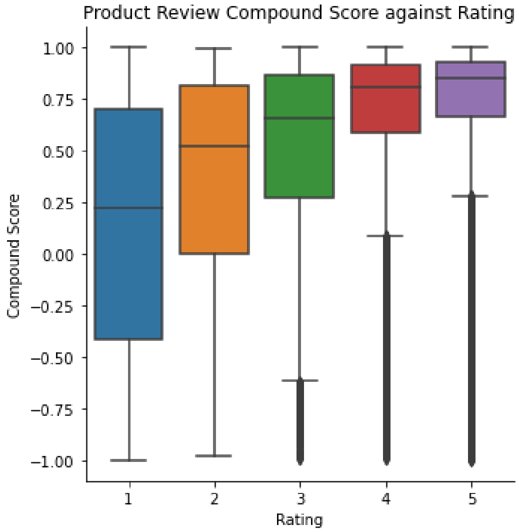
$$\hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}$$

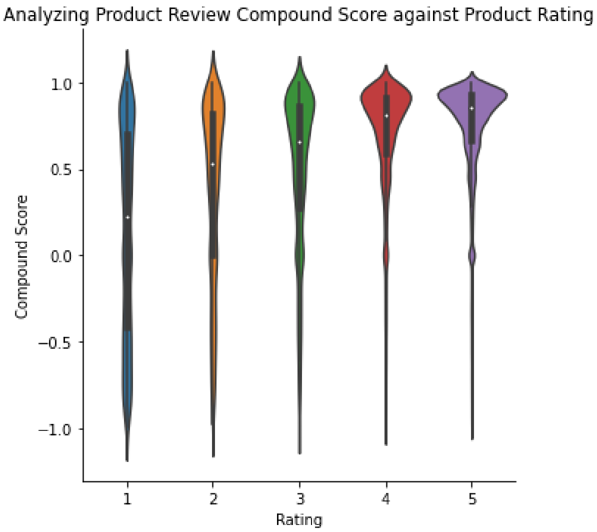
Y_1 is the number of 5 stars reviews that are positive, and Y_2 is the number of under 5 star reviews that are positive. We have $\hat{p} = \frac{6373+1890}{6761+3239} = 0.8263$ and

$$Z = \frac{0.9426 - 0.5835}{\sqrt{0.8263(1-0.8263)(0.0004566)}} = 44.3587$$

Since 44.3587 > 1.96, we reject our null hypothesis at a = 0.05 level. We can then conclude that, although reviews below 5 stars are having more than half being classified as positive, at a = 0.05 level, the proportion of positive reviews in 5 star reviews are statistically significantly different from the proportion of positive reviews in reviews under 5 stars.

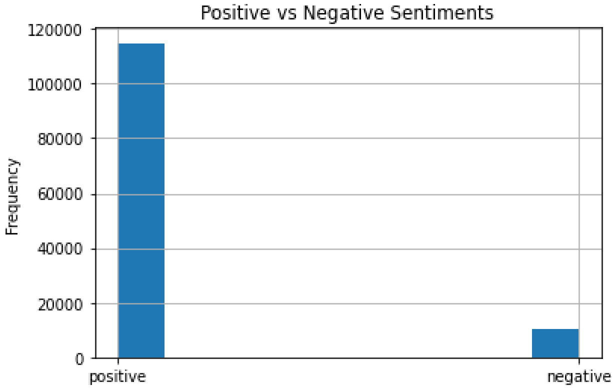
NLTK VADER

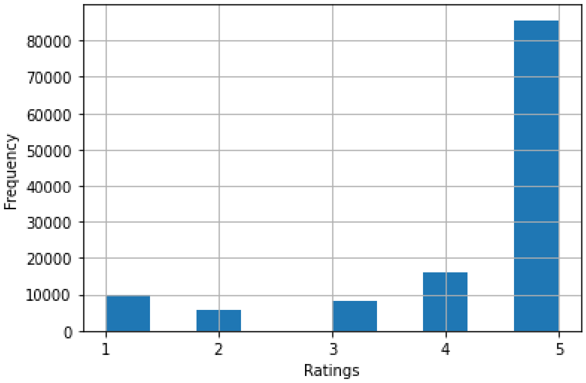




Sentiment	Count	Proportion
Positive	114762	0.918096
Negative	10238	0.081904
Total	125000	1.00

Looking at the boxplot and catplot for the VADER compound score on the product reviews, we can see that most of the reviews for all ratings cluster towards positive scores (>0). In the case of 5-star ratings, the reviews were assigned a completely positive score, with the median score lying around ~0.80. In the case of 1-star ratings, the scores seem to be more spread out, however, the median score still lies closer to the positive end of the spectrum. We inspect this further by analyzing the count and proportion of positive and negative sentiments in the reviews. Over 91% of the reviews in our dataset are classified as positive and only about 8% are classified as negative. Just based on these, our hypothesis is rejected. There seem to be positive reviews associated with all types of ratings.





Majority of the products in our sample have positive reviews. We inspect this further by analyzing the frequency of ratings in our dataset. Upon further investigation, it becomes evident that the dataset is grossly imbalanced with a majority of the product reviews having 5 star ratings.

Rating	Count	Count		Proportion	
		Positive	Negative	Positive	Negative
1	9740	6018	3722	0.6179	0.3821
2	5611	4318	1293	0.7696	0.2304
3	8151	7029	1122	0.8623	0.1377
4	15882	14985	897	0.9435	0.0565
5	85616	82412	3204	0.9626	0.0374

This table breaks down the product review star rating by positive/negative classification in the NLTK VADER model. As we can see, all across the ratings, the majority of the reviews were classified as positive. 1 star rating reviews have the highest negative classification of all the ratings at 38%, but those are still less compared to the positive classifications in the 1 star review category. For all other star rating categories (2+ stars), very minimal number of reviews were classified as negative, with the lowest proportion being about 3% for the 5 star rating group.

Conclusion

Our conclusions from models vary a bit and we will highlight the ones that have some of them. First, the NLTK VADER model is a robust model which is easy to interpret and use. Based on the results from NLTK, it is safe to say that our original hypothesis does not hold true. The Amazon product review scale is quite complex and can not be condensed into a binary representation of positive and negative. Due to the imbalance in the data, further analysis is required with a different dataset which has a relatively more equal distribution across all star ratings.

For the GPT-3 model, it does show that reviews with 5 stars are more associated with positive sentiment, with 94.26% of the 5 star reviews being classified as positive. However, GPT-3 also classified more than half of the reviews under 5 stars as positive. Therefore, our first half of the hypothesis seems supported by GPT-3, but the second half of hypothesis seems not supported by the result of GPT-3. One should notice that although more than half of the reviews under 5 stars are being classified as positive, the proportion of under five star reviews being classified as positive are statistically different from the proportion of five star reviews being classified as positive.

The BERT model on the other hand produced results that placed 3 and 4 stars somewhere distinctly between 2 and 5 star rating which means they are not always positive nor always negative and the statistical tests we performed kind of confirmed them. Reviews have a very peculiar nature of text, people when they review write pros and cons and express their summary of sentiment in the star rating. And the language models we used were probably of quite general nature, not specifically trained on reviews, more specifically reviews of ecommerce products. Maybe one of these steps would help achieve some more interesting and closer results.

References

- Text pre-preprocessing (Link)
- Data Source for Amazon reviews (https://jmcauley.ucsd.edu/data/amazon/)
- Language Models and few-shot learners (https://arxiv.org/abs/2005.14165)
- BERT Pre-training paper (arXiv:1810.04805 [cs.CL])
- A BERT tutorial (Link)
- Google Notebook to Implement BERT (Link)
- IMDB Reviews data for Training BERT Model(https://ai.stanford.edu/~amaas/data/sentiment/)