Amazon Sentiment Analysis

By Srivatsa Puranam and Rushil Patel

Project Overview

With the world of Natural Language Processing (NLP) achieving more and more popularity over the last 10 years, understanding and analyzing human language has never been easier. One of the most important and impactful applications of Natural Language Processing is sentiment analysis, which opens the doors for a large range of practical applications.

One of the largest potential applications of sentiment analysis is a company's ability to notice trends in their customers' opinions on the products they sell. The analysis of customer reviews can reveal trends in customer preferences over time, recurring themes in customer feedback, etc.

Our project's objective was to develop a sentiment analysis tracker for various reviews of products that were purchased online from Amazon, the largest online store that made roughly 143.3 billion dollars in the first quarter of 2024. Initially, we planned to web scrape data directly from the Amazon website and use those reviews to generate sentiments. We started to use web scraping early on this project, but there were many challenges and errors that made it so that we could not continue to web scrape data from Amazon directly. Therefore, we had to adjust our project accordingly.

We found a dataset on Kaggle that included essential information such as the date, review text, summary, and rating for a variety of products. This dataset provided all the necessary elements to proceed with our sentiment analysis, making it a suitable alternative to our original plan.

Based on our initial proposal, we had identified two datasets that listed words commonly associated with positive and negative sentiments. Using these Kaggle datasets allowed us to analyze the reviews from the initial dataset with amazon reviews, and generate sentiment analysis.

This adjustment in our project approach allowed us to effectively carry out the sentiment analysis, fulfilling the project's goals despite the challenges faced with the original plan to scrape data from the Amazon website.

Conducting Sentiment Analysis

For the first part of the project, we had to import all the necessary libraries in order to perform the data analysis that was a major part of the project. This included importing the pandas library and setting their stop words to english words. Once this was done, the "amazon-reviews.csv" file was read as a dataframe with the columns indicating date, summary, review, and rating. The main focus for sentiment analysis was on the review, so the summary column was dropped. Any of the NA values in the review columns were later dropped from the dataframe, in order to maximize the sentiment analysis that we were trying to implement.

In order to analyze the text, we knew that when examining the file, there were many extra, random characters that had to be removed. Therefore any extra characters and stop words were deleted from the reviews so that our sentiment analysis can properly examine the words, in accordance with what he had discussed in class as well. Here is an example of the ten results from our dataframe after we had removed the stop words.

	review	rating	date
0	perfect new parents able keep track babys feed...	5.0	2013-07-16
1	book life saver helpful able go back track tre...	5.0	2013-06-29
2	helps know exactly babies day gone mother law ...	5.0	2014-03-19
3	bought times older son bought newborn super ea...	5.0	2013-08-17
4	wanted alternative printing daily log sheets n...	4.0	2014-04-01
5	great basics wish space write things bigger lo...	4.0	2014-05-10
6	3 month old son spend half days mother half ne...	5.0	2013-07-17
7	book perfect im first time new mom book made e...	5.0	2013-01-27
8	wanted love pretty expensive months worth cale...	3.0	2014-04-22
9	baby tracker brand books absolute best tracker...	5.0	2013-11-19

Next, in order to actually perform sentiment analysis, we used files that we had in place which stored which words were deemed to be considered “positive” and which words were considered as “negative.” Some examples of words in our positive words file are: 'gaining', 'enthusiastic', 'enviable', 'reassurance', 'comely'. Some examples of words in our negative file are: 'debase', 'dimmer', 'indulge', 'sever', 'insensitively'.

Now that these words are set in a “positive_words” and “negative_words” dictionary, we were able to use this to examine reviews, and determine which reviews can be deemed as having a positive or negative sentiment. The logic that we used was that if a certain word was in the positive_dictionary, then its positive score count would increase by one. If the word was in the negative_dictionary, then its negative score count would increase by one. The sentiment score would then be gained by subtracting the negative score count from the positive score count. Based on the sentiment score using this logic, if the score is greater than 0, then the sentiment is positive. If the score is less than 0, then the sentiment is negative. If the score is zero, then the sentiment is considered to be neutral.

Here is the output once the above sentiment_score and sentiment type logic is applied.

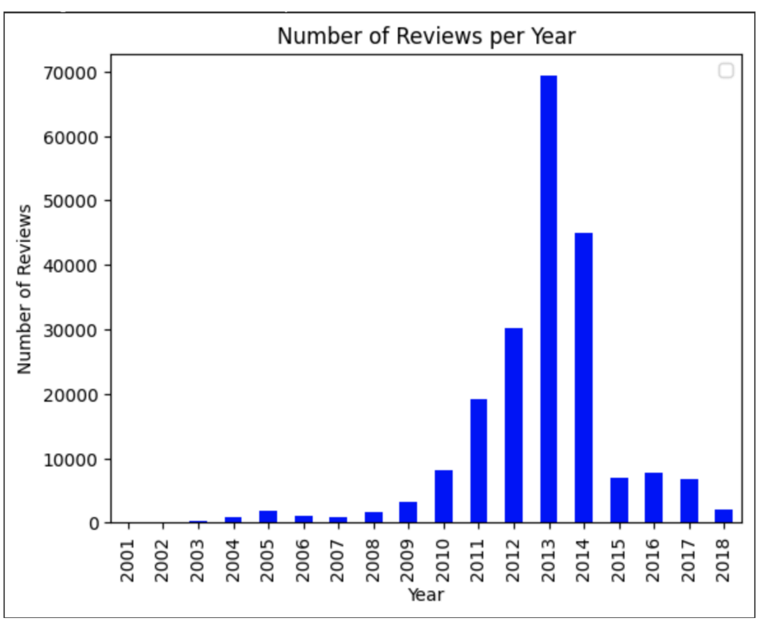
	review	rating	date	sentiment_score	sentiment
0	perfect new parents able keep track babys feed...	5	2013-07-16	3	positive
1	book life saver helpful able go back track tre...	5	2013-06-29	3	positive
2	helps know exactly babies day gone mother law ...	5	2014-03-19	2	positive
3	bought times older son bought newborn super ea...	5	2013-08-17	8	positive
4	wanted alternative printing daily log sheets n...	4	2014-04-01	3	positive
...	...	...	...	...	...
205326	great expected thank	5	2014-07-20	2	positive
205327	ive thinking trying nanoweb strings bit put hi...	5	2014-07-02	9	positive
205328	tried coated strings past including elixirs ne...	4	2014-07-22	5	positive
205329	well made elixir developed taylor guitars stri...	4	2014-07-01	9	positive
205330	strings really quite good wouldnt call perfect...	4	2014-07-16	5	positive

205251 rows × 5 columns

Results and Analysis

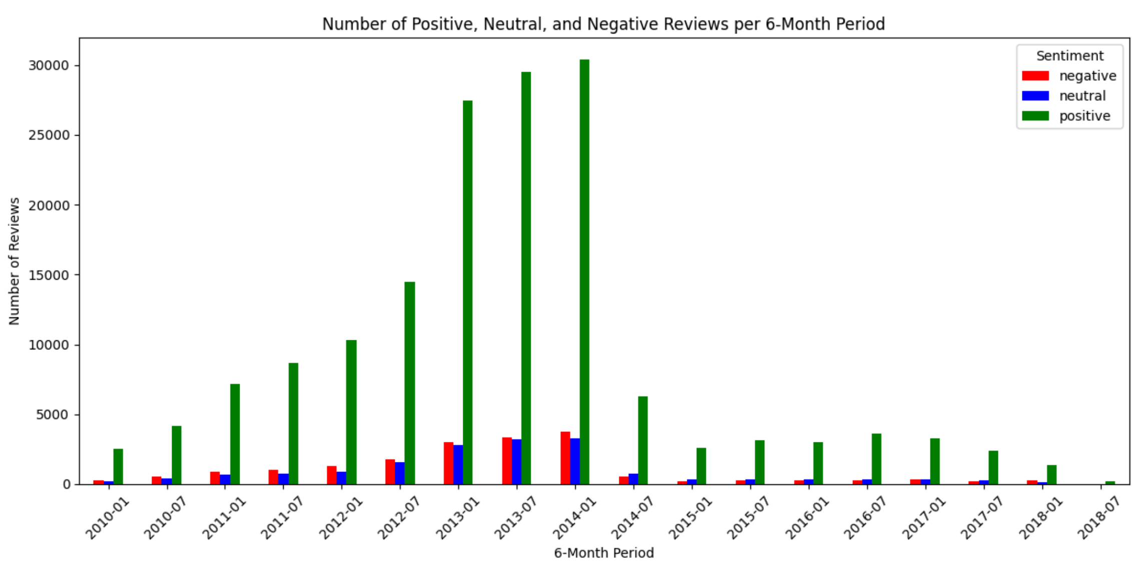
The results of the sentiment analysis proved very surprising for us, as many of our expectations were subverted. The first thing we looked at was the number of reviews that were provided in the dataset every year and the overall average number of reviews per year, as this provides a solid basis on which we can interpret the rest of the data to come. When finding and plotting the number of reviews per year and the average number of reviews per year, we obtained the following information and bar graph:

Average Number of Reviews per Year: 11402.833333333334



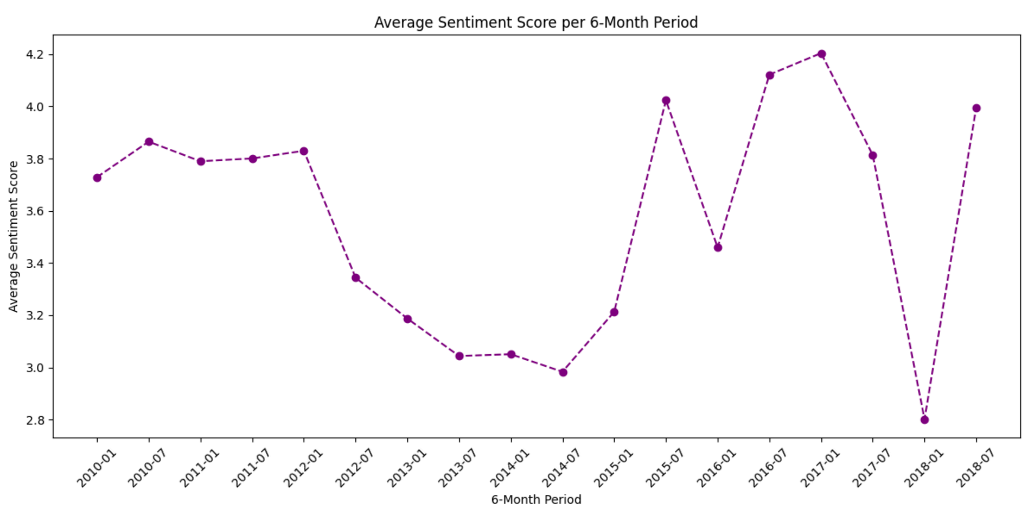
From this information it is clear that the average number of reviews per year is very much skewed due to the relatively small number of reviews in the dataset in the years before 2011. This information will affect our analysis, especially when we analyze the obtained data on a month by month basis, as the data obtained from a certain time period is more significant for current times if it is more recent and if it has more data points. Taking the information from the above bar graph into account, we decided that going forward in the analysis we would not use the years from 2001 to 2009, as those years have such a small amount of data that it would not be practical to use it to derive any significant results from our future analyses.

Having obtained this information, a natural question followed, as we wanted to recognize the relative trends in positive, neutral, and negative reviews every year. In order to do this we made another color coded bar graph to visualize the relative amounts of positive, negative, and neutral reviews, and we decided to visualize 6 months at a time so that we could also see yearly trends in the data.



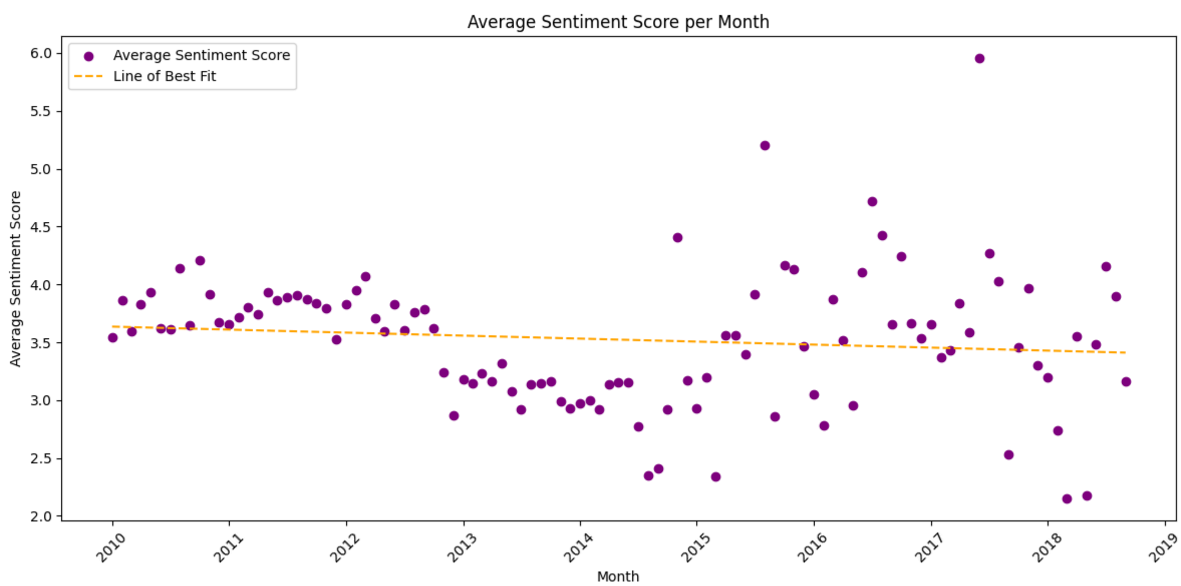
Upon analyzing this data, we recognized some extremely intriguing patterns, both in terms of the number of reviews throughout a year and the relative number of positive and negative reviews every year. The first thing that we noticed was that in most cases, there are more reviews in general in the second half of the year than in the first half of the year, and this holds true when looking individually at the positive reviews and negative reviews as well. Regarding the relative amount of positive reviews and negative reviews, we noticed something that was surprising. The amount of positive reviews in every case far exceeded the amount of negative reviews. This was surprising to us, as we expected the opposite since we thought that people would take to the internet and take the effort to write a review only when frustrated with the product. This expectation was further proven to be wrong when looking at the average sentiment score for positive reviews and negative reviews. The average sentiment score for a review that has a negative sentiment is: -2.106899207746479 and the average sentiment score for a review that has a positive sentiment is: 4.299331979012177. This means that the average positive review is roughly twice as positive as the average negative review is negative.

This trend in the average sentiment score can be seen when plotting the overall average sentiment score for every year:



In every single year in the dataset, we can see that the average sentiment score is positive with it being roughly 2.8 at its lowest.

When using sentiment analysis to take a deeper look at this data, we are also able to understand more about the change in average sentiment over time. In order to take a closer look at what we could learn, we made a scatter plot that had each point as the average sentiment score of every month. After plotting this, we plotted a line of best fit to give us a sense of how the average sentiment has changed over the years, as the average sentiment score over multiple years can give us an understanding of the nature of the people's opinions and sentiment towards the products in amazon over time. The previous bar graphs either gave us an understanding of the relative amounts of positive sentiments and negative sentiments, which did not account for the magnitude of the sentiment score in the positive and negative reviews or it just showed us the overall sentiment for every year rather than show us the overall trend. This graph on the other hand was able to give us this understanding that was previously lacking:



With this scatterplot and line of best fit, we can see that even with all of the fluctuations in terms of the number of reviews, the magnitude of the sentiment of reviews (regardless of if they are positive or negative), etc. the general sentiment of the reviews in Amazon overall, regardless of the time of year seem to be relatively the same at all times, as the average sentiment always remained slightly positive with an extremely slight negative slope. This ultimately shows that the consumer experience with the products sold on Amazon seem to be rather consistent over time.

Conclusion

By graphing various statistics such as the number of reviews per year, number of positive, negative, and neutral reviews per 6 month period, and average sentiment score per month, we were able to gain valuable insights into how sentiment trends have progressed over time. In summary, the positive comments and the negative comments can be understood by the following:

The 10 most positive reviews as well as a summary of the positive sentiment scores:

	review	rating	date	sentiment_score	sentiment	year	6_month_period	year_month	month
0	see update written spectra released s1 spectra...	4	2013-03-30	86	positive	2013	2013-01	2013-03	3
1	let start saying really like device number rea...	4	2015-07-29	81	positive	2015	2015-07	2015-07	7
2	let start saying really like device number rea...	4	2015-07-29	81	positive	2015	2015-07	2015-07	7
3	loreal dermablend professional line cosmetics ...	5	2017-07-03	73	positive	2017	2017-07	2017-07	7
4	loreal dermablend professional line cosmetics ...	5	2017-07-03	73	positive	2017	2017-07	2017-07	7
5	loreal dermablend professional line cosmetics ...	5	2017-07-03	73	positive	2017	2017-07	2017-07	7
6	loreal dermablend professional line cosmetics ...	5	2017-07-03	73	positive	2017	2017-07	2017-07	7
7	constructionthis amp certainly looks sharp nea...	5	2011-02-10	72	positive	2011	2011-01	2011-02	2
8	warning depth review want quick summary see bo...	4	2016-07-11	71	positive	2016	2016-07	2016-07	7
9	ive using bready year half love always getting...	5	2014-04-14	67	positive	2014	2014-01	2014-04	4

We can see that the review with the highest sentiment score had a score of 86, with the top 10 sentiment scores ranging from around 70-85.

```
Total number of positive sentiments:  160474
The average rating for a review that has positive sentiment is:  4.395808666824532
The standard deviation rating for a review that has positive sentiment is:  0.9610526378067489
```

The 10 most negative reviews as well as a summary of the negative sentiment scores:

	review	rating	date	sentiment_score	sentiment	year	6_month_period	year_month	month
195302	given tripp trapp chair gift sadly bought tihs...	1	2011-03-26	-17	negative	2011	2011-01	2011-03	3
195303	rating 35 4i already britax bagile stroller li...	4	2013-10-05	-17	negative	2013	2013-07	2013-10	10
195304	researched lot monitors like humidifiers doesn...	1	2012-01-06	-18	negative	2012	2012-01	2012-01	1
195305	purchased orbit infant stroller system year ha...	1	2010-08-24	-19	negative	2010	2010-07	2010-08	8
195306	one cases cheaper betteri always read reviews ...	1	2011-12-27	-19	negative	2011	2011-07	2011-12	12
195307	ten stars yes problem product gets 10 stars fi...	5	2015-12-09	-19	negative	2015	2015-07	2015-12	12
195308	bassinet overpriced ridiculous could would eit...	2	2011-01-15	-21	negative	2011	2011-01	2011-01	1
195309	update review later wanted post case anyone is...	4	2012-10-19	-21	negative	2012	2012-07	2012-10	10
195310	downgrade 1o star returning back brookmays sel...	1	2012-01-03	-23	negative	2012	2012-01	2012-01	1
195311	pros1 sensitive microphone hear baby breathing...	5	2013-12-08	-26	negative	2013	2013-07	2013-12	12

The review with the lowest sentiment score had a score of -17, with the lowest ten sentiment scores ranging from around -25 to -15.

```
Total number of negative sentiments:  18176
The average rating for a review that has negative sentiment is:  3.2994608274647885
The standard deviation rating for a review that has negative sentiment is:  1.4976374740361458
```

Based on this data, we can see that it closely matches what is expected in reality because the average rating for a positive sentiment is approximately around 4.3958, which is greater than the average rating for a negative sentiment which is approximately around 3.299946. Therefore, there is a correlation between a higher rating and a higher sentiment score, which makes sense as if a user is happy with a certain product, they would write positively about it while also ensuring they give a higher review rating. It is also noteworthy that the standard deviation is less for positive sentiments due to the fact that there are a greater number of positive sentiments. It is known in statistics that the standard deviation decreases as n (the number of samples) increases.

From this project, we were able to learn a variety of concepts that can be further applied and used as we pursue our Computer Science major with an interest in data science. We were able to learn how to use various libraries in order to preprocess data, analyze data, and even learned how to visually showcase our data through graphs and charts. This allowed us to perform the appropriate analyses and make varying conclusions as mentioned above.