Amazon Reviews Sentiment Analysis

Milestone: Project report

Group 1

Nirmit Sachde

Ashmita Pal

857-492-8683

857-654-7663

sachde.n@northeastern.edu

pal.ashm@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Nirmit Sachde

Signature of Student 2: Ashmita Pal

Submission Date: 03/15/2024

Problem Setting

In the era of digital commerce, comprehending consumer sentiments and behaviors holds utmost importance for businesses. Platforms like Amazon provide vast datasets of consumer reviews, offering insights into customer preferences and attitudes. Yet, the challenge lies in efficiently analyzing and interpreting this data to drive informed decision-making.

Problem Definition

Sentiment Analysis: The objective is to assess consumer sentiments towards various products based on the provided reviews. By leveraging natural language processing (NLP) techniques, the aim is to categorize sentiments as positive, negative, or neutral, enabling businesses to gauge product perception accurately.

Data Sources

https://archive.org/details/amazon_reviews_dump

Data Description

The dataset available to address the problems being addressed consists of Amazon reviews data. Here's a detailed description:

Dataset Name: Amazon Reviews Dataset

Number of Columns: 15

Number of Rows: The dataset contains millions of rows for 12 categories of products, we have chosen around 1 lakh random rows of each category. So the final size of the dataset that we use is 1.1 million+ rows.

Sample of Variable Names:

marketplace: The marketplace where the review was posted (e.g., US).

customer_id: Unique identifier for the customer who posted the review.

review_id: Unique identifier for the review.

product_id: Unique identifier for the product being reviewed.

product_parent: Identifier for the parent product (if applicable).

product_title: Title of the product being reviewed.

product_category: Category of the product (e.g., Apparel).
star_rating: The rating given by the customer (on a scale of 1 to 5).
helpful_votes: Number of helpful votes the review received.
total_votes: Total number of votes the review received (helpful and unhelpful combined).
vine: Indicates if the review was part of the Amazon Vine program (Y/N).
verified_purchase: Indicates if the purchase was verified (Y/N).
review_headline: Headline of the review.
review_body: The main text/content of the review.
review_date: Date when the review was posted.
This dataset provides comprehensive information about customer reviews on Amazon, including details about the products, customer feedback, ratings, and other relevant attributes. It can be leveraged for various analyses and applications in understanding customer sentiment, product performance, and market trends.

Project Planning

- Key Milestones
    - Data Preprocessing: Clean and prepare the dataset for subsequent analysis.
    - Sentiment Analysis: Utilize NLP techniques to analyze sentiment polarity in the reviews.
    - Model Evaluation: Assess the performance of sentiment analysis and customer segmentation models

- Challenges
    - Data Preprocessing: Addressing challenges such as missing values, text normalization on a large dataset.
    - Sentiment Analysis: Handling nuances in language and context to accurately infer sentiment from reviews.
    - Customer Segmentation: Determining the optimal number of clusters to ensure meaningful segmentation and actionable insights.

Data Mining Tasks: The primary objective is to prepare the textual data for natural language processing (NLP) tasks. Various preprocessing functions are defined to handle tasks like lowercasing, stripping whitespaces, removing special characters, punctuation, HTML tags, and unescaping HTML entities. These functions ensure that the text data is clean and standardized before further analysis. Additionally, stopwords removal and lemmatization are integrated into the pipeline using spaCy, a powerful NLP library. The core function, run_batch_text_normalization_pipeline, orchestrates the entire process, applying the predefined preprocessing steps to batches of text efficiently. The resulting processed data, stored in a DataFrame named amazon_reviews, is now ready for advanced analytical tasks such as sentiment analysis or topic modeling.

Data Mining Models/Methods:

Vader Analysis

We conducted sentiment analysis on Amazon reviews using the VADER sentiment analysis tool. Since our data is unlabelled, we assumed that VADER gives us 100% accuracy, and we labeled the sentiment of each review using VADER. Initially, functions were defined to calculate the compound sentiment score and assign sentiment labels based on predefined thresholds. Then, a function was implemented to compute sentiment scores and labels in batches, leveraging the parallel processing capabilities of tqdm. This function iterated over the reviews, calculated their sentiment scores, assigned labels, and aggregated the results into lists. Finally, the sentiment scores and labels were added to the DataFrame, associating each score and label with its corresponding review. This approach facilitated efficient sentiment analysis of large datasets and provided valuable insights into the sentiment distribution of Amazon reviews.

In addition to the sentiment analysis conducted using the VADER sentiment analysis tool, an alternative approach utilizing the NLTK library was explored. NLTK was imported, and the SentimentIntensityAnalyzer from nltk.sentiment was employed for sentiment analysis and a new column named 'nltk_sentiment' was added to the Amazon reviews DataFrame to store sentiment scores calculated using NLTK's polarity_scores method applied to each review text. However, since the results obtained from NLTK and VADER were found to be identical, VADER was chosen for sentiment analysis due to its simplicity and efficiency.

product_category	star_rating	helpful_votes	review_body	normalized_review_body	review_body_compound_score	review_body_sentiment_label	nltk_sentiment
Electronics	4.0	6.0	I didn't want to spend a lot of money for a sm...	nt want spend lot money small portable radio s...	0.3549	positive	0.3549
Electronics	5.0	0.0	Solid earbuds. The mic is fine for calls. Th...	solid earbuds mic fine calls cord sturdier...	0.3400	positive	0.3400
Electronics	5.0	0.0	Great	great	0.6249	positive	0.6249
Electronics	5.0	0.0	Heavy gauge on the cable makes it feel like a ...	heavy gauge cable makes feel like heavy duty a...	0.7717	positive	0.7717
Electronics	1.0	0.0	Bought this for regular cable and only receive...	bought regular cable received channel set 5 ft...	0.6908	positive	0.6908

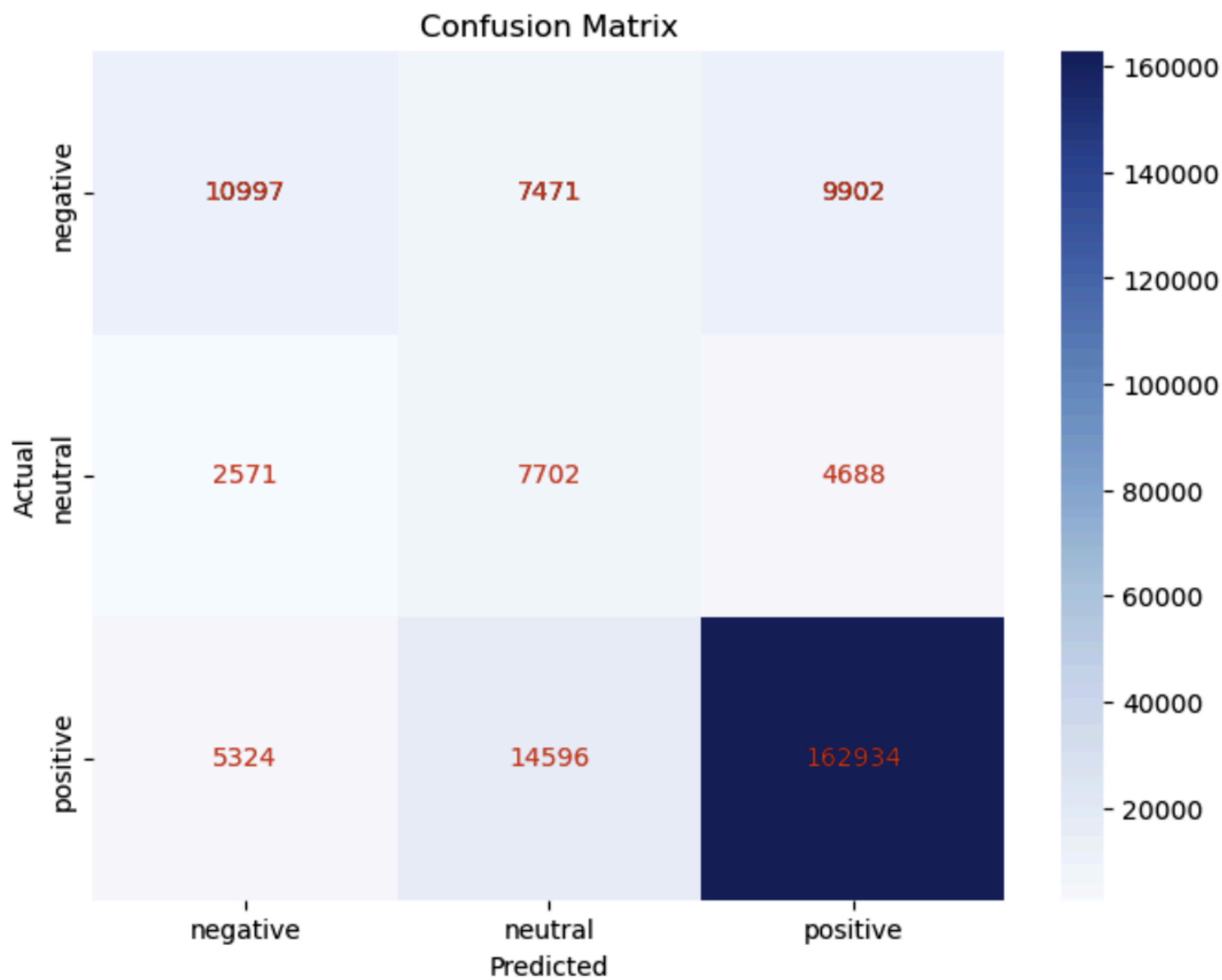
TextBlob

For sentiment analysis using TextBlob, a widely used natural language processing library, we applied it to the Amazon reviews dataset. Initially, we imported the TextBlob module. Subsequently, a new column titled 'textblob_polarity' was appended to the amazon_reviews DataFrame to hold the polarity scores computed through TextBlob's sentiment analysis. This process involved employing a lambda function to process each review text, generating a TextBlob object for analysis, and retrieving its sentiment polarity attribute via the sentiment.polarity method.

To categorize polarity scores obtained from TextBlob sentiment analysis into sentiment labels, we implemented a function named map_to_sentiment. This function accepted a polarity score as input and assigned a sentiment label based on predefined thresholds: scores greater than 0.05 were labeled as 'positive', scores less than -0.05 were labeled as 'negative', and scores between -0.05 and 0.05 were considered 'neutral'. This operation resulted in the creation of a new column titled 'textblob_sentiment', where each polarity score was mapped to its corresponding sentiment label based on the defined thresholds.

Performance Evaluation

```
Accuracy: 80.30%
```



```
Classification Report for TextBlob Sentiment Analysis:
              precision    recall  f1-score   support

    negative       0.58      0.39      0.47     28370
     neutral       0.26      0.51      0.34     14961
    positive       0.92      0.89      0.90    182854

    accuracy                           0.80    226185
   macro avg       0.59      0.60      0.57    226185
weighted avg       0.83      0.80      0.81    226185
```

Some misclassified rows:

I have had the Body Back Buddy for a while now and really like it. But what really motivated my review was that over the weekend I really wrenched up my back playing sports and using the Body Back Buddy really helped loosen up my overly tight muscles. This morning I could hardly tie my shoes, now I am feeling so much better. I have foam rollers, lacrosse balls and stretch bands and nothing worked as well as this device (all useful though). The only downside is it is a bit big, so travelling with it is difficult, but it is the size that allows you to get the leverage you need to work on your back.
Actual: positive **Predicted:** neutral

These TF glasses are awesome, truly awesome. They are stylish enough to be trendy but subtle enough to remain classy. Unfortunately, mine arrived with a cracked demo lens. I know it's only the demo lens but the crack it where the nose piece touches the frame rim. I don't know whether this is a defect in this specific pair, a known problem with this design, or simply a fluke due to handling & shipping. Regardless, I decided not to take a chance on getting my RX filled because of the damaged demo lens.

Pity.
Actual: negative **Predicted:** positive

My Dad got this last year and it was very nice.
He had lot of hearing issues in his ear and this one worked great.

Pros:
NIce reception and no voice issues.

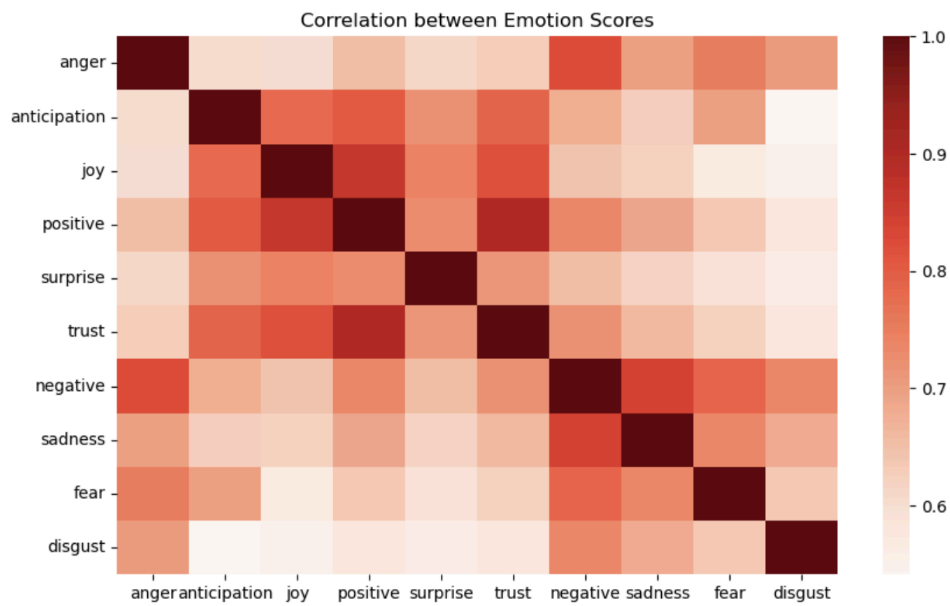
Cons:
Size is too big, for Indians who has small ear, it does not sit correctly.
Actual: positive **Predicted:** neutral

bought as a shower gift, it was well made and the colors were as expected.
Actual: positive **Predicted:** negative

This watch looks nicer on the web than in person i would not recommend anyone buying this watch unless you have money to waste it's made of plastic like material
Actual: positive **Predicted:** negative

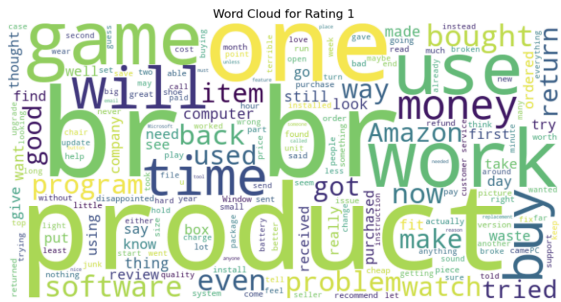
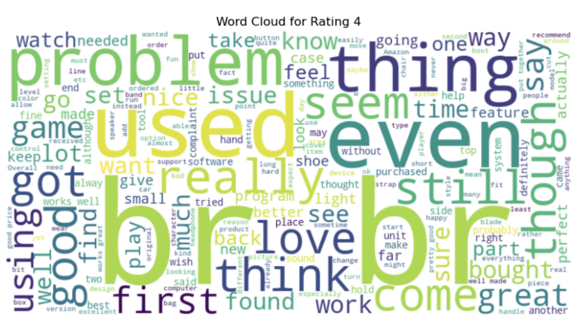
NRCLex

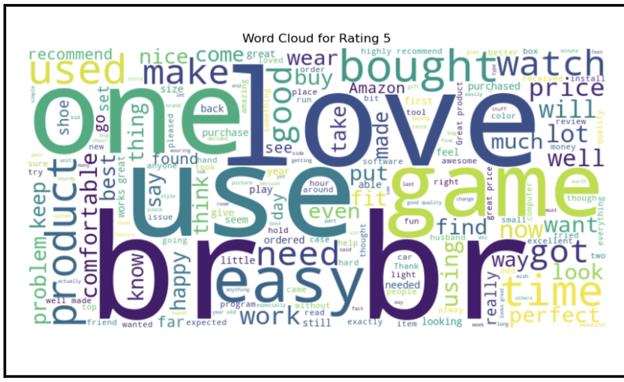
We utilize the NRCLex library to conduct emotion analysis on the Amazon reviews dataset. We define a function named get_emotion_scores(text) to compute emotion scores for each review by initializing an NRCLex object with the review text and extracting the raw emotion scores. Subsequently, we apply this function to each review text in the 'review_body' column of the amazon_reviews DataFrame, storing the computed emotion scores in a new column named 'emotion'. This process enhances the dataset by adding insights into the emotional content of the Amazon reviews, facilitating deeper analysis and understanding of customer sentiments.

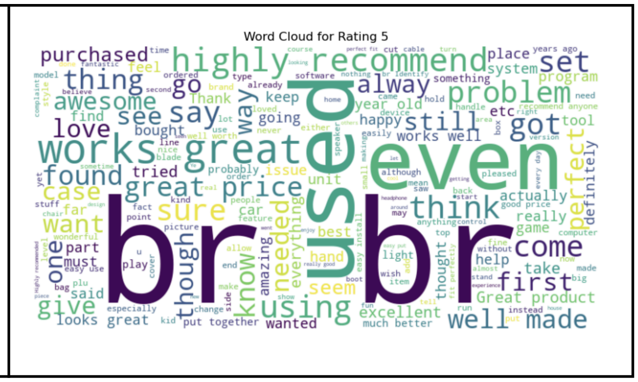


Our visualization depicts the correlation among emotion scores derived from Amazon reviews. The heatmap reveals insights into the connections among different emotions, with deeper shades of red indicating stronger correlations. This visualization enhances our comprehension of emotion relationships within the dataset, enabling more profound analysis of customer sentiments. Notably, we observe a significant correlation between anger and negative sentiments, or negative sentiments and sadness, affirming the accuracy of our analysis conducted using NRCLex.

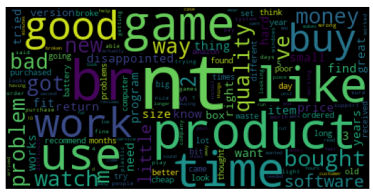
Wordcloud analysis

Before removing Additional stopwords	After removing Additional stopwords
	
	
	
	





Word clouds of the normalized review bodies belonging to each sentiment label category



Positive keywords: good, great, perfect, love, quality

Negative keywords: disappointed, problem, time, money

Neutral keywords: works, expected, bought, described

We create word clouds to visually illustrate the most frequently occurring words within positive, negative, and neutral reviews extracted from the Amazon reviews dataset. First, we aggregate the review texts corresponding to each sentiment category ('positive', 'negative', 'neutral') into separate strings: positive_reviews, negative_reviews, and neutral_reviews, respectively.

Unique products and product categories in the dataset¶

Total number of unique products in dataset: 463528

Total number of unique product categories in dataset: 12

1. Electronics
2. Automotive
3. Tools
4. Shoes
5. Video Games
6. Grocery
7. Personal_Care_Appliances
8. Sports
9. Apparel
10. Software
11. Furniture
12. Watches

Aggregation statistics via grouping by product category

	star_rating	review_body_compound_score	helpful_votes
product_category			
Apparel	4.106968	0.569039	0.947642
Automotive	4.247274	0.484083	0.988600
Electronics	4.038221	0.500570	1.844653
Furniture	4.084768	0.556738	2.507663
Grocery	4.313400	0.581397	1.593695
Personal_Care_Appliances	3.977236	0.466586	3.351520
Shoes	4.240021	0.602486	0.879789
Software	3.563895	0.447955	4.332179
Sports	4.228284	0.532467	1.389989
Tools	4.266768	0.485176	1.812968
Video Games	4.056032	0.547894	2.234979
Watches	4.136516	0.557433	1.183937

The dataset displays a notable positive bias across most product categories, evident from consistently high mean positive sentiment compound scores and predominantly high star ratings, mostly exceeding 4.0. The only exception to this trend is the personal care appliances product category group, which exhibits a slightly lower mean star rating of 3.97. Regarding the mean number of helpful votes, it appears to be relatively low across the board, indicating limited usefulness for analysis purposes.

Aggregation statistics via grouping by product category and the sentiment label itself

product_category	review_body_sentiment_label	star_rating	review_body_compound_score	helpful_votes
Apparel	negative	2.290520	-0.439895	1.267812
	neutral	3.241231	0.000111	0.387528
	positive	4.375392	0.726091	0.964739
Automotive	negative	2.733950	-0.450599	1.379504
	neutral	3.845878	0.000033	0.479689
	positive	4.525374	0.682180	0.985078
Electronics	negative	2.385250	-0.492692	2.014030
	neutral	3.528069	-0.000016	0.626418
	positive	4.380706	0.724458	1.928011
Furniture	negative	2.223412	-0.508465	2.859008
	neutral	3.356875	0.000018	1.430810
	positive	4.410084	0.750995	2.516327
Grocery	negative	2.502551	-0.476221	2.688886
	neutral	3.817118	0.000081	0.813405
	positive	4.549133	0.738553	1.523326
Personal_Care_Appliances	negative	2.524866	-0.502016	3.475134
	neutral	3.446792	0.000334	1.341329
	positive	4.334238	0.713589	3.483996
Shoes	negative	2.537019	-0.466932	1.227629
	neutral	3.509752	0.000135	0.436830
	positive	4.462434	0.752236	0.875886

This DataFrame presents a clearer picture of how sentiment scores and labels correlate with the star rating of reviews. Referring to the Vader documentation for compound scores, where scores greater than or equal to 0.05 indicate positive sentiment, scores between -0.05 and 0.05 indicate neutral sentiment, and scores less than or equal to -0.05 indicate negative sentiment, we observe that the mean compound scores align logically with these categories. Negative sentiment groups have mean star ratings around 2.5, neutral sentiment groups have mean star ratings around 3.5, and positive sentiment groups have mean star ratings around 4.5. This alignment suggests that our sentiment analysis accurately reflects the expected ratings associated with different sentiment categories, validating the reliability of our sentiment generation process. The correlation between a review's sentiment label/score and the final star rating assigned by customers for products listed on their online store page is evident.

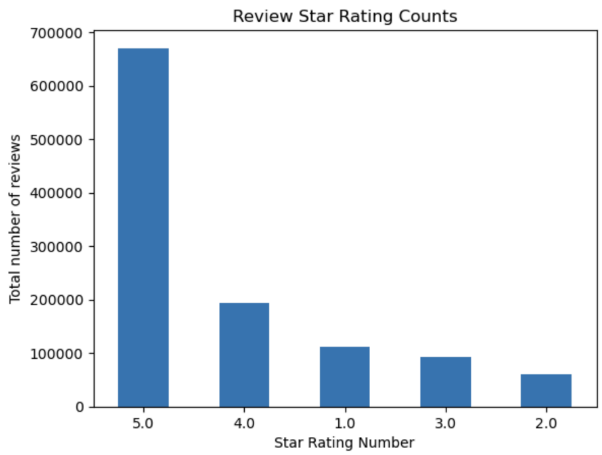
Find the products with both the highest star rating and sentiment compound scores¶

			star_rating	review_body_compound_score
product_title	product_category	review_body_sentiment_label		
Corel PaintShop Pro X4 Ultimate [Old Version]	Software	positive	5.0	1.0000
Corel Photo and Video Pro X4 Ultimate Bundle [Old Version]	Software	positive	5.0	1.0000
Halo: The Master Chief Collection	Video Games	positive	5.0	1.0000
Numi Organic Herbal Tea, Tea Bags	Grocery	positive	5.0	1.0000
PaintShop Pro X5	Software	positive	5.0	1.0000
...	...	...	...	...
Lotus Domino R5 System Administration Curriculum CBT Training CDs	Software	positive	5.0	0.9998
Madden NFL 13	Video Games	positive	5.0	0.9998
Marantz AV8801 11.2 Channel Home Theater Pre-Amplifier/Processor (Discontinued by Manufacturer)	Electronics	positive	5.0	0.9998
Mass Effect 2	Video Games	positive	5.0	0.9998
Math Missions: The Race to Spectacle City Arcade Grades K-2 [OLD VERSION]	Software	positive	5.0	0.9998

100 rows × 2 columns

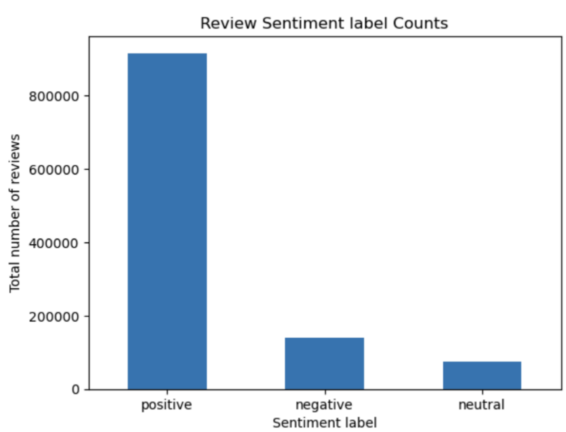
Based on the top 100 rows sorted by star rating and compound score, it's evident that electronics, video games, and software emerge as dominant product categories, receiving high ratings and praise from consumers.

Bar graph of the total number of reviews per star rating category (1.0-5.0, as whole numbers)



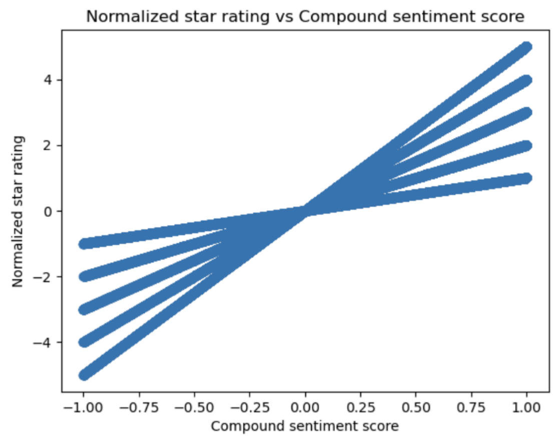
Our previous calculations reveal that the majority of the dataset consists of reviews with a star rating of 4 or higher. This prevalence of high star ratings contributes to the bias towards positive sentiments observed in the dataset.

Bar graph of total number of reviews for each sentiment label category



Once more, the dataset confirms a positive bias in the reviews, where the star rating is closely linked to both the compound score and its corresponding sentiment label. This reaffirms the connection between the perceived sentiment of reviews and the ratings they receive, highlighting the influence of sentiment on overall ratings.

Is there a linear relationship between star rating and the compounding sentiment score?



To effectively analyze the relationship between star ratings and compound sentiment scores, we initially consider that star ratings are discrete variables ranging from 1.0 to 5.0, with whole numbers only. Plotting them directly against compound scores in a scatterplot would be naive and incorrect due to the discreteness of star ratings. To address this limitation, we propose normalizing or transforming the star ratings into a continuous variable. One approach is to calculate the product of the star rating and the compound score, resulting in a new series of floating-point numbers ranging from -5 to 5. This normalized star rating can then be interpreted, with values closer to -5 indicating highly negative sentiments and those closer to 5 representing highly positive sentiments.

Summary

Data Preprocessing: The dataset was cleaned and prepared for analysis, addressing challenges such as missing values, text normalization, and removal of special characters and HTML tags.

Sentiment Analysis Models:

1. VADER Analysis: Utilized VADER sentiment analysis tool due to its simplicity and efficiency. Provided accurate sentiment labels for reviews.
2. TextBlob: Applied TextBlob for sentiment analysis, but results were identical to VADER. TextBlob's polarity scores were categorized into positive, negative, and neutral sentiments.
3. NRCLex: Conducted emotion analysis using NRCLex library, providing insights into emotional content of reviews.

Performance Evaluation: Identified misclassified reviews and evaluated the performance of sentiment analysis methods. Despite some misclassifications, overall sentiment analysis yielded accurate results.

In addition to the sentiment analysis models mentioned, a train-test split was applied to the TextBlob model for sentiment analysis. This step involved dividing the dataset into a training set, used to train the TextBlob model, and a test set, used to evaluate its performance. Despite this additional step, the results obtained from TextBlob were found to be identical to those from VADER, with both models effectively categorizing the sentiment polarity of reviews into positive, negative, and neutral categories.

Visualization and Analysis:

1. Generated word clouds to visualize frequently occurring words in positive, negative, and neutral reviews.
2. Correlation heatmaps depicted relationships among different emotions, affirming the accuracy of emotion analysis.

3. Analyzed sentiment distribution, emotion scores, and the relationship between star ratings and sentiment scores through bar graphs and scatter plots.

Impact of Project Outcome:

Provided business insights such as customer feedback analysis, trend identification, enhanced customer experience, operational improvements, and strategic planning derived from sentiment analysis.

Conclusion

In conclusion, the sentiment analysis of Amazon reviews offers valuable insights into consumer perceptions, preferences, and attitudes towards various products. Through meticulous data preprocessing and the application of advanced NLP techniques, we were able to categorize sentiments as positive, negative, or neutral accurately.

Despite encountering challenges such as misclassified reviews, the overall performance of sentiment analysis models, particularly VADER, proved to be effective in capturing the sentiment polarity of reviews. Additionally, the exploration of emotion analysis using NRCLex provided deeper insights into the emotional content of customer feedback.

This project underscores the significance of sentiment analysis as a powerful tool for extracting insights from large volumes of customer feedback data. By understanding and responding to consumer sentiments effectively, businesses can drive innovation, foster customer loyalty, and achieve sustained growth in today's dynamic marketplace.