Sentiment Analysis of Amazon Products Reviews

Capstone 1 Milestone Report

Outline

- ❏ Problem statement and clients
- ❏ Description of dataset
  - Methods for data cleaning and wrangling
- ❏ Exploratory data analysis
  - Data storytelling
  - Summary of findings
  - Inferential statistical data analysis
- ❏ Conclusion

Problem Statement and Clients

E-commerce has revolutionized today's shopping experience. Online customer reviews and ratings of every product/brand/service help the consumers to make a smart and informed decision before buying a product or service. Customer reviews are also beneficial for manufacturers to improve their products and services. Retail websites like Amazon.com allow users to submit their opinion in the form of numerical stars (from 1-5) or comments about a product. These star ratings provide excellent labels for training ML models to predict the sentiment of text related to a particular product or line of products. This allows sellers to get an overview of sentiment for a large number of unlabelled reviews and in some cases can and help sellers understand how customers are feeling about their product. In my capstone 1 project, I propose building an ML model to evaluate the positive and negative sentiment of an Amazon.com product review.

Sentiment analysis of product reviews has diverse applications. E-commerce companies such as Amazon.com, eBay.com, etc and technology companies (Apple, Microsoft, etc) can use it to predict what people think about their product or market trend. Social media companies can also be used to study the sentiment of social conversations.

Description of Dataset

For this project, I will use the baby products, grocery_foods and call phones reviews datasets from Amazon.com. The dataset is made available by Julian McAuley, UCSD and is available by request. The dataset contains information about reviewer ID, asin (product ID), reviewer name, helpful (helpfulness rating of the review), review text, overall (rating of the product), summary (summary of the review), review time. For this project, the mostly used features are review text and rating. The datasets were downloaded as the JSON file and loaded into Pandas DataFrame in the jupyter notebook.

Amazon Review Data Link: http://jmcauley.ucsd.edu/data/amazon/

Methods for data cleaning and wrangling

After I load the datasets into Pandas DataFrame I did the following steps to clean and wrangle the datasets:

- df.shape and df.columns attribute to find out the number of rows and columns, and column names respectively
- Renamed some of the columns using df.rename() function.
- Counted the missing values using isnull() and info() methods
- Filled the missing values of the reviewer's name with 'No Name' using fillna() method
- Checked for duplicate reviews using duplicated() method and drop the duplicate values using drop_duplicates() method.

I pre-processed the review text using the following methods

- Removed short reviews (less than 3 words)
- Removed non-English reviews using langdetect and nltk libraries
- Removed punctuations from the review text
- Removed stopwords from the review text using nltk stopwords module
- Tokenization (converting a sentence into a list of words) of review text using word_tokenize from nltk.tokenize
- Made charts for distributions of ratings and review length/% caps.

Exploratory data analysis

In this exploration of the Amazon product reviews dataset, I will analyze the features of review text with respect to ratings. For this project, I will use three product categories: baby products, grocery and gourmet foods, and cellphones dataset.

These are the questions I will look in the data:

- What does the distribution of data based on review categories (or star ratings) and distribution of review length look like?
- What does the distribution of the number of ratings in different product categories look like?
- How do the reviews change across high and low star ratings based on review length and percentage of uppercase words per review for different product categories?
- What are the most frequently used words for high and low ratings?

Data storytelling

I have plotted the distribution of ratings for the three different product categories and noticed that most of the reviews are five stars in all three categories (Fig.1). I also see that cell phones tend to have more ratings of 1 than the other categories.

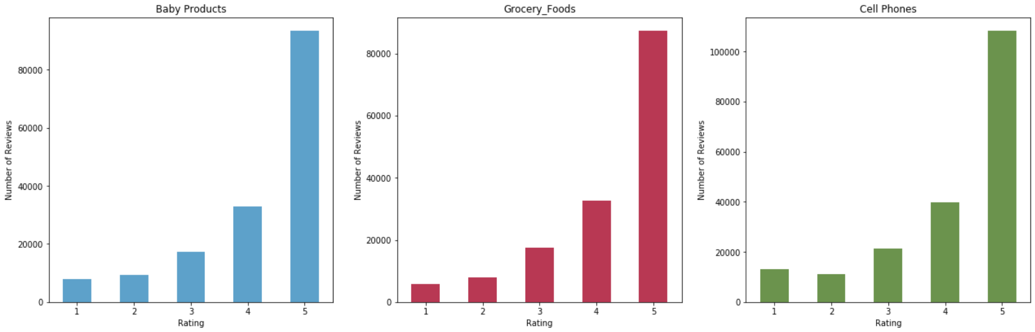


Fig 1. Distribution of rating by product categories

From the distribution of review length by product categories (Fig. 2), we can see the cell phone reviews are shorter in length compared to the other two product categories. About 74.8% of cell phone reviews are of length less than equal to 100, whereas baby products and grocery food categories have about 56% and 53% of reviews of length less than or equal to 100 respectively.

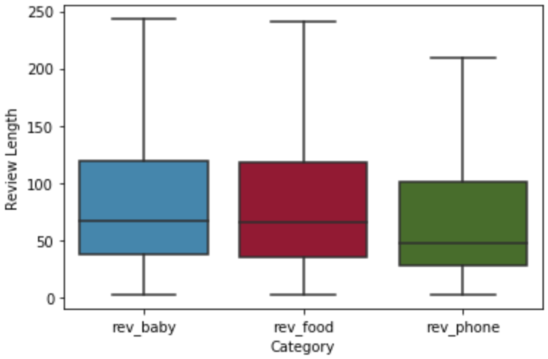


Fig 2. Distribution of review length by product categories

In all of the three product categories, five star rated reviews are found to be the shortest length followed by one star rated reviews (Fig.3). Two, three and four star rated reviews are longer and show different trends in different products categories. In the case of baby products and cell phones categories four star rated reviews are longer and in grocery_foods three star rated reviews are longer. From this exploration we can see that, when people are highly satisfied or dissatisfied with the product, they write very short reviews and, in other cases, they have to explain where and which part of the product they are unsatisfied with hence the reviews are longer. As we can see an example of a short 5 star review below:

Example of 5 star review with less than 5 words:
'Love this product'
'Great little headset.'
'Nice Case. Great Price.'

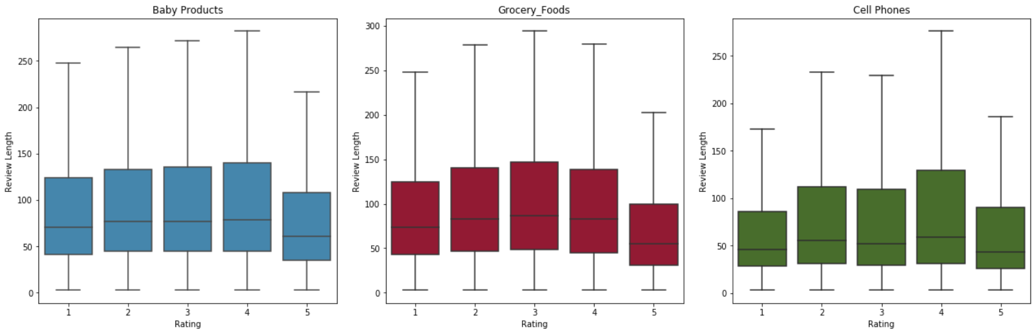


Fig 3. Distribution of Review Length by Rating

The distribution of % capitals by ratings (Fig 4.) shows that the reviews corresponding to 1-star and 5-star ratings have more uppercase letters and the three star rated reviews have least upper case letters. It might be due to highly satisfied or dissatisfied customers writing reviews of short sentences which contain more uppercase letters.

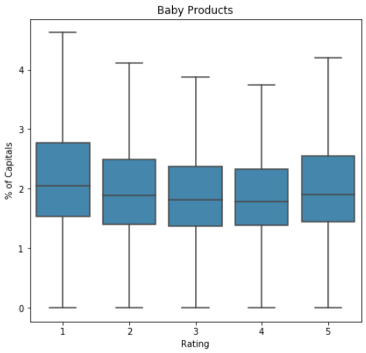


Fig 4. Distribution of % of Capitals by Rating

The % of capitals show bimodal distribution in the three product categories (Fig. 5), one peak centered at around 20% and the other peak centered at around 70% of capitals.

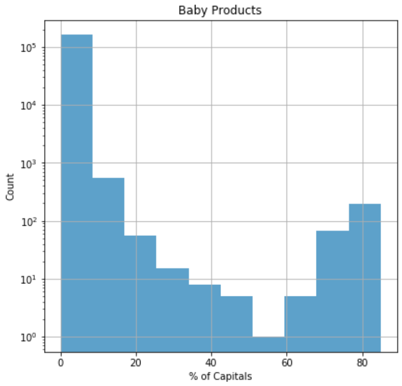


Fig 5. Distribution of % of Capitals

In fig.6 we have plotted the average number of reviews >= 60% capital letters grouped by rating. 1 and 5 star reviews are more likely to contain reviews with >= 60% capital letters across product categories then other star ratings.

Example of a reviews with >= 60% capitals:
'AMAZING SOUND, MOVIE THEATER QUALITY' (5-star)
"COULDN'T DOWNLOAD ANYTHING" (1-star)

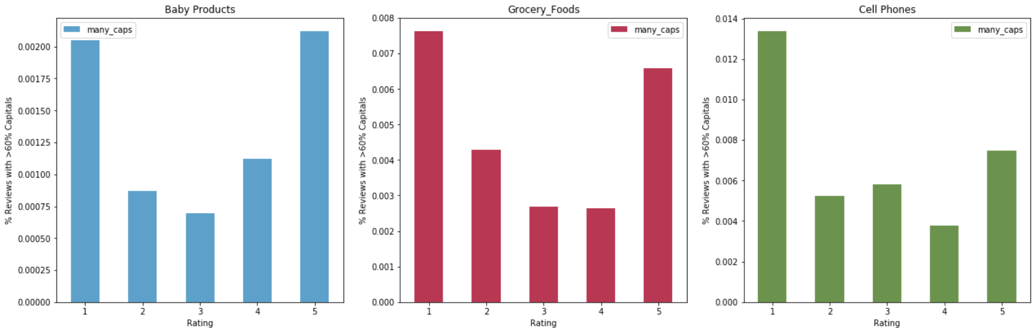
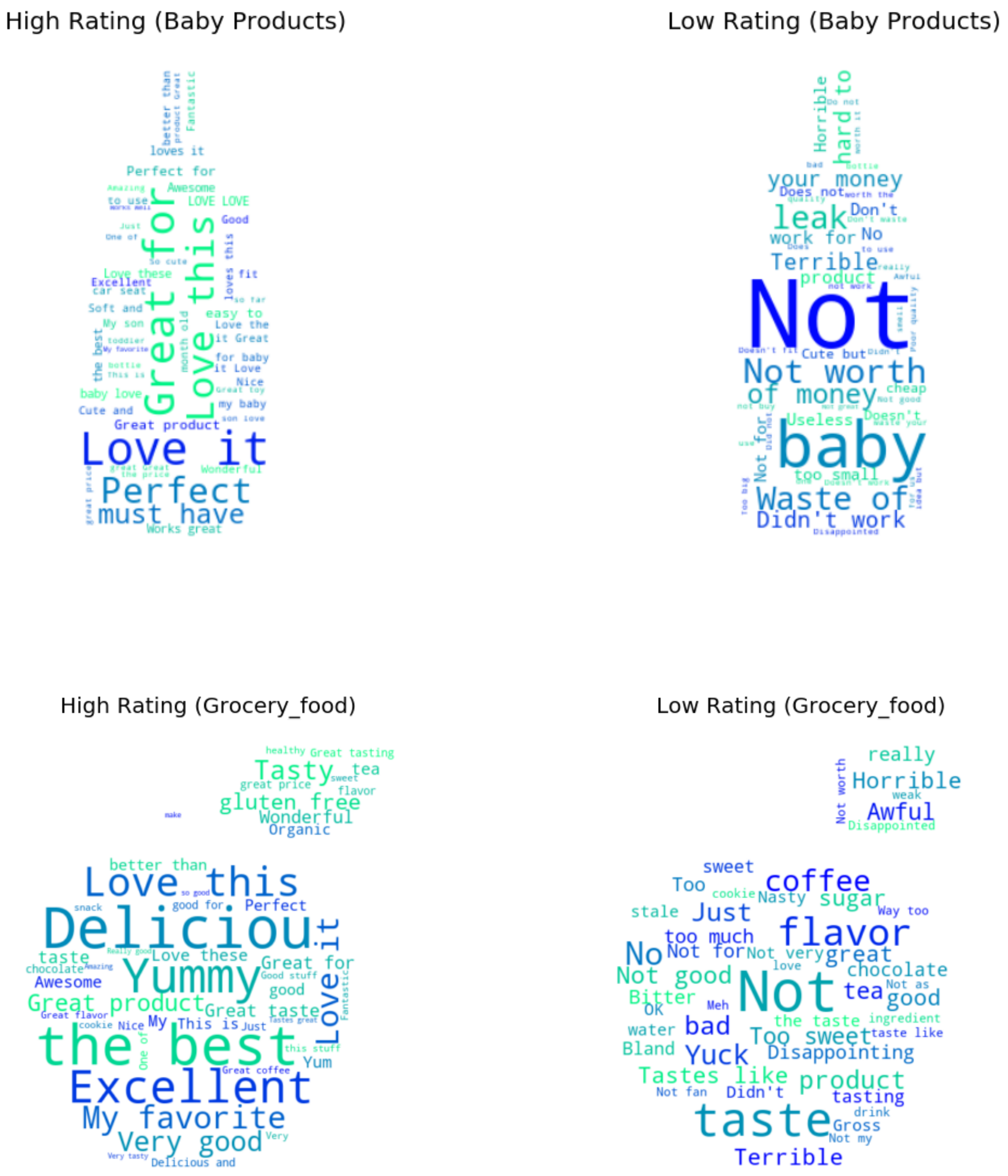


Fig.6 Distribution of % of capitals (> 60%) by rating

The word clouds from review summaries are plotted by high and low ratings. In this case I have taken 5 stars ratings as high and 1- 2 stars as low ratings. In the word clouds the font size of a word indicates its frequency and importance in the review. The words used more often in the reviews are greater in font size and darker in color. The size of the word decreases as the frequency of the word in the reviews decreases. As we can see from the word cloud of high rated grocery_foods summary, the words like ‘Delicious’, ‘Yummy’, ‘Testy’, ‘Excellent’, ‘the best’, ‘Love this’ looks bigger and darker as they were used more frequently used in the reviews. On the other hand in the word cloud of low rated reviews the words like ‘Not good’, ‘Terrible taste’, ‘Disappointing’, ‘Yuck’ are the most used words.


High Rating (Baby Products)

Low Rating (Baby Products)

High Rating (Grocery_food)

Low Rating (Grocery_food)

High Rating Summary (phone)

This is
great price
Works great
great Great
the best
it Great
Very good
the price
case for
Love this
Excellent
screen protector
case Love
Perfect
Great case
easy to
Awesome
Love it
Nice case
Great for
Nice
Great product
this case
Very nice
Amazing
Five Star

Low Rating Summary (phone)

Doesn't work
Does not
case
Not good
Disappointed
not work
product
Horrible
Cheap
Terrible
phone
Do not buy
not fit
Not worth
Don't
bad
Junk
battery

Summary of findings from Exploratory Data Analysis

From exploratory data analysis we found in all of the three product categories:

- Most of the reviews are 5 star
- Five and one star rated reviews are found to be shorter
- Reviews corresponding to 1-star and 5-star ratings have more uppercase letters and three star rated reviews have least upper case letters.
- The % of capitals show bimodal distribution, one peak centered at around 20% and the other peak centered at around 70% of capitals.
- 1 and 5 star reviews are more likely to contain >= 60% capital letters across product categories than other star ratings.

Inferential statistical data analysis

For the statistical data analysis all the ratings are divided into two groups which are 5-star and not_5 star (1, 2, 3, and 4 star rating). The reviews associated with 5-star rating are compared with lower rated reviews (not_5 star) based on the word counts and percent of uppercase letters present per review. Statistical comparison of these features between 5-star and other lower rated reviews are described below.

1. *Comparison of review length between 5-star and not_5 star rating*

To compare the review length (measured by word count) for 5-star and not5-star rating, an independent two-sample t-test for unequal variance is used. Below are the hypotheses of interest.

Null Hypothesis: The review length for 5-star and other rated reviews are the same.
Alternate Hypothesis: The review length for 5-star and other rated reviews are different.

Baby Products: t = 41.398 , p = 0.000
Grocery_Foods: t = 58.596, p = 0.000
Cell Phones: t = 23.899, p = 0.000

In all the three product categories, since p_value is less than 0.05 we reject the null hypothesis and conclude that the word counts for 5-star rated reviews are significantly shorter than all 1-star through 4-star (not5) rated reviews.

2. *Comparison of %caps between 5-star and not_5 star rating*

An independent two-sample t-test for unequal variance is used to compare the % caps between 5-star and other lower rated reviews.

Null Hypothesis: The % of capitals for 5-star rated reviews are the same as other rated reviews.
Alternate Hypothesis: The % of capitals for 5-star rated reviews are different from the other rated reviews.

Baby Products: t = 12.134, p = 0.000
Grocery_Foods: t = 13.586, p = 0.000
Cell Phones: t = 12.302, p = 0.000

In all of the three product categories, since p_value is less than 0.05 we reject the null hypothesis and conclude that the % of capitals for 5-star rated reviews are higher than the other rated reviews.

Conclusion

Review length for 5-star rated reviews are shorter than other rated reviews and the % of capitals for 5-star rated reviews are higher than the other rated reviews.