Sentiment Analysis: Amazon Reviews

by Ketaki Nagarkar

Introduction:

In recent years I have increasingly relied on e-commerce websites such as Amazon.com and Wayfair for shopping. I have bought everything from cell phone screen protectors to kayaks online. Reviews on websites are a good way of evaluating a product. Additionally, social media platforms such as facebook and twitter have also been a great source for getting product opinion and feedback. According to Forbes.com (forbes.com), due to the COVID 19 pandemic, e-commerce grew by 77% in May of 2020, compared to the previous year. Additionally, an increasing number of consumers got comfortable with the idea of online shopping leading to a shift in consumer shopping behaviour. With buyers moving their attention to online digital channels, this creates a good opportunity for product companies to capture shoppers interests

For this project, Amazon Reviews dataset for cell phone accessories was used. The main goal for modeling was to accurately predict a positive or negative sentiment of a product based on a user's review. Data was analysed to understand predictive words for a sentiment and visualise interesting trends.

Use case:

Sentiment analysis modeling based on the amazon dataset can be used to gain valuable customer insight regarding user sentiment of a product from different social media platforms such as twitter, reddit and Facebook. This will be helpful to drive business strategies and profits. Three product types were closely looked at, namely cell phone power chargers, cell phone screen protectors and cell phone dashboard holders.

Data Extraction and Preprocessing:

The review data for *Cell phones and Accessories* was obtained from an open-source amazon reviews 2016 dataset (Amazon Dataset 2016). The dataset was in csv format, and this was read into a pandas dataframe. A random 10% sample dataset was used as the project dataset consisting of approximately 500,000 rows.This data set consisted of different variables such as reviewer name, time of day of review. product asin, a rating column and two text columns, namely, the complete review, and the review summary.

Initially the review data was analysed to detect which languages were used and how many of the reviews were in english. The vast majority of the reviews were in english. Before further processing, only the data with english language text was taken and the rest was removed.

The text information was initially preprocessed. Text was converted to lowercase, numbers and punctuations were extracted, stop words were removed, and the text was lemmatized.

EDA:

The sentiment rating column has 5 unique ratings numbered 1 through 5.

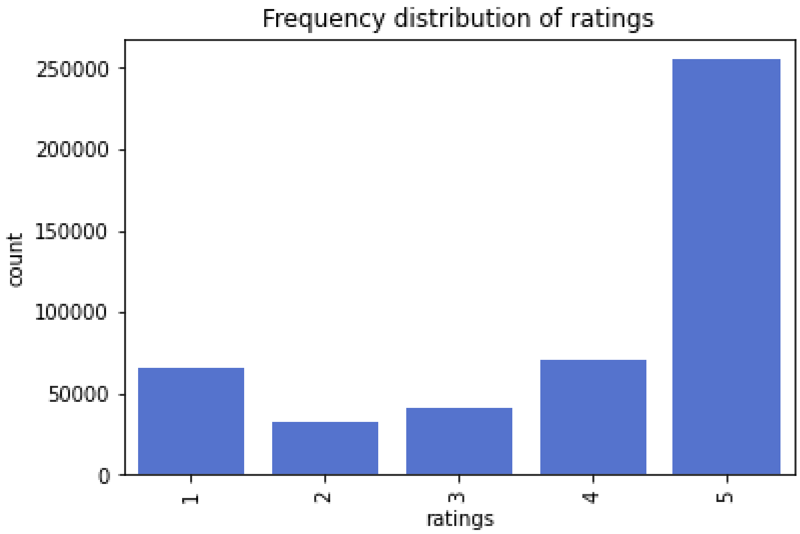


Figure 1: Rating frequency distribution

The frequency distribution shows that 5 has the highest count compared to 1 through 4. For modeling 5 was considered in a separate class for positive sentiment and 1-4 as class for negative sentiment. When looking at the two classes (5,1-4), the classes look somewhat balanced with 56% being a rating of 5 and approximately 44% being a rating of 1-4.

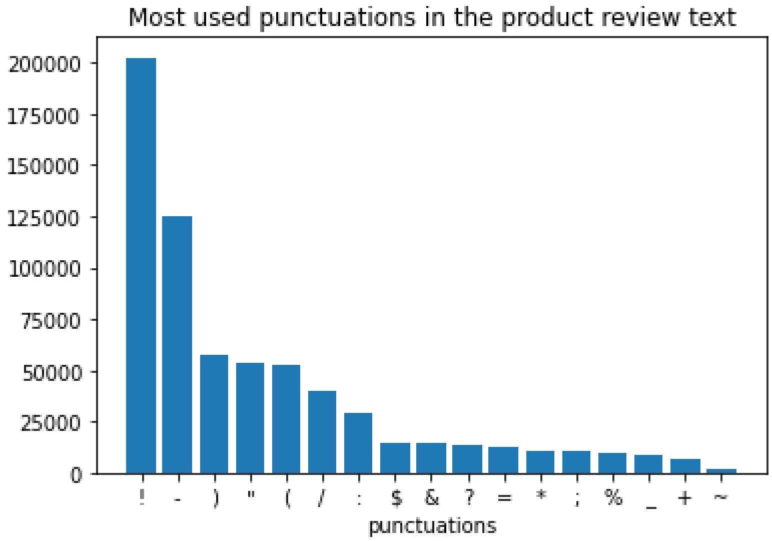
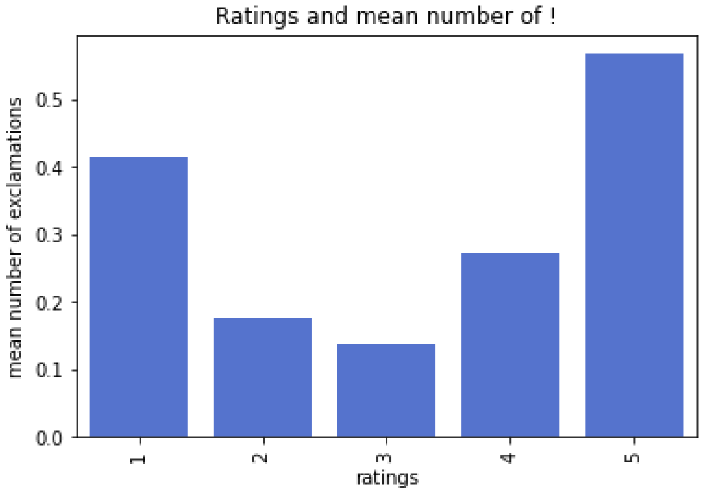


Figure 2: Commonly used punctuations

First, I tried to uncover some characteristics in the text that would be potentially important for sentiment prediction. Punctuations from the review text were extracted into a seperate column. Similar to stopwords, common punctuations not relevant to sentiment, such as a period or comma, were removed and the remainder punctuations were examined.

Exclamation stands out as a very often used punctuation compared to other punctuations such as a $ sign or ? A closer look at the text which used exclamation(!) and question marks (?) gives a better idea of how those ratings fared.



From figure 3 we can deduce that ratings of 1 or 5, i.e. strong negative and positive sentiments resulted in higher mean number of exclamations in the reviews. One way ANOVA test amongst the 5 rating groups for number of exclamations yields a very low p value (<.001)

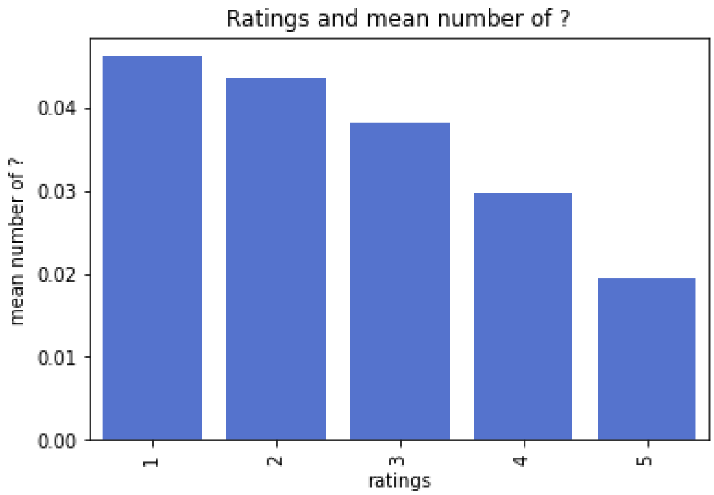
Figure 3: Influence of exclamations on ratings

Example of a review with a positive sentiment and high number of punctuations:

It's a very cute phone case but, it doesn't have a lip. It so far is protective,the picture doesn't come off or fade. I LOVE THIS PHONE CAESE SO MUCH!!

Example of a review with a negative sentiment and high number of punctuations:

Don't know what they've done with the latest/greatest version of the Blue Parrot but it's junk ! Noise cancelling capability is good but the sound is so garbled it sound like people are talking to you with they're mouth full of oatmeal. My older model Blue Parrot was impressive in it's sound quality. DO NOT BUY THIS THING IT"S JUNK !!



A higher use of question marks was associated with a lower rating. An independent t test on the *5* and *4* rated groups was statistically significant, yielding a p value < 0.01.

Figure 4: Influence of question marks on ratings

Example of a review with a positive sentiment and less number of question marks:

I will not belabor the technical aspects of this phone--but I will sing higest praises to Nokia for making it available in the U.S. Now, if Americans would avail themselves of the data functions of this phone and not just the voice, AT&T, T-Mobile and Cingular "Internet" dl/ul prices would be cheaper.

This is, simply put, the best phone on the market to date! The camera is great, that it shoots short videos is amazing, and couple it with the very comfortable Jabra Bluetooth headset, and you have a true deal that even P.T. Barnum himself would be envious of.

There are other phones with more vivid colors or larger screens, etc., but you get so much more phone than any other is capable of today (in the US). You can't go wrong--and it's free! How can you beat it? You can't.

Example of a review with negative sentiment and high number of question marks:

I purchased the goods on the 3 rd September 2012 and to date I not received the goods, no the 7th October 2012?????????????????????????

Why

Malcolm

Examining most predictive words:

In the project dataset, I wanted to examine which words were most predictive of a positive or negative sentiment of *cell phone and accessory* products, and specifically of 3 particular kinds of products I was interested in most.

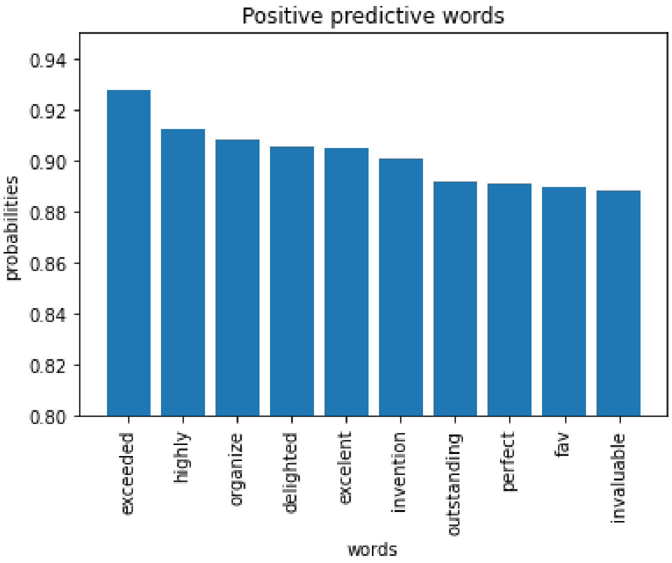
Steps to identify strongly predictive words:

- Set X as the review Text cleaned, y as the binary class for ratings
- use TFIDF vectorizer to fit and transform the X feature set, train a multinomial naive bayes classifier using the vectorized X set and y
- Create a data set from the vectorized X features such that each row has exactly one feature(word). This is represented by the identity matrix.

```
x = np.eye(tfidf_train.shape[1])
```

- Use the trained classifier to make predictions on this matrix
- Sort the rows by predicted probabilities, and pick the top and bottom K rows

Examining highest predictive words for a positive and negative sentiment in the cell phones and accessories dataset



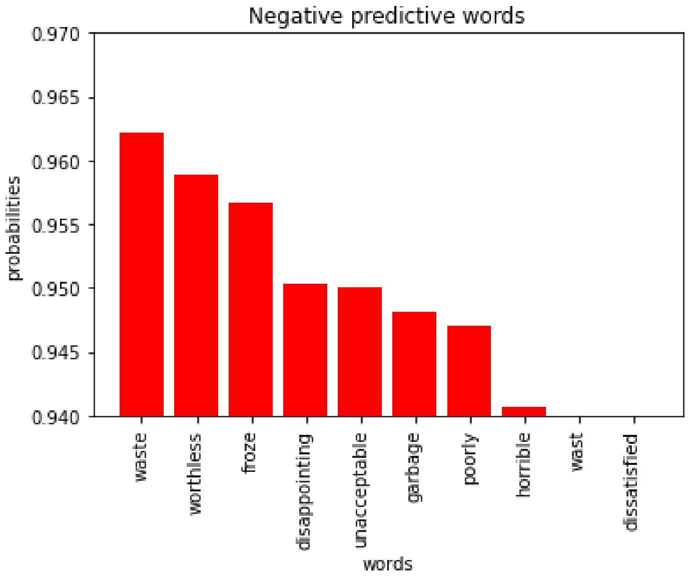


Figure 5: Positive predictive words

Figure 6: Negative predictive words

From Figure 5 , the 3 words most predictive of a good *cell phone & accessory* are *‘exceeded’, ‘highly’ and ‘organize’* and from Figure 6 we can see that words most predictive of a negative sentiment are *'waste', 'worthless', and 'froze'*. Some interesting words predictive of a positive sentiment are *'invention','invaluable' and 'organize'*

Next I examined predictive words for a cell phone *power charger*, *screen protector and dashboard holder,* this time I separated out reviews belonging to each of the 3 products and followed the steps outlined above to get the most predictive words.

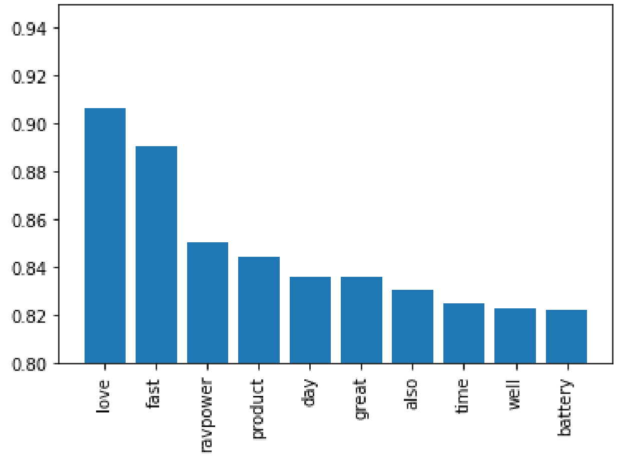


Figure 7: Power Charger positive

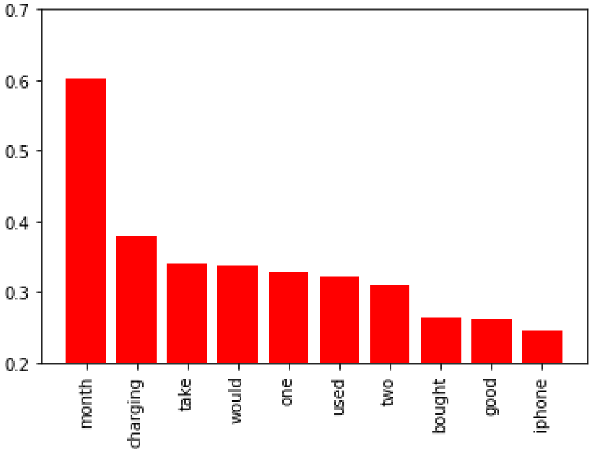


Figure 8: Power Charger negative

Some interesting negative predictive words are *take* and *month*, and *iphone* for the Ravpower charger

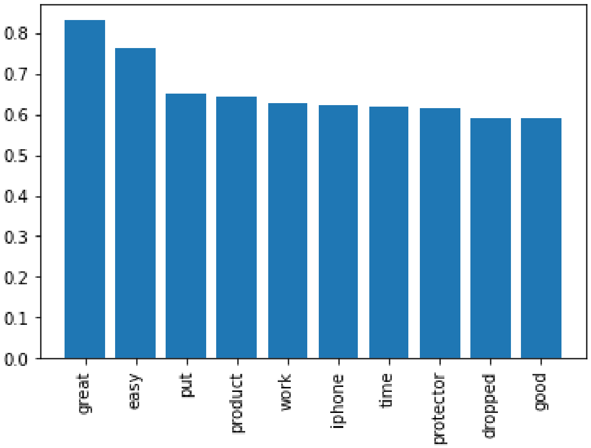


Figure 9: Screen Protector positive

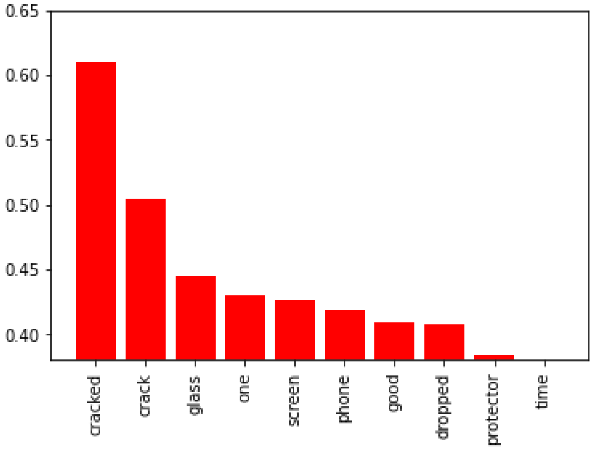


Figure 10: Screen protector negative

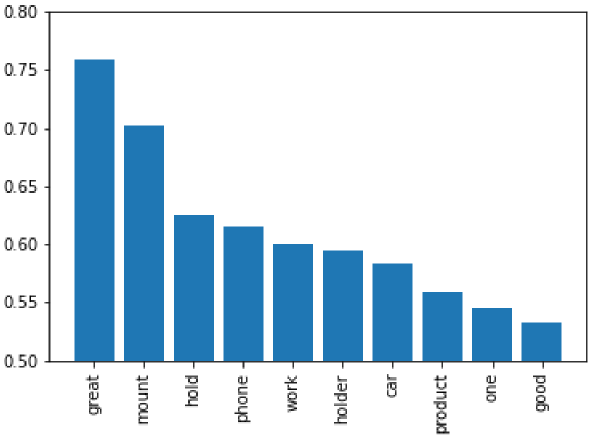


Figure 11: Dashboard Holder positive

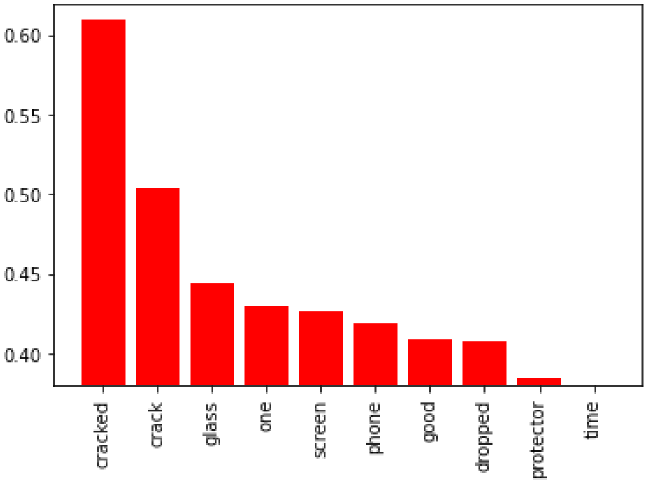


Figure 12: Dashboard Holder negative

EDA results:

- Products had a rating of 5 more often than a rating of 4 or lower
- Use of punctuations such as exclamation marks are predictive of a strong negative or positive sentiment
- A higher use of ? is associated with lower ratings
- For cell phone accessories, some interesting words predictive of a positive sentiment are *'invention','invaluable' and 'organize'*
- For the RavPower charger for a cell phone some interesting predictive words for negative sentiment were *'take','month' and 'iphone'*

Modeling:

Sentiment rating column had ratings from 1 to 5. Ratings were converted to a binary class for target variable prediction, with a rating of 5 indicative of positive class :1, and rating of 4 or below as a negative class: 0

A random subset of 46,000 rows from the project dataset was used for modeling, The cleaned review text was used as a text feature. Additionally, number of words and number of sentences, number of exclamations in the review Text were included as numerical features.

Initially the CountVectorizer + Multinomial naive bayes model was searched for best min_df and compared with TFIDF Vectorizer + Multinomial naive with best min_df. The CountVectorizer performed slightly better, and this was used as the vectorizer for the rest of the modeling with a min_df=4.

Three models were assessed, following steps were performed:

- Training and Test data was split
- From the training and test sets, text and numerical features were extracted
- CountVectorizer was fit using Text features from the training set, then the text features from the test set was fitted and transformed, these were then converted into a dense array
- Numerical features from the training and test set were transformed used MinMaxScaler()
- Before fitting the models, transformed text and numerical features were then concatenated into the respective Training and Test sets.
- Multinomial naive bayes, Random Forest Classifier, and Logistic Regression classifier were initially grid searched for best parameters, and the best parameters were then applied to the final models
- For model assessment, two metrics were taken into consideration, overall balanced accuracy and f1 score for positive sentiment.

Metrics for thresholding and model evaluation based on the business case:

Scenario 1: When a new product is introduced into the market, businesses will be interested in an overall perspective of how well the product is performing. This information can be derived based on consumer reviews that are both positive and negative. In that scenario, correct prediction of both classes (i.e. positive and negative) is important, and the metric of choice for evaluating the model will be a Balanced Accuracy score.

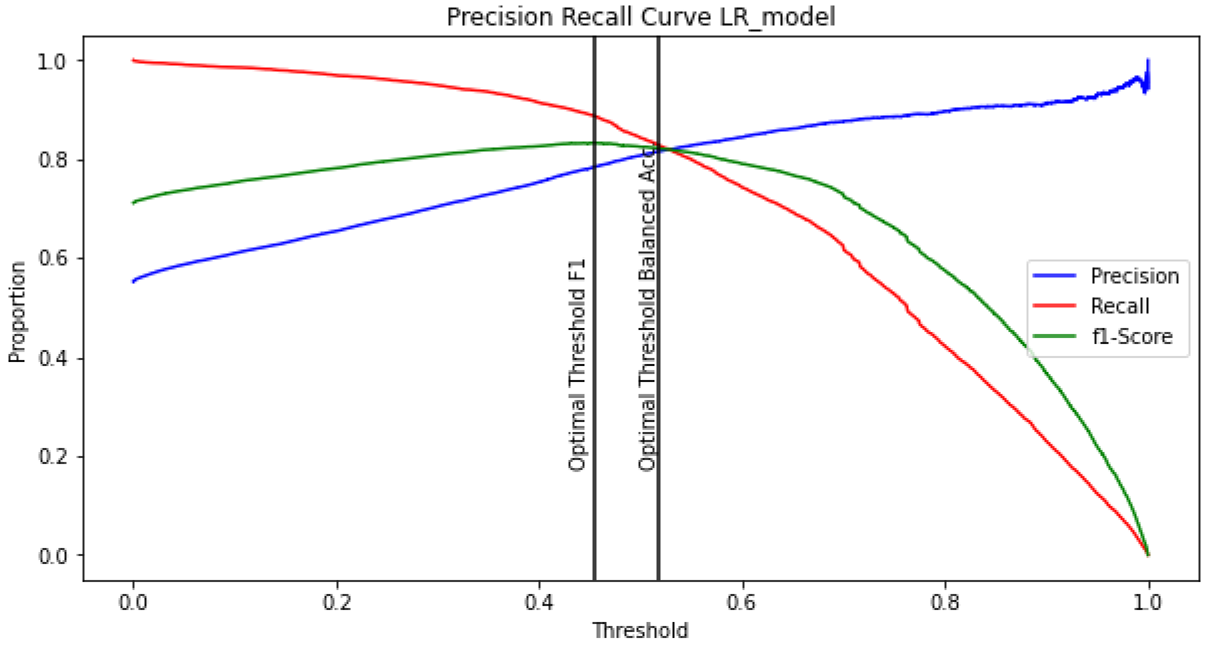
In the table below I have compared 3 different machine learning models, and obtained their best Balanced Accuracy score (BAS) for each model's best threshold based on the BAS.

model	balanced accuracy	best threshold	parameters	best model for balanced accuracy
Multinomial Naive Bayes	0.79819	0.565	alpha=0.9	
Random Forest Classifier	0.79766	0.499	bootstrap=False,min_samples_leaf=2,min_samples_split=10,n_estimators=300	
Logistic Regression Classifier	0.79933	0.518	penalty='l2',C=0.1	Logistic Regression Classifier

Scenario 2: From a consumer marketing perspective, it will be important to understand which reviews are indicative of a positive sentiment, and understand from the reviews what qualities about the product made it stand out from the competition. Marketing can highlight in their advertisements certain qualities of the product based on the positive feedback from the consumer. Businesses can also use this to understand what makes their competitor's product appealing to the consumer as well. To assess the positive sentiment reviews, I will use the f1 score (for a balance between precision and recall) for the positive class as a metric for thresholding and model evaluation.

model	f1 score	best threshold	parameters	best model for f1 score
Multinomial Naive Bayes	0.82959	0.515	alpha=0.9	
Random Forest Classifier	0.83159	0.432	bootstrap=False,min_samples_leaf=2,min_samples_split=10,n_estimators=300	
Logistic Regression Classifier	0.83269	0.454	penalty='l2',C=0.1	Logistic Regression Classifier

Results:



When evaluating models on both business case scenarios and metrics, Logistic Regression Classifier performs better compared to the other two models with a higher **Balanced Accuracy** score of **0.79933** at an optimal threshold of 0.518 and a higher **f1 score** of **0.83269** for the positive class at an optimal threshold of 0.454

Discussion and conclusion:

The analysis and modeling of sentiment analysis on these Amazon reviews could be applicable to many different business scenarios if applied to that appropriate data for that business. On the electronics data examined here specifically, the EDA showed some interesting insights - such as the fact that exclamation points indicate a strong positive or negative sentiment. In finding the most predictive words, I found further insights that could help direct business decisions. For example, for the RavPower charger product, 'iphone' came up as one of the words predictive of negative sentiment, indicating there may be some modifications needed to target iPhone customers.

Review for a RavPower charger with the word 'iphone'
'IPhone adapter not fit'

Using the modeling steps taken here, companies could gather data from their products. Predictive word data can be scanned to know which products are being talked about/trending and what consumers think of them.

Next Steps:

For further study, additional kinds of punctuations, and special characters such as emojis can be explored to gain insights into their influence on sentiments. Furthermore, ensemble models combining models built on reviews for products generally, products in that particular product category, and products for that specific product might yield improved predictive power.

Other product type review sets can be studied. The findings across groups of products should be compared for similarities and differences of how certain features influence sentiment.