Women's E-Commerce Clothing Reviews

Sentiment Analysis

Tina Huang 2021-07-03

Online shopping becomes more and more popular especially during the coronavirus pandemic. With eCommerce steadily growing over the past few years. Online consumer feedback has also grown in volume. More than ever customers have something to say with the products after purchasing.

According to the Harvard Business Review, "*Companies can boost profits by almost 100% by retaining just 5% more of their customers.*" Running a successful business by making a customer regularly supports the business is the key. It' s important to explore the data of customer review and improve products and services from the feedback. **Sentiment analysis** is a good way to gauge customers satisfaction. It helps us understand if the customers like our products or services.

In this project, I want to answer the following questions:

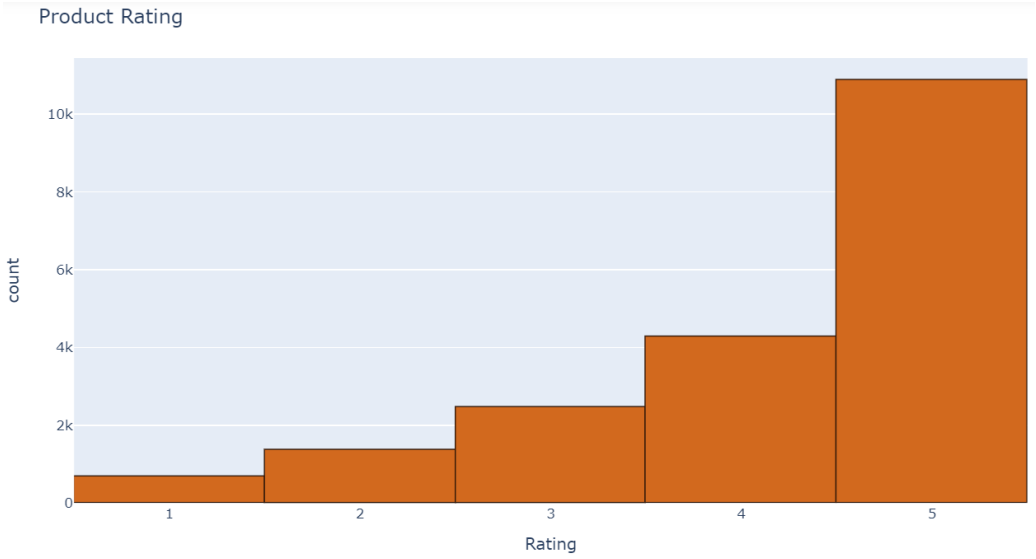
1. **How to improve products based on customers reviews?**
2. **How do customers respond to the products?**
3. **What are the trends in the review data?**

Data at a glance

This is a Women's Clothing E-Commerce dataset revolving around the reviews written by customers. Checking the head of the dataframe:

	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name
0	767	33	NaN	Absolutely wonderful - silky and sexy and comf...	4	1	0	Initmates	Intimate	Intimates
1	1080	34	NaN	Love this dress! it's sooo pretty. i happene...	5	1	4	General	Dresses	Dresses
2	1077	60	Some major design flaws	I had such high hopes for this dress and reall...	3	0	0	General	Dresses	Dresses
3	1049	50	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	1	0	General Petite	Bottoms	Pants
4	847	47	Flattering shirt	This shirt is very flattering to all due to th...	5	1	6	General	Tops	Blouses

When the company launches a new product, we could look at the variable “Rating” to see if majority of the customer ratings are positive or negative. From the plot, we can see that most of the customer rating is positive. The result leads me to believe that most text reviews will be positive, too.



What are the most frequently used words in the reviews?

I created a word cloud to see the most frequently used words in the reviews. Before that, I set a stop words list and remove stop words from the data. Stop words are the words in any language which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. The word cloud finally looks like this:

Some popular words that can be observed here include “dress,” “love,” “like,” “size,” and “fit.” From the words like “love” and “like”, I know that customers usually have positive reviews. It also indicates that dress might be a popular product in the company. Besides, customers care about the size of the product. They talk about if the clothes fit properly.

Sentiment classification

I classified all reviews with “Rating” > 3 as positive (+1). All reviews with “Rating” < 3 were classified as negative (-1). Reviews with “Rating” = 3 were dropped, because they were neutral. The sentiment classification model only classified positive and negative reviews. The data frame with sentiment index looks like this:

	Title	Review Text	Rating	sentiment
3	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5	1
4	Flattering shirt	This shirt is very flattering to all due to th...	5	1
5	Not for the very petite	I love tracy reese dresses, but this one is no...	2	-1
6	Cagrcoal shimmer fun	I aded this in my basket at hte last mintue to...	5	1
7	Shimmer, surprisingly goes with lots	I ordered this in carbon for store pick up, an...	4	1
8	Flattering	I love this dress. i usually get an xs but it ...	5	1
9	Such a fun dress!	I'm 5"5' and 125 lbs. i ordered the s petite t...	5	1
12	Perfect!!!	More and more i find myself reliant on the rev...	5	1
13	Runs big	Bought the black xs to go under the larkspur m...	5	1
15	Nice, but not for my body	I took these out of the package and wanted the...	4	1

The words “beautiful,” “cute” and “great” initially appeared in the negative sentiment word cloud, despite being positive words. This is probably because they were used in a negative context, such as “not great.” Because of this, I removed those words from the word cloud.

After classifying reviews into positive and negative, I built word clouds for each group.

Word cloud — Positive Sentiment

Some popular words that can be observed here include “dress,” “love,” “perfect,” “pretty,” and “lovely.” These words are positive, also indicating that reviews rating 4-star or 5-star in the dataset express a positive sentiment.

Word cloud — Negative Sentiment

On the other hand, we can see some popular words such as “disappointed,” “wanted,” “love,” “fit” and “dress” from the reviews rating 1-star or 2-star in the dataset.

Although I removed some words like “beautiful” and “cute”, we can still observe some positive words from negative sentiment. I was curious about why the words like “wanted” and “love” appeared on the negative word cloud with a high proportion. Aren’t they positive words? So, I filtered “Title” contained “wanted” or “love” and sentiment equaled -1.

There are 72 out of 2051 negative reviews contain the words “wanted” or “love.” It’ s 3.51%. We can see that most of the reviews contain the words “wanted” and “love” in a comment simultaneously. Customers wanted to love the product, but something stopped them to do this.

	Title	Review Text	Rating	sentiment
68	I wanted to love this top	I really loved this top online and wanted to l...	2	-1
603	Wanted to love but sadly will return	These run very small!! they are also short, al...	1	-1
806	Wanted to love it - but	I wanted to love this. however, the fit was fu...	2	-1
880	So wanted to love it	Ditto what the first reviewer said, unfortunat...	2	-1
1061	Farm animal pullover	Wanted to love this sweater for its beautiful ...	2	-1
1063	I wanted to love this sweater but	This is a beautiful sweater with deep rich col...	2	-1
1257	Wanted to love	I wanted to love this top, but it unfortunatel...	2	-1
1300	Wanted to love	I bought two, one in white and one in blue pri...	2	-1
1316	I wanted to love it	I really like the look of the top when it is p...	2	-1
1331	Wanted to love it	An absolutely gorgeous, but poorly fitting blo...	2	-1
1347	Not what i wanted	I am reasonably petite, but i like sweaters ve...	2	-1

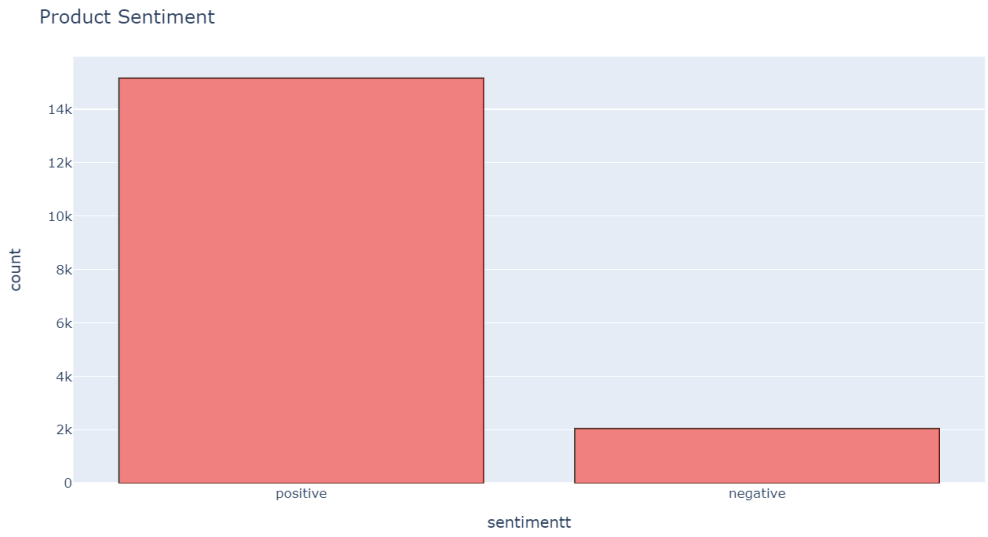
There are 183 out of 2051 negative reviews contain the words “fit,” “big,” and “huge.” It’ s 8.92%. From the number, we know that “size” is a big problem that the company needs to address. Customers could not get the right information about the

size before purchasing. The company should provide the detail of size for each product and make an accurate recommendation for customers.

	Title	Review Text	Rating	sentiment
118	Awkward fitting	First, the fabric is beautiful and lovely for ...	2	-1
334	Not a good fit	This top was way too short (i'm only 5'1) and ...	1	-1
499	Cute concept fits weird	The fun colors drew me to this but it sure fit...	2	-1
518	Weird color and fit	Color is not like photo and fit doesn't work i...	2	-1
669	Weird fit	I tried this on in the store and was disappoin...	2	-1
...	...	...	...	...
22657	Strange fit	The pattern and fabric are lovely, but the sle...	2	-1
22803	Drab color weird fit	Couldn't make it work, even belted. ivory colo...	2	-1
22844	Sweater is huge	Had to return. ordered an xs which usually fit...	2	-1
23016	Large fit	I was really excited to get this dress and to ...	2	-1
23405	Very big scratchy	I'm petite, and i knew buying this that it wou...	2	-1

183 rows × 6 columns

The distribution of reviews with sentiment across the dataset:



Overall, positive reviews are much more than negative reviews.

Training the Model

Finally, I built the sentiment analysis model to make a prediction on whether the review is positive or negative.

Data cleaning

I used the title data to come up with predictions. First, I remove all punctuation from the data.

Split the data frame

The new data frame only have two columns — “Title” (the predictor variable), and “sentiment” (the predicted variable).

	Title	sentiment
3	My favorite buy	1
4	Flattering shirt	1
5	Not for the very petite	-1
6	Cagrcoal shimmer fun	1
7	Shimmer surprisingly goes with lots	1

Then, I split the data frame into training and testing data. 80% of the data was used for training, and 20% was used for testing.

CountVectorizer from the Scikit-learn

Since the logistic regression algorithm cannot understand text, I used a count vectorizer from the Scikit-learn library. It is used to transform a given text into a vector based on the frequency (count) of each word that occurs in the entire text.

Logistic regression

I built a simple logistic regression model and trained the data on it. I also made predictions using the model.

Test the accuracy of the model

A confusion matrix that looks like this:

```
array([[ 222,   47],
       [ 228, 3182]], dtype=int64)
```

From the matrix, we show the model 450 samples where someone wrote a negative review. However, the model can only identify about 1/2 of them. (222/ (222+228) = 49%) On the other hand, we show the model 3229 samples where someone wrote a positive review, the model can catch 98.5% (3182/ (47+3182)) of them, which is surprisingly good.

The classification report:

	precision	recall	f1-score	support
-1	0.49	0.83	0.62	269
1	0.99	0.93	0.96	3410
accuracy			0.93	3679
macro avg	0.74	0.88	0.79	3679
weighted avg	0.95	0.93	0.93	3679

The overall accuracy of the model on the test data is approximately 93%, which means of all predictions, 93% of them are correct. The model is fairly good to predict if a review is positive or negative.

Appendix

Description of Data

1. Data source

   This is a Women's Clothing E-Commerce dataset revolving around the reviews written by customers.

2. Data type

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 23486 entries, 0 to 23485
Data columns (total 10 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   Clothing ID              23486 non-null  int64 
 1   Age                      23486 non-null  int64 
 2   Title                    19676 non-null  object
 3   Review Text              22641 non-null  object
 4   Rating                   23486 non-null  int64 
 5   Recommended IND          23486 non-null  int64 
 6   Positive Feedback Count  23486 non-null  int64 
 7   Division Name            23472 non-null  object
 8   Department Name          23472 non-null  object
 9   Class Name               23472 non-null  object
dtypes: int64(5), object(5)
memory usage: 2.0+ MB
```

Documentation of Cleaning or Manipulation of Data

1. Identify data type.
2. Identify unique id and find no duplicates.
3. Create smaller data frame with data columns in need.
4. Identify missing data and drop rows with missing values.
5. Check statistic summary for "Rating" column.
6. Classify reviews into "positive" and "negative."
7. Filter reviews by the popular words and count the number of the reviews.
8. Order the data with sentiment.
9. Split data into training and testing data.
10. Fit model with Logistic Regression

Reference

1. Hacking The Product Rating Race With Sentiment Analysis (2021)
2. A Beginner's Guide to Sentiment Analysis with Python