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.
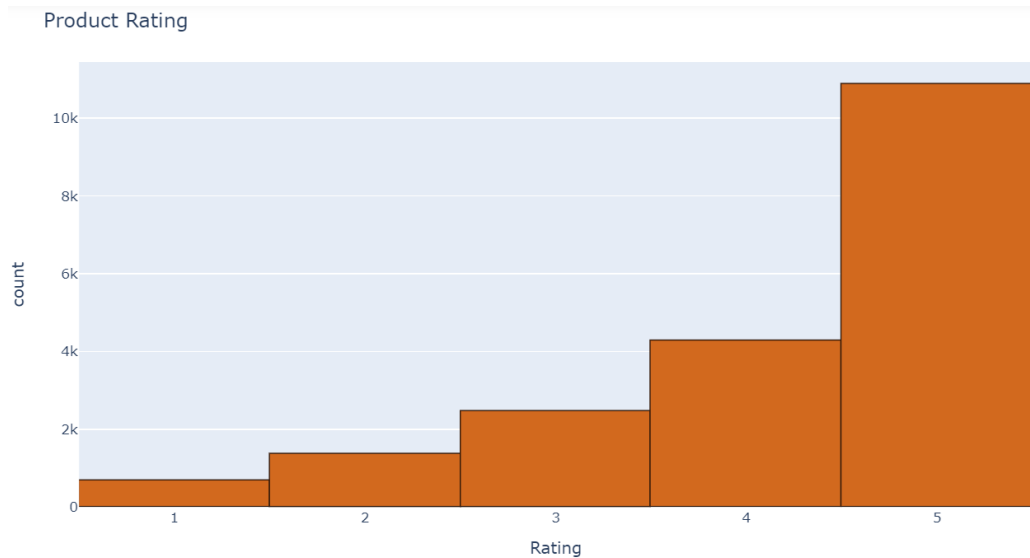
Product Rating

Exploratory Analysis

What products/items do customers complain about the most?

I wanted to find products that customers were not satisfied with the most. After plotting the horizontal bar graph for column "Class Name" with rating equals 1 or 2, we can see that **Top-3 products** with lower rating are **Dresses, Knits, Blouses.** The women's clothing shop should give priority to these three products and work on their improvements.

Class Name
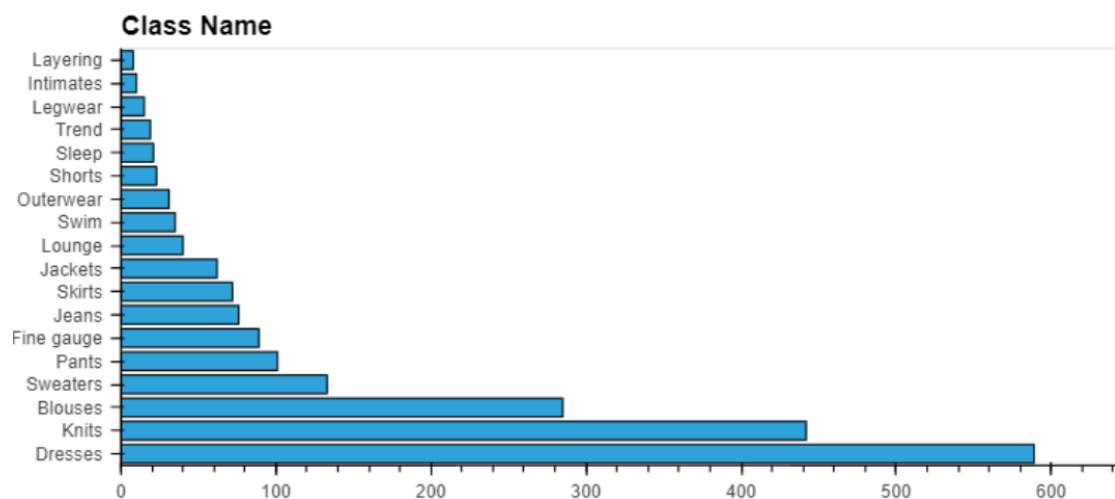
To understand which items with most dissatisfaction, I plotted the horizontal bar graph for column "Clothing ID" with rating equals 1 or 2. **Top-10 items** to improve with Clothing IDs: **1078, 862, 868, 1094, 829, 867, 1095, 1110, 1081, 1077**.
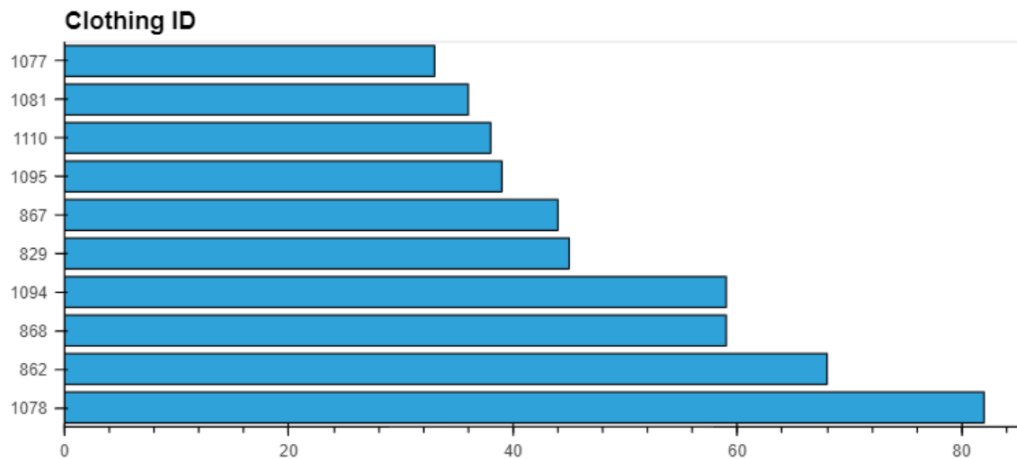
What are the most frequently used words in the reviews?

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," "love" and "like" appeared on the negative word cloud with a high
proportion. Aren't they positive words? So, I filtered "Title" contained "wanted,"
"love" or "like" and sentiment equaled -1.

There are 169 out of 2051 negative reviews contain the words "wanted," "love" or
"like." It' s 8.24%. We can see that most of the reviews contain the words "wanted,"
"love" or "like" in a comment simultaneously. Customers wanted to love the product,
but something stopped them to do this. Some reviews contain the words "looks like."
In this case, customers had no sentiment for the word "like."

	Title	Review Text	Rating	sentiment
22	Not what it looks like	First of all, this is not pullover styling. th...	2	-1
68	I wanted to love this top	I really loved this top online and wanted to l...	2	-1
241	Loved this then washed it twice and it fell apart	I loved this dress from the moment i tried it ...	1	-1
581	Do not like the front	The back of this t-shirt looks cute but the fr...	2	-1
603	Wanted to love but sadly will return	These run very small!! they are also short, al...	1	-1
793	If only it looked like the photo	This sweater was a big let down. i am 5'2" so ...	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
1306	Fits like maternity	I wanted to love this, was so excited when i o...	2	-1

Have a better understanding of customer experience

To improve products based on customers reviews, I analyzed the negative reviews with different issues. I categorized the issues by size, quality, and color:

Size issue

There are 243 out of 2051 negative reviews contain the words **"fit," "big," or "huge."** It's 11.85%. 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
88	Huge	Really cute piece, but it's huge. i ordered an...	2	-1
97	Huge	I was very excited to order this top in red xs...	1	-1
118	Awkward fitting	First, the fabric is beautiful and lovely for ...	2	-1
185	Fits strange, flimsy material	I was worried about this item when i ordered i...	2	-1
334	Not a good fit	This top was way too short (i'm only 5'1) and ...	1	-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

243 rows × 4 columns

Quality issue

There are 155 out of 2051 negative reviews contain the words **"quality" or "fabric."** It's 7.56%. There are many customers complain about quality and fabric. The company need to **communicate with the manufacturer to upgrade the quality of products and examine clothing materials**.

|	Title	Review Text	Rating	sentiment
110	Poor quality	This is so thin and poor quality. especially f...	1	-1
114	Disappointing quality	The design/shape of the dress are quite flatte...	2	-1
389	Bad quality	I don't normally review my purchases, but i wa...	1	-1
475	Poor quality	I bought this in the white, size m (140lbs., 5...	2	-1
1052	Quality & cut	My perspective is as a sewer so i'm more picky...	1	-1
...	...	...	...	...
23031	Nice fabric poor design	This dress has a beautiful fabric, i really ha...	2	-1
23255	Looks great poor quality	Pants are overpriced and very uncomfortable. t...	1	-1
23274	Cute but poor quality	I'm 5'7" 155lbs and ordered a large. it fit, b...	1	-1
23287	Too much fabric shoulders wont stay down	This is a gorgeous print, but i found it too f...	2	-1
23437	Not typical of retailer quality	This dress is so cute in the photo and fit tru...	2	-1

155 rows × 6 columns

Color issue

From the negative reviews which contain the words **"color,"** we can find that there are two main problems for the color:

1. **Color is not as pictured on the website.**
2. **Discolored after one wash.**

The company should **make the pictures look like the products** and **specify the color correctly** on the website to avoid misleading. Besides, the company need to **communicate with the manufacturer for the discolored issue** and fix the problem.

	Title	Review Text	Rating	sentiment
518	Weird color and fit	Color is not like photo and fit doesn't work i...	2	-1
1618	Great quality and color but boxy and ill-fitting	I wanted to love this sweater. the photo of th...	2	-1
2675	Color run	I love the way these shorts fit and look; very...	1	-1
3552	Way too big/colors off	Finally received this in the mail today and wa...	2	-1
5450	Orange color is pretty, quality is poor	I ordered the orange in size xl. the fit on me...	2	-1
6055	Color transfers to other clothes!	I really loved this vest online and decided to...	2	-1
9589	Colors are more intense than they appear	I purchased this dress for my graduation. the ...	2	-1
9765	Color not as shown	The dress is very pretty, but the sequins are ...	2	-1
9778	Color is not as it appears in the photo	I was very surprised to see such dark blue seq...	2	-1
9857	Color bleeds and discolors the pant.	This is a problem i have had with the pilcro h...	2	-1
9867	Not colorfast!	I'm so bummed! i absolutely loved the pair of ...	2	-1
10084	Sloppy looking but pretty colors	I ordered this style in the blue, red, yellow ...	2	-1
10793	Super cute on, colors completely faded	Love this dress! wore it last night for a musi...	2	-1
11570	Weird color	Not flattering and not the color it appears on...	2	-1
12110	Color not as pictured	I returned the dress because the color was not...	2	-1
12504	Color not as pictured on website	I wish i had read the other review about the c...	2	-1
12642	Beautiful suede, but colors are not correct.	Just received these and they fit perfectly, bu...	2	-1
13172	A colorful big box	Since this sold out so fast i ordered an xsp w...	1	-1
13586	Color not as pictured	I had high hopes for this top based on the pho...	1	-1
13594	Colors look nothing like online photos:(	When i saw this top online, i thought i'd love...	1	-1
13832	Wrong color and too big	I was so excited for this dress to come. i am ...	2	-1
14504	Very thin material. nice colors	I agree with the other reviewer about the grea...	1	-1
15978	Color became spotty after washing	I bought a pair of these in stone and after th...	1	-1
15990	Not colorfast at all!	Bought the gold pair and dripped something on ...	1	-1
15992	Color ruins easily	I bought a pair of these a month ago and first...	1	-1
16014	Discolored after one wash.	Similar to others here, these pants were ruine...	1	-1
16773	Cream colored	The top is more of a cream color than an ivory...	2	-1
17566	Did not like the fit, but beautiful colors	I really wanted to love this shirt! the colors...	2	-1
18478	Grey not color as pictured on website	I wanted to love this top, but when i received...	2	-1
18599	Huge jacket, wash color is not as pictured	Like one of the previous reviewers, when my ja...	1	-1

The distribution of reviews with sentiment across the dataset:
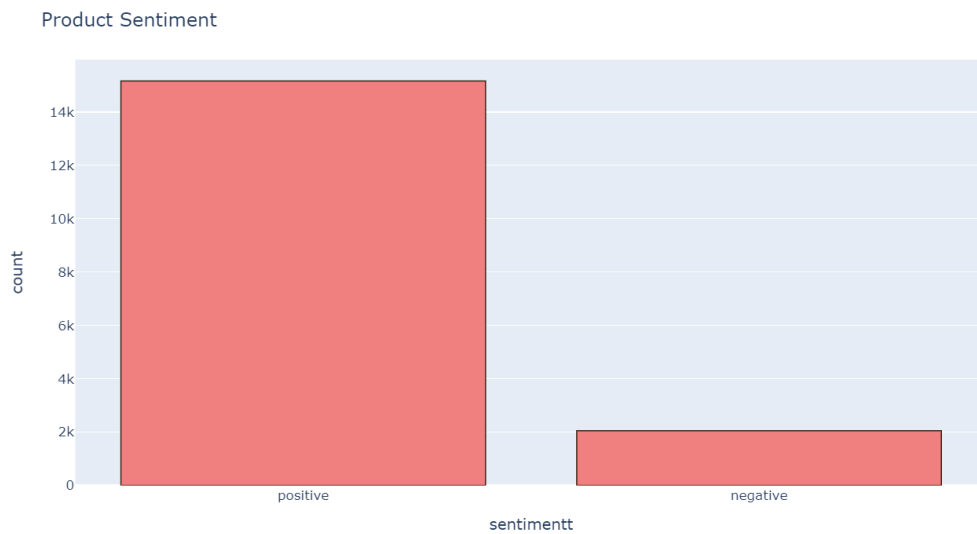
Overall, positive reviews are much more than negative reviews.

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       0.52      0.80      0.63       282
           1       0.98      0.94      0.96      3378

    accuracy                           0.93      3660
   macro avg       0.75      0.87      0.80      3660
weighted avg       0.95      0.93      0.93      3660
```

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.

Conclusion

1. **Top-3** products with lower rating are **Dresses, Knits, Blouses**.
2. **Top-10** items to improve with Clothing IDs: **1078, 862, 868, 1094, 829, 867, 1095, 1110, 1081, 1077**.
3. "**Size**" is a big problem that the company needs to address. The company should provide the detail of size for each product and make an accurate recommendation for customers.
4. There are many customers complain about **quality and fabric**. The company need to communicate with the manufacturer to upgrade the quality of products and examine clothing materials.
5. For **color issue**, the company should make the pictures look like the products and specify the color correctly on the website to avoid misleading. Besides, the

company need to communicate with the manufacturer for the discolored issue and fix the problem.

6. We can make a prediction on whether the review is positive or negative by **Logistic Regression model** to gain more information.

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