Women's e-Commerce Sentiment Analysis

Barbara Payne
Bellevue University
DSC680-T302
Summer 2021

Abstract

Within the last year alone, e-commerce sales have skyrocketed. Online sales were popular due to people not having to leave the house and the convenience factor. But with COVID-19, people were forced to stay home for some time. Physical stores were closed to the public, which led to people buying more online. In 2020, online sales reached $791.7 billion, which was up 32.4% from the previous year. This made 2020, the highest annual online sales growth of any year. Even this year, in the first quarter of 2021, sales have increased online to 39%, which was higher than the first quarter of 2020 (*Covid's Impact on Online Shopping*, 2021).

A big part of the e-commerce world is reviews. Reviews really help make or break a product. Machine learning can help determine if a product is doing great or if it is doing poorly based on reviews customers leave by preforming a sentiment analysis. Sentiment analysis can identify which products are getting positive reviews versus negative reviews based on the text that is given in the review. This can really help businesses as the machine learning can go through each review, not an actual person. It can quickly identify which reviews are positive and which are negative, which can help determine which products are doing better than others.

Introduction

E-Commerce

Online retail has grown so much within the last year. Due to COVID-19 people were forced to stay home and business locations had to temporarily close. The need for an online presence grew rapidly if businesses wanted to make money. Since people could not physically go to the store, they were inclined to do their shopping online. Online shopping has greatly increased where in 2020 it increased close to 33% reaching over $790 billion.

Though places have begun to open, online sales continue to rise. Some of the online retail may go towards travel and live entertainment as more opens up but shopping online is here to stay. It is predicted that e-commerce sales will grow 13.7% this year reaching over $900 billion (Droesch, 2021). This is almost double of the amount made in 2019, before the pandemic. There is just much more convenience to shopping online than going to the store. With technology it is at our fingertips, there is so much we can do within our homes.

Sentiment Analysis

A disadvantage of online shopping is not being able to see the product in person. On e-commerce website, people can leave their opinions on the product by leaving a review. These reviews can help potential buyers make their decision on whether or not they want to make the purchase. The reviews can also let the business know which products are more popular and which products may need to be reexamined.

Machine learning can help businesses determine which products are doing better than others by preforming a sentiment analysis. A sentiment analysis is the process of detecting positive and negative views within text. It is greatly used in customer feedback to monitor brand image and product sentiment (*Sentiment Analysis: The Go-To Guide*, n.d.). Rather than having a person read through all the text and having them determine which are positive reviews and which are negative, this machine learning opportunity can do it for them.

Sentiment analysis is used to determine the tone behind the text. It is estimated that e-commerce companies can loose $10 billion worth in revenue per year if they are unable to capitalize on sentiment analysis (*Customer Sentiment Analysis: A crucial need in E-Commerce*, 2021). There are many benefits to it as well. It can enhance the overall customer experience. By

giving a place for people to voice their opinion and then do something about it makes the customer feel heard. There is a lot of information that can be gained from a sentiment analysis that are really beneficial to businesses.

Dataset

https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

The dataset I chose can be found on Kaggle through the above link. I found that this dataset had a lot of variables that could be worked with, if wanting to go past just doing a sentiment analysis. The data does come a real source, but to keep anonymity, the name of the store has been taken out. Some information about the review is given, but most important is more information on the product that is being reviewed. This can help identify which departments are doing better than others and could pinpoint the star items in each department, as well as which items need to be reevaluated.

Methods

This dataset contains different variables with corresponding values. The first method conducted will be Exploratory Data Analysis (EDA). The data must be examined for any outliers and what could be considered bad data. Bad data here would be considered to be reviews with no text. A review without any text is useless as there is no information that can be gained from it. Though it may have a rating with it, the text is key to preforming a sentiment analysis.

The sentiment analysis itself will be the main method used for this project. Within the sentiment analysis, the text will need to be cleaned to ensure that is can be performed correctly. Some methods that will be used will be the removal of stop words, removal of

punctuation, and stemming. I can then do a word cloud to gather the most common words used in both the positive and negative reviews.

Lastly, if time permits, I could also do some analysis on the reviews themselves with the other variables within the dataset. I could see which item had the most positive reviews and which item had the most negative reviews. I could also look at the greater department and see which departments are more popular over the others.

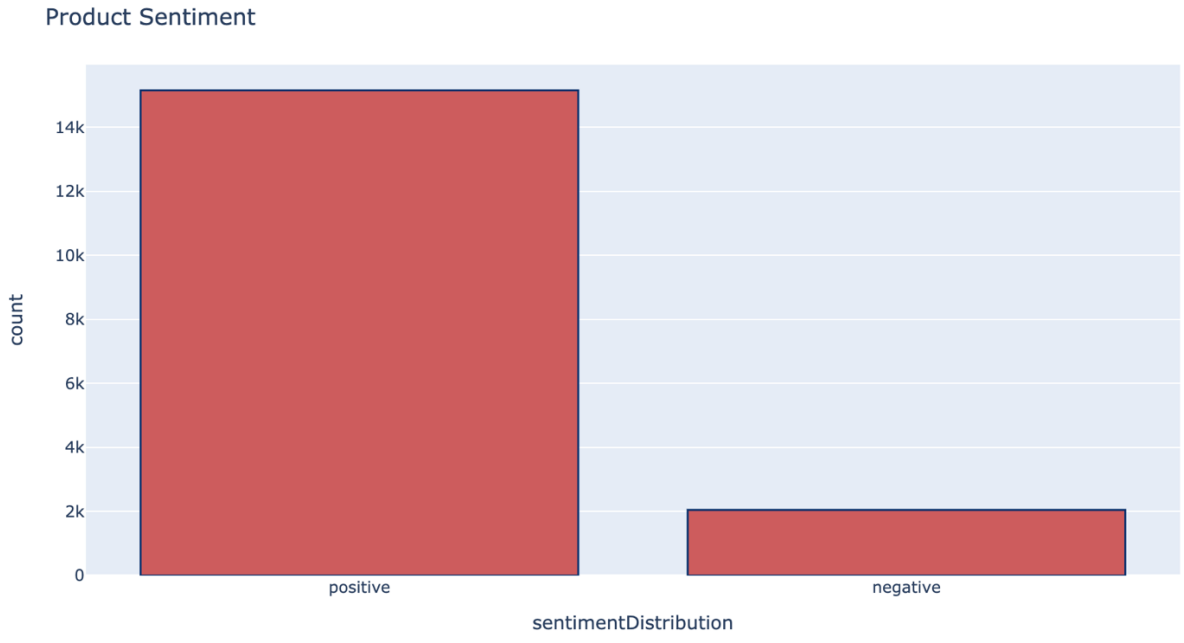
Results

Within in the EDA phase, I found that there were quite a few reviews that did not have any text value. There was also some that did not have any text in the title of the text. I decided to remove these from the dataset as they provided me little value within the sentiment analysis. I did do a word cloud of positive and negative words that appeared most often in the reviews. Some of the more common words that appeared in the original word clouds for both positive and negative were "love", "dress", "cute", "great", and "beautiful". After removing these words, my word clouds were as follows:

Positive Word Cloud

Negative Word Cloud

I found that more customers were willing to leave a positive review than a negative review. In fact, there were only 691 reviews that were a rating of 1 (the lowest) where there were over 10,000 reviews of 5 (the highest). This can definitely skew the data for any predictive analysis as there is an extreme number of higher reviews than there is lower. This could mean that not everyone who has bought the product left a review. Some people do find that leaving a review can be tiresome and do not want to be bothered with it.



Sentiment Analysis – Final Results

I was pretty successful in my sentiment analysis. I did a Bag of Words approach to get a count of how many times each word was used. I then used logistic regression to train and test my model. I had an accuracy at 93%, which is very high and proved that the sentiment analysis went very well. Also, I had done a confusion matrix to get the number of true positives, which were 3,247. I am very happy with my sentiment analysis results and the research that was conducted.

Other Analysis – Final Results

I wanted to conduct some more analysis on the data itself. I wanted to see which departments had the highest and lowest ratings and then be able to pinpoint what specific items had the highest and lowest ratings. The dresses class showed that it had the most ratings, so I decided to do some analysis on that department alone. When it came to finding which item had the highest rating it was clothing ID 1078. However, it appeared that clothing 1078 also had the lowest rating scores as well. This proved that my analysis was not successful. There were a lot more 5-star ratings than the 1-star ratings, which was skewing the data anyway.

Conclusion

Sentiment analysis are crucial for online businesses. There is so much more that a sentiment analysis can assist with besides reviews. It can also help with the overall brand of the businesses. If there are reviews out there of the brand overall, a sentiment analysis can be performed to get an idea of how their customers view them.

Image and how the business is perceived is everything as it can really make or break a company. There are so many different benefits to preforming sentiment analysis. Some of

these benefits include developing an insightful business strategy, understanding the customers, measuring the marketing campaign, finding influencers, and boosting customer service (Marta, 2021). Without customers in e-commerce, there would be no business.

My sentiment analysis was very successful at 93% accuracy, with other analysis not as successful. The overall goal was to do a sentiment analysis, so I am happy with how that had turned out. I believe if I had a different dataset that had a more evenly distributed reviews, the other analysis would have been more successful.

References

COVID's impact on online shopping. Digital Commerce 360. (2021, June 16). https://www.digitalcommerce360.com/article/coronavirus-impact-online-retail/.

Customer Sentiment Analysis: A crucial need in E-Commerce. Digital Analytics Consulting Company | nabler.com. (2021, June 30). https://www.nabler.com/bi-consulting/articles/customer-sentiment-analysis-in-e-commerce-data-initiatives/.

Droesch, B. (2021, May 17). *How will the pandemic affect US ecommerce sales in 2021?* Insider Intelligence. https://www.emarketer.com/content/how-will-pandemic-affect-us-ecommerce-sales-2021.

Marta. (2021, April 16). *The Benefits of Sentiment Analysis*. Brand24 Blog. https://brand24.com/blog/the-benefits-of-sentiment-analysis/.

Sentiment Analysis: The Go-To Guide. MonkeyLearn. (n.d.). https://monkeylearn.com/sentiment-analysis/.

10 Questions

1. Why did you want to do a sentiment analysis?
2. Why did you choose the dataset you chose?
3. What were you hoping to gain from this project?
4. What challenges did you face while doing the project?
5. What went right in the project?
6. Is there anything you would have done differently?
7. Why did you choose the methods within the sentiment analysis?
8. Was there any data that was excluded from the dataset?
9. How much time did you spend on the analysis?
10. What you know now, what would you tell your past self when starting the project?