How can Fast Fashion Brands Utilize Online Clothing Reviews?

Breonna Moore

Project Design

I used a dataset from Kaggle analyze clothing reviews. Insights gained from this analysis can be used by clothing brands to understand trends, decrease return rates, improve their products quickly, and protect their brands.

The classes used to describe the reviews' sentiments are 1, which signifies a negative sentiment, and 5, which signifies a positive sentiment. Modeling for this project was performed using Logistic Regression (LR) and Complement Naive Bayes (CNB). Both models were cross validated across 10 folds. I found that LR was the best model for predicting to use by comparing the F1 scores. After adjusting thresholds and tuning parameters, LR produced a recall of 0.95 for both the positive and the negative sentiments, meaning that of all the reviews, I was able to predict 95% of them correctly for both classes. Recall was the metric of choice to optimize the LR model because I wanted to ensure that I was able to predict the negative sentiment as correctly as possible in order to give the brands a topic modeling on the negative sentiment reviews that was accurate.

I focused the unsupervised natural language processing on the negative sentiment reviews to find actionable insights to improve brands products. I tried both Non-negative Matrix Factorization (NMF) and Latent Semantic Analysis (LSA) to do topic modeling. I looked at a range of 5 to 50 components. I found that NMF of 40 components gave me the most interpretable topics. Of those 40 topics, I found 5 topics that were very well described by their terms. They fell into the categories of quality, design, fit, flaws, and price. These are all actionable categories on negative sentiment reviews that a brand could further look into and change.

Tools

- Python
    - Data Collection: Pandas
    - Data Analysis: Scikit-learn, NumPy
    - Data Visualization: Matplotlib, Seaborn, WordCloud
- Microsoft PowerPoint, Microsoft Word

Data

The dataset from Kaggle was collected from real businesses and contained 23,486 clothing reviews. I created features from the clothing reviews using term frequency-inverse document frequency (TF-IDF) which converted the words into numerical data.

Originally the data's rating ranged from 1 to 5, with 5 being the highest rating a customer could give a piece of clothing, but I divided the dataset down to just the 1's and the 5's to do a sentiment analysis to classify the positive and negative sentiment reviews. This created a dataset of 11,549 rows of imbalanced data.

Initially the logistic regression model was mislabeling many positive sentiment reviews as negative sentiment reviews (NSRs). This mislabeling caused the predicted NSRs to have more actual positive sentiment than actual NSRs, which then caused the topic modeling to not make sense for NSRs. To remedy the class imbalance in this dataset, I used SMOTE Oversampling on the NSRs because they are the minority class. I also adjusted the thresholds of the classes. This improved the prediction of NSRs which in turn improved the topic modeling of the NSRs.

What I Would Improve Upon Next Time

I tried to use clustering on the topics by using the elbow method to find the correct number of clusters as well as KMeans Clustering, but the results were not favorable. I would like to try more clustering methods to see if I could get the topics to fall into clusters that matched the 5 categories that I found in the topics. I also think once I have set a path for my project, it would be good for me to stay the course. I changed my project topic three times and changed my entire design and goal multiple times as well based on the advice of others that meant well. Seeking guidance of those more experienced is necessary when learning something new, but I also think that at some point, you have to be comfortable making the decisions on your own. This can be hard to do when you are new to a field and have little intuition about the best methods, but it is something that I would like to improve.