Amazon Review Data (2018) – Pet Supplies

Sentiment Analysis & Text Classification Project

Enkhchimeg Tsendnyam

List of Appendices

Background/History

As more and more businesses are incorporating or exclusively offering online products and services, the product reviews have become critical method of receiving customer feedbacks. Consumers are posting reviews directly on companies' products pages in real time. With the vast amount of consumer reviews, this creates an opportunity for everyone to see how the market reacts to a specific product type. I myself have been mostly a brick-and-mortar shopper for most of my life until the pandemic hit in 2019. For the last few years, Amazon has been one of main online shopping sources where I buy some of my food, books and pet-supplies. A little background history on Amazon: Jeff Bezos founded Amazon from his garage in Bellevue, Washington in 1994. Initially an online marketplace for books, it has expanded into multitude of product categories such as several media (books, DVDs, music CDs, videotapes and software), apparel, baby products, electronics, beauty products, gourmet food, groceries, health and personal-care items, industrial & scientific supplies, kitchen items, jewelry, watches, lawn and garden items, musical instruments, sporting goods, tools, automotive items, and toys & games. "*It also has multiple subsidiaries including Amazon Web Services (cloud computing), Zoox (autonomous vehicles), Kuiper Systems (satellite Internet), Amazon Lab126 (computer hardware R&D) Its other subsidiaries include Ring, Twitch, IMDb, and Whole Foods Market.*" Today, Amazon is one of the Big Five, "MAMAA" companies in the U.S. information technology industry, along with Meta(Facebook), Apple, Microsoft, and Alphabet (Google).

Business Problem

As an avid Amazon shopper, sometimes when I'm looking for a specific product, I spend hours reading the description, ingredients as well as customer reviews & ratings on products. This can be daunting and very time-consuming task especially if there are so many products available that meet the requirements I'm looking for. For example, this winter, I spent a good few days looking for a water resistant, non-slippery and warm winter boots for my dog before making my decision. Only, if there was an easy way to analyze product reviews and there is. Sentiment analysis is a natural language processing(NLP) technique, to determine whether data is positive, negative or neutral. In this project, I will attempt to do a sentiment analysis & text classification on Amazon Pet Supplies Review Data using python.

Data Explanation

The dataset that's being utilized for this project is Dr. Jianmo Ni's (professor at University of California San Diego) **"Amazon Review Data (2018)"** dataset from Github(https://nijianmo.github.io/amazon/index.html). The full dataset has over 230 million number of reviews and the dataset includes data from May 1996 to October 2018. I will be working on a smaller sub-set, per-category data on "Pet Supplies Review" . Pet Supplies Review dataset comes in the json.gz file format and it includes some but not all of the metadata. It includes: ID of the reviewer, ID of the product, name of the reviewer, helpful votes of the review, text of the review, rating of the product, summary of the review, and time of the review. This

dataset has over 150,000 samples, which I will only use 10% of for pre-processing and modeling.

```
#Display the data first 5 rows using head()
df.head(5)
```

	reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime
0	A14CK12J7C7JRK	1223000893	Consumer in NorCal	[0, 0]	I purchased the Trilogy with hoping my two cat...	3	Nice Distraction for my cats for about 15 minutes	1294790400	01 12, 2011
1	A39QHP5WLON5HV	1223000893	Melodee Placial	[0, 0]	There are usually one or more of my cats watch...	5	Entertaining for my cats	1379116800	09 14, 2013
2	A2CR37UY3VR7BN	1223000893	Michelle Ashbery	[0, 0]	I bought the triliogy and have tested out all ...	4	Entertaining	1355875200	12 19, 2012
3	A2A4COGL9VW2HY	1223000893	Michelle P	[2, 2]	My female kitty could care less about these vi...	4	Happy to have them	1305158400	05 12, 2011
4	A2UBQA85NIGLHA	1223000893	Tim Isenhour "Timbo"	[6, 7]	If I had gotten just volume two, I would have ...	3	You really only need vol 2	1330905600	03 5, 2012

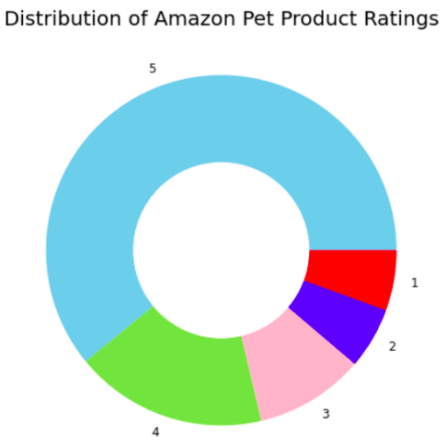
```
#Check the dimension of the table
print("The dimension of the table is: ", df.shape)
```

```
The dimension of the table is:  (157836, 9)
```

Some of the research questions that I will explore and attempt to answer are:

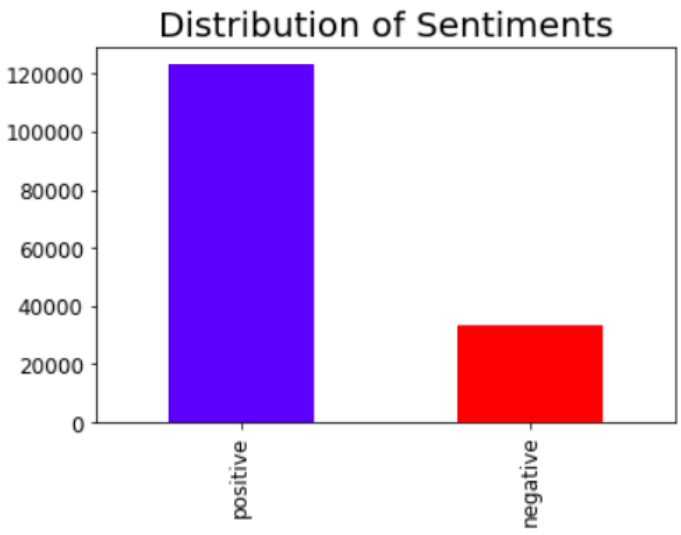
- What are customers are saying about Amazon Pet Supplies?
- How are customers are rating Amazon Pet Supplies?
- What portion of the reviews are positive, negative, or neutral?
- What are the most common words in the reviews?
- How accurate is my model on classifying product reviews?

Methods/Analysis



Prior to working on data preparation, I was able to do a small data exploration where you can see how customers rated amazon pet supplies. As you can see on this chart, majority of the people rated amazon pet supplies with 5 stars. Based on the overall rating, I created an additional column called “sentiment” where it marked rating 4 or above as “positive” and

anything below as that as “negative. Distribution of this newly created column is below, and you can see this is a little bit imbalanced data.

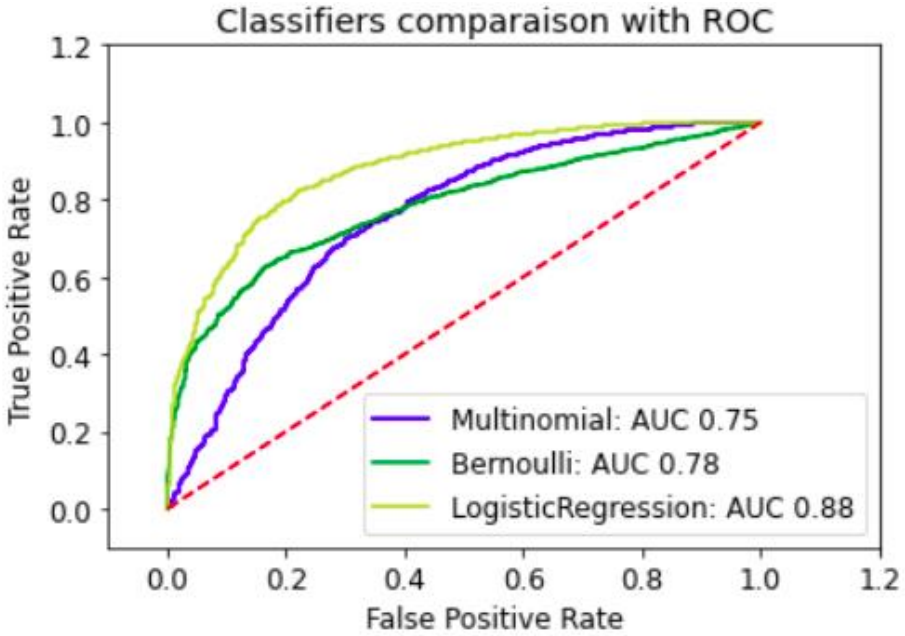


The data preparation methods for this project are focused on pre-processing the text data, specifically the review Text column by cleaning, transforming, and extracting features. Initial data cleaning steps included removing null values, converting the text to lowercase, removing punctuations, special characters, and stop-words. Once the initial cleaning steps are done, I was able to visualize the most common words in the reviews.

Then, I further pre-processed only 10% of the sample data or about 15649 rows by using methods such as tokenization and lemmatization. Tokenization splits the sentences into into single works which is needed to generate feature extraction process. After splitting my data into train and test data, I built Countervector, and TFIDF vector for each tokenized and lemmatized words. After all the features were generated, I focused on building a few different Naïve Bayes Classifiers such as Bernouli, Multinomial as well as Logistic Regression using sklearn library. A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task.

Model Evaluation

Since our problem is a binary classification (positive or negative), I used a ROC AUC(Receiver Operator Characteristic Area Under the Curve) score to evaluate the model. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR). After testing out the test samples on all three classifiers, Logistic regression came out as the best algorithm in terms of performance accuracy with 85%.



Even though looked at Precision, Recall and F-1 Score for my classifiers, I didn't use it in model my evaluation. Since accuracy is simply a ratio of correctly predicted observation to the total observations, is the most intuitive performance measure metric in this case.

Limitations/Challenges

As I predicted before, I did run into some computing with large amount of data. Even after only using 10 percent of my sample data, I was not able to use planned Long Short-term Memory – LTSM modeling technique. Even though LTSM could have been a better technique for my project, I was still able to use other binary classification modeling techniques. Since this was one of my first sentiment analysis & text classification project, I had somewhat a challenging time with pre-processing steps.

Future Uses/Additional Applications

In the future, on projects like this, I would like to make LTSM work. Also, if I will be continuing to learn more about pre-processing steps and techniques to make the models better. If I were to redo this specific project, I would even further under sample my dataset to have more balanced data.

Ethical Assessment

Every time there is customer data involved, it's sensitive in nature and there is a number of ethical considerations, laws & regulations to consider. I know I have a reviewer name column in my dataset, but since I don't have any other metadata that is specific to the customer, it was ok, as I will be able to tie the data back to the specific customer.

References:

Ni, J. (2018). *Amazon Review Data (2018)*. Amazon review data. Retrieved January 30, 2022, from https://nijianmo.github.io/amazon/index.html

“Amazon Review - Machine Learning Project¶.” Amazon-Review-Classification, https://t-lanigan.github.io/amazon-review-classifier/.

Bhattacharyya, S., Goled, S., Bhorayal, R., & Choudhary, A. (2021, February 2). How to implement LSTM RNN network for sentiment analysis. Analytics India Magazine. Retrieved January 30, 2022, from https://analyticsindiamag.com/how-to-implement-lstm-rnn-network-for-sentiment-analysis/

Roy, A. (2020, July 12). A guide to text classification and sentiment analysis. Medium. Retrieved January 30, 2022, from https://towardsdatascience.com/a-guide-to-text-classification-and-sentiment-analysis-2ab021796317

Wikimedia Foundation. (2022, February 6). Amazon (company). Wikipedia. Retrieved February 7, 2022, from https://en.wikipedia.org/wiki/Amazon_(company)