Capstone 2 – Final Report

Conducting NLP on Amazon Store Review Data

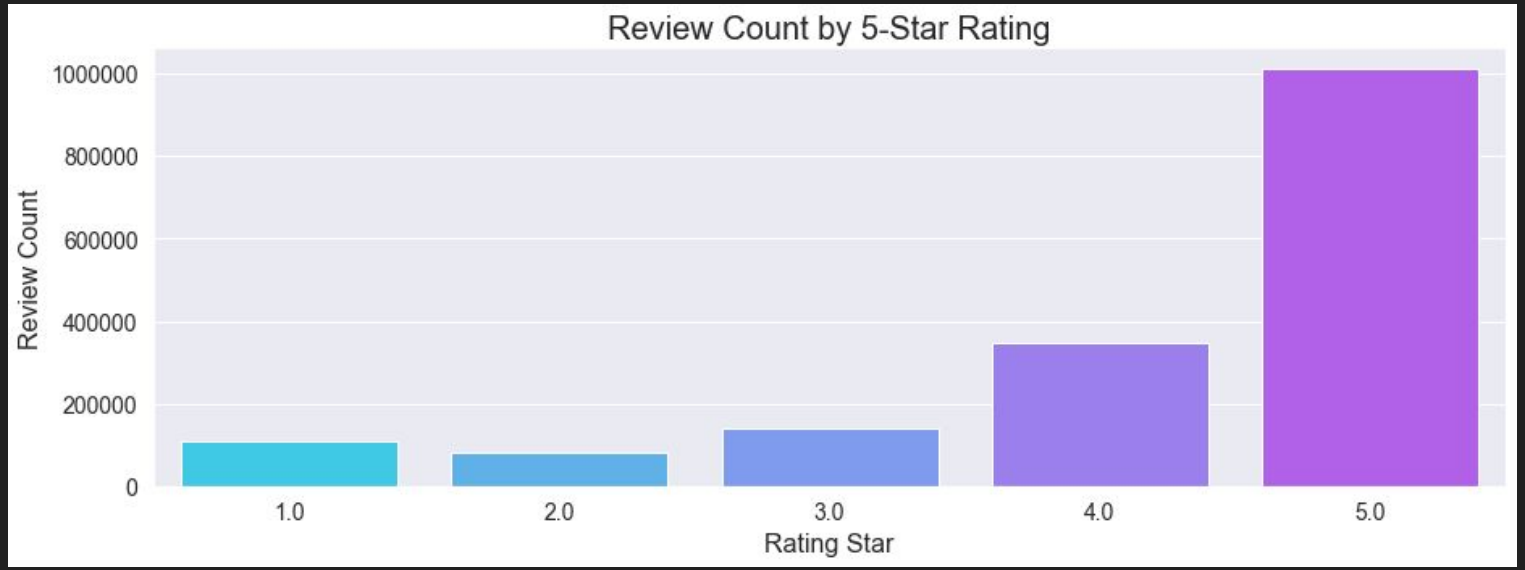
Author: Zach Palamara

Project Aim and Background

- Goal was to develop Machine learning models using Natural Language Processing (NLP)
- Used Amazon store review data to predict review sentiment
    - Dataset from “Electronics”

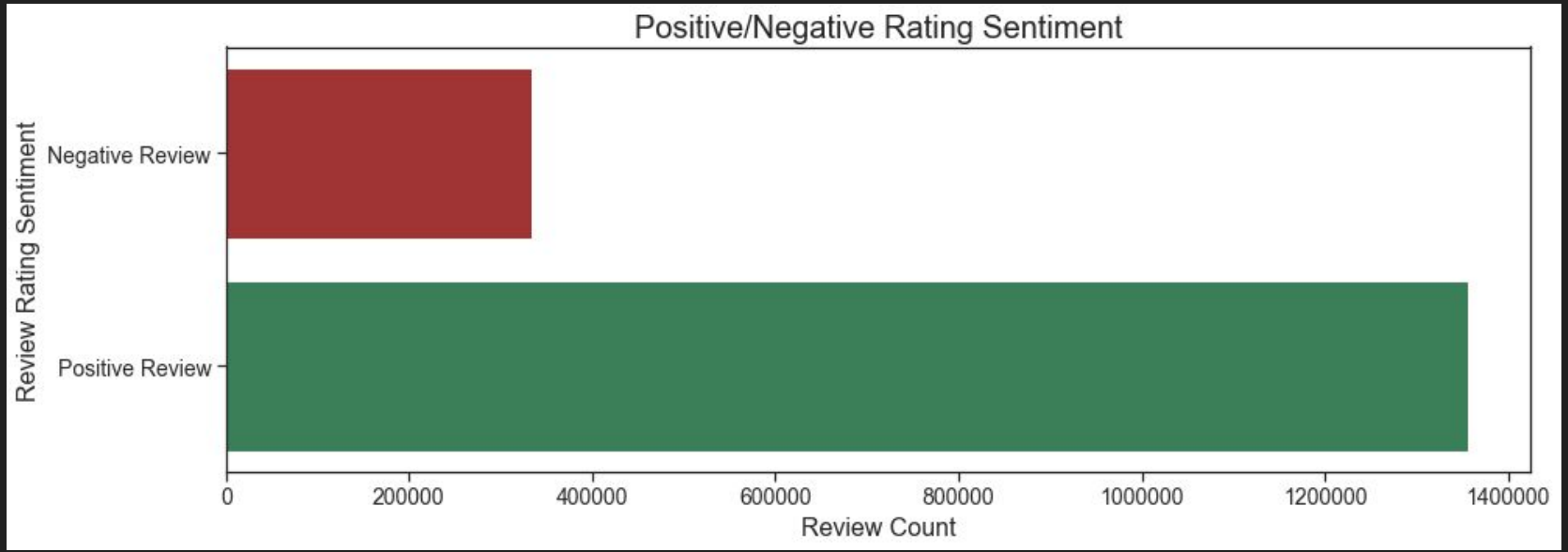
Initial Findings

- The first step I took in examining the data was to look at a simple count plot of reviews broken out by their respective 5-star rating.



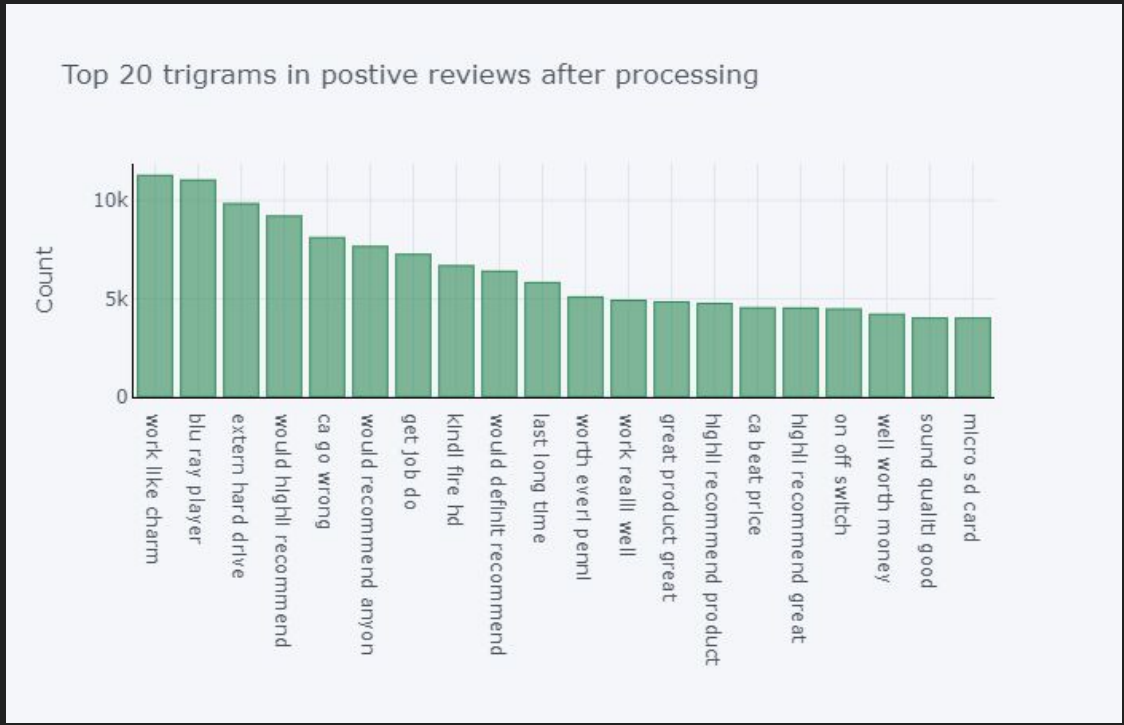
Review Sentiment

- I created a new binary target feature, which represents 1,2 and 3 star reviews as 0 or “Negative” and 4 and 5 star reviews as 1 or “Positive”.

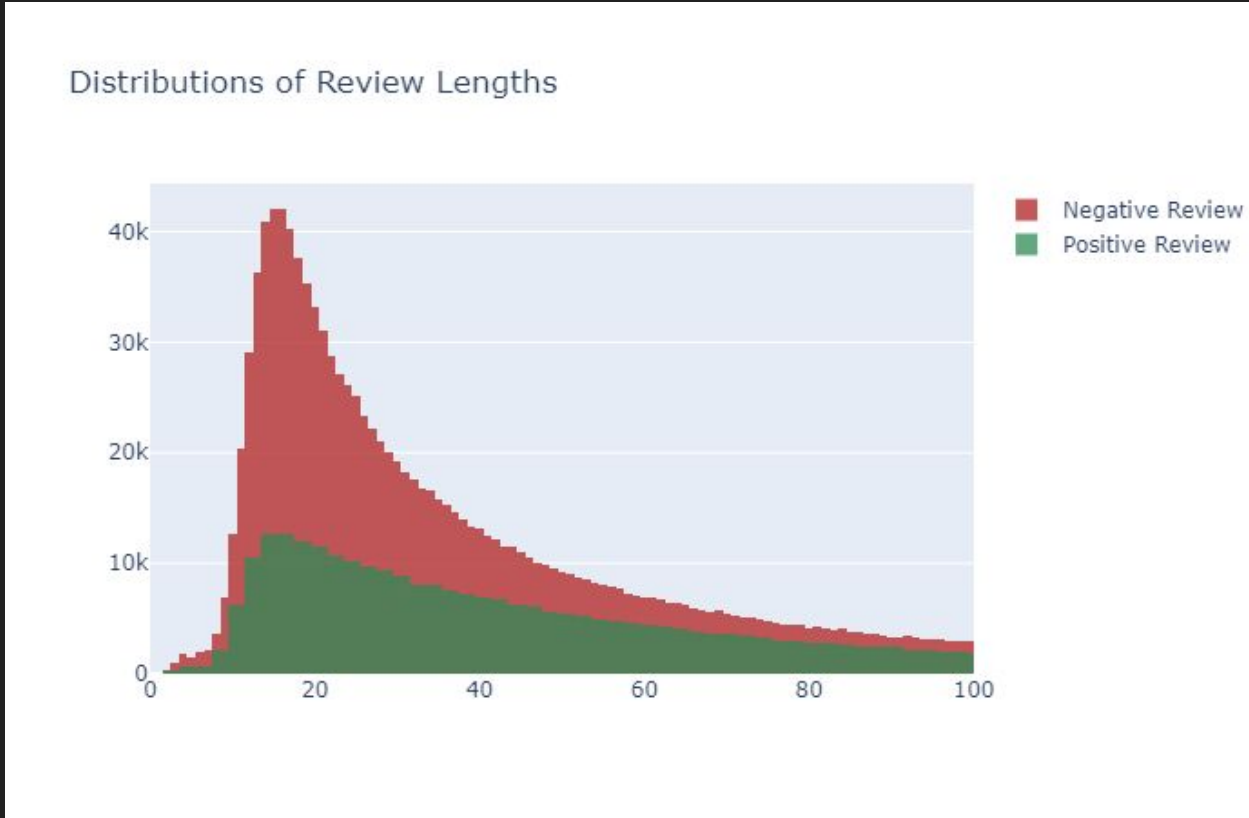


Text Feature Engineering

1. Remove Stop Words
2. Tokenization
    a. CountVectorizer()
3. Lemmatization
4. Stemming
5. N-Gram Features



Review Length

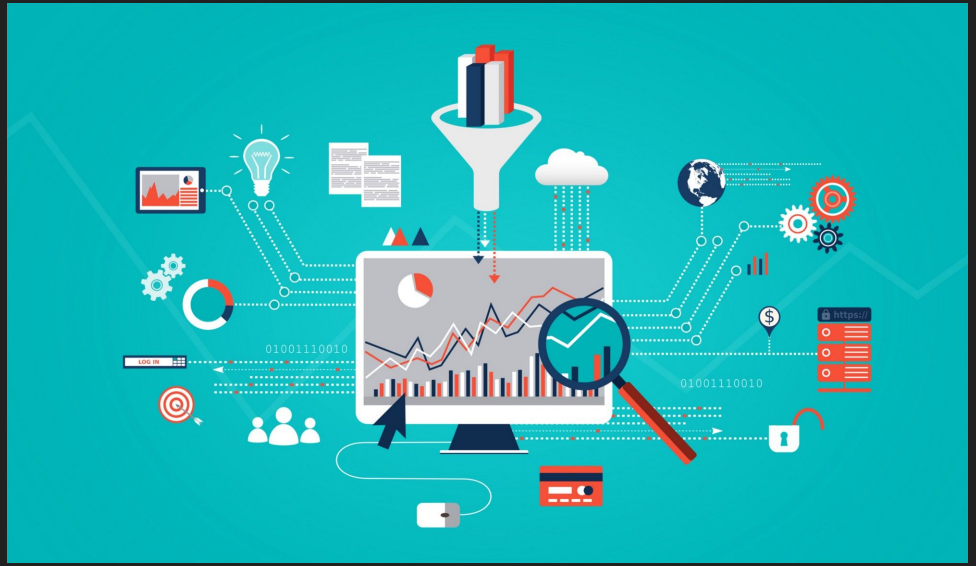


Word Cloud

Based on to 20 Bigrams

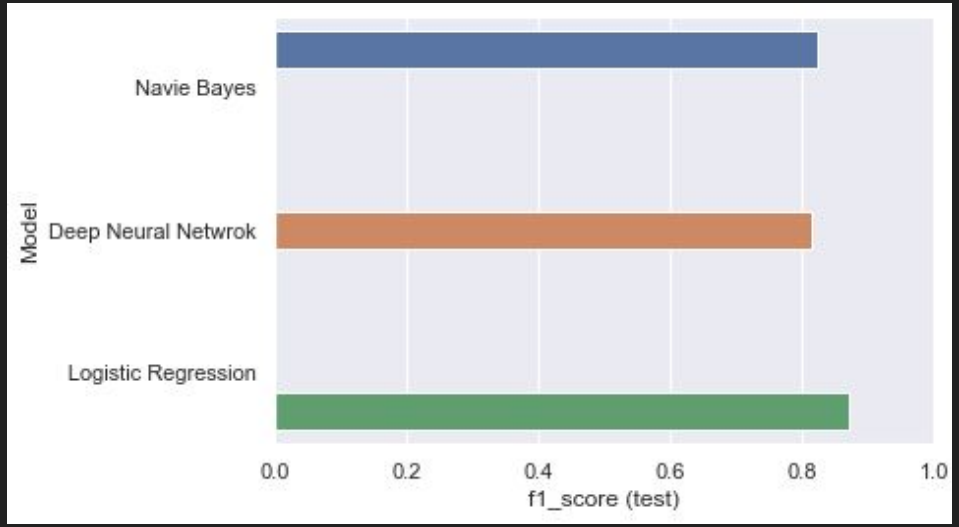
Model Preprocessing

1. Bootstrapping Samples
2. Shuffling Dataset
3. Text Feature Vectorization
4. Split Training and Test Set



Machine Learning

Model	F1 - Score	Wall Time
Naive Bayes	0.8225	1min 26s
Deep Neural Network	0.8134	~ 24 hrs (for full data set)
Logistic Regression	0.8721	14 min 46s



Conclusion

- Based on the results, it seems like Logistic Regression may be the best classifier for this particular type of review data.
- Not only did it have a much better F1-Score than both the Naive Bayes model and DNN, it took a reasonable time to fit the model.
- However, the results for the DNN are based on only a 3% sample of the entire dataset.

Considerations for Future Development

- Moving forward I would like to be able to train the DNN on all the review data, but since the dataset is so large I will likely need a GPU to accelerate the training process. It’s quite possible that the DNN may perform better after training on all of the data.

- Additionally, I would like to develop models utilizing the numerical features within this dataset.