Sentiment Analysis of Amazon Reviews

Project Report – A Capstone Element for Springboard’s Data Science program

Table of Contents

[1. Introduction 3](#_Toc123186923)

[1.1 Objective 3](#_Toc123186924)

[1.2 Significance 3](#_Toc123186925)

[2. Data 4](#_Toc123186926)

[2.1 Amazon Reviews for Sentiment Analysis 4](#_Toc123186927)

[3. Data Cleaning and Wrangling 5](#_Toc123186928)

[3.1 Data Extraction 5](#_Toc123186929)

[3.2 Data Cleaning 5](#_Toc123186930)

[4. Data Modeling 6](#_Toc123186931)

[4.1 Multinomial NB 6](#_Toc123186932)

[4.2 Logistic Regression 6](#_Toc123186933)

[4.3 BERT 6](#_Toc123186934)

[5. Conclusion and Recommendations 7](#_Toc123186935)

[Works Cited 8](#_Toc123186936)

# 1. Introduction

Amazon.com, Inc is an American multinational technology company focusing on e-commerce, cloud computing, online advertising, digital streaming, and artificial intelligence. (Wikipedia, 2022) The company allows users to leave reviews of products that they have purchased through the website, which are placed on a 1 to 5 scale of stars. .

## Objective

The objective of this analysis is to:

* Better understand what keywords may affect the probability of a certain star count
* Provide an understanding of how Amazon reviews may be biased towards a higher or lower ranking
* Create a model to predict the number of stars a review was given based off the textual review itself.

The code for this project can be found [here](https://github.com/evanwmeeks/Springboard/blob/78af10b9457709710223bd9b901e9bb952708d48/Capstone%203/Amazon_Reviews_for_Sentiment_Analysis.ipynb)

## 1.2 Significance

By thoroughly understanding the dataset, we will identify elements that affect the star rating of a review and can then be used to predict how many starts a reviewer gave. This NLP model can then be used to predict what future review ratings may be based off of other websites or word-of-mouth.

# 2. Data

## 2.1 Amazon Reviews for Sentiment Analysis

This dataset was posted to Kaggle by Adam Bittlingmayer to Kaggle, an online Data Science / Machine Learning community and dataset repository. The original dataset can be found [here](https://www.kaggle.com/datasets/bittlingmayer/amazonreviews)

The dataset contained two files: “test.ft.txt.bz2” and “train.ft.txt.bz2”. The reviews contained in each file have been prelabeled as “ \_\_label\_\_1” for 1- and 2-star reviews, and “\_\_label\_\_2” for 4- and 5-star reviews. 3-star reviews were omitted from the data collection phase as they represent a neutral data point. The training dataset contains 3.6 million reviews, with the test dataset containing 400,000.

# 3. Data Cleaning and Wrangling

The purpose of Data Cleaning and wrangling steps are:

- To ensure that all features are of the correct format and type.

- To ensure unnecessary elements like stop words, punctuation and URLS are removed.

- To prepare the dataset for natural language processing.

## 3.1 Data Extraction

The original files were compressed using the BZip2 compression algorithm. The files had to be read into the notebook using python’s BZIP package. The files were then read in line-by-line, decoded, and split into their labels and reviews. Due to the sheer size of this dataset, a method was built to subset the data to a smaller size for easier/faster processing, a check to export the processed data for faster retrieval, and a check to verify the existence of exported data before reprocessing the entire set. A custom CSV reader was also built to allow for quick reimportation of exported data.

## 3.2 Data Cleaning

With the nature of Natural Language Processing, data cleaning is primarily focused on the removal of stop words. Stop words represent the most common words in any language (articles, prepositions, pronouns, conjunctions, etc.) that add little to no value to the text. These are removed to provide higher focus to more important information while simultaneously decreasing file size, allowing for faster processing and training.

Hyperlinks, whitespace, and numbers provide a similarly low amount of value to the data and are subsequently removed.

# 4. Data Modeling

Once cleaned, the data was then modeled using Multinomial Naive Bayes algorithm, Logistic Regression and BERT .

## 4.1 Multinomial NB

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output. (Shriram, 2022)

The Multinomial algorithim was implemented and trained using the cleaned training data and produced an accuracy of 84.9%.

## 4.2 Logistic Regression

Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, (IBM, 2022 )

After training, Logistic Regression produced an accuracy of 90.0%

## 4.3 BERT

BERT and other Transformer encoder architectures have been wildly successful on a variety of tasks in NLP (natural language processing). They compute vector-space representations of natural language that are suitable for use in deep learning models. The BERT family of models uses the Transformer encoder architecture to process each token of input text in the full context of all tokens before and after, hence the name: Bidirectional Encoder Representations from Transformers. (TensorFlow, 2022)

For this project, the training data itself was first split into train and test sets using SciKit Learn’s train\_test\_split. From there, the BERT model’s preprocessing elements and layers were built before being fit to the model.

Because of the sheer amount of data, it was impossible to process the entire dataset through BERT. A Subset of only 10,000 records was selected, and resulted in an accuracy score of 76.3%

# 5. Conclusion and Recommendations

In this project, we examined the sentiment of reviews across Amazon.com’s website to predict whether the review was positive or negative in nature. Through this analysis, a model was built that accurately predicted the sentiment 90% of the time. Using this model, it is possible to predict future reviews to better understand trends in regard to specific products, categories, or the retail market as a whole. The next recommended steps would be:

1. Further train the BERT model on a larger subset of the data to verify the model’s accuracy and prevent overfitting.
2. Develop the most successful model into a deployable app that can be used by nontechnical team members to analyze reviews.
3. Consider retraining on data that labels each review with its actual star count so that a model can be trained to accurately predict the individual review’s star count, not just whether the review was positive or negative.

# Works Cited

IBM. (2022 , 12 28). *What is logistic regression?* Retrieved from IBM: https://www.ibm.com/topics/logistic-regression

Shriram. (2022, 10 3). *Multinomial Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2023*. Retrieved from UpGrad: https://www.upgrad.com/blog/multinomial-naive-bayes-explained/

TensorFlow. (2022, 12 28). *Classify text with BERT* . Retrieved from TensorFlow: https://www.tensorflow.org/text/tutorials/classify\_text\_with\_bert

Wikipedia. (2022, December 26). *Amazon (company)*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Amazon\_(company)#Product\_reviews