Final Project

IE7500 Statistical Learning for Engineering

A sentiment analysis study for Amazon reviews

Tran, Dat – tran.dat@northeastern.edu

Bai, Xitong - bai.xit@northeastern.edu

Abstract

In this project, we apply sentiment analysis to examine customer reviews for Amazon products across various categories, including groceries, books, electronics, and more. Our goal is to understand customer satisfaction with Amazon's products. We extract textual features from our dataset's corpus to serve as input for two tasks: a classification task and a regression task.

For the classification task, we predict whether a review is positive or negative. For the regression task, we predict the rating associated with each review. Our models demonstrate robust performance, achieving 90% accuracy for the binary classification task and a mean absolute error of 1 for the rating prediction, where the rating ranges from 1 to 5.

Introduction

Overview:

This study focuses on customer reviews with the aim of assessing sentiment and predicting ratings to aid in product enhancement and customer service improvement at Amazon. Our approach involves exploring sentiment analysis tasks to address two key objectives:

1. Determining whether a textual review is classified as positive or negative. For this purpose, reviews with ratings of 1-3 stars are categorized as negative, while those with ratings of 4-5 stars are considered positive.
2. Predicting the rating of a review, which ranges from 1 to 5.

The overarching goal of this project is to evaluate customer satisfaction levels with Amazon's offerings. By providing valuable insights into sentiment and ratings, we aim to assist Amazon and other stakeholders in enhancing product quality and delivering an improved customer experience.

Motivation:

Our motivations include gaining practical experience in data analysis and machine learning techniques, developing skills in natural language processing and sentiment analysis, and contributing to real-world applications that have tangible impacts on business operations. Additionally, we are driven by the opportunity to apply theoretical knowledge gained in coursework to a meaningful project with the potential to address practical challenges.

Our Approach Outline:

In this project, we have chosen to employ a combination of Naive Bayes and logistic regression approaches for sentiment analysis due to their effectiveness in handling text data and classification tasks. Naive Bayes is well-suited for sentiment analysis tasks due to its simplicity and ability to handle large feature spaces efficiently, making it a suitable choice for processing textual data. Logistic regression, on the other hand, provides a probabilistic framework for binary classification tasks, allowing us to model the relationship between features and sentiment labels with flexibility. By utilizing both approaches, we aim to compare their performance and explore the strengths and limitations of each method in the context of sentiment analysis for customer reviews.

Cite Corpus:

Amazon Reviews 2023. (2023). Amazon Reviews 2023 dataset [Dataset]. Retrieved from Dataset Link

Background

The Amazon review dataset is a big dataset consisting of many categories of products.

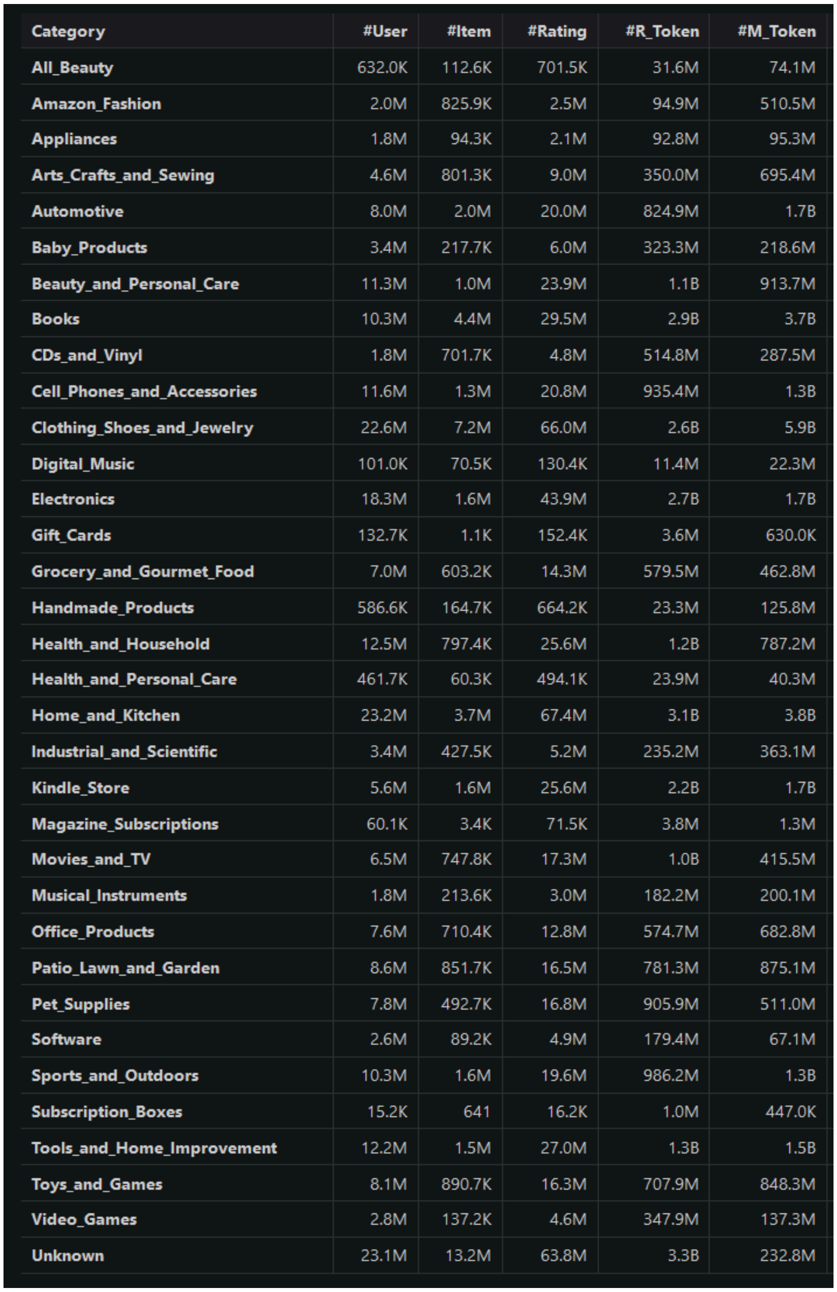
Upendra Singh et al. (2022) [1] have made an analysis project using various combinations of voice components and deep learning. The suggested module focuses on identifying sentences as ‵Positive‵, ‵Neutral‵, ‵Negative‵, or ‵Indifferent‵. Their automation achieves a maximum validation accuracy of 79.83% when using Fast Text as word embedding and the Multi-channel Convolution Neural Network.

In another work Tan et al. [2] explored a variety of methodologies and advanced deep learning neural network architectures such as RNN (Recurrent Neural Network), and transformer-based models such as BERT and GPT.

Approach

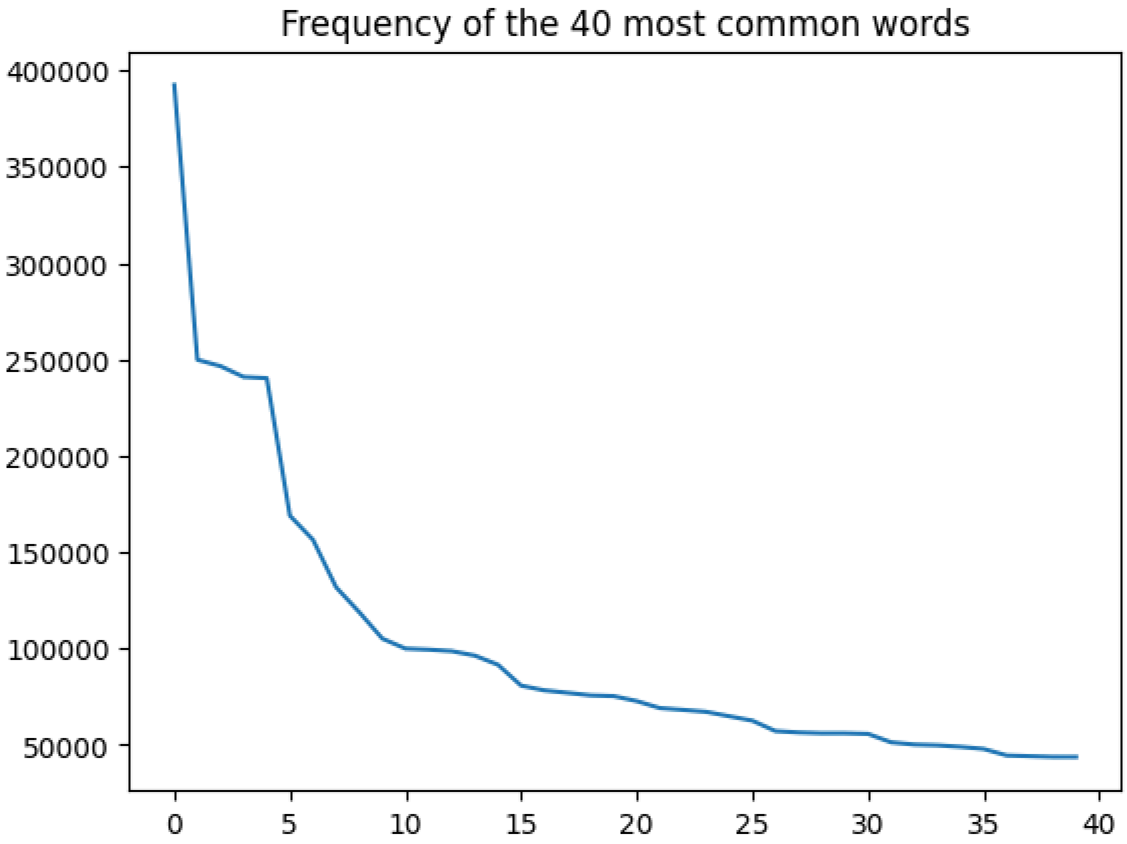
Data Preprocessing

The dataset contains reviews spanning from 2000 to 2023. The review content is packed into many categories as shown below.



Category	#User	#Item	#Rating	#R_Token	#M_Token
All_Beauty	632.0K	112.6K	701.5K	31.6M	74.1M
Amazon_Fashion	2.0M	825.9K	2.5M	94.9M	510.5M
Appliances	1.8M	94.3K	2.1M	92.8M	95.3M
Arts_Crafts_and_Sewing	4.6M	801.3K	9.0M	350.0M	695.4M
Automotive	8.0M	2.0M	20.0M	824.9M	1.7B
Baby_Products	3.4M	217.7K	6.0M	323.3M	218.6M
Beauty_and_Personal_Care	11.3M	1.0M	23.9M	1.1B	913.7M
Books	10.3M	4.4M	29.5M	2.9B	3.7B
CDs_and_Vinyl	1.8M	701.7K	4.8M	514.8M	287.5M
Cell_Phones_and_Accessories	11.6M	1.3M	20.8M	935.4M	1.3B
Clothing_Shoes_and_Jewelry	22.6M	7.2M	66.0M	2.6B	5.9B
Digital_Music	101.0K	70.5K	130.4K	11.4M	22.3M
Electronics	18.3M	1.6M	43.9M	2.7B	1.7B
Gift_Cards	132.7K	1.1K	152.4K	3.6M	630.0K
Grocery_and_Gourmet_Food	7.0M	603.2K	14.3M	579.5M	462.8M
Handmade_Products	586.6K	164.7K	664.2K	23.3M	125.8M
Health_and_Household	12.5M	797.4K	25.6M	1.2B	787.2M
Health_and_Personal_Care	461.7K	60.3K	494.1K	23.9M	40.3M
Home_and_Kitchen	23.2M	3.7M	67.4M	3.1B	3.8B
Industrial_and_Scientific	3.4M	427.5K	5.2M	235.2M	363.1M
Kindle_Store	5.6M	1.6M	25.6M	2.2B	1.7B
Magazine_Subscriptions	60.1K	3.4K	71.5K	3.8M	1.3M
Movies_and_TV	6.5M	747.8K	17.3M	1.0B	415.5M
Musical_Instruments	1.8M	213.6K	3.0M	182.2M	200.1M
Office_Products	7.6M	710.4K	12.8M	574.7M	682.8M
Patio_Lawn_and_Garden	8.6M	851.7K	16.5M	781.3M	875.1M
Pet_Supplies	7.8M	492.7K	16.8M	905.9M	511.0M
Software	2.6M	89.2K	4.9M	179.4M	67.1M
Sports_and_Outdoors	10.3M	1.6M	19.6M	986.2M	1.3B
Subscription_Boxes	15.2K	641	16.2K	1.0M	447.0K
Tools_and_Home_Improvement	12.2M	1.5M	27.0M	1.3B	1.5B
Toys_and_Games	8.1M	890.7K	16.3M	707.9M	848.3M
Video_Games	2.8M	137.2K	4.6M	347.9M	137.3M
Unknown	23.1M	13.2M	63.8M	3.3B	232.8M

In this project, we perform sentiment analysis on 5 diverse categories: All Beauty, Appliances, Handmade Products, Industrial and Scientific, and Musical Instruments. These categories were selected because they represent a wide range of product types and cover a significant portion of the available categories within our dataset, which comprises 30 categories in total. By analyzing sentiment across these diverse categories, we aim to gain insights that are applicable across various product domains and provide valuable feedback to improve customer satisfaction and product quality.



Term frequency from the corpus of All Beauty category with 500000 reviews

Originally, we thought about making a general model. But it is not feasible to do so, because the amount of data is too large and we just cannot train the model as we run out of memory. Also, a specialized model for each category will make more sense and have better performance than a general model.

We limit the data size used to train each model to only 500k reviews per category, which contains the latest reviews by date. We also filter to only verified buyers, for authentic review.

Then we run the textual data through the following pipeline:

- Tokenization
- Filtering out stopwords
- Lemmatization

Our textual data consists of 2 features: title and body text. Since title is usually indicative of the classification and rating, we apply a weight to the title feature. And using this weight, we combine both textual content of "title" and "text" into a single text feature "combined".

We now have a processed corpus ready for us to extract features from.

Feature extraction

We use term frequency (TF) (i.e. Bag of Words) and term frequency-inverse document frequency (TF-IDF) as input for our models.

Binary classification

We define positive class as reviews with 4-5 stars and negative class as reviews with 1-3 stars. Then we solve the problem as a binary classification task.

Models used: Logistic regression, Naive Bayes.

For each model, we run it with both the TF and TF-IDF features

Regression

We predict the star rating of the new review.

Model used: Linear Regression, Recurrent Neural Network.

We run the Linear Regression model with the TF-IDF feature.

For the RNN model, we need to transform the text data into embeddings of dense vector representations, then feed these embeddings into the LSTM layers. The LSTM outputs are passed through a fully-connected layer, then a dropout layer for regularization, and finally a single output for value prediction. The detail design is as following:

- Embedding layer
- 64 LSTMs layer
- 32 LSTMs layer
- 32 Dense layer, ReLU activation
- Dropout layer, p = 0.5
- Output layer

The model is compiled and trained with the following configuration

- Optimizer: Adam
- Loss function: Mean squared error
- Stop early with the loss of validation set does not improve for 3 consecutive steps.
- Batch size: 32
- Trained for 3 epochs

Result

Binary classification

The training process takes a few seconds.

We evaluate the binary classification task using the mean accuracy score, which is the average f1-score across the 2 classes.

Category	Naive Bayes		Logistic Regression	
	TF	TF-IDF	TF	TF-IDF
All Beauty	0.88	0.86	0.89	0.89
Appliances	0.89	0.87	0.90	0.90
Handmade Products	0.92	0.91	0.94	0.94
Industrial and Scientific	0.88	0.85	0.89	0.89
Musical Instruments	0.88	0.86	0.90	0.9

Regression

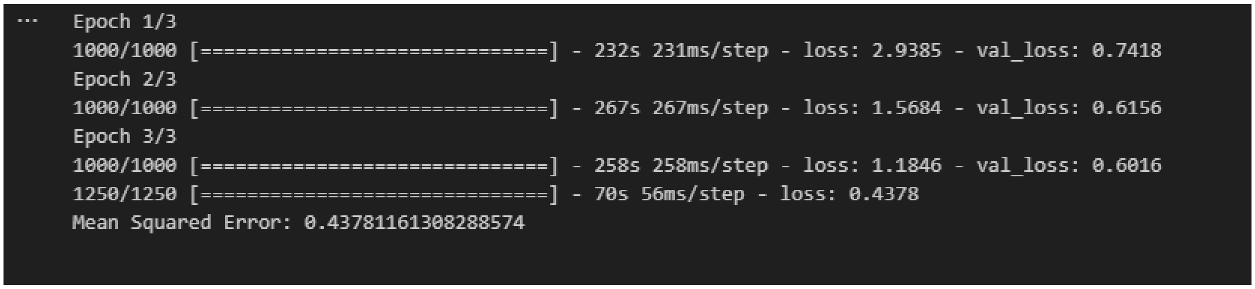
The training process takes a long time, compared to the classification task.

For linear regression, it takes 1 to 1.5 minutes.

For the regression task, we use root mean square error and mean absolute error to measure the performance of the Linear Regression model.

Category	RMSE	MAE
All Beauty	1.1203	0.7677
Appliances	1.167	0.764
Handmade Products	0.8359	0.5083
Industrial and Scientific	1.2081	0.803
Musical Instruments	1.1759	0.7537

The recurrent neural network model takes a very long time to train. We train the model in only 1 category: Handmade Products and only take 50000 review records instead of 500000 like other models. Even then, it still takes 13 min to train the model for only 3 epochs.



```
...   Epoch 1/3
      1000/1000 [==============================] - 232s 231ms/step - loss: 2.9385 - val_loss: 0.7418
      Epoch 2/3
      1000/1000 [==============================] - 267s 267ms/step - loss: 1.5684 - val_loss: 0.6156
      Epoch 3/3
      1000/1000 [==============================] - 258s 258ms/step - loss: 1.1846 - val_loss: 0.6016
      1250/1250 [==============================] - 70s 56ms/step - loss: 0.4378
      Mean Squared Error: 0.43781161308288574
```

It can be seen that the loss function has not converged yet. But the model has already achieved lower loss than the linear regression model.

It achieves a RMSE of 0.7609 and a MAE of 0.5551 on the test set.

We can anticipate that the RNN model can have a much better performance with more training time.

Conclusion

This sentiment analysis study on Amazon reviews has provided valuable insights into customer satisfaction across various product categories. By employing a combination of Naive Bayes and logistic regression models, we have successfully classified customer reviews into positive and negative sentiments and predicted their corresponding ratings with high accuracy.

The study’s results demonstrate the effectiveness of machine learning techniques in processing and analyzing large volumes of textual data. The use of term frequency and term frequency-inverse document frequency as features has proven to be effective in capturing the essence of customer sentiments.

It can be concluded that the binary classification task can be trained in a very short time for a large amount of data, in relation to the regression task, which requires a much longer time to train. Simple learning models like Naive Bayes and Logistic Regression have proven to be effective for the task.

In conclusion, the project has achieved its objectives of assessing customer sentiment and predicting ratings, thereby offering actionable insights for Amazon to enhance its offerings. Future work could explore the integration of more sophisticated deep learning models and the inclusion of additional features such as semantic analysis to further refine the accuracy of sentiment classification and rating predictions.

Reference

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9764295/

“Towards improving e-commerce customer review analysis for sentiment detection”, Upendra Singh, Anant Saraswat, Hiteshwar Kumar Azad, Kumar Abhishek, and S Shitharth

[2] http://cs229.stanford.edu/proj2018/report/122.pdf

Tan, Wanliang, Xinyu Wang, and Xinyu Xu. "Sentiment analysis for Amazon reviews." *International Conference*. 2018.