AIT 526
Project Report

Sentiment Analysis of Amazon Online Purchase Ratings

Team 2

1. Fangxin Zhang
2. Muhammad Hassan
3. Shirinithi Thiruppathi
4. Tewodros Tamene

Dr. Duoduo Liao
December 2, 2021

Table of Contents

Abstract

As more customers rely on online purchases, the importance of user ratings is growing. To identify which is the greatest product to buy, a consumer must read thousands of reviews of similar products prior to making a purchase, which can be time-consuming. Customers' Amazon reviews are chaotic and unstructured, exposing underlying patterns that, when correctly mined with the right analysis tools, can revolutionize a company's products and service development.

Sentiment Analysis is one of the techniques that can be implemented to make sense of all the rating data. This project will focus on Sentiment Analysis on all unstructured content leveraging Amazon's online purchase rating metadata. The project work is based on the Software category of the metadata to automatically categorize and make sense of the customer ratings. This enables Amazon to gain a better understanding of customer behavior and demand for its products and services.

The project task includes extensive text processing, exploratory analysis, time series analysis, and popularity-based recommender system. The Sentiment model utilizes the TF-IDF method and is built on classification machine learning algorithms such as Logistic Regression, Decision Tree, and KNN to compare performance metrics. Additionally, TextBlob is used to find the polarity measure of each rating. The model categorizes the customer ratings into positive, negative, and neutral which can be further analyzed to make recommendations.

Of the machine leaning sentiment analysis models, the Logistic Regression model gave the highest performance. Based on the findings of the project, narratives are created, and recommendations are made. This will help decision-makers and other stakeholders improve or change the business's direction as well as improve customer satisfaction with Amazon products and services.

Keywords: Amazon, Sentiment Analysis, Machine Learning, TF-IDF, Time Series Analysis

1 Introduction

1.1 Background

Understanding what customers' opinions and emotions are about a product or a purchase is an important aspect for a business to grow and evolve to retain a competitive edge in its respective industry. However, the sentiment and emotion that is shared by customers tend to generate an enormous amount of data that requires the use of advanced analytical tools to decipher and utilize. The concept of deciphering customers' opinions is known as Sentiment Analysis.

Sentiment Analysis is a Natural Language Processing (NLP) technique that is focused on analyzing text to identify and extract subjective information in source material. This helps a business, small or big, to understand and monitor the social sentiment of their brand, product or service and conduct in-depth market research based on sentiments of customers, while monitoring online conversations. Additionally, it can also aid in improving the overall development of the business.

The world of e-commerce has grown significantly over the years, and this project will focus on the company, Amazon, that has managed to dominate the e-commerce world and its Sentiment Analysis. Conducting Sentiment Analysis is, however, not an easy task since Amazon is an exceptionally large platform that generates an enormous amount of data from customer reviews. This raw data needs to pass through the necessary analytics processes to be mined and used as a potential business transformative resource. Proper analysis of these huge raw data results in an increase in profits and poor analysis will see a decline in both services and profits. Sentiment analysis and the use of NLP is one technology that can transform the raw sentiment data into something meaningful and usable. Additionally, Sentiment analysis is the most impactful when conducted over time, allowing businesses to analyze changes in sentiment in response to specific events, such as the global pandemic.

1.2 Related work

Several studies have been conducted on Amazon reviews in recent times and below are some examples.

- The piece of research article by (Haque, Saber, & Shah, 2018) has chosen 3 categories of product reviews; Electronics, Musicals, Cellphones & Accessories to calculate performance metrics such as; Accuracy, Precision, Recall and F1 Score based on supervised learning algorithms to determine which algorithm produces the greatest accuracy.
- The piece of research article by (Aljuhani & Alghamdi, 2019) has chosen the category Mobile Phones to conduct their research on Sentiment Analysis of Amazon reviews by using various feature extraction approaches such as Bag-of-words with (Bigram, Trigram), TF-IDF with (Unigram, Bigram, Trigram), word2vec, word2vec with Bigram, and glove. Different types of machine learning classifiers, such as Logistic Regression, Naive Bayes, Stochastic Gradient Decent and deep learning algorithms such as Convolutional Neural Networks (CNN) were also used to calculate the respective performance metrics.

- The piece of research article by (Bhatt, Patel, Chheda, & Gawande, 2015) focusses on developing an interactive dashboard that aids in visualizing the customer reviews in a graphical model.

1.3 Project Objective

The main objective of this project is to utilize NLP through Sentiment Analysis of Amazon's online purchase ratings dataset to find trends, anomalies, and other useful information to make product and business recommendations and to transform the overall service of the company regarding the Software category. To do this, feature extraction approaches such TextBlob and TF-IDF are used. Several types of classification algorithms such as Logistic Regression, Decision Tree, SVC, and Random Forest are also used to calculate and compare accuracy. Furthermore, time series analysis of the sentiment is also one aspect of the project. The project also aims to give recommendations to decision-makers and other stakeholders based on the findings of the Sentiment Analysis and other analysis techniques that will be conducted throughout the project work.

2 Data Acquisition

2.1 Overview

The main dataset that will be used for the project is the Amazon Metadata Product Dataset, obtained from GitHub. The dataset is massive with approximately 233.1 million reviews, in JSON format. It includes reviews from May 1996 to October 2018. The following figure shows the overall data content. Since this project focuses on the category Software as seen that it has more than 400 thousand reviews.

Amazon Fashion	reviews (883,636 reviews)	metadata (186,637 products)
All Beauty	reviews (371,345 reviews)	metadata (32,992 products)
Appliances	reviews (602,777 reviews)	metadata (30,459 products)
Arts, Crafts and Sewing	reviews (2,875,917 reviews)	metadata (303,426 products)
Automotive	reviews (7,990,166 reviews)	metadata (932,019 products)
Books	reviews (51,311,621 reviews)	metadata (2,935,525 products)
CDs and Vinyl	reviews (4,543,369 reviews)	metadata (544,442 products)
Cell Phones and Accessories	reviews (10,063,255 reviews)	metadata (590,269 products)
Clothing Shoes and Jewelry	reviews (32,292,099 reviews)	metadata (2,685,059 products)
Digital Music	reviews (1,584,082 reviews)	metadata (465,392 products)
Electronics	reviews (20,994,353 reviews)	metadata (786,868 products)
Gift Cards	reviews (147,194 reviews)	metadata (1,548 products)
Grocery and Gourmet Food	reviews (5,074,160 reviews)	metadata (287,209 products)
Home and Kitchen	reviews (21,928,568 reviews)	metadata (1,301,225 products)
Industrial and Scientific	reviews (1,758,333 reviews)	metadata (167,524 products)
Kindle Store	reviews (5,722,988 reviews)	metadata (493,859 products)
Luxury Beauty	reviews (574,628 reviews)	metadata (12,308 products)
Magazine Subscriptions	reviews (89,689 reviews)	metadata (3,493 products)
Movies and TV	reviews (8,765,568 reviews)	metadata (203,970 products)
Musical Instruments	reviews (1,512,530 reviews)	metadata (120,400 products)
Office Products	reviews (5,581,313 reviews)	metadata (315,644 products)
Patio, Lawn and Garden	reviews (5,236,058 reviews)	metadata (279,697 products)
Pet Supplies	reviews (6,542,483 reviews)	metadata (206,141 products)
Prime Pantry	reviews (471,614 reviews)	metadata (10,815 products)
Software	reviews (459,436 reviews)	metadata (26,815 products)
Sports and Outdoors	reviews (12,980,837 reviews)	metadata (962,876 products)
Tools and Home Improvement	reviews (9,015,203 reviews)	metadata (571,982 products)
Toys and Games	reviews (8,201,231 reviews)	metadata (634,414 products)
Video Games	reviews (2,565,349 reviews)	metadata (84,893 products)

Figure 1: Overview of Amazon purchase data

2.2 Data Attributes Descriptions

The Amazon Metadata Product Dataset consists of the following fields.

- ReviewerID - ID of the reviewer
- Asin - Product ID
- ReviewerName - Name of the respective reviewer
- Vote - Number of users who found the review made by the reviewer useful
- ReviewText - The review made by reviewer in the form of a text
- Overall - The rating of the product by the reviewer
- Summary - The summary of the review
- UnixReviewTime - Time of review made by reviewer in Unix format
- ReviewTime - Time of review made by reviewer in raw format
- Image - Images posted by reviewer upon receiving the product in hand

The following figure shows the data types of each field in the dataset.

```
overall           float64
verified             bool
reviewTime         object
reviewerID         object
asin               object
style              object
reviewerName       object
reviewText         object
summary            object
unixReviewTime      int64
vote               object
image              object
dtype: object
```

Figure 2: Variable data types

2.3 Other Data Sources

The records used in this study go from 1996 to 2018. Initially, we planned to incorporate a time series analysis of product reviews following the COVID-19 pandemic, spanning the years 2018 to 2021. The data, however, was not available. As a result, this project work has no other data sources besides the one mentioned in the previous section.

2.4 Project Approach

For this project to develop, we will have to follow the following steps.

1. Import the dataset
2. Preprocess and clean- Delete the record if review text or review score equal to null values, and remove the punctuation, stop words and so on.

3. Create 'sentiment' column - This is an important preprocessing phase; we are deciding the outcome column (sentiment of review) based on the overall score. If the score is greater than 3, we take that as positive and if the value is less than 3 it is negative If it is equal to 3, we take that as neutral sentiment.
4. Train-test splitting (75:25) - Using 75% of data as train data, and 25% data as test data.
5. Model Building: Sentiment Analysis - Model selection: select the best performing model. Firstly, using both TextBlob and TF-IDF to analyze the sentiment, considering all the classification algorithms including Logistic Regression, Decision Tree & KNN to check the performance, and choose a best way to analyze the sentiment.
6. Build the story generation and visualization from reviews - How the number of reviews changes by year, which product has the most positive reviews, how the review rate has changed by year, sentiment polarity distribution for the product.
7. Making recommendations and conclude based on findings.

2.5 Data Preprocessing

Preprocessing refers to the process of cleaning and organizing raw data in order to make it appropriate for model development. This procedure entails removing duplicate records from the dataset as well as dealing with missing values and encoding. Preprocessing processes such as removing punctuation, special characters, digits, whitespaces, and stopwords, which are typical in NLP projects, are also undertaken. In order to preserve consistency and complete the stemming process, all upper-case letters must be converted to lower-case ones. All the preprocessing tasks are covered in detail in the subsections that follow.

2.5.1 Overview of the Software data

The first step was to obtain the dataset based on the SOFTWARE category. The dataset is a json file and has several fields. The following figure shows the data frame created in python.

```
df.head()
```

	overall	verified	reviewTime	reviewerID	asin	style	reviewerName	reviewText	summary	unixReviewTime	vote	image
0	4.0	True	03 11, 2014	A240ORQ2LF9LUI	0077613252	{'Format:': ' Loose Leaf'}	Michelle W	The materials arrived early and were in excell...	Material Great	1394496000	NaN	NaN
1	4.0	True	02 23, 2014	A1YCCU0YRLS0FE	0077613252	{'Format:': ' Loose Leaf'}	Rosalind White Ames	I am really enjoying this book with the worksh...	Health	1393113600	NaN	NaN
2	1.0	True	02 17, 2014	A1BJHRQDYVAY2J	0077613252	{'Format:': ' Loose Leaf'}	Allan R. Baker	IF YOU ARE TAKING THIS CLASS DON"T WASTE YOUR ...	ARE YOU KIDING ME?	1392595200	7	NaN
3	3.0	True	02 17, 2014	APRDVZ6QBIQXT	0077613252	{'Format:': ' Loose Leaf'}	Lucy	This book was missing pages!!! Important pages...	missing pages!!	1392595200	3	NaN
4	5.0	False	10 14, 2013	A2JZTTBSLS1QXV	0077775473	NaN	Albert V.	I have used LearnSmart and can officially say ...	Best study product out there!	1381708800	NaN	NaN

Figure 3: Data frame head

Statistical description of variables in the data is also a good method that can be used for overview of the dataset. The following figure shows the summary of the statistical description of the 'overall' field in the data which contains the rating of each product.

```
count    459436.000000
mean          3.570175
std           1.626662
min           1.000000
25%           2.000000
50%           4.000000
75%           5.000000
max           5.000000
Name: overall, dtype: float64
```

Figure 4: Statistical description of the overall rating values

2.5.2 Handling Missing Values

When working with a dataset, one of the most important steps is to deal with missing values. There are some missing values, as can be seen in the image below, and dealing with the missing values is the first step. The fields with missing values are shown in the figure below. The variables 'vote', 'picture', and 'style' have the most missing values but that's acceptable since in this project, these variables are not necessary in conducting a Sentiment Analysis.

```
#Checking for null values

process_reviews.isnull().sum()
```

```
overall               0
verified              0
reviewTime            0
reviewerID            0
asin                  0
style            225035
reviewerName         24
reviewText           66
summary              56
unixReviewTime        0
vote             331583
image            457928
dtype: int64
```

Figure 5: Missing values

The records that have missing values for the 'reviewText' and 'summary' were dropped whereas for the remaining, the NaN value was replaced by 'Missing'. The following figure shows that all the missing values have been handled.

```
process_reviews.isnull().sum()
```

```
overall           0
verified          0
reviewTime        0
reviewerID        0
asin              0
style             0
reviewerName      0
reviewText        0
summary           0
unixReviewTime    0
vote              0
image             0
dtype: int64
```

Figure 6: Handled columns with no missing values

2.5.3 Text Preprocessing

Text preprocessing is an essential step of any sentiment analysis task. Removing punctuation Is the first step in this process. Next step is removing stop words. However, the general nltk stop words method cannot be implemented since removing words like 'not', 'wouldn't' and so on would lead to a misleading analysis. That is, if we remove such words, it will contradict the sentiment. Instead, creating a list of stop words that if removed wouldn't have an impact on contradicting the target variable were chosen.

In this project, content into numerical feature vectors using the Bag of Words strategy. To implement the Bag of Words strategy, we will use SciKit-Learn's CountVectorizer to perform the following:

- Text preprocessing:
- Tokenization (breaking sentences into words)
- Stopwords (filtering "the", "are", etc)
- Occurrence counting (builds a dictionary of features from integer indices with word occurrences)
- Feature Vector (converts the dictionary of text documents into a feature vector)

2.6 Exploratory Data Analysis

Exploratory data analysis is an important step in the data science development cycle. Exploratory data analysis is utilized to analyze data sets to summarize their main characteristics, by using visualizations and statistical methods. Visualizations at this step help to see the overall picture of the data and the target variable. Additionally, it is possible to see underlying facts of the sentiment people share on the SOFTWARE category before developing a sentiment analysis model. Some of the exploratory data analysis tasks that were performed in this project are shown below.

The following is the distribution of the overall sentiment column which ranges from 1 to 5. The largest score is 5 followed by 1 and 4.

```
process_reviews['overall'].value_counts()
```

```
5.0    212374
1.0    102528
4.0     73586
3.0     39390
2.0     31442
Name: overall, dtype: int64
```

Figure 7: Count of the overall rating values

The above information can be visualized as shown in the the following figure.

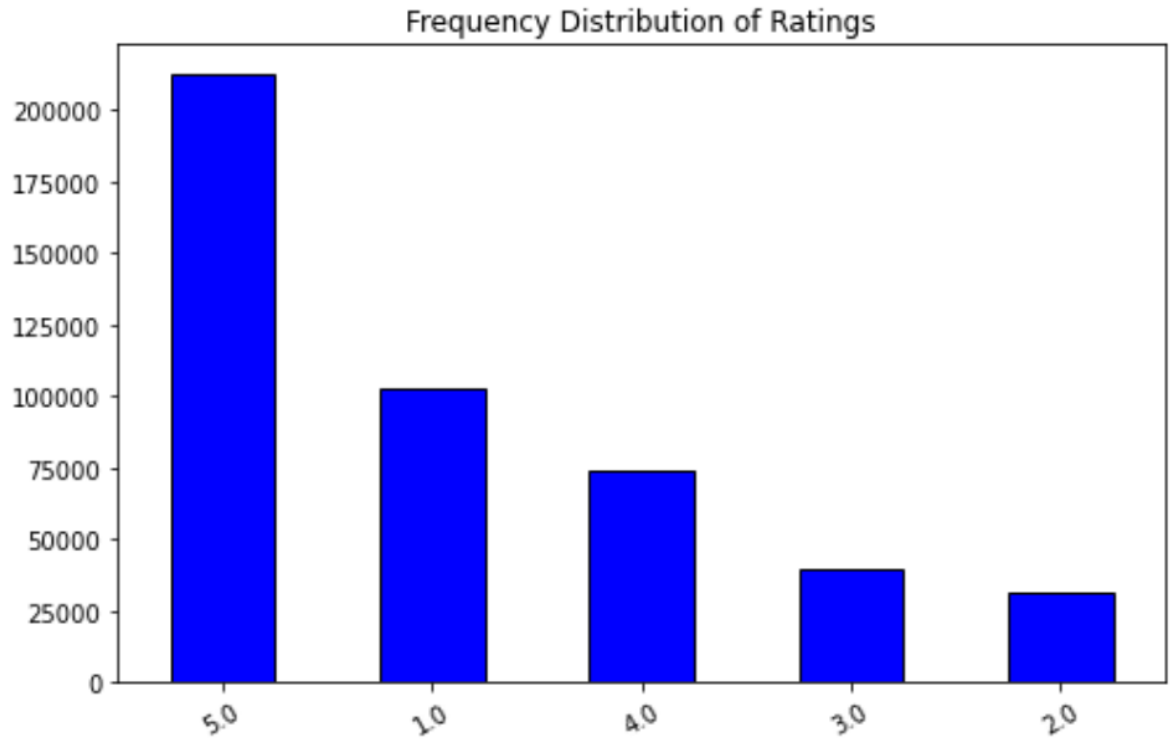

Figure 8: Frequency distribution of rating values

Performing further exploratory analysis, it is possible to see the uniqueness of some of the attributes in the dataset. For instance, the total number of ratings, reviews and products can be calculated. The following figure shows this.

```
Total data
-----------------------------

Total No of Ratings  : 459436
Total No of Reviwers : 375147
Total No of Products : 21663
```

Figure 9: Data characteristics

An analysis of which customers have given the most ratings can help to further explore the data. The following figure shows the top 5 customers and number of ratings they have given.

```
reviewerID
A5JLAU2ARJ0BO      73
A680RUE1FDO8B      71
A225G2TFM76GYX     69
A3W4D8XOGLWUN5     68
A15S4XW3CRISZ5     66
```

Figure 10: Top 5 reviewers

2.7 Feature Engineering

Feature engineering is the process of selecting and transforming variables when creating statistical or machine learning models. This process is essential when working with a large dataset such as the one to be used in this process. Feature engineering techniques are used to create fields that do not exist in the raw data but are necessary for the model development. For instance, a new attribute for categorizing the ratings will be needed for the model development step.

The first task in this section of the project is creating a new field called 'sentiment' which is the target variable for the classification task. The ratings have an overall score from 1 to 5. 1 to 2 is categorized as negative, 3 is categorized as neutral, and 4 to 5 is categorized as positive rating.

	overall	verified	reviewerID	asin	style	reviewerName	unixReviewTime	vote	image	reviews	sentiment	date	year
0	4.0	True	A240ORQ2LF9LUI	0077613252	{'Format:': ' Loose Leaf'}	Michelle W	1394496000	Missing	Missing	The materials arrived early and were in excell...	Positive	03 11	2014
1	4.0	True	A1YCCU0YRLS0FE	0077613252	{'Format:': ' Loose Leaf'}	Rosalind White Ames	1393113600	Missing	Missing	I am really enjoying this book with the worksh...	Positive	02 23	2014
2	1.0	True	A1BJHRQDYVAY2J	0077613252	{'Format:': ' Loose Leaf'}	Allan R. Baker	1392595200	7	Missing	IF YOU ARE TAKING THIS CLASS DON"T WASTE YOUR ...	Negative	02 17	2014
3	3.0	True	APRDVZ6QBIQXT	0077613252	{'Format:': ' Loose Leaf'}	Lucy	1392595200	3	Missing	This book was missing pages!!! Important pages...	Neutral	02 17	2014
4	5.0	False	A2JZTTBSLS1QXV	0077775473	Missing	Albert V.	1381708800	Missing	Missing	I have used LearnSmart and can officially say ...	Positive	10 14	2013

Figure 11: Data frame after creating sentiment column

The following figure shows the distribution of the sentiment column with the positive, neutral, and negative labels.

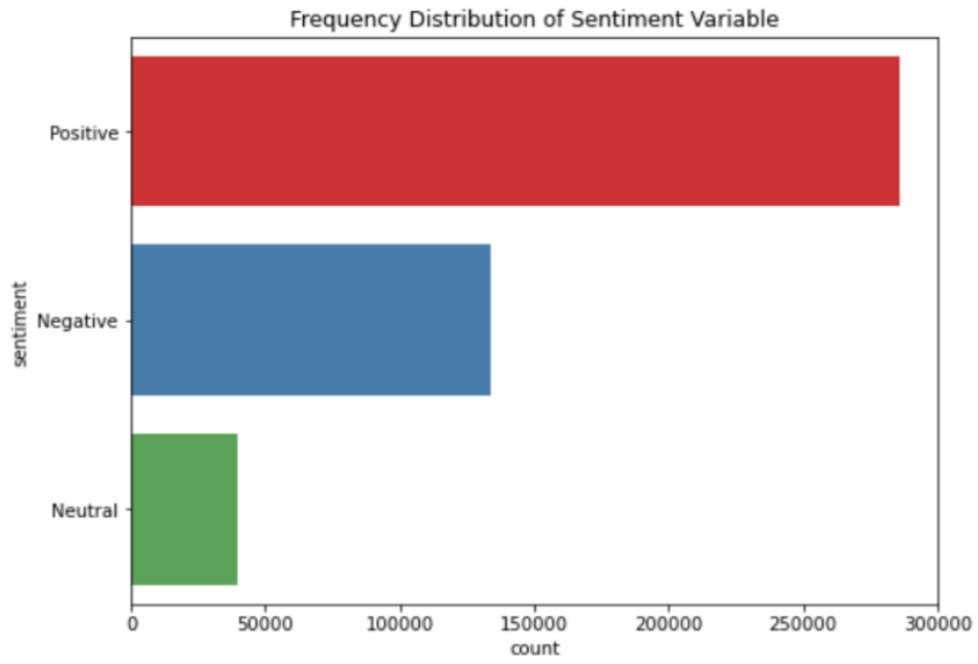

Figure 12: Frequency distribution of sentiment variable

3 Analytics and Model Development

The amazon review is unstructured and unorganized, so we need to use sentiment analysis to make sense of all this unstructured text by automatically tagging it, which helps to understand the customer behavior and needs on a company's products and services. Generally, the feedback provided by a customer on a product can be categorized into 3 ways; Positive, Negative, and Neutral. Interpreting customer feedback through product reviews helps companies evaluate how satisfied the customers are with their products. This method processes huge amounts of data in an efficient and cost-effective way.

3.1 Sentiment Analysis

For performing sentiment analysis, the NLTK package of the Python (Natural Language Toolkit, 2021) is used. NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum. And for us, it is a good way to access the Wordnet. WordNet is a large lexical database of English developed by Princeton University (Princeton University, 2021). Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms, each expressing a distinct concept.

A package named Text Blob is good to use to perform sentiment analysis (Loria, 2020). It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more

3.1.1 TextBlob

When calculating sentiment for a single word, TextBlob takes average for the entire text. For heteronym words, Textblob does not negotiate with different meanings. In other words, only the most common meaning of a word in the entire text is taken into consideration.

In this study, polarity is used as the main sentiment classifier. Polarity: Polarity is a float which lies in the range of [-1,1] where 1 refers to a positive statement and -1 refers to a negative statement.

Before developing the sentiment model, some features need to be engineered. It is important to create polarity, review length and word count features. Textblob is used to create the polarity feature which will be used to figure out the rate of sentiment. In this case it is between [-1,1] where -1 is negative and 1 is positive polarity. The review length feature is the length of the review which includes each letter and space, and word length is the measures how many words are there in review. The following figure shows all the engineered features.

reviews	sentiment	year	month	day	polarity	review_len	word_count
materials arrived early excellent condition ho...	Positive	2014	03	11	0.339744	120	16
really enjoying book worksheets make review go...	Positive	2014	02	23	0.250000	98	13
if taking class dont waste money called book b...	Negative	2014	02	17	-0.094231	176	29
book missing pages important pages couldnt ans...	Neutral	2014	02	17	0.100000	97	13
used learnsmart officially say amazing study t...	Positive	2013	10	14	0.333333	394	54

Figure 13: Added columns of polarity, review length, and word count

Having created the polarity, rating, text length, and text word count columns, we can see the sentiment polarity distribution.

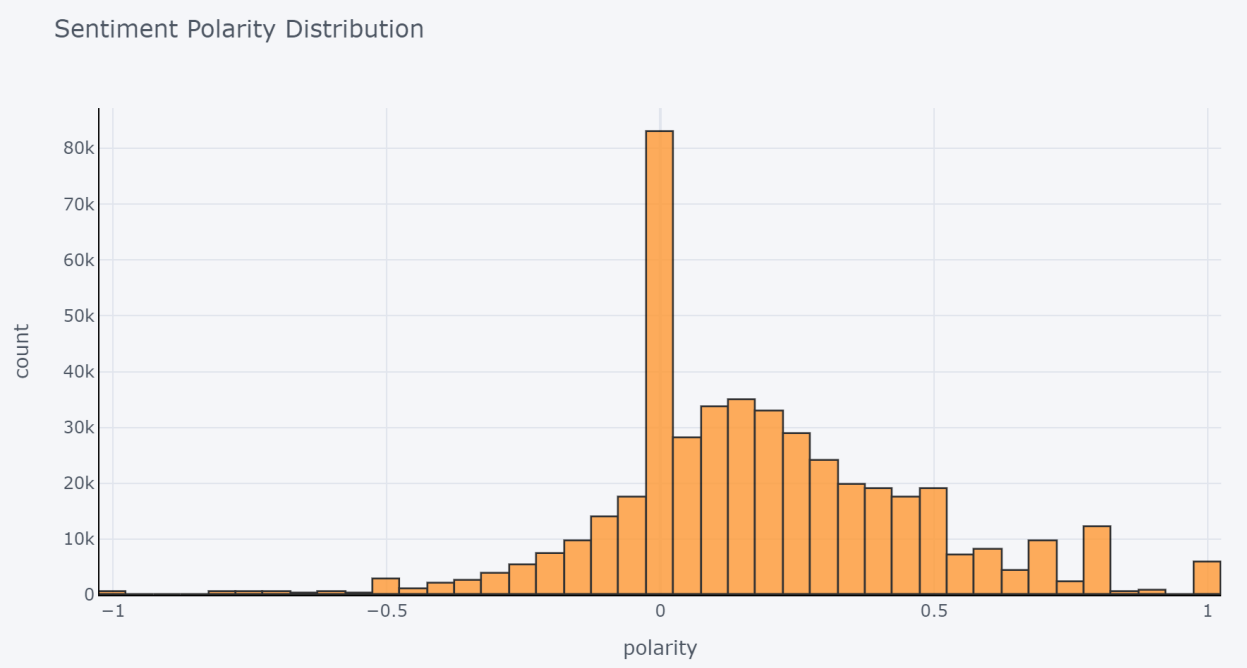

Figure 14: Sentiment polarity distribution

From the above, we can see that there are several positive polarities compared to the negative polarities. But the neutral polarity is also large. On the other hand, the polarity distributions assure the number of positive reviews, and the graph is normally distributed but not standard normal. The following is the review rating distribution.

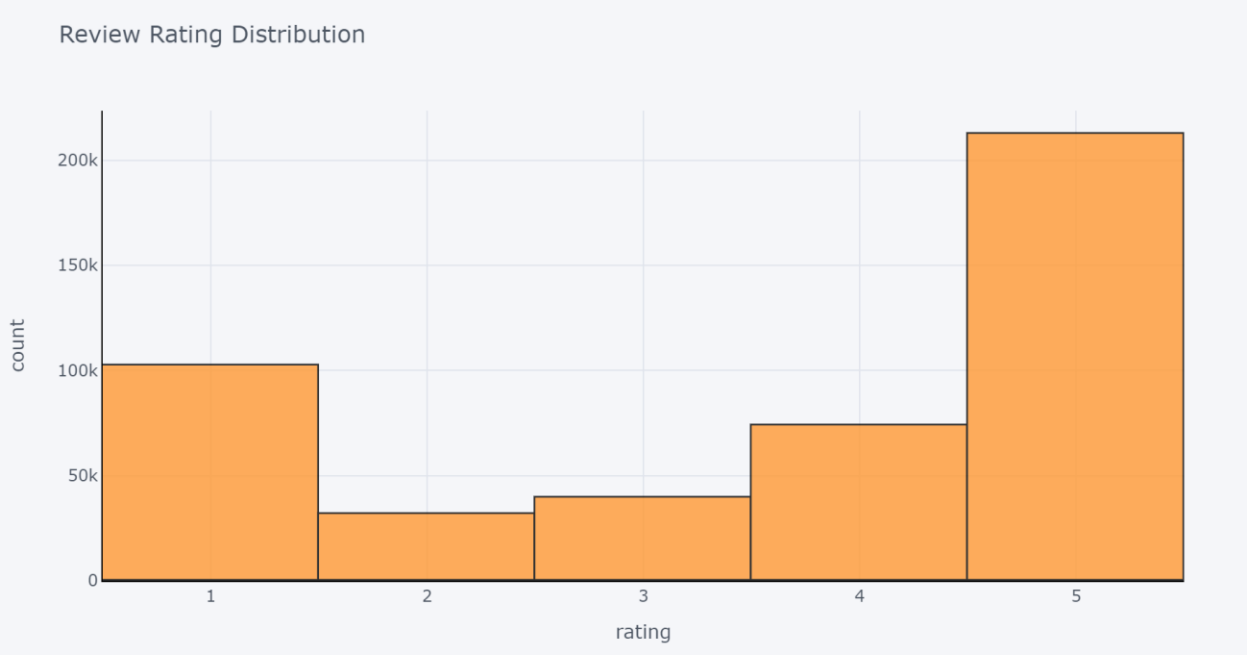

Figure 15: Review rating distribution

From the graph, we can see that there is a large 5 rating, followed by 1 rating. Coming to text length graph, we have the following distribution for the review text length.

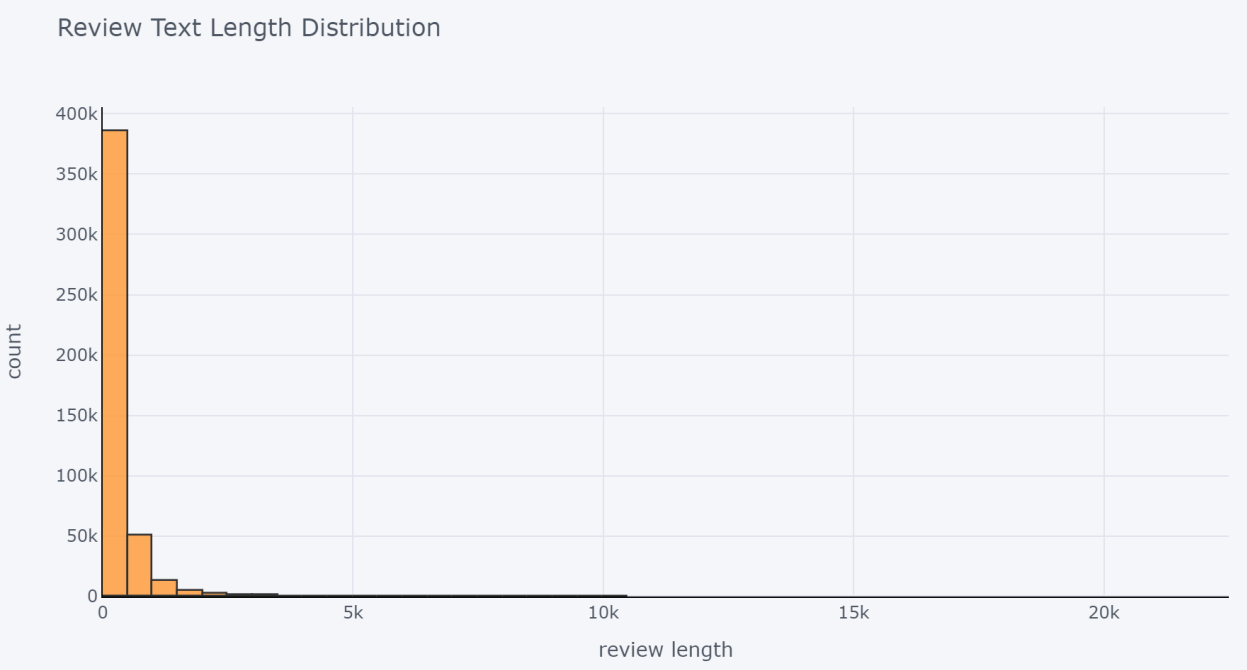

Figure 16: Review text length distribution

As can be seen from the above graph, the graph has a right-skewed distribution.

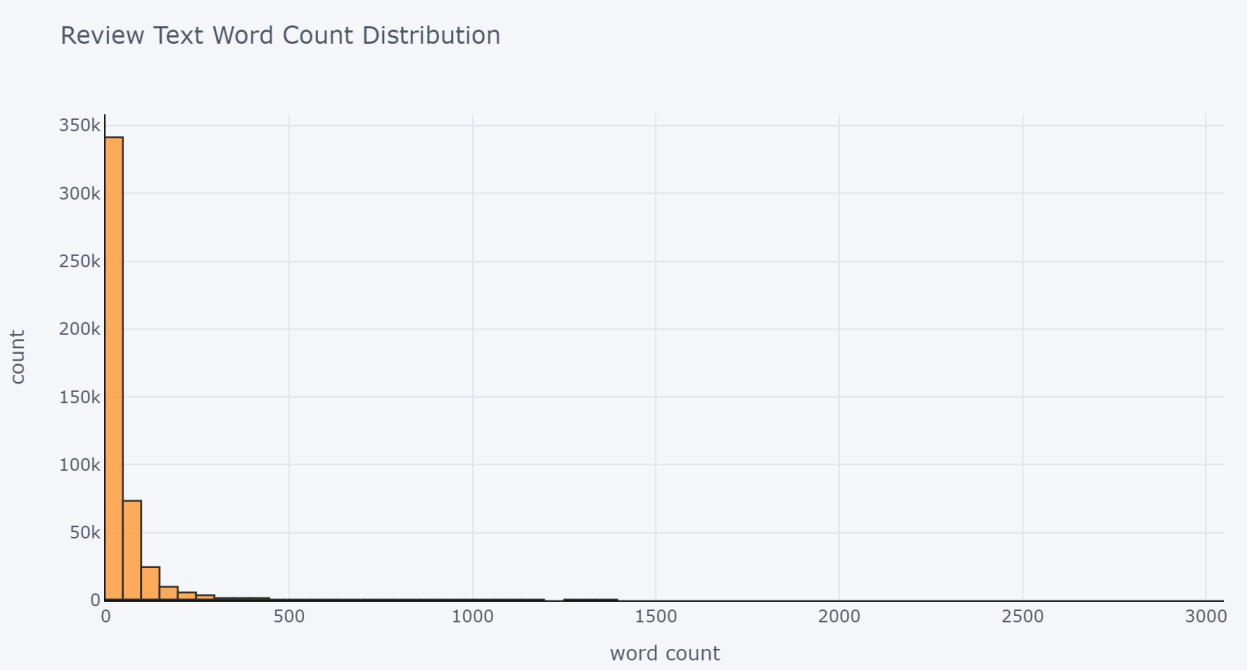

Figure 17: Review text word count distribution

3.1.2 Text Analysis

Text analysis is an important aspect of the overall sentiment analysis task. N-grams can be used to analyze the text based on its sentiment variable. Monogram, bigram, and trigram methods can be used to visualize the most frequent words based on sentiments. The following figure shows the most frequent one-word in review based on sentiment.

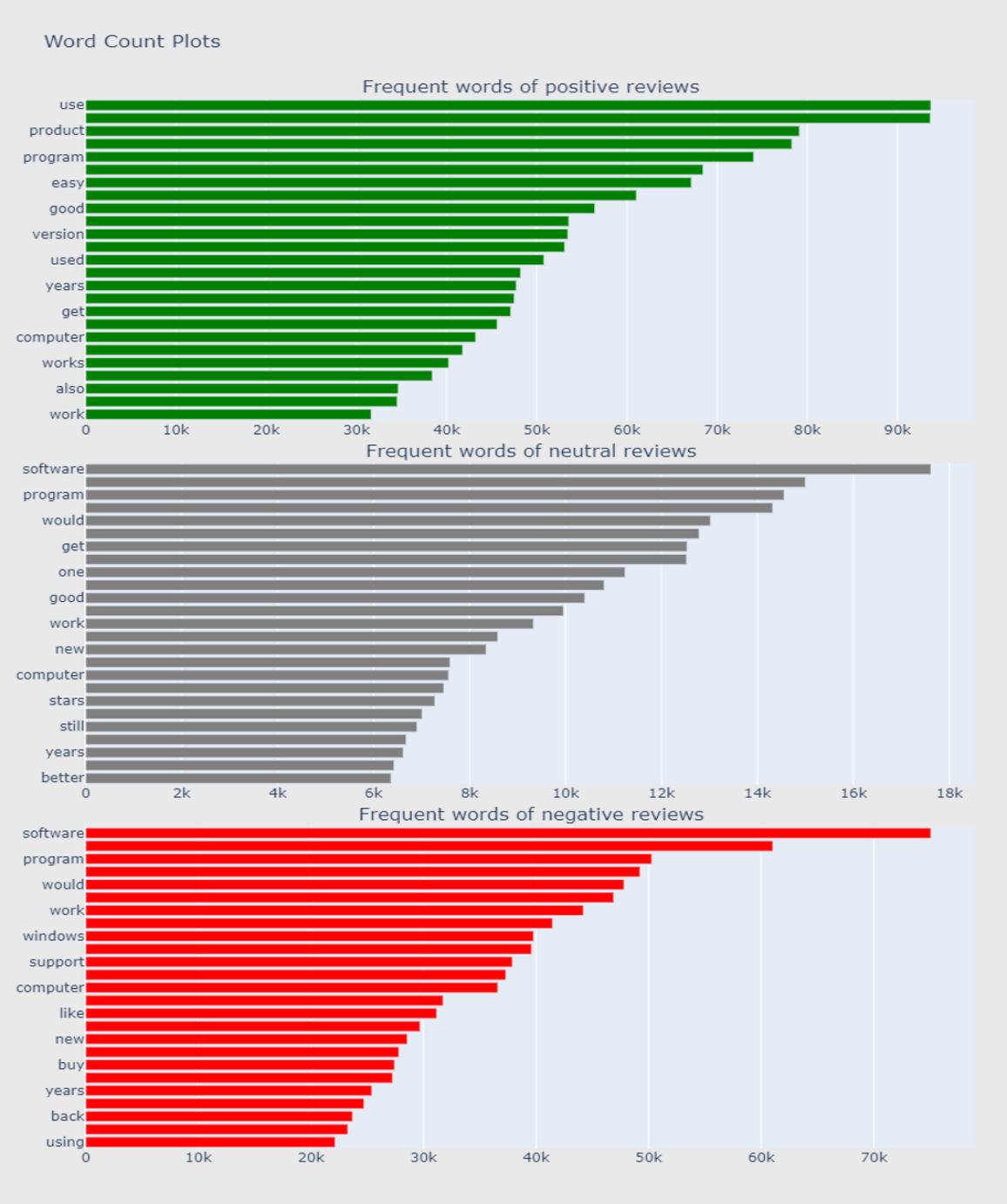

Figure 18: Monogram word count plot

The following figure shows the most frequent two words in reviews based on sentiments.

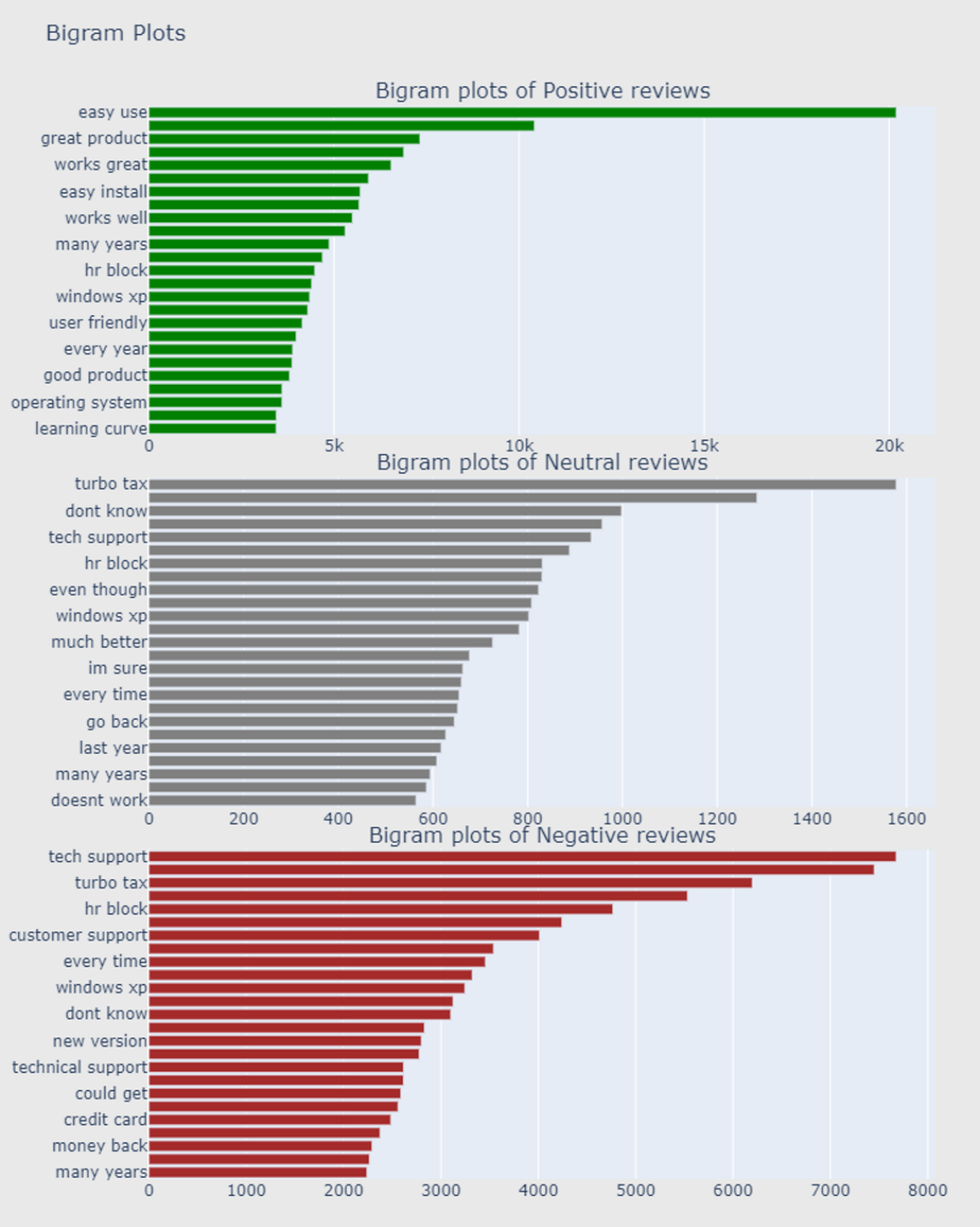

Figure 19: Bigram word count plot

The following figure shows the most frequent three words in reviews based on sentiments.

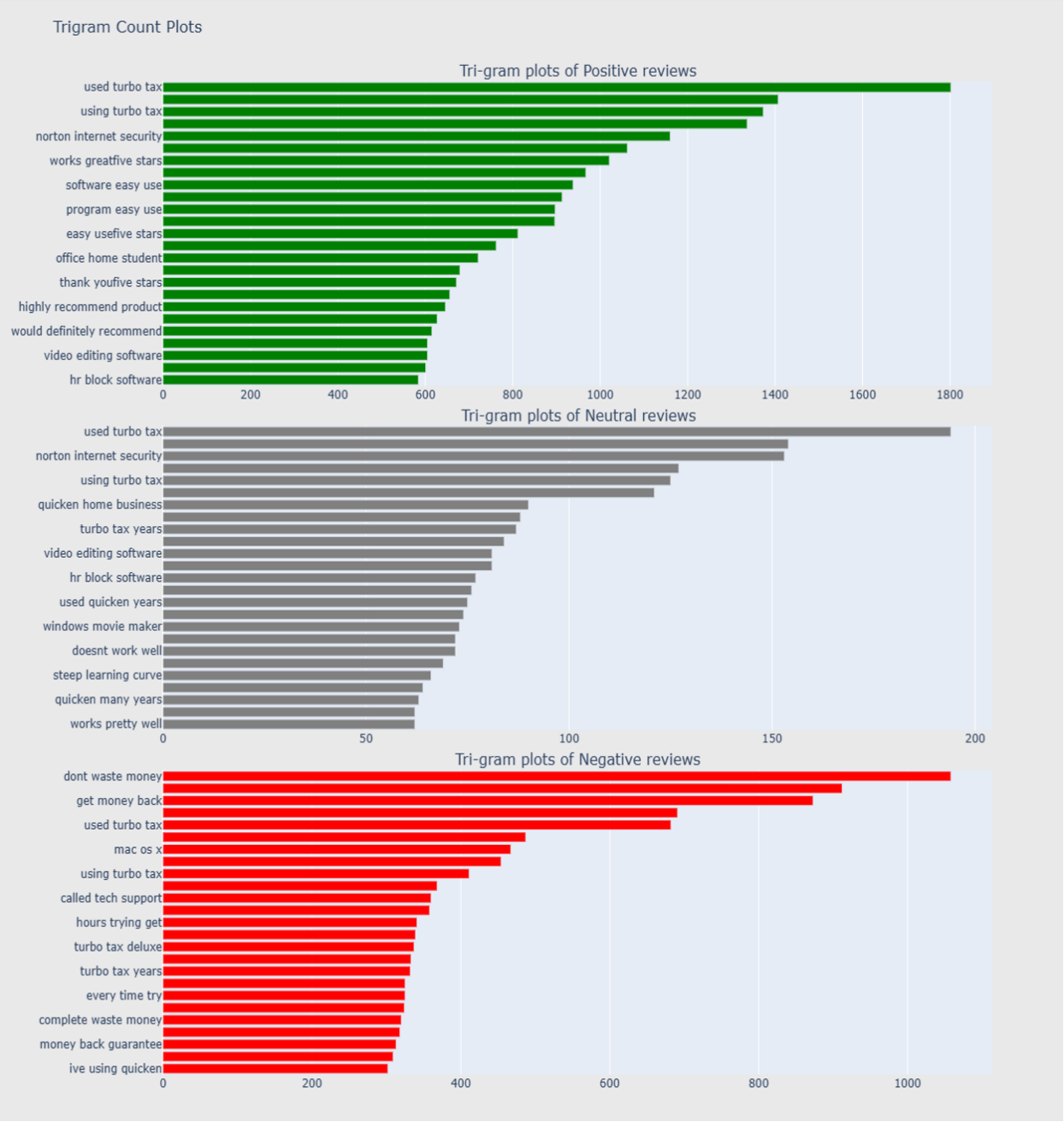

Figure 20: Trigram word count plot

3.1.3 Word Cloud

Word clouds are useful visualizations to tell the story behind the data. The following word cloud visualizations are on the positive, negative, and neutral reviews. For positive reviews we have the following word cloud.

Figure 21: Word clouds for positive reviews

For neutral reviews we have the following word cloud.

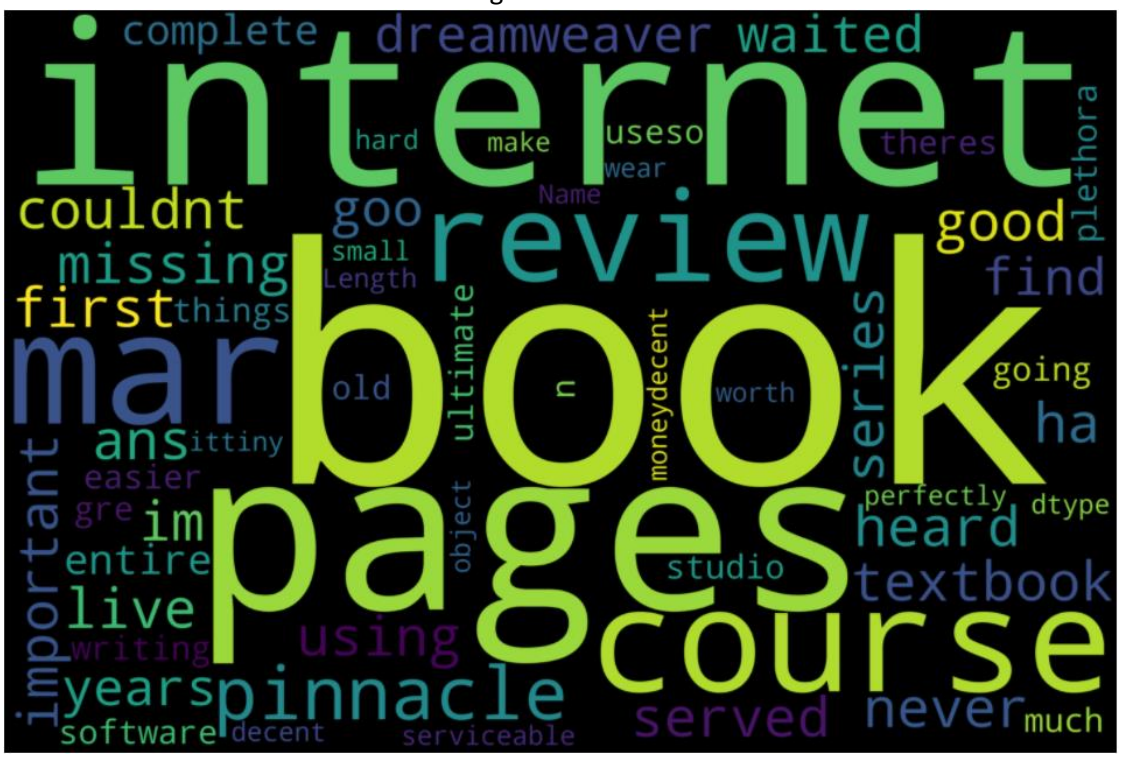

Figure 23: Word clouds for neutral reviews

For negative reviews we have the following word cloud.

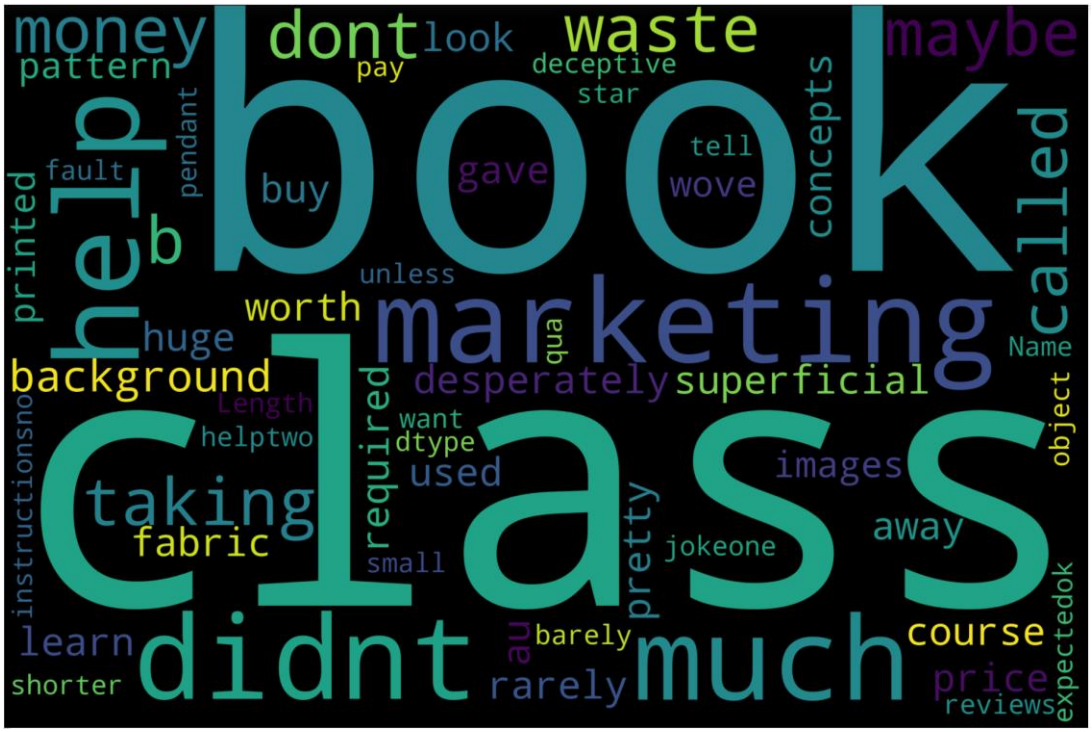

Figure 23: Word clouds for negative reviews

3.1.4 Targeted Variable

First, encoding the target variable/sentiment is a necessary step. 0, 1, and 2 were assigned as labels of the targeted variable. In other words, for the classification task, the targeted variable has to be changed into numbers from the positive, negative, and neutral labels. To do this, the encoding method can be utilized by using label encoder. After doing this, 0, 1, and 2 are assigned as negative, neutral, and positive. The following figure shows the distribution of the targeted variable.

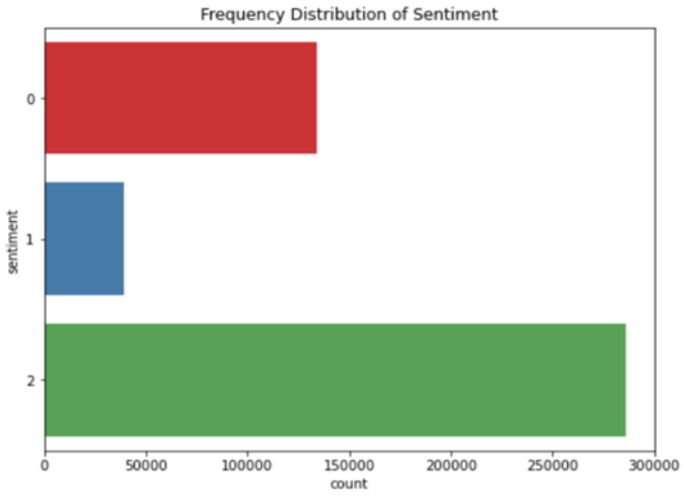

Figure 24: Frequency distribution of class labels

As can be seen from the above graph, there is an imbalance in the target variables with more positive labels. In order to build a more robust model with no overfitting balancing this imbalance is important. Python's SMOTE library is used to balance the classes of the targeted variable.

3.1.5 Stemming

Another method utilized in this step is stemming. Stemming is a method of deriving the root word from the inflected word. Here we extract the reviews and convert the words into reviews to its root word. The root words do not need to carry a semantic meaning which is why stemming is used. This is done by first extracting the reviews column data and creating a data frame based on this. The following shows the extracted data frame.

	reviews
0	materials arrived early excellent condition ho...
1	really enjoying book worksheets make review go...
2	if taking class dont waste money called book b...
3	book missing pages important pages couldnt ans...
4	used learnsmart officially say amazing study t...

Figure 25: Review text after stemming

The stemming is performed on the above data frame and as an example display, the following is the output from of the third corpus.

```
corpus[3]
```

```
'book miss page import page couldnt answer test question never happen beforemiss page'
```

Figure 26: Stemmed text

3.1.6 TF-IDF

Before developing the sentiment analysis model, it is necessary to convert the review texts into vector formation as computers cannot understand words and their sentiment. As stated in the proposal, the TF-TDF method will be used to convert the texts.

TF-IDF stands for term frequency — Inverse document frequency and is a technique that is used to quantify a word in documents, we generally compute a weight to each word which signifies the importance of the word in the document and corpus.

With longer documents, we typically see higher average count values on words that carry very little meaning, this will overshadow shorter documents that have lower average counts with same frequencies, as a result, we will use Tf-IDF Transformer to reduce this redundancy:

- Term Frequencies (TF) divides number of occurrences for each word by total number of words
- Term Frequency times Inverse Document Frequency (Tf-IDF) downscales the weights of each word.

$$tf\ idf\ (t, d, D) = tf\ (t, d)\ .\ idf\ (t, D)$$

Where:

$$tf\ (t, d) = log\ (1 + freq\ (t, d))$$

$$idf\ (t, D) = log\ (\ \frac{N}{count\ (d \in D{:}t \in d)}\ )$$

Figure 27: TF-IDF function

Then, we can use a classification method to get the results. The following figure shows the process with TextBlob and TF-IDF.

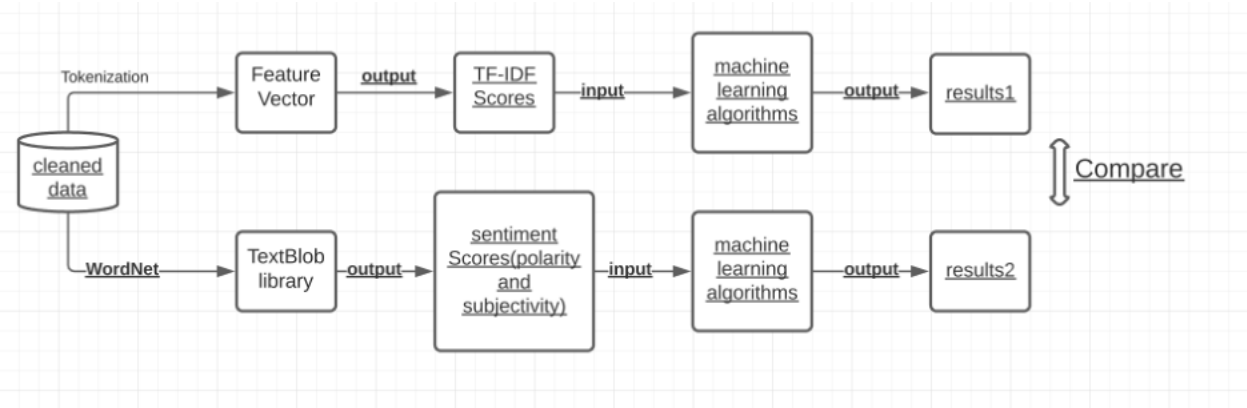

Figure 28: How to process data with TextBlob and TF-IDF

This method is a widely used technique in Information retrieval and text mining. In this process, we choose to split as bigram (two words) and consider their combined weight by taking only the top 5000 words from the reviews. The top 5000 are chosen because of difficulty with the computing capacity of the local machine to handle such computation fast.

3.2 Model Development

After finishing the preprocessing, exploratory analysis, feature engineering and extraction, and text analysis steps, we moved in to developing the model. For the model development, the 75/25 train-test split was followed.

Classification algorithms including Logistic Regression, Decision Tree, and KNN are used to check the performance, and compare with the accuracy and select the best machine learning based sentiment model. Finally, the best model was chosen based on performance.

3.2.1 Initial Model

For the model, X is the TF-IDF score of each review, and Y is the encoded sentiment classes. For this we only used the top 5000 features of the TF-IDF score. The model is first run by using cross validation to have select the best model. The machine learning models algorithms that are used to train the model are Logistic Regression, Decision Tree, and KNN.

3.2.2 Model Selection

The model selection was between Logistic Regression, Decision Tree, and KNN Machine Learning models. Accuracy was used as the main performance measure. As can be seen from the table below, Logistic Regression model outperformed the rest of the models with an accuracy of 80%, while the Decision Tree and KNN models resulted in an accuracy of 72% and 67% respectively. Thus the Logistic Regression model was selected for further hyperparameter tuning tasks to improve performance. The following table shows the accuracy of the three machine learning models.

Model	Accuracy
Logistic Regression	80%
Decision Tree	72%
KNN	67%

Table 1: Model accuracy comparison

3.2.3 Hyperparameter Tuning

Implementing additional techniques is necessary to improve performance. This Is done by performing hyperparameter tuning of the base machine learning model. For hyperparameter tuning, the GridSearchCV method is used. This can increase the accuracy by doing a GridSearchCV on our model's hyperparameters. GridSearchCV is a cross-validation technique for tuning a Machine Learning model. The objective of this process is to find the most optimal parameters. For this, regularization parameters and penalty for parameter tuning is used.

After developing the pipeline, the parameters that increase performance were determined C = 2.5595 and random state of 0. Using this parameter and plugging in the Logistic Regression model, the performance of the model was not improved. This is because the features selected for TF-IDF are small and increasing this feature greatly increases the run time of model training process.

3.3 Visualizations and Analysis

In this section, some visualizations from the final model and findings are discussed. In this step, visualizations from the final sentiment analysis model are presented. Additionally, recommendations are provided based on the results of the sentiment analysis model.

3.3.1 Classification Metrix

In this section the classification performance metrics of the Logistic Regression model are show. The classification report which contains the accuracy, precision, and recall values of the target variable, the confusion matrix, and ROC curve are shown as follows.

The classification report of the Logistic regression model is summarized below. From the classification report we can see that Positive and Negative categories have a better performance while the neutral category is comparatively lower.

Sentiment	Accuracy	Precision	Recall
Negative-0	70%	74%	66%
Neutral-1	64%	58%	71%
Positive-2	74%	79%	69%

Table 2: Classification report

The confusion matrix of the final Logistic Regression model is shown below. As can be seen from the confusion matrix, there are several instances of misclassification for all three categories. This can be improved by doing further hyperparameter tuning and increasing the feature size on the TF-IDF X input to the model.

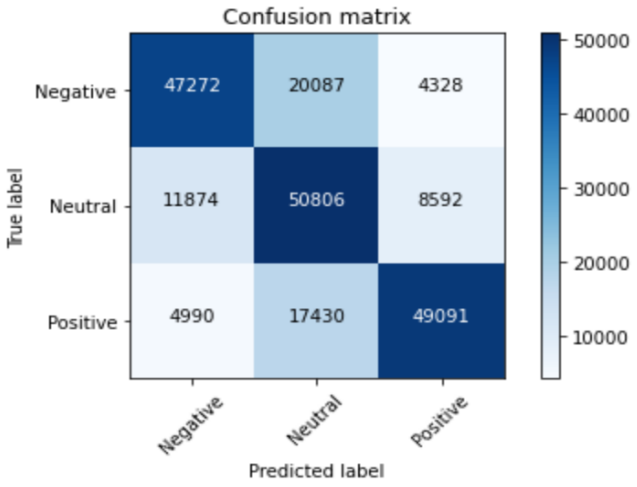

Figure 29: Confusion matrix

The ROC curve of the three Logistic Regression model is shown below. From the graph we can see the performance of the model for each target label which are separated by color.

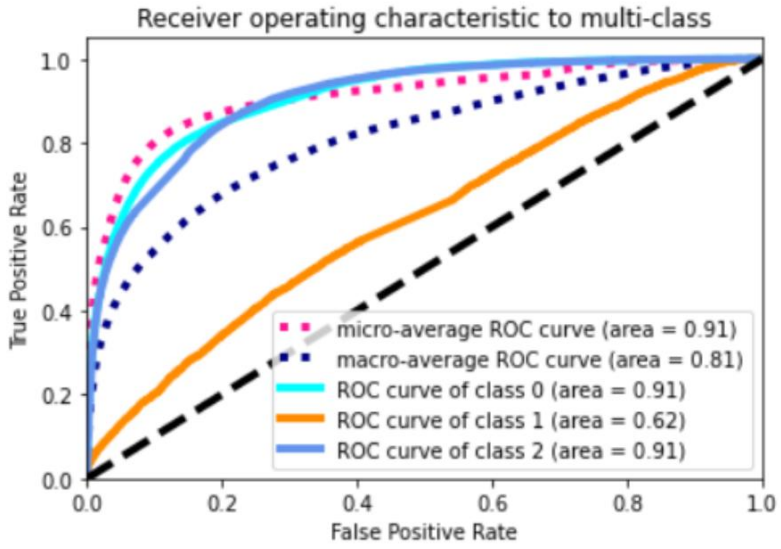

Figure 30: ROC curve

3.4 Time Series Analysis

Time series analysis regarding the sentiment of users over time is used to investigate the sentiment over time. This can help how certain decisions made over time can have had a positive impact on sales and overall service. The time series analysis could provide useful insights into how poor the sales are for the products with poor reviews, which in turn could potentially help Amazon remove the item from its catalogue and increase the inventory of items that has been doing well over the years to meet customers' demands and to remain competitive as the number one e-commerce website.

Time series analysis not only helps with the tracking of sentiment over time, but it also helps with predicting future sentiment of Amazon products with the help of machine learning techniques such as Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN).

The following visualization of the sentiment count based on year can be used a time series analysis on the distribution if the sentiment categories through time.

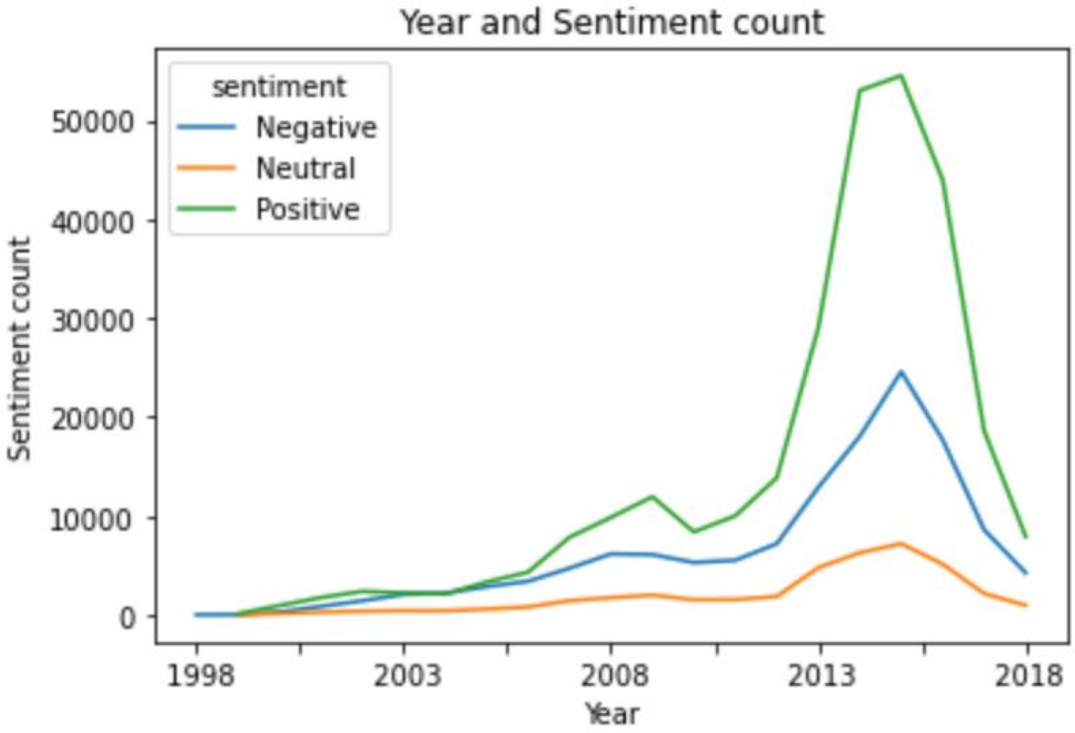

Figure 31: Year and sentiment visualization

The visualization shown above is a time series visualization that shows the trend of positive, negative, and neutral reviews from 1998 to 2018. In this graph we can see that the highest ratings of software products were recorded between 2014 and 2015. Additionally, we can see that there is a steady trend of neutral reviews over the years. We can also observe that there is a sharp decrease in all three categories of reviews. This can be further demonstrated in the following graph.

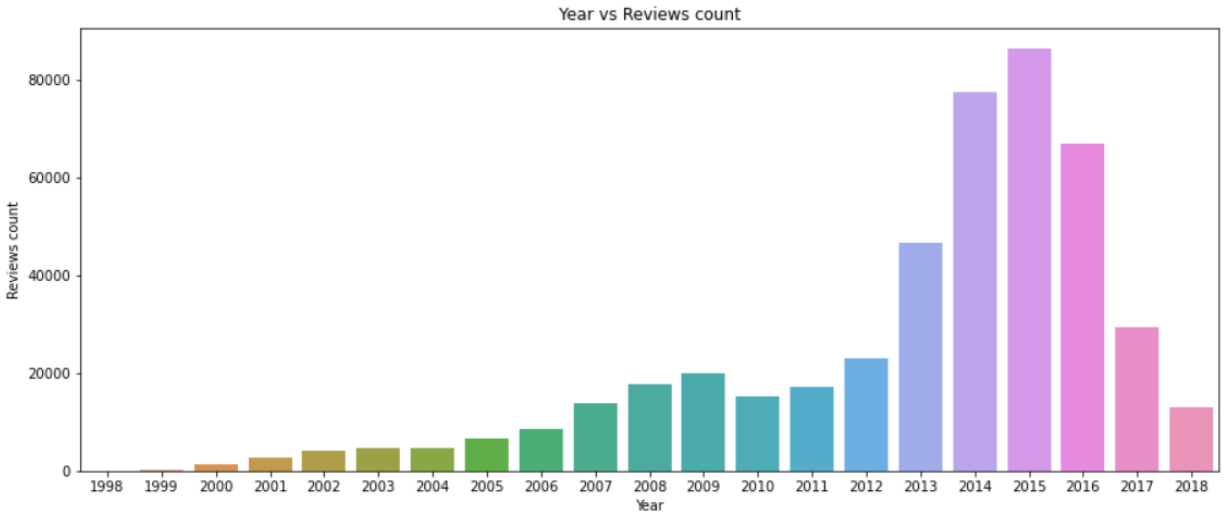

Figure 32: Year and reviews visualization

As is demonstrated by the graph above, there is a steady increase in Software product reviews from 2012 until it reaches its peak in 2015. This is due to the fact that Amazon was expanding its market during this time and many advents in the Software realm from startups to big corporation were being introduced and expanded as a business commodity.

Another time-series visualization that can be demonstrated is the monthly distribution of reviews. According to the graph below, the number of reviews is highest in the first three months of the year, with the month of March having the highest count, and reviews tend to decrease in the middle of the year, with the month of June having the lowest review count until they slightly increase by the end of the year.

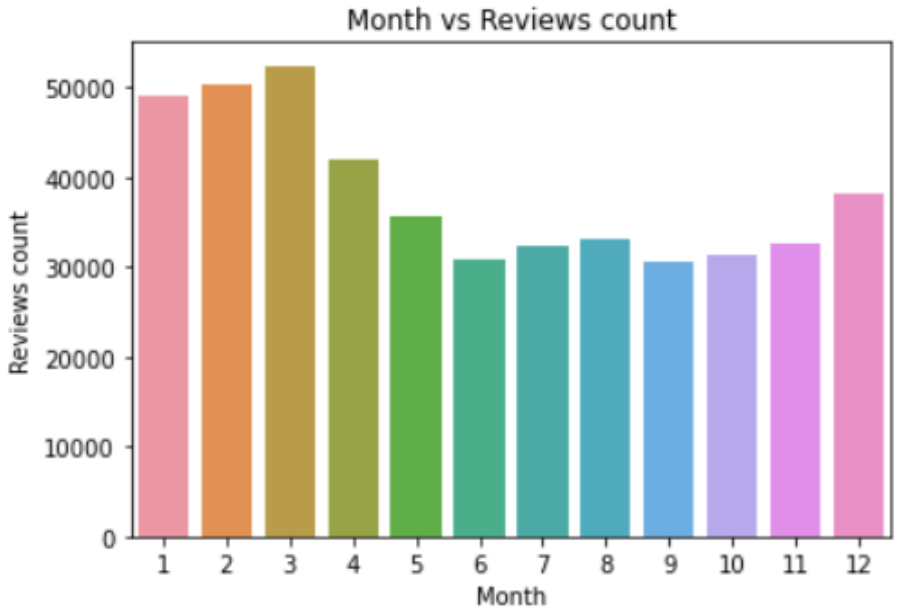

Figure 33: Month and reviews visualization

Additionally, it is possible to look at the monthly aspect of the time-series analysis with respect to the three sentiment categories. From the graph below, we can see that positive reviews are the highest in the month of March and lowest in June. On the other hand, the Negative reviews are at their highest in the month of January whereas lowest negative reviews are recorded in the month of September. Neutral reviews show an almost uniform distribution with the highest neutral reviews being recorded in March whereas the lowest being recorded in June.

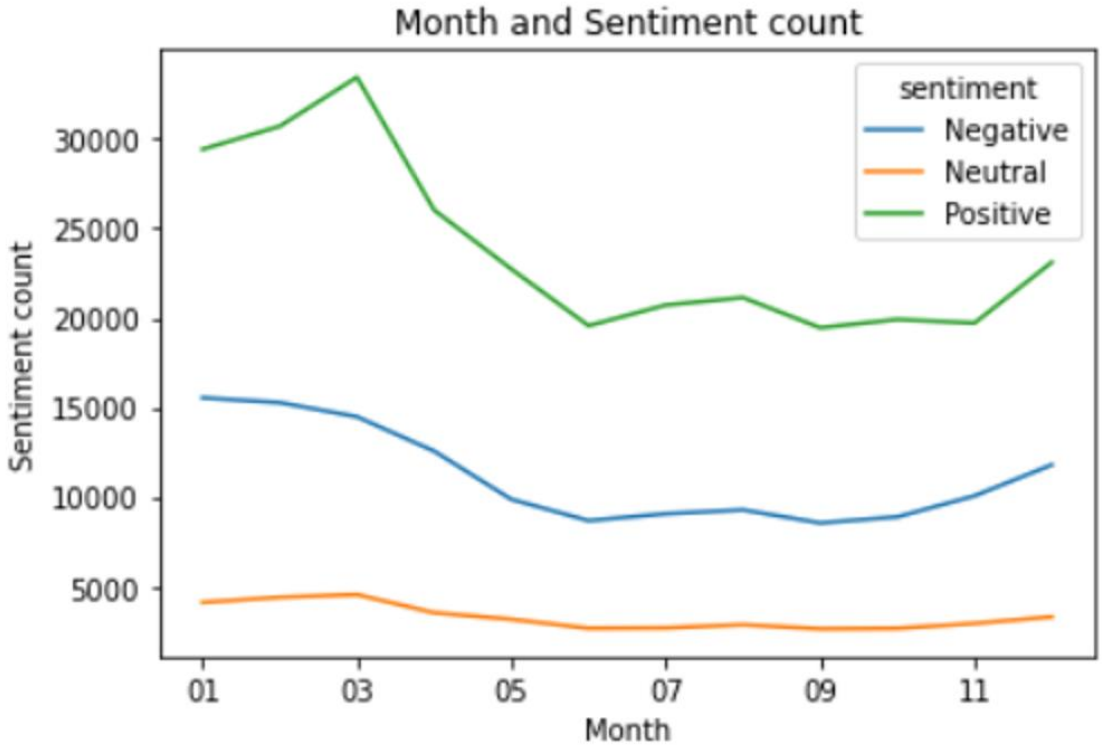

Figure 34: Month and sentiment visualization

Finally, it is possible to see the daily review distribution. The following graph shows a uniform distribution of the review count. But there is a large drop at the end of the month.

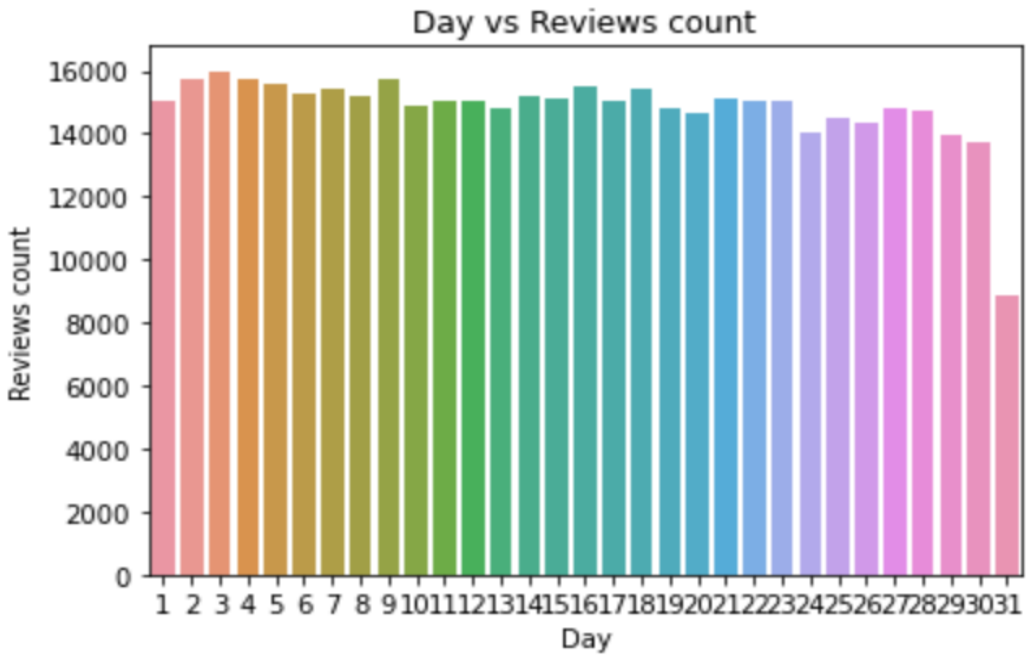

Figure 35: Day and reviews visualization

3.5 Recommender Analysis

Recommender systems are increasingly becoming an essential part of companies. Recommender system creates a similarity between the user and items and exploits the similarity between customer/item to make recommendations. These systems can help customers find the right products, increase customer engagement, help item providers to deliver items to the right customer, and helps make contents more personalized. In this project, a popularity-based analysis is conducted. Popularity-based recommendation systems are based on trends. Sentiment analysis and other techniques can be used to give a popularity-based recommendation.

For this analysis, products which have more than 50 ratings are selected. The following shows the top five products with the respective product number/asin and average rating.

```
asin
B000050ZRE    4.937908
B00SX73LIK    4.930769
B000EORV8Q    4.929012
B0001FS9NE    4.915309
B0000AZJY6    4.897924
Name: overall, dtype: float64
```

Figure 36: Top 5 popular products with average rating

The following shows the total number of ratings per product for the top five.

```
asin
B00UB76290    8994
B00CTTEKJW    7939
B00NG7JVSQ    6395
B00H9A60O4    4730
B00E6LJ2SA    4048
Name: overall, dtype: int64
```

Figure 37: Total number of ratings per product for the top five

The following figure shows the top 30 most popular products based on the given ratings.

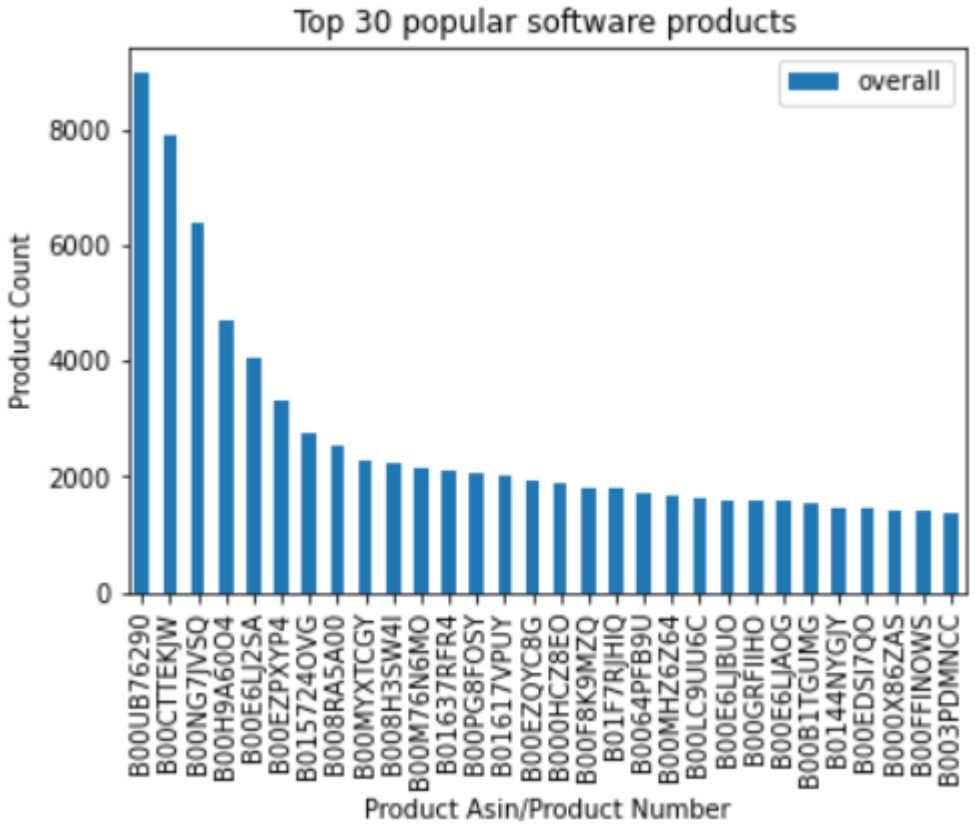

Figure 38: Top 30 most popular products

Based on the results shown in the above analysis, it is possible to give customers recommendation. This is helpful both to customers and for the growth of the overall business.

4 Impact of the Project

With massive volumes of data being created every second, businesses must adapt and improve big data solutions. For instance, with the use of Sentiment analysis, the raw data can be mined and turned into an impactful resource for the company. The implementation of this project has demonstrated this and would help the company understand its customers and potentially grow if implemented. This project has demonstrated the above-mentioned points with extensive analysis and modeling of Amazon's customer ratings. This project can be utilized as a starting point for other to develop even more advanced and robust mechanism of sentiment analysis of such data.

5 Conclusion

In this project, Sentiment Analysis techniques were utilized to categorize the unstructured Amazon rating metadata and make sense of it. We were able to build different models that included machine learning based classification models. The Logistic Regression model was the best model and can be further improved by conducting additional feature engineering & text preprocessing techniques. Also, model performance can be improved by implementing further hyperparameter tuning techniques for all machine learning models. On the other hand, the recommender system allows a company like Amazon to recommend products based on sentiment of previous buyers. Additionally, Sentiment based Time-series analysis is important for the growth of a company. Overall, the project allowed us to implement NLP and Machine Learning techniques on a real-world problem.

Regarding future work, we would like to make use of advanced machine learning algorithms to obtain a higher performance, conduct a study of additional Amazon online market categories and make comparisons, create a dashboard to visualize results, conduct an analysis beyond the rating-based Sentiment Analysis and explore further into the extraction of narratives from the review text.

6 References

Aljuhani, S. A., & Alghamdi, N. S. (2019). A Comparison of Sentiment Analysis Methods on Amazon Reviews of Mobile Phones. *International Journal of Advanced Computer Science and Applications*, 608-617.

Bhatt, A., Patel, A., Chheda, H., & Gawande, K. (2015). Amazon Review Classification and Sentiment Analysis. International Journal of Computer Science and Information Technologies, 5107-5110.

Devin, J., Chang, M., Lee, K., Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. Retrieved from https://arxiv.org/abs/1810.04805

Gokce, E. (2020). Sentiment analysis on Amazon reviews. Towards data science. Retrieved from https://towardsdatascience.com/sentiment-analysis-on-amazon-reviews-45cd169447ac

Haque, T. U., Saber, N. N., & Shah, F. M. (2018). Sentiment Analysis on Large Scale Amazon Product Reviews. *IEEE International Conference on Innovative Research and Development (ICIRD)*.

Loria, S. (2020). TextBlob: Simplified Text Processing. Retrieved from https://textblob.readthedocs.io/en/dev/

McKinsey & Company. (2021). US consumer sentiment and behaviors during the coronavirus crisis. Retrieved from https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/survey-us-consumer-sentiment-during-the-coronavirus-crisis

Natural Language Toolkit. (2021). NLTK documentation. Retrieved from https://www.nltk.org

Ni, J. (2018). Amazon review data. Retrieved from https://nijianmo.github.io/amazon/

Princeton University. (2021). WordNet. Retrieved from https://wordnet.princeton.edu