E-commerce Sentiment Analysis

Description

In this project, I aim to perform sentiment analysis using a dataset from an e-commerce domain. The dataset consists of over 34,000 consumer reviews for Amazon brand products. It includes attributes such as brand, categories, primary categories, reviews.title, reviews.text, and sentiment labels (Positive, Negative, Neutral). My goal is to predict the sentiment or satisfaction of a purchase based on various features and review text.

Dataset

The dataset provides a valuable resource for understanding sentiment and satisfaction levels in e-commerce. It contains a wide range of consumer reviews, covering different products and categories. The reviews are accompanied by relevant attributes and sentiment labels, enabling the development of sentiment analysis models.

Project Tasks

Week 1

In the first week, I focused on tackling the class imbalance problem in the dataset and gaining insights through exploratory data analysis. The tasks included:

- Conducting an exploratory data analysis (EDA) to understand the characteristics of positive, negative, and neutral reviews.
- Checking the class count for each sentiment class to identify any class imbalance issues.
- Converting the reviews into TF-IDF scores, a technique to represent textual data numerically.
- Training a multinomial Naive Bayes classifier and observing the impact of class imbalance on the classification results.
- Tackling the class imbalance problem through oversampling or undersampling techniques.
- Evaluating the models using precision, recall, F1-score, and AUC-ROC curve, with a focus on the F1-Score as the evaluation criteria.

Week 2

In the second week, I delved into model selection and advanced techniques to improve sentiment classification. The tasks included:

- Applying multi-class Support Vector Machines (SVM) and neural networks for sentiment classification.
- Exploring ensemble techniques such as combining XGBoost with oversampled multinomial Naive Bayes.
- Engineering a feature called sentiment score and incorporating it into the models to evaluate its impact on performance and gain insights.
- Applying Long Short-Term Memory (LSTM) neural networks, a type of recurrent neural network, to capture sequential information in the reviews.
- Comparing the accuracy of neural networks with traditional machine learning algorithms.

- Determining the best settings for LSTM and experimenting with Gated Recurrent Units (GRU) to classify reviews as positive, negative, or neutral using techniques like Grid Search, Cross-Validation, and Random Search.

Moreover, I explored topic modeling techniques to gain insights into different aspects of the products and analyze clusters of similar reviews. Techniques like Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) were used for topic modeling.

Screenshots

Screenshots of relevant plots, classification results, topic clusters, and any other visual representations have been added to the README document to illustrate the analysis and findings.

Dataset

```
# Load the train_data CSV file
train_data_path = 'Ecommerce/train_data.csv'
train_data = pd.read_csv(train_data_path)
train_data.head()
```

	name	brand	categories	primaryCategories	reviews.date	reviews.text	reviews.title	sentiment
0	All-New Fire HD 8 Tablet, 8" HD Display, Wi-Fi...	Amazon	Electronics,iPad & Tablets,All Tablets,Fire Ta...	Electronics	2016-12-26T00:00:00.000Z	Purchased on Black FridayPros - Great Price (e...	Powerful tablet	Positive
1	Amazon - Echo Plus w/ Built-In Hub - Silver	Amazon	Amazon Echo,Smart Home,Networking,Home & Tools...	Electronics,Hardware	2018-01-17T00:00:00.000Z	I purchased two Amazon in Echo Plus and two do...	Amazon Echo Plus AWESOME	Positive
2	Amazon Echo Show Alexa-enabled Bluetooth Speak...	Amazon	Amazon Echo,Virtual Assistant Speakers,Electro...	Electronics,Hardware	2017-12-20T00:00:00.000Z	Just an average Alexa option. Does show a few ...	Average	Neutral
3	Fire HD 10 Tablet, 10.1 HD Display, Wi-Fi, 16 ...	Amazon	eBook Readers,Fire Tablets,Electronics Feature...	Office Supplies,Electronics	2017-08-04T00:00:00.000Z	very good product. Exactly what I wanted, and ...	Greattttttt	Positive
4	Brand New Amazon Kindle Fire 16gb 7" Ips Displ...	Amazon	Computers/Tablets & Networking,Tablets & eBook...	Electronics	2017-01-23T00:00:00.000Z	This is the 3rd one I've purchased. I've bough...	Very durable!	Positive

Class Imbalance Problem - EDA

```
Positive Review Example:
Purchased on Black FridayPros - Great Price (even off sale)Very powerful and fast with quad core processors Amazing soundWell
builtCons -Amazon ads, Amazon need this to subsidize the tablet and will remove the adds if you pay them $15.Inability to acce
ss other apps except the ones from Amazon. There is a way which I was able to accomplish to add the Google Play storeNet this
is a great tablet for the money
Negative Review Example:
was cheap, can not run chrome stuff, returned to store.
Neutral Review Example:
Just an average Alexa option. Does show a few things on screen but still limited.
Class Count:
Positive    3749
Neutral      158
Negative      93
Name: sentiment, dtype: int64
```

Topic Modeling with LDA

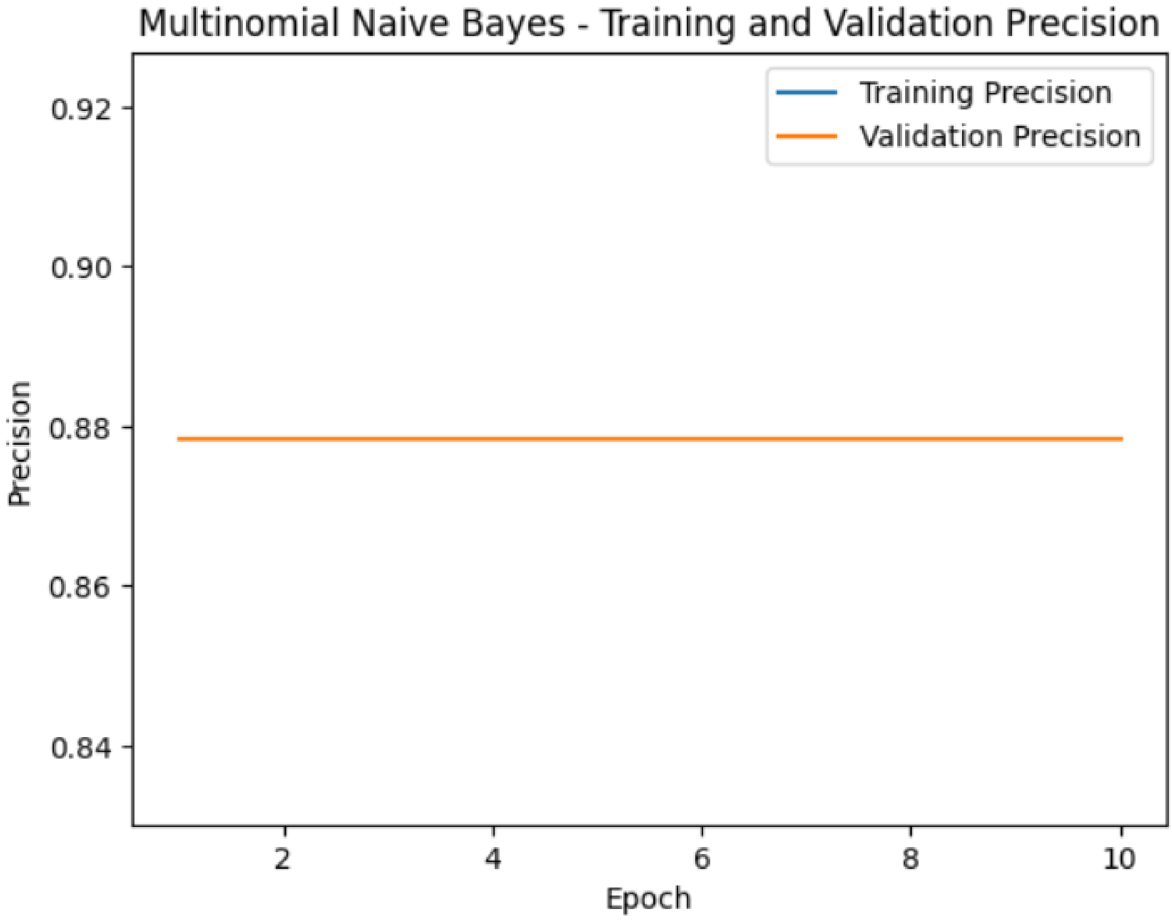
```
Topic 1: keeps, does, product, young, amazon, busy, good, love, expectations, tablet
Topic 2: echo, alexa, love, great, music, home, use, easy, sound, product
Topic 3: kindle, tablet, bought, love, gift, great, got, christmas, loves, read
Topic 4: old, year, loves, tablet, bought, grandson, easy, daughter, great, son
Topic 5: tablet, great, use, easy, price, good, kindle, kids, love, bought
```

Topic Modeling with NMF

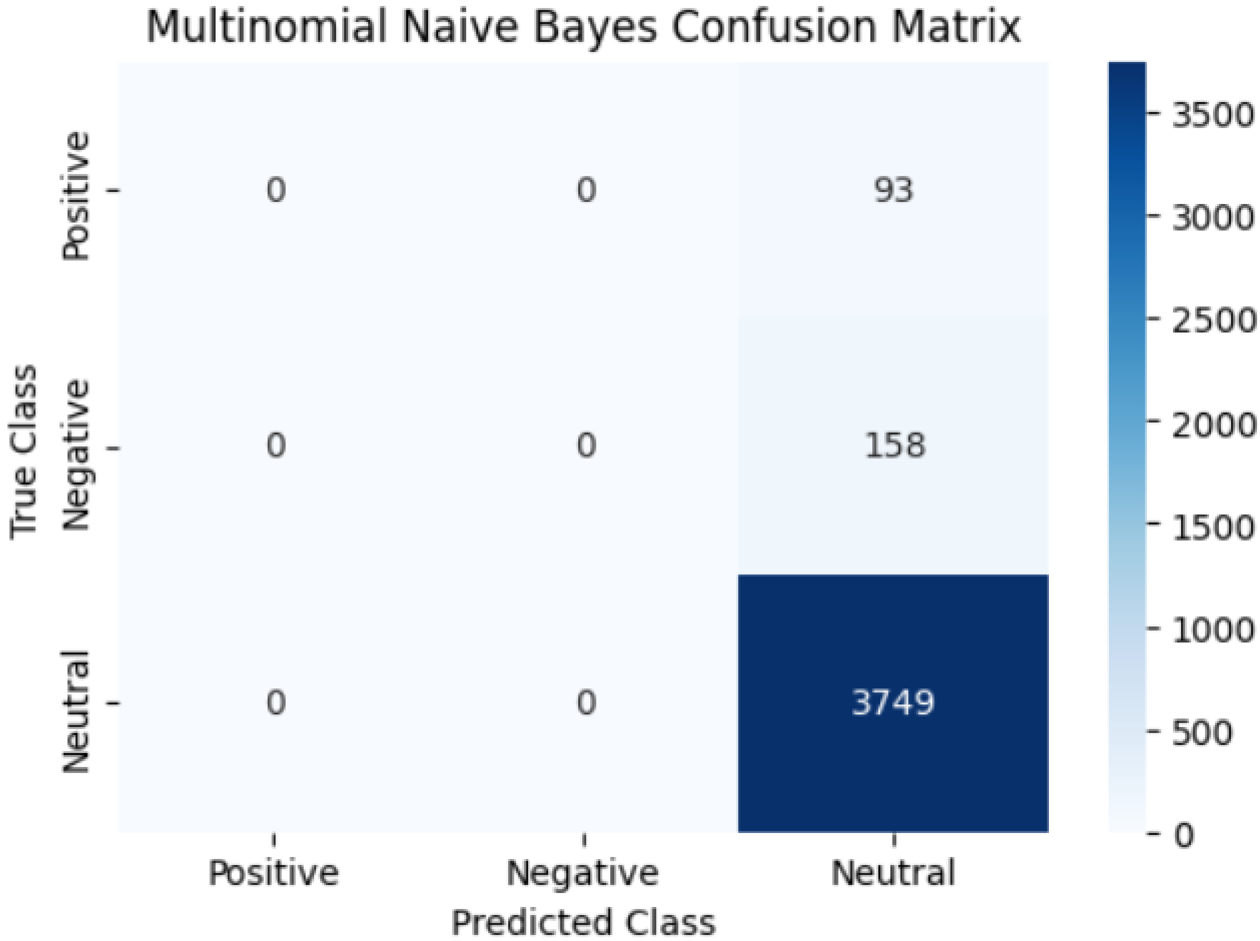
```
Topic 1: love, kindle, kids, books, read, reading, new, games, awesome, size
Topic 2: loves, old, bought, year, daughter, tablet, gift, grandson, son, christmas
Topic 3: great, tablet, price, good, works, product, kids, recommend, apps, buy
Topic 4: easy, use, set, product, setup, fun, recommend, super, lightweight, navigate
Topic 5: echo, alexa, music, home, amazon, plus, like, smart, sound, screen
```

Multinomial Naive Bayes

```
Epoch number 0 , training precision = 0.8784375625 , validation precision = 0.8784375625
Epoch number 1 , training precision = 0.8784375625 , validation precision = 0.8784375625
Epoch number 2 , training precision = 0.8784375625 , validation precision = 0.8784375625
Epoch number 3 , training precision = 0.8784375625 , validation precision = 0.8784375625
Epoch number 4 , training precision = 0.8784375625 , validation precision = 0.8784375625
Epoch number 5 , training precision = 0.8784375625 , validation precision = 0.8784375625
Epoch number 6 , training precision = 0.8784375625 , validation precision = 0.8784375625
Epoch number 7 , training precision = 0.8784375625 , validation precision = 0.8784375625
Epoch number 8 , training precision = 0.8784375625 , validation precision = 0.8784375625
Epoch number 9 , training precision = 0.8784375625 , validation precision = 0.8784375625
```

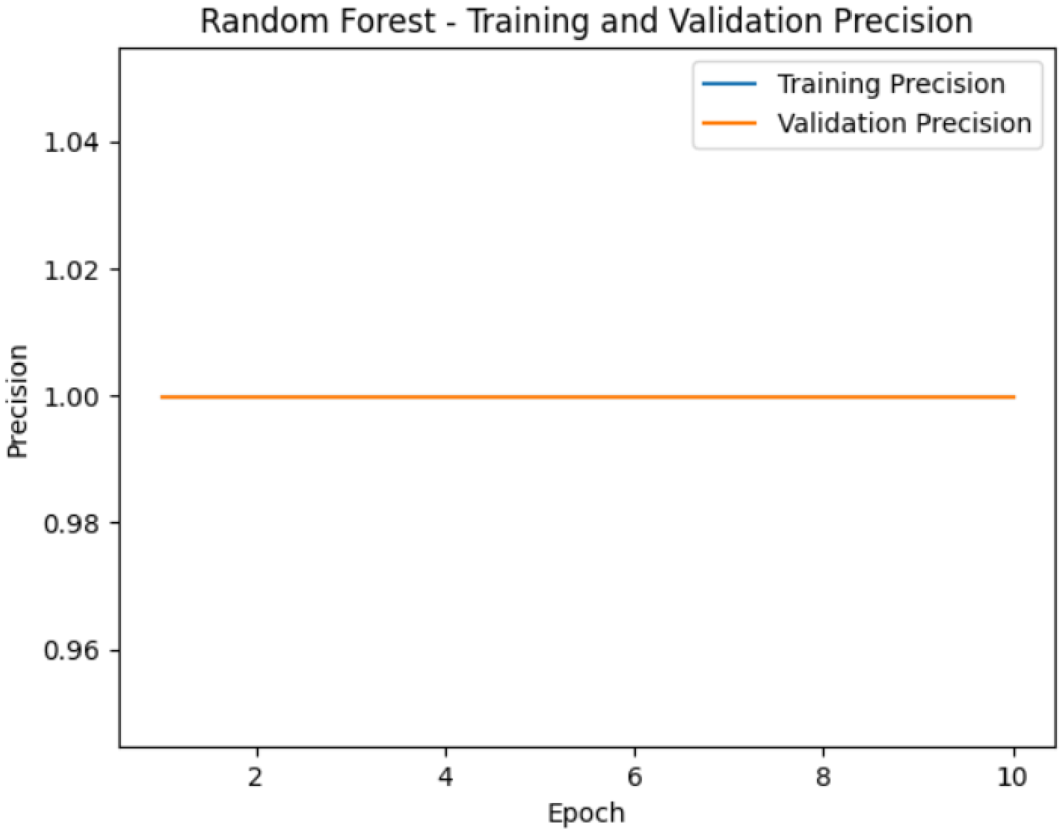



```
Multinomial Naive Bayes Classifier Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        93
           1       0.00      0.00      0.00       158
           2       0.94      1.00      0.97      3749

    accuracy                           0.94      4000
   macro avg       0.31      0.33      0.32      4000
weighted avg       0.88      0.94      0.91      4000
```

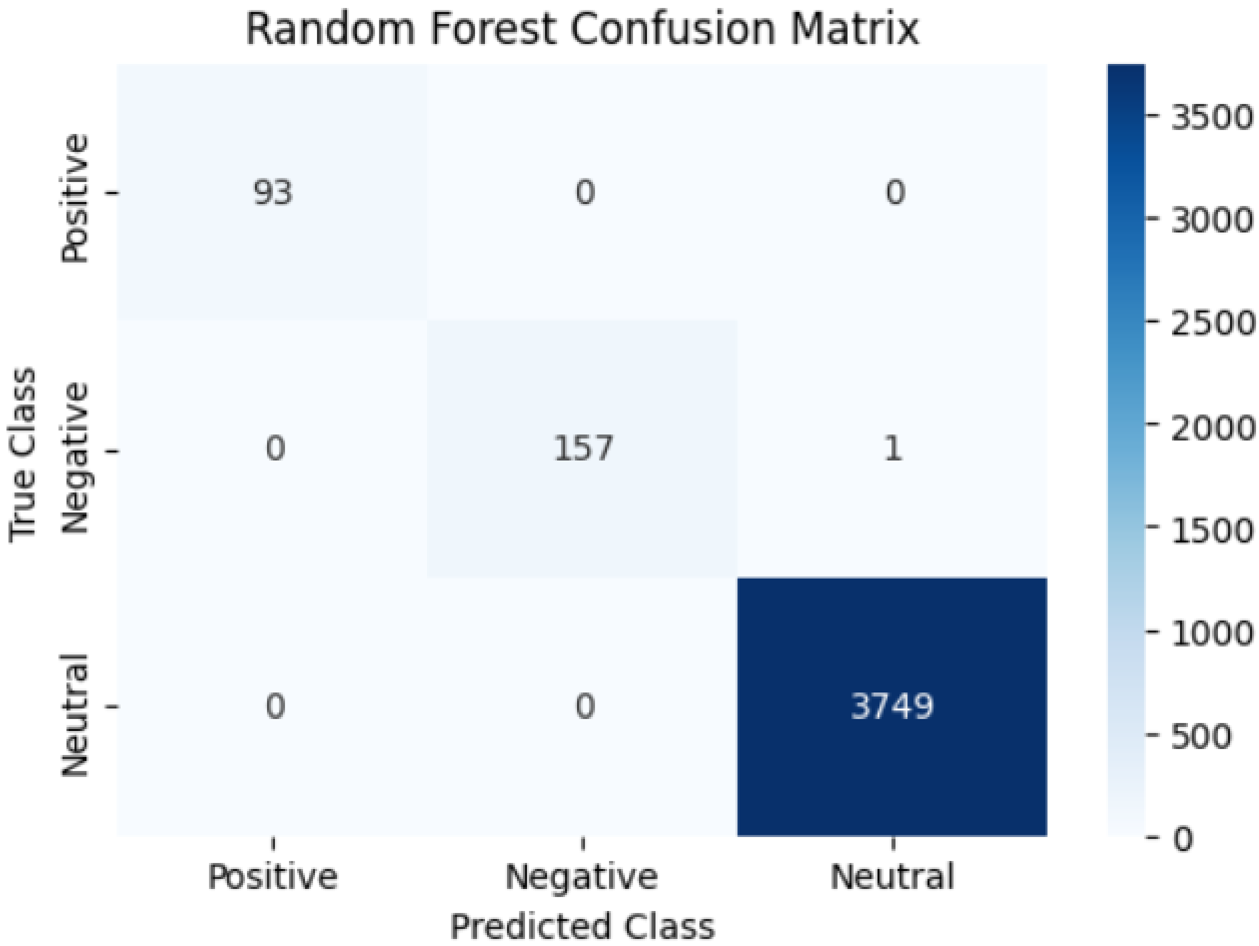


Random Forest

```
Epoch number 0 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 1 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 2 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 3 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 4 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 5 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 6 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 7 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 8 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 9 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
```

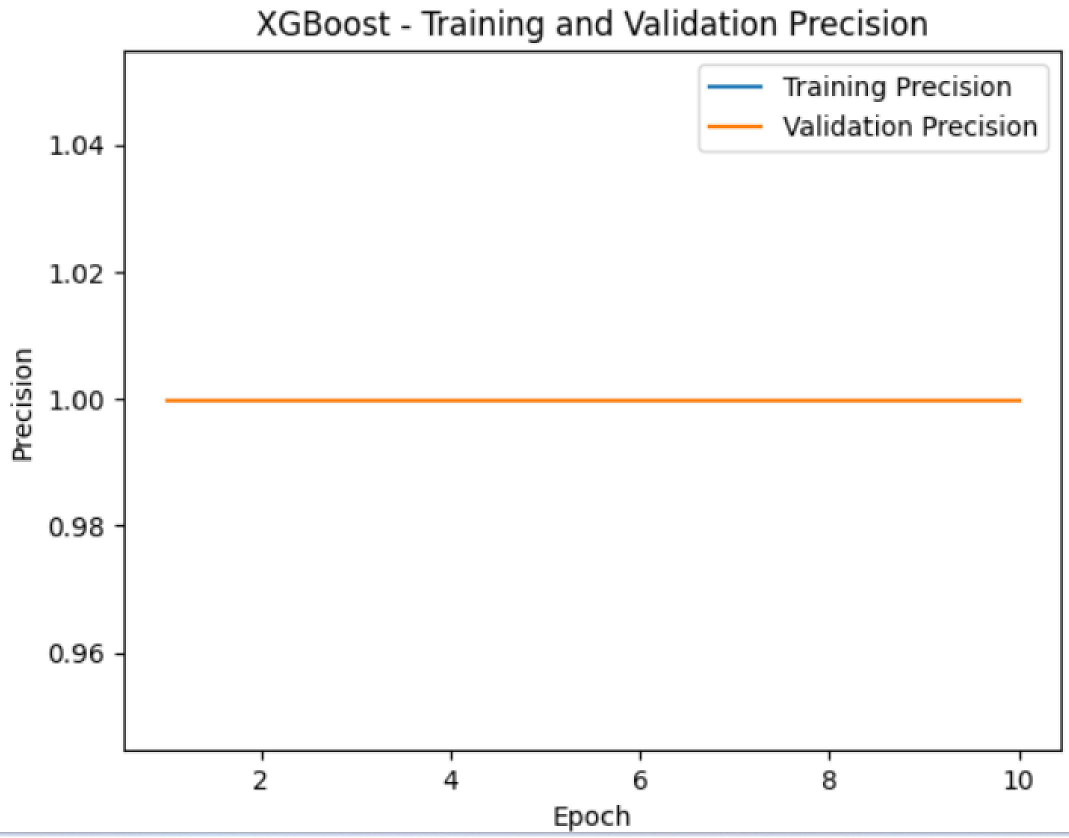



```
Random Forest Classifier Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        93
           1       1.00      0.99      1.00       158
           2       1.00      1.00      1.00      3749

    accuracy                           1.00      4000
   macro avg       1.00      1.00      1.00      4000
weighted avg       1.00      1.00      1.00      4000
```

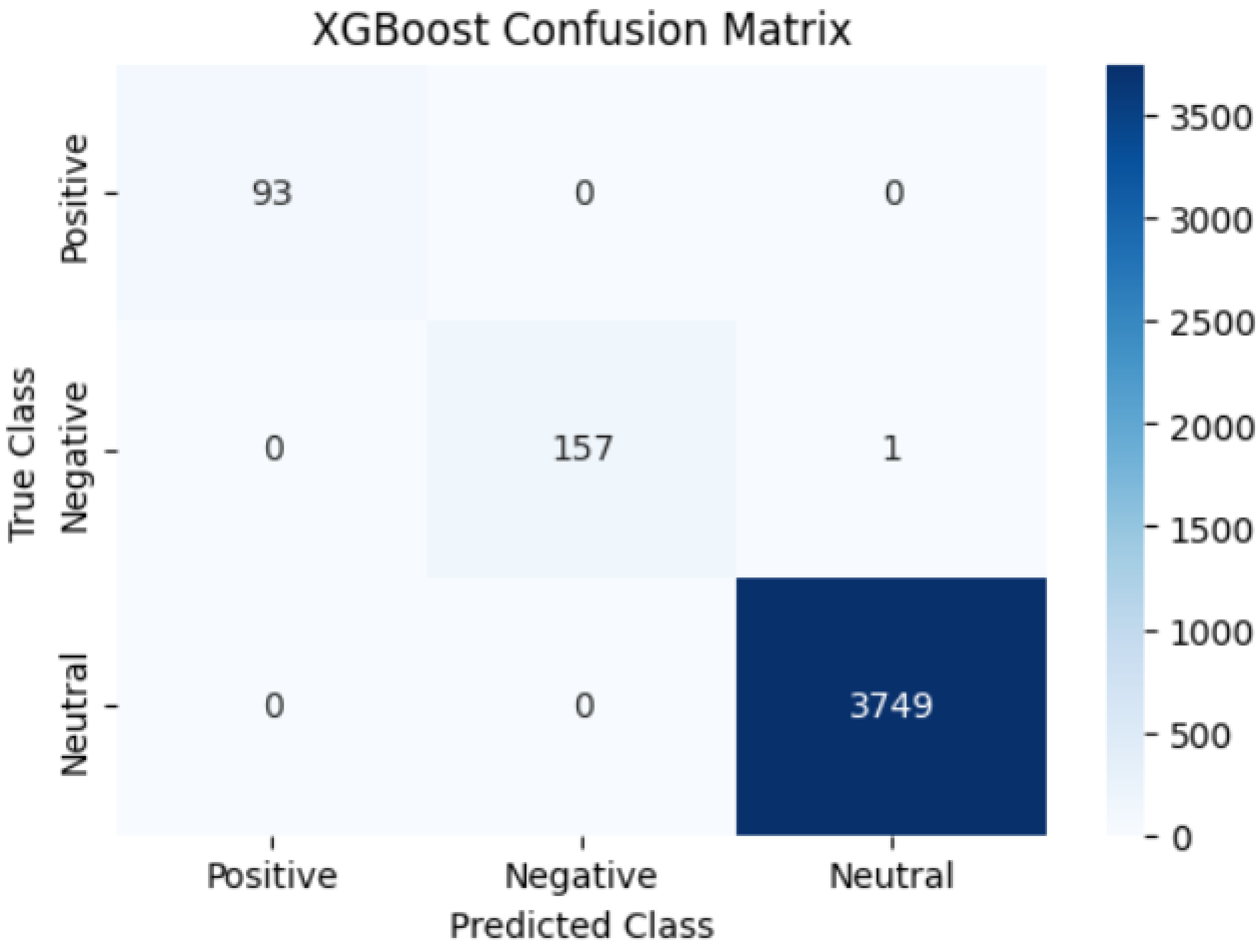

Random Forest Confusion Matrix
True Class
Positive
Negative
Neutral
93
0
0
0
157
1
0
0
3749
Positive
Negative
Neutral
Predicted Class
3500
3000
2500
2000
1500
1000
500
0

XGBoost

```
Epoch number 0 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 1 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 2 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 3 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 4 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 5 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 6 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 7 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 8 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
Epoch number 9 , training precision = 0.9997500666666668 , validation precision = 0.9997500666666668
```



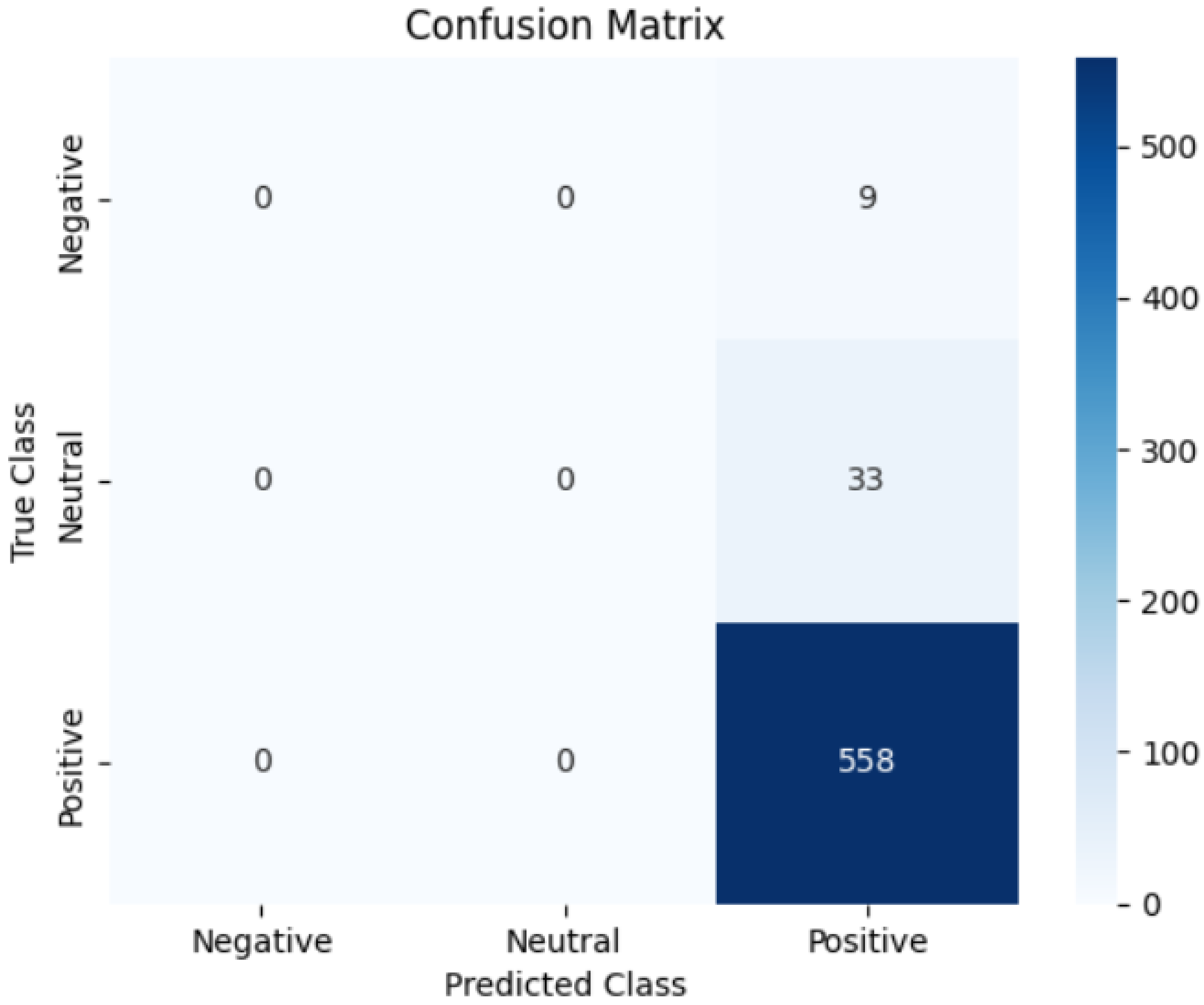

```
XGBoost Classifier Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        93
           1       1.00      0.99      1.00       158
           2       1.00      1.00      1.00      3749

    accuracy                           1.00      4000
   macro avg       1.00      1.00      1.00      4000
weighted avg       1.00      1.00      1.00      4000
```



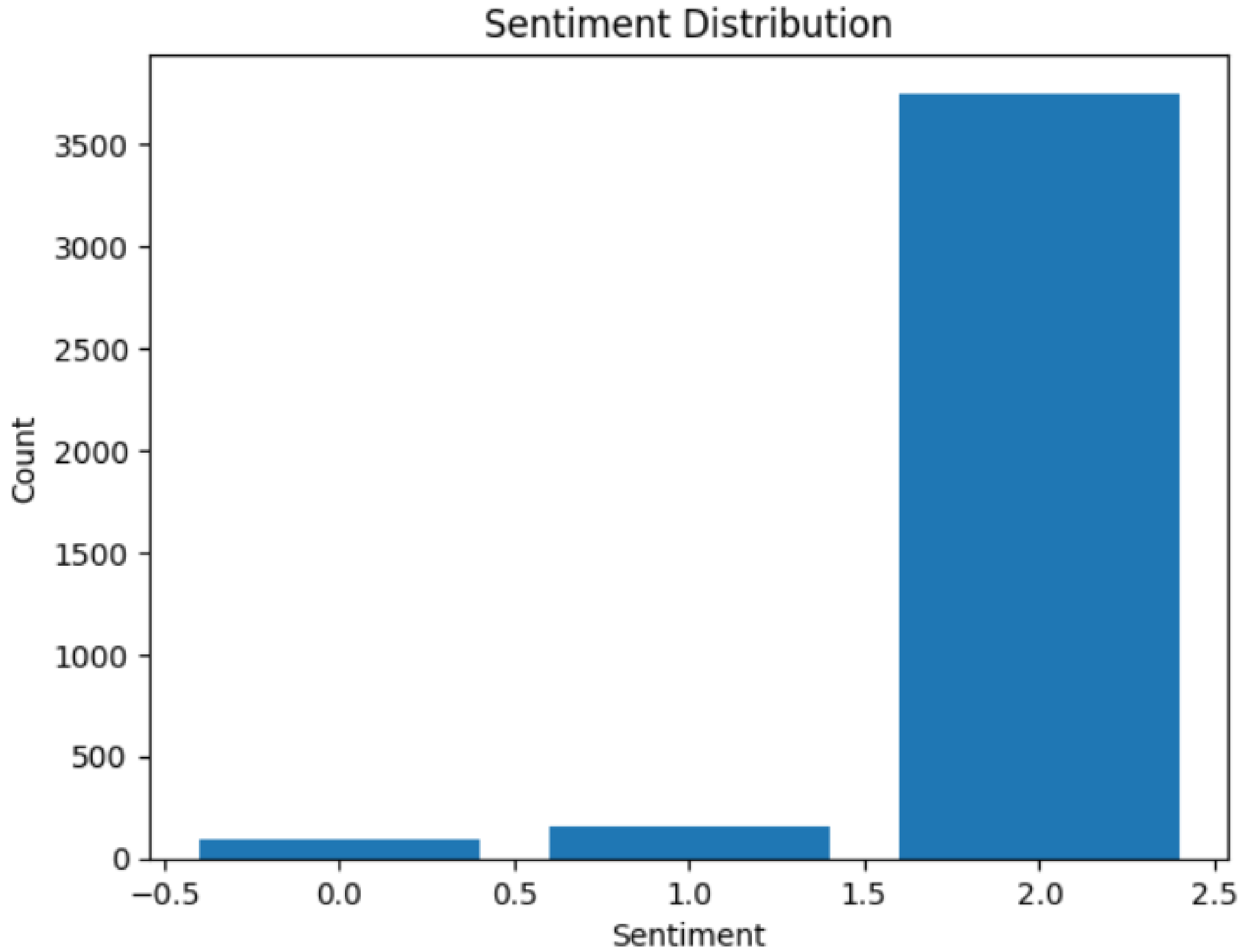
Neural Network model

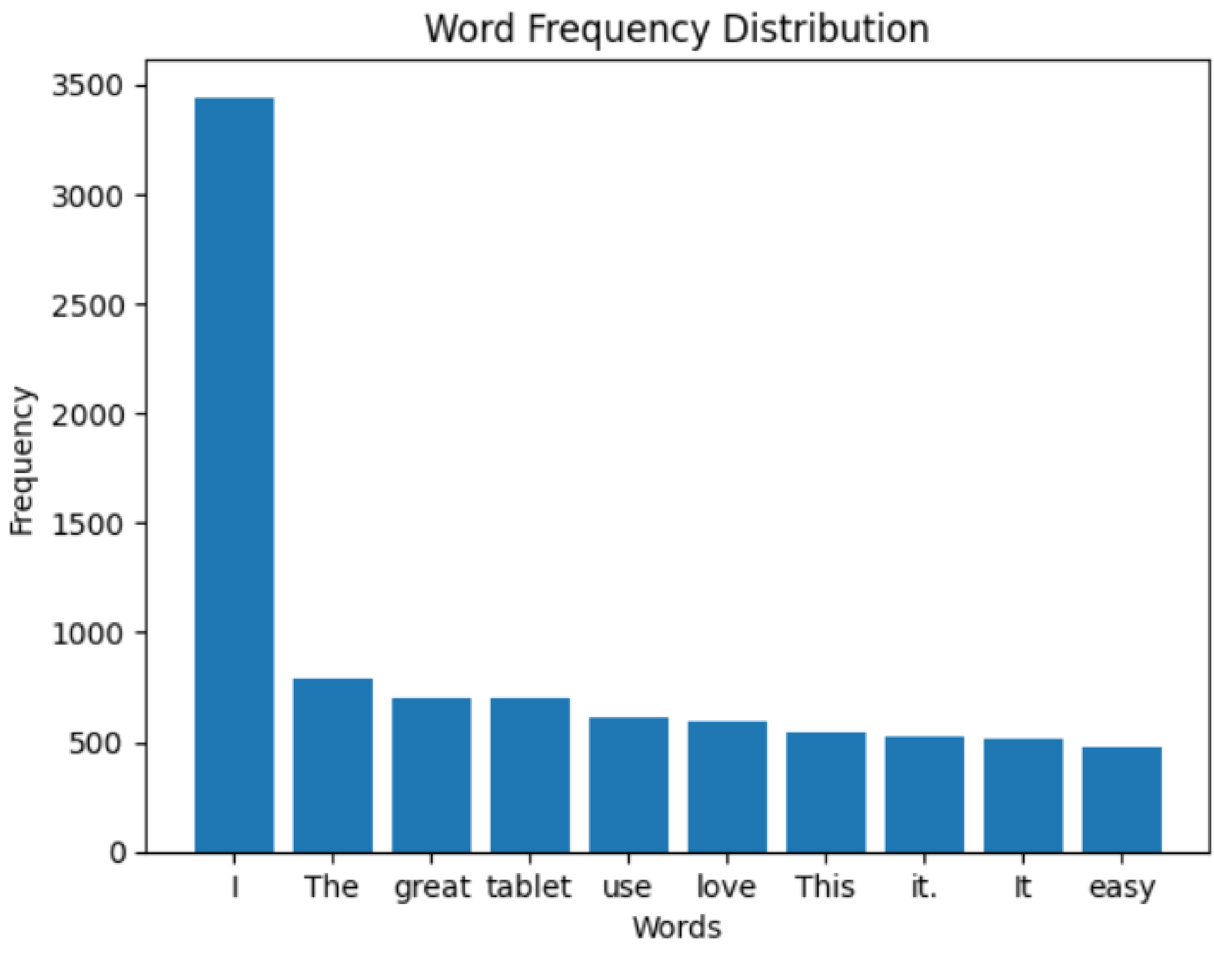

Training and Validation Accuracy
Training Accuracy
Validation Accuracy
0.98
0.97
0.96
0.95
0.94
0.93
Accuracy
0
1
2
3
4
5
Epoch

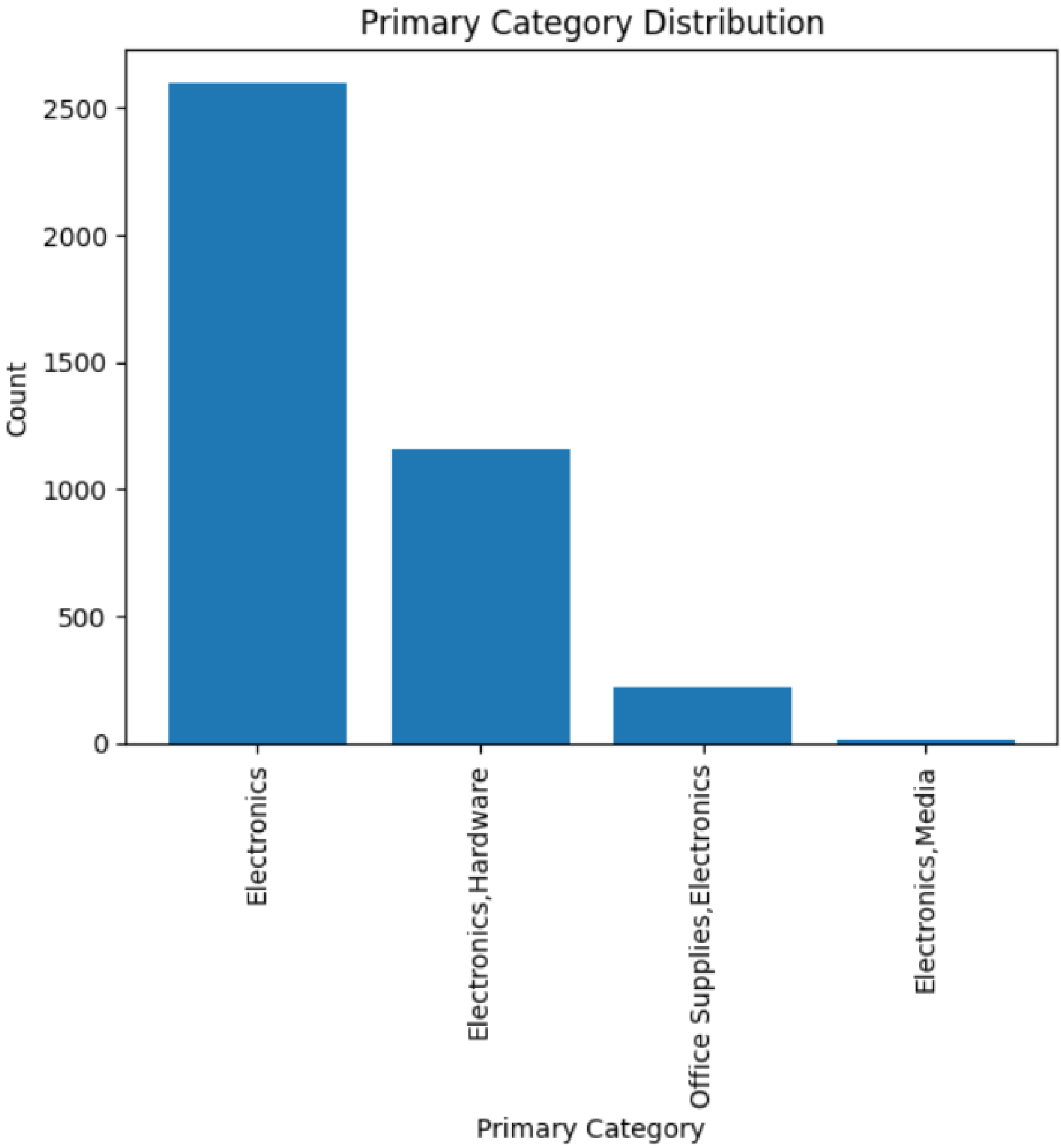
	precision	recall	f1-score	support
0	1.00	0.00	0.00	9
1	1.00	0.00	0.00	33
2	0.93	1.00	0.96	558
accuracy			0.93	600
macro avg	0.98	0.33	0.32	600
weighted avg	0.93	0.93	0.90	600

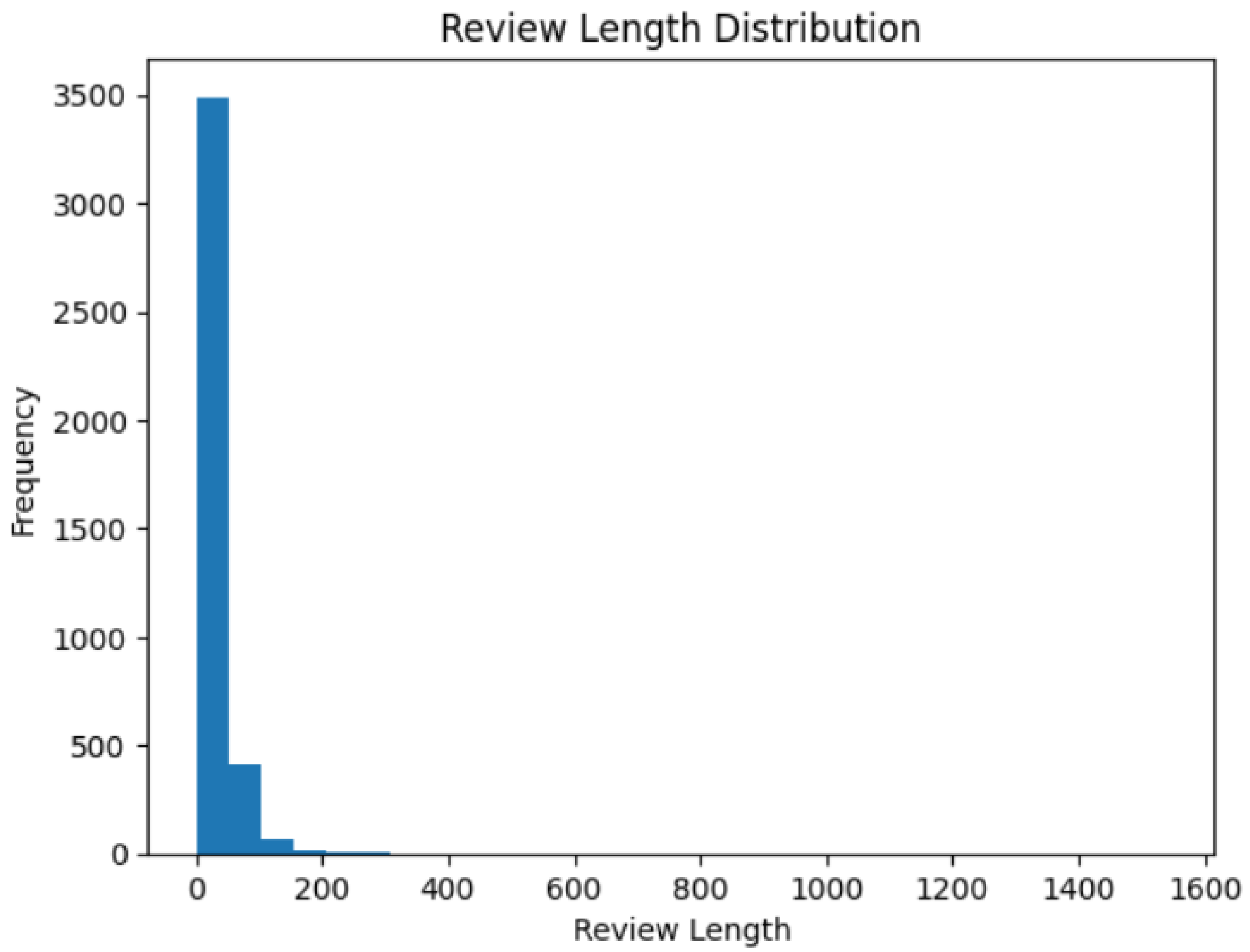


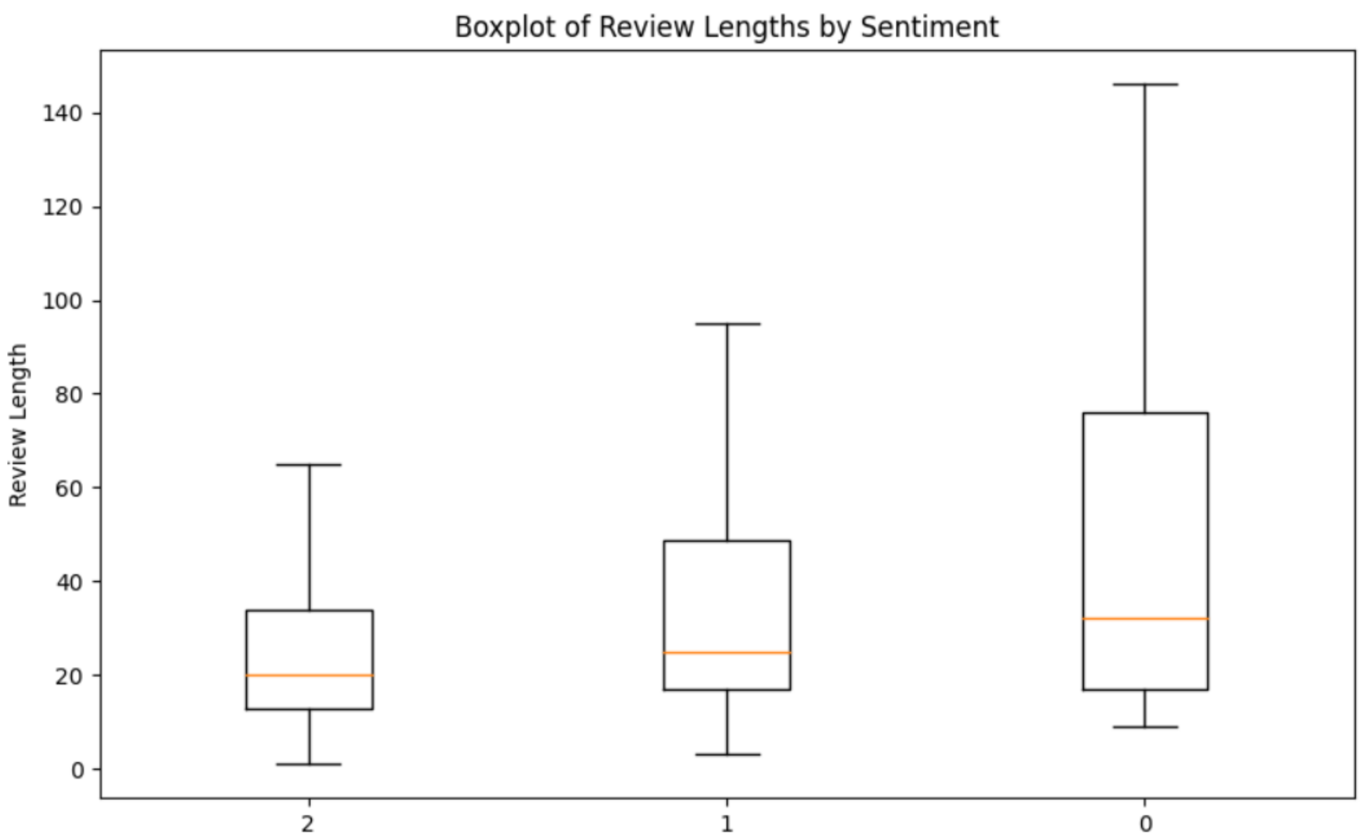
Visualization


Sentiment Distribution
Count
Sentiment
3500
3000
2500
2000
1500
1000
500
0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5


Word Frequency Distribution
Frequency
Words
3500
3000
2500
2000
1500
1000
500
0
I
The
great
tablet
use
love
This
it.
It
easy


Primary Category Distribution
Count
2500
2000
1500
1000
500
0
Electronics
Electronics,Hardware
Office Supplies,Electronics
Electronics,Media
Primary Category


Review Length Distribution
Frequency
0
500
1000
1500
2000
2500
3000
3500
0
200
400
600
800
1000
1200
1400
1600
Review Length


Boxplot of Review Lengths by Sentiment
Review Length
0
20
40
60
80
100
120
140
2
1
0