Sentiment Analysis – Amazon Fine Food Reviews

Introduction

It has been observed in recent years the significant increase in research to understand the sentiment in textual resources.[1] One of such research is termed Sentiment Analysis (also known as emotion AI or opinion mining)[2]. Sentiment Analysis is a Natural Language Processing (NLP) Technique that mines and extracts important information from text or images to determine the sentiments of the source material. [3] This technique helps businesses better understand the social sentiment and customer response behind their service, brand, or product by analysing their reviews and social media posts to determine customer satisfaction and dissatisfaction.[4] The sentiments expressed in blogs and news can also employed to predict market movements.[5] Due to recent advances in Deep Learning (DL), algorithms' ability to analyse text has significantly improved.[6]

Advanced artificial intelligence techniques are an effective way to conduct in-depth research. Customers provide their opinions and feedback on products and services on different social media platforms and forums,[7] therefore, classifying customer opinions and reviews regarding a brand is extremely important. Owing to the recent advancement of recurrent neural networks and its capacity to construct whole-sentence representations based on sentence structure, Deep Learning has begun to surpass all other approaches.

Background

There have already been published several outstanding research papers on this topic (e.g. unsupervised sentiment neuron), by Stanford (e.g., Sentiment Analysis) and OpenAi.[8] In a research by Xu Yun et al from Stanford University, the researchers utilized supervised learning algorithms, like, perceptron algorithm, naive bayes and supporting vector machine to predict some Yelp reviews rating on the dataset. The researchers utilized hold out cross validation, 70% data for training data and 30% data as testing data. The precision and recall values were decided using various classifiers.

In the paper Amazon Reviews, business analytics with sentiment analysis[9], in order to construct a model, the researchers extracted the sentiments from the reviews and examined the results[10]; they claimed to have achieved high accuracy with the tool, using Multinomial Naive Bayesian(MNB) and support vector machine as the primary classifiers.[11]

Similar to the above papers, in my project I will be using Naive Bayesian, Supporting Vector Machine as well as Deep-Learning approaches to examine the reviews rating from the Amazon Fine Foods Reviews dataset from Kaggle[12]. After building the models, I will analyse the models' accuracy to better understand how these algorithms function with sentiment analysis tasks.

---

1 Tan, W., Wang, X. & Xu, X. (2018) Sentiment Analysis for Amazon Reviews. Cs229.stanford.edu. Available online: https://cs229.stanford.edu/proj2018/report/122.pdf [Accessed 8/7/2022].

2 *Sentiment Analysis*, 2022b). https://en.wikipedia.org/w/index.php?title=Sentiment_analysis&oldid=1086472834.

3 A. Mohan (-09-13T16:20:47.945Z) *Sentiment Analysis on Amazon Food Reviews: From EDA to Deployment.* Available online: https://towardsdatascience.com/sentiment-analysis-on-amazon-food-reviews-from-eda-to-deployment-f985c417b0c [Accessed May 18,2022].

4 N. Shrestha and F. Nasoz, 'Deep Learning Sentiment Analysis of Amazon.Com Reviews and Ratings'. *International Journal on Soft Computing, Artificial Intelligence and Applications,* 8, 1, Feb 28, (2019), 1-15.

5 Shrestha and Nasoz, 'Deep Learning Sentiment Analysis of Amazon.Com Reviews and Ratings'. , 1-15

6 Mohan *Sentiment Analysis on Amazon Food Reviews: From EDA to Deployment.*

7 N. Selvaraj (-09-12T18:01:49.662Z) *A Beginner's Guide to Sentiment Analysis with Python.* Available online: https://towardsdatascience.com/a-beginners-guide-to-sentiment-analysis-in-python-95e354ea84f6 [Accessed May 18,2022].

8 From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews

9 M. S. Elli and Y. Wang, 'Amazon Reviews, Business Analytics with Sentiment Analysis'. (2015).

10 Elli and Wang, 'Amazon Reviews, Business Analytics with Sentiment Analysis'.

11 Elli and Wang, 'Amazon Reviews, Business Analytics with Sentiment Analysis'.

12 *Amazon Fine Food Reviews* Available online: https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews

Objectives

The main objective is to extract the sentiments from the customer reviews and examine the rate of recurrence of the sentiments. After this task, a second objective is to build, train and test an ML model to classify the customer reviews into positive and negative sentiments.[13] Following are the objectives to be achieved:

- Evaluate the sentiments created from the customer reviews[14]
- Examine the relationship between customer reviews as regards to different products.[15]
- Increase the classification accuracy. [16]

The Dataset

The Amazon Fine Foods Reviews dataset[17] comprises reviews of fine foods from amazon.[18] There are over 500,000 reviews in the data, covering more than 10 years up to October 2012. Product and user information, ratings, and a plain text review are all included in the Reviews dataset.[19]

The dataset contains the following columns:[20]

1. Product Id: product identifier
2. User Id: user identifier
3. Profile Name: user's profile name
4. Helpfulness Numerator: Number of customers who considered the review to be helpful
5. Helpfulness Denominator: Number of customers who indicated whether they found the review helpful or not
6. Score: review rating between 1 and 5
7. Time: timestamp of the review
8. Summary: review summary
9. Text: customer review

Methodology

The Sentiment Analysis methodology is reliant on machine learning algorithms and NLP to establish the sentiments of online reviews.[21] In this project the focus will be to analyse the sentiments on Amazon Fine Foods Reviews.[22] On the Amazon platform, customers are able to review and rate products on a scale of five.[23] The range of 1 to 5 star ratings stand for very negative, negative, neutral, positive, very positive experiences by the customers.[24] Should a customer give 5 stars, it means that they had a very positive experience with the product received, however, if the rating is 1 star, this means that the customer had a very negative experience.[25]

---

[13] A. R. Hamdallah, *Amazon Reviews using Sentiment Analysis* . (Rochester Institute of Technology. 15 December). [Accessed Aug 7, 2022]],
[14] A. R. Hamdallah, *Amazon Reviews using Sentiment Analysis* . (Rochester Institute of Technology. 15 December). [Accessed Aug 7, 2022]],
15 A. R. Hamdallah, Amazon Reviews using Sentiment Analysis . (Rochester Institute of Technology. 15 December). [Accessed Aug 7, 2022]],
16 A. R. Hamdallah, Amazon Reviews using Sentiment Analysis . (Rochester Institute of Technology. 15 December). [Accessed Aug 7, 2022]],
[17] Amazon Fine Food Reviews. (n.d.) Kaggle.com. Available online: https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews [Accessed 7/5/2022].
[18] Amazon Fine Food Reviews. (n.d.) Kaggle.com. Available online: https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews [Accessed 7/5/2022].
19 Amazon Fine Food Reviews. (n.d.) Kaggle.com. Available online: https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews [Accessed 7/5/2022].
20 A. Mohan (-09-13T16:20:47.945Z) *Sentiment Analysis on Amazon Food Reviews: From EDA to Deployment.* Available online: https://towardsdatascience.com/sentiment-analysis-on-amazon-food-reviews-from-eda-to-deployment-f985c417b0c [Accessed Aug 7,2022].
21 A. R. Hamdallah, *Amazon Reviews using Sentiment Analysis* . (Rochester Institute of Technology. 15 December). [Accessed 9 August 2022], .
22 A. R. Hamdallah, Amazon Reviews using Sentiment Analysis . (Rochester Institute of Technology. 15 December). [Accessed Aug 7, 2022]],
23 A. R. Hamdallah, Amazon Reviews using Sentiment Analysis . (Rochester Institute of Technology. 15 December). [Accessed Aug 7, 2022]],
24 Ummara Ahmed Chauhan et al., 'A Comprehensive Analysis of Adverb Types for Mining User Sentiments on Amazon Product Reviews'. *World Wide Web,* 23, 1811-1829, (2020).
25 A. R. Hamdallah, Amazon Reviews using Sentiment Analysis . (Rochester Institute of Technology. 15 December). [Accessed Aug 7, 2022]],

There are several essential steps to follow in sentiment analysis, to accomplish high-quality results:[26]

- Understand the data: Firstly, have a look at the data to understand what it is and what type of parameters are in the data[27]
- Data Cleaning: Null values and empty entries, non-relevant data and duplicates are all removed from the dataset. This ensures that a clean dataset is analysed.
- Exploratory data analysis: EDA is a significant process as it ensures that the model building and evaluation is easy and without errors.
- Pre-processing: Unstructured text data can be pre-processed in several ways so that computers can understand it for analysis. Ensure the text is lowercase, remove punctuations, html tags, alphanumeric words, words with repeated characters, and stopwords.[28] The text frequency will be visualised using WordCloud.
- Sentiment classification: The text data is then categorised into two sentiments - positive and negative.
- The results will be displayed on bar charts, pie charts, and line charts, for better understanding of the results and online reviews trends.[29]
- Model Evaluation — The model performance will then be evaluated, and predictions will be made.

Dataset Claims

As a first step, I checked that the claimed data information is accurate:

568,454 reviews
256,059 users
74,258 products
260 users with > 50 reviews
Reviews from Oct 1999 - Oct 2012
Number of Attributes/Columns in data: 10

This table confirms the dataset information.

Attribute	Claim	Actual	Match
Number of reviews	568454	568454	YES
Distinct users	256059	256059	YES
Distinct products	74258	74258	YES
Users with more than 50 reviews	260	260	YES
Reviews Time Range	Oct 1999 - Oct 2012	Oct 1999 - Oct 2012	YES

No. of Datapoints : 568454
No. of Features : 10

Data Cleaning and EDA

In order to ensure that the data is clean and ready for analysis, I check for any missing values. I can see that there are 27 missing values under “Summary” and there are 16 missing values under “ProfileName”.

	Total	Percentage
Summary	27	0.004750
ProfileName	16	0.002815

On further examination, although ProfileName have these missing values, the corresponding Userid is present in the dataset, which is enough for identifying the review writer. Although there are missing values under Summary, the corresponding “Text” is present, which may provide more information about the review context. Also, it can be

26 A. R. Hamdallah, Amazon Reviews using Sentiment Analysis . (Rochester Institute of Technology. 15 December). [Accessed Aug 7, 2022]],
27 Das, M. (2020) Sentiment Analysis on Amazon Fine Food Reviews by using Linear Machine Learning Models. International Journal for Research in Applied Science and Engineering Technology, 8(9), 675-678. [Accessed Aug 7, 2022],

28 M. Kosaka (-11-23T22:13:57.299Z) *Cleaning & Preprocessing Text Data for Sentiment Analysis.* Available online: https://towardsdatascience.com/cleaning-preprocessing-text-data-for-sentiment-analysis-382a41f150d6 [Accessed Aug 7,2022].
29 A. R. Hamdallah, Amazon Reviews using Sentiment Analysis . (Rochester Institute of Technology. 15 December). [Accessed Aug 9, 2022],

noted that all of reviews corresponding to the missing Summary values, are duplicated – everything, except the productid is duplicated.

An example of a duplicated review:

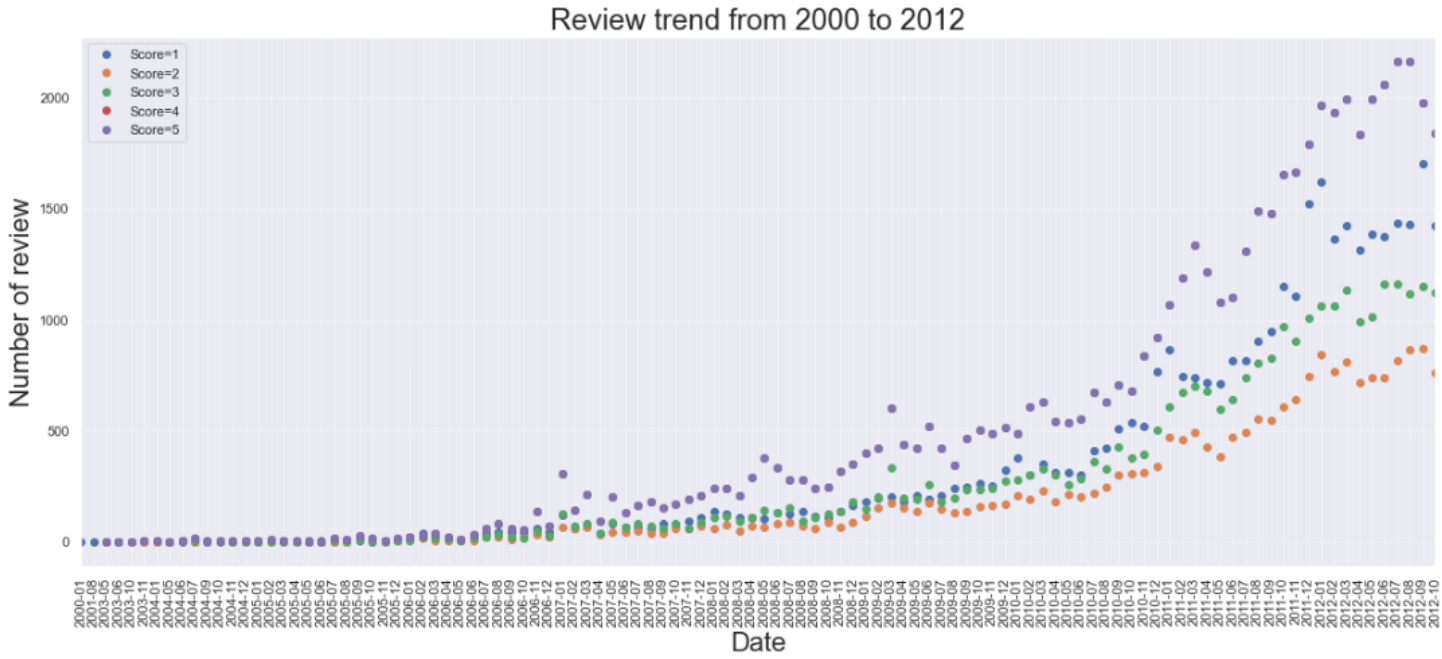
	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
33958	33959	B00412W76S	A3TJPSWY2HE4BS	S. Layton "homeschool blogger"	1	24	2	1173312000	NaN	I only used two maybe three tea bags and got p...
40548	40549	B00020HHRW	A3TJPSWY2HE4BS	S. Layton "homeschool blogger"	1	24	2	1173312000	NaN	I only used two maybe three tea bags and got p...
101106	101107	B0014B0HWK	A3TJPSWY2HE4BS	S. Layton "homeschool blogger"	1	24	2	1173312000	NaN	I only used two maybe three tea bags and got p...

The duplicated reviews are deleted, with only the first one being kept for analysis. This reduced the data to around 69%.

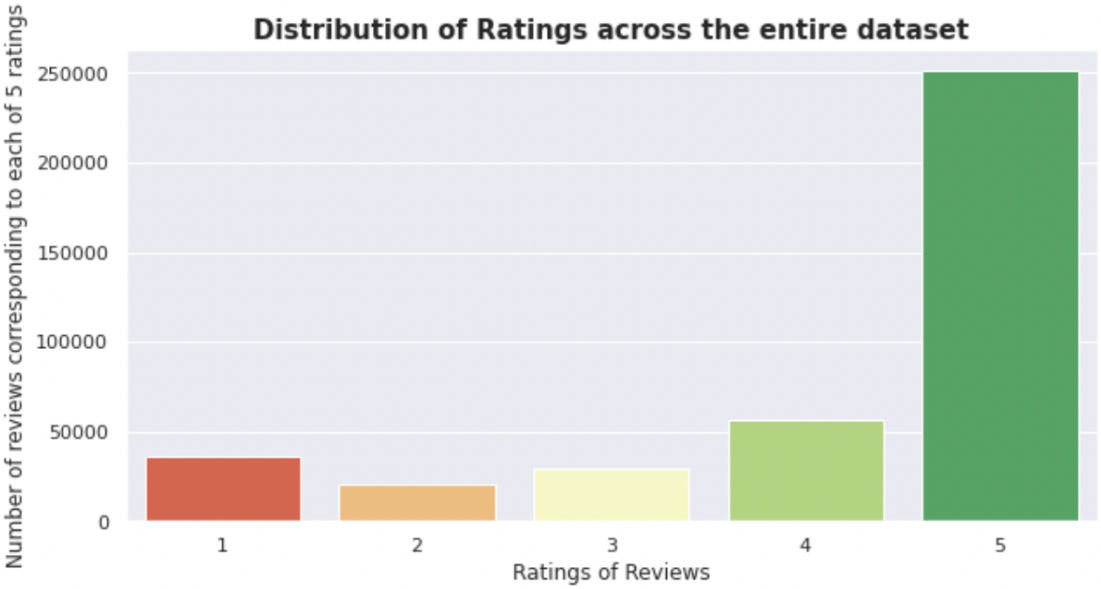
	Count	Percentage of Total
Duplicate Reviews	174521	30.70098899823029
Original/Remaining Reviews	393933	69.29901100176971

Analysing the reviews over the years

Looking at the review trend, the number of reviews is consistent from 2001 up to 2006. However, the number of reviews increased, especially the number of 5-star ratings. This could be due to the increase of customers on the platform.[30] This is confirmed by the distribution of ratings(scores) across the dataset.

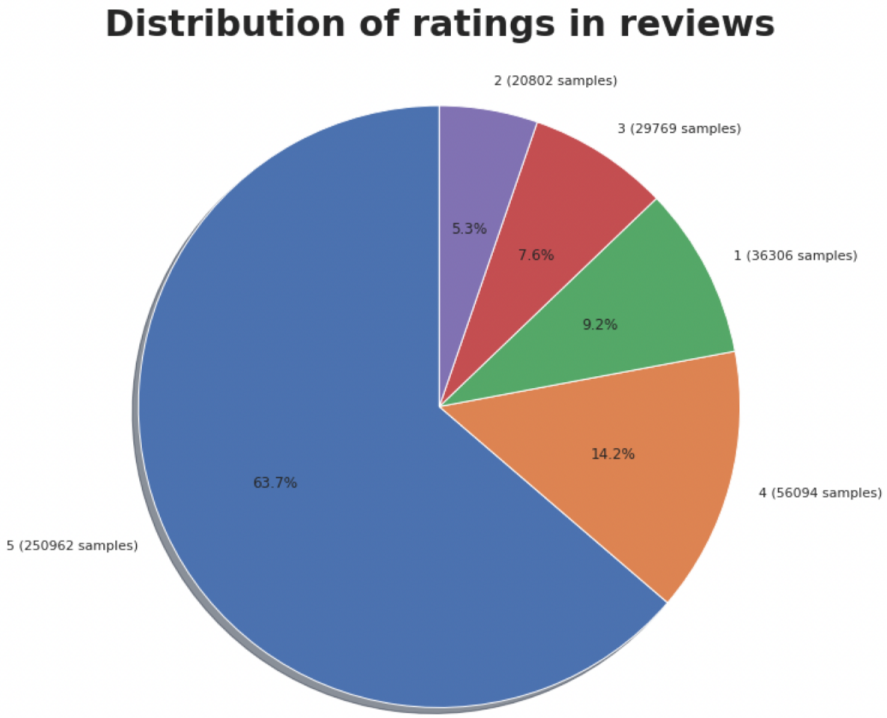


30 Mohan *Sentiment Analysis on Amazon Food Reviews: From EDA to Deployment.*

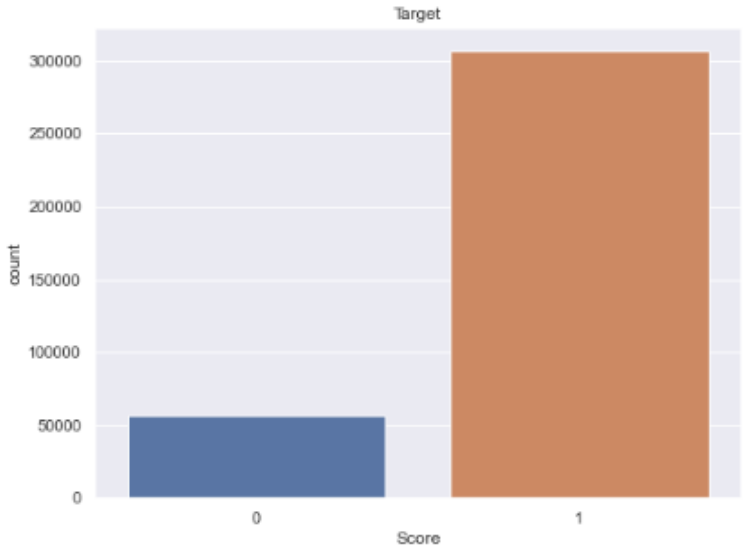



```
1     36306
2     20802
3     29769
4     56094
5    250962
Name: Score, dtype: int64
```

Analysing the Distribution of rating in the reviews, I observed that there are 77.9% of positive reviews with ratings higher than 3 (4 and 5), while there are 14.5% of negative reviews having ratings lower than 3 (1 and 2). The remaining 7.6% are reviews with a 3-star rating. From this, I can note that most of the customers are satisfied with their products.

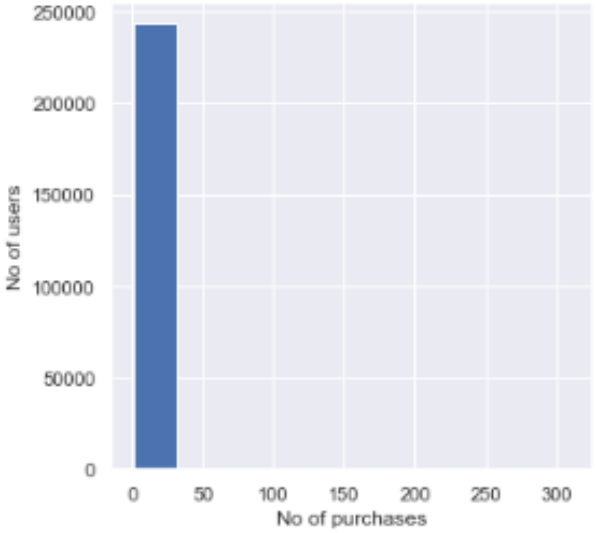


For further analysis and to avoid any confusion, I created a target variable from the review ratings. I filtered out and discarded all neutral ratings (reviews 3-star ratings) and assigned all ratings above 3 to the positive sentiment class and all ratings below 3 to the negative sentiment class. Just as detailed above, there are more positive than negative scores.



Analysing the Productid and Userid

Analysing the number of products bought by the customers, I can note that majority of the users purchased only 1 item; the maximum number of items purchased by one customer was 310. The helpfulness numerator should not be greater than the helpfulness denominator, however, there seem to be some inconsistences with the dataset showing the helpfulness numerator being greater than the helpfulness denominator. After some examination, the datapoints causing the inconsistency were discarded.




```
count    243412.000000
mean          1.496081
std           2.537677
min           1.000000
25%           1.000000
50%           1.000000
75%           1.000000
max         310.000000
Name: No_of_products_purchased, dtype: float64
```

Data Pre-processing
Text Data

After the above EDA, the data needs to be pre-processed for further analysis, model building and prediction.[31] In this pre-processing stage, the following will be carried out:

- Removal of html tags
- Removal of all and any punctuations and special characters (, or . or # etc.)
- Removal of alpha-numeric words and ensure that words are in English.
- Ensure that the words are lowercase
- Removal of stopwords.

Once the pre-processing is done, visualised a WordCloud of the top words with the highest frequency in the text data to gain a better understanding.

```
There are 93076587 words in the combination of all review.
```


```
CPU times: user 34.9 s, sys: 3.27 s, total: 38.1 s
Wall time: 39.4 s
```

The words "even", "think", "well", "however", "know" in the reviews have a high frequency (big size among the word counts) of occurrence. This means that the words occur more times than other words in the reviews. A WordCloud of the top words with the highest frequency in the positive reviews was also created: I can see the words: "highly recommend", "much better", "great product", "loved", "enjoy", and "perfect". However, there are negative words such as "problem", still noted in the WordCloud.

31 Mohan, A. (2020a) Sentiment analysis on amazon food reviews: From EDA to deployment. Available online: https://towardsdatascience.com/sentiment-analysis-on-amazon-food-reviews-from-eda-to-deployment-f985c417b0c [Accessed Aug 7, 2022].

```
There are 77034691 words in the combination of all positive reviews.
```

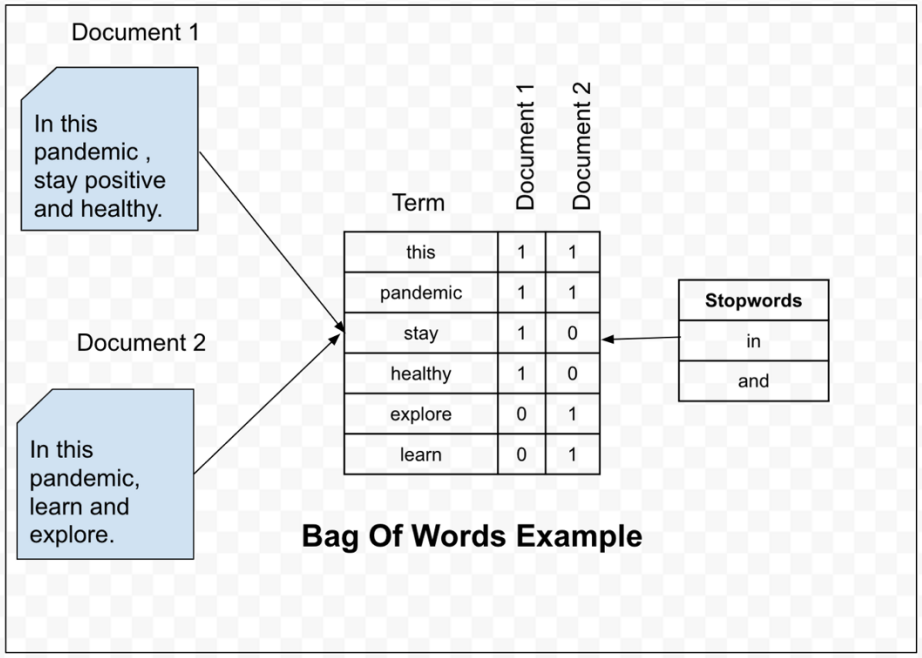

```
CPU times: user 29.4 s, sys: 1.81 s, total: 31.2 s
Wall time: 31.5 s
```

After the pre-processing is done, I ca proceed with the next step in the analysis. I split the data into a train and text data. Before splitting the data, I sort the data based on time as this can influence the reviews.[32]

32 Mohan, A. (2020a) Sentiment analysis on amazon food reviews: From EDA to deployment. Available online: https://towardsdatascience.com/sentiment-analysis-on-amazon-food-reviews-from-eda-to-deployment-f985c417b0c [Accessed Aug 7, 2022].

Bag Of Words (BOW), Language modelling (Unigram) and Word2Vec

This is a common method used in text classification and sentiment analysis. Text, whether a sentence or document, is presented as a bag of words, regardless of grammar or order but keeping the range. The frequency of each word if then utilized as a feature for a classifier training.[33] Various measures of illustrating the text can be calculated after it has been transformed into a BOW.[34] Term frequency or the number of times a term appears in the text, is the most common characteristic computed from the BOW.[35] Term frequencies are weighted based on the inverse of document frequency, or TF-IDF, as a method of normalising them.[36]



(Shrestha, Nikita 2020)

There are different types of language models.[37] A model consisting of a sequence oof n-words is known as an n-gram.[38] unigram model is a model centred on single words. “Model” is a unigram (n = 1), “fruit juice” is a bigram (n = 2), “Bag of Words” is a trigram (n = 3).[39]

With Word2Vec, embeddings are constructed from the trained text from a corpus[40] The embeddings maintain both semantic and syntactic patterns within the text while as well as allowing for the discovery of new similarities and relationships based on vector math. [41] Applying the scikit-learn CountVectorizer(),[42] I created a BOW vectorization, TF-IDF vectorization, average Word2Vec, and TF-IDF Word2Vec techniques.[43] These techniques were then applied

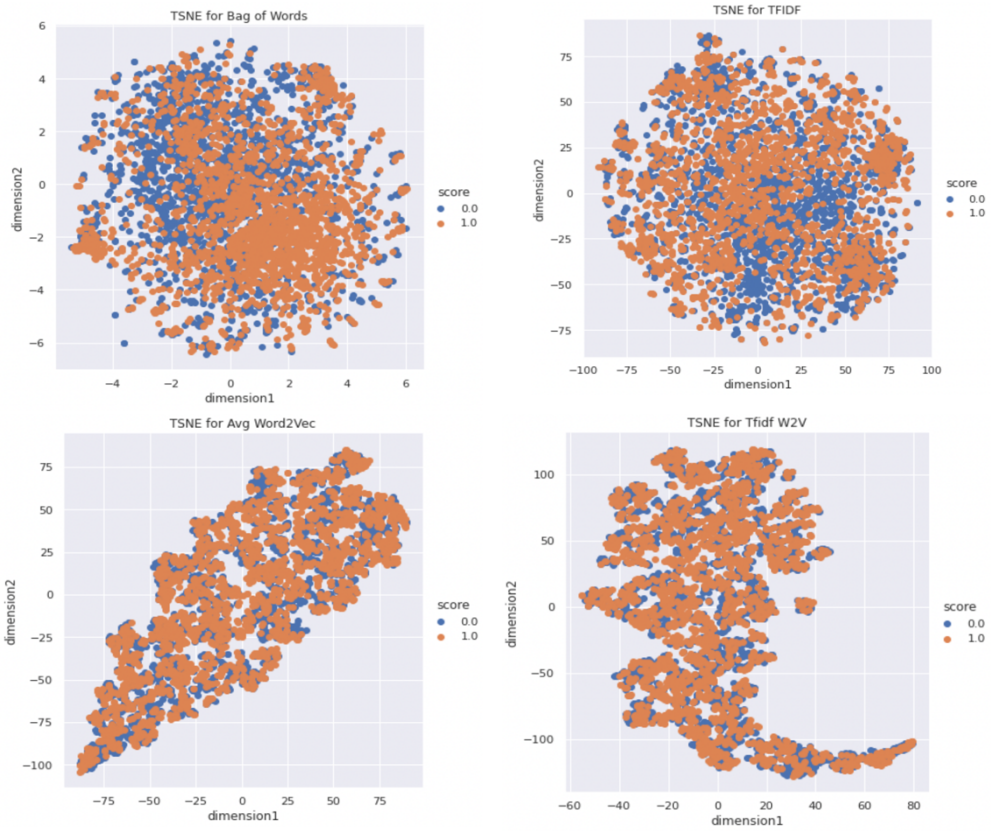
33 *Bag-of-Words Model* (-05-11T07:59:26Z) Available online: https://en.wikipedia.org/w/index.php?title=Bag-of-words_model&oldid=1087243933 [Accessed Aug 8,2022].
34 *Bag-of-Words Model*
35 *Bag-of-Words Model*
36 *Bag-of-Words Model*
37 Nguyen, K. (2020) *N-gram language model*. Medium. Available online: https://medium.com/mti-technology/n-gram-language-model-b7c2fc322799 [Accessed 8/8/2022].
38 Nguyen, K. (2020) *N-gram language model*. Medium. Available online: https://medium.com/mti-technology/n-gram-language-model-b7c2fc322799 [Accessed 8/8/2022].
39 Nguyen, K. (2020) *N-gram language model*. Medium. Available online: https://medium.com/mti-technology/n-gram-language-model-b7c2fc322799 [Accessed 8/8/2022].
40 Chen, A. (2020) An Overview of Word2Vec, Where It Shines, and Where It Didn’t. Medium. Available online: https://medium.com/analytics-vidhya/an-overview-of-word2vec-where-it-shines-and-where-it-didnt-cb671b68a614 [Accessed 5/8/2022].
41 Chen, A. (2020) An Overview of Word2Vec, Where It Shines, and Where It Didn’t. Medium. Available online: https://medium.com/analytics-vidhya/an-overview-of-word2vec-where-it-shines-and-where-it-didnt-cb671b68a614 [Accessed 5/8/2022].
42 sklearn.feature_extraction.text.CountVectorizer. (n.d.) scikit-learn. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html [Accessed 3/8/2022].
43 Mohan, A. (2020b) Sentiment analysis on amazon food reviews: From EDA to deployment. Available online: https://towardsdatascience.com/sentiment-analysis-on-amazon-food-reviews-from-eda-to-deployment-f985c417b0c [Accessed May 18, 2022].

for featuring the text and saved as separate vectors.[44] The unigram method was used for the BOW and TF-IDF and trained the models.[45]

T-distributed stochastic neighbour embedding – TSNE is a method for exploring and visualising high dimensional data.[46] It is a non-linear dimensionality reduction method for embedding data in high dimensionality.[47] Using TSNE, I visualised the data by:

1. Running TSNE at different iterations, while keeping the perplexity constant.[48]
2. After finding the most stable iteration, I ran TSNE again with different iterations in order to get an improved outcome.[49]
3. I ran the TSNE again with the same, after obtaining a stable outcome.[50]

However, after conducting the above, it can be noted that the TSNE was not able to clearly separate the positive and negative reviews/points.[51]



[44] Mohan, A. (2020b) Sentiment analysis on amazon food reviews: From EDA to deployment. Available online: https://towardsdatascience.com/sentiment-analysis-on-amazon-food-reviews-from-eda-to-deployment-f985c417b0c [Accessed May 18, 2022].
[45] Mohan, A. (2020b) Sentiment analysis on amazon food reviews: From EDA to deployment. Available online: https://towardsdatascience.com/sentiment-analysis-on-amazon-food-reviews-from-eda-to-deployment-f985c417b0c [Accessed May 18, 2022].
[46] t-distributed stochastic neighbor embedding - Wikipedia. (n.d.) En.wikipedia.org. Available online: https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding [Accessed 3/8/2022].
[47] t-distributed stochastic neighbor embedding - Wikipedia. (n.d.) En.wikipedia.org. Available online: https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding [Accessed 3/8/2022].
[48] Mohan, A. (2020b) Sentiment analysis on amazon food reviews: From EDA to deployment. Available online: https://towardsdatascience.com/sentiment-analysis-on-amazon-food-reviews-from-eda-to-deployment-f985c417b0c [Accessed May 18, 2022].
[49] Mohan, A. (2020b) Sentiment analysis on amazon food reviews: From EDA to deployment. Available online: https://towardsdatascience.com/sentiment-analysis-on-amazon-food-reviews-from-eda-to-deployment-f985c417b0c [Accessed May 18, 2022].
[50] Mohan, A. (2020b) Sentiment analysis on amazon food reviews: From EDA to deployment. Available online: https://towardsdatascience.com/sentiment-analysis-on-amazon-food-reviews-from-eda-to-deployment-f985c417b0c [Accessed May 18, 2022].
[51] Mohan, A. (2020b) Sentiment analysis on amazon food reviews: From EDA to deployment. Available online: https://towardsdatascience.com/sentiment-analysis-on-amazon-food-reviews-from-eda-to-deployment-f985c417b0c [Accessed May 18, 2022].

Experiments

Classification is important in machine learning and data mining.[52] The objective in classification is to build a classifier with a collection of training samples with class labels[53]

Naive Bayes

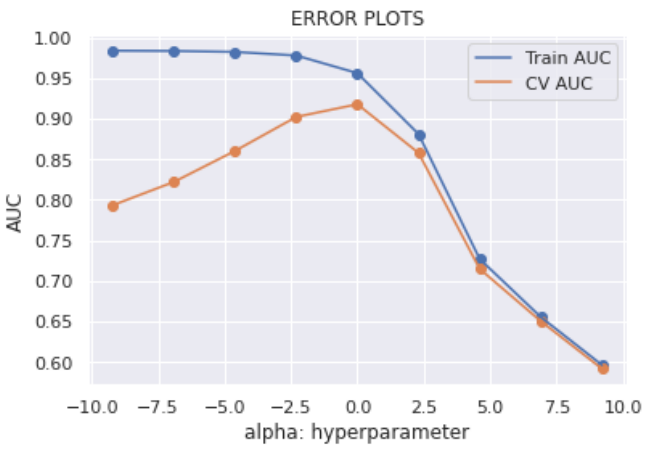
Naive Bayes is one of the most common supervised learning algorithms for classification problems. It assumes conditional independence of the attributes given the value of the class variable:[54]

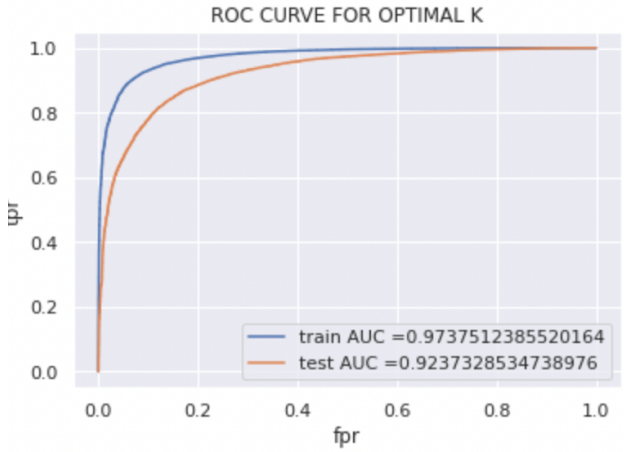
$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

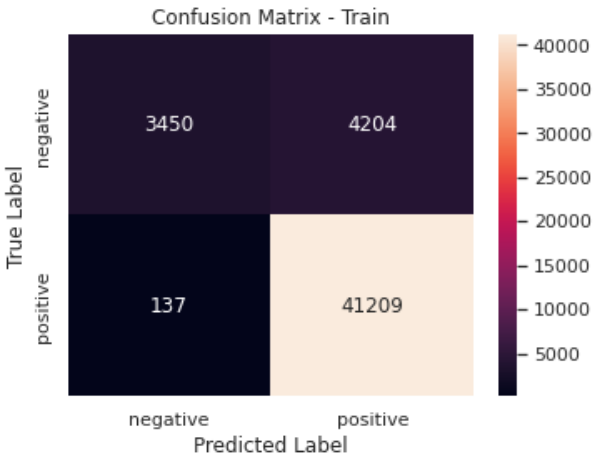
$$\Downarrow$$

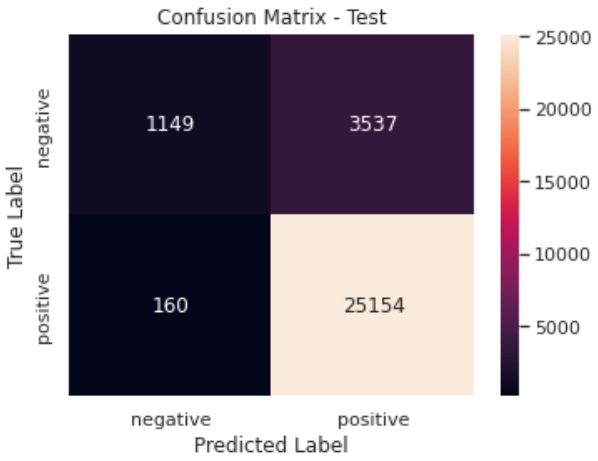
$$y = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

I created a Naïve Bayes model as a baseline model for evaluation. As I used manual cross-validation, I split the data into a train, test and cv set and applied multinominal Naïve Bayes to the Bag of Word and TF-IDF features.









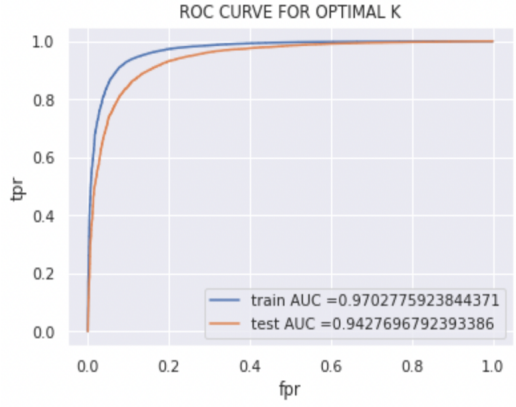
[52] Zhang, H. (2004) *The Optimality of Naive Bayes*. Academia.edu. Available online: https://www.academia.edu/5833043/The_Optimality_of_Naive_Bayes [Accessed 6/8/2022].
[53] Zhang, H. (2004) *The Optimality of Naive Bayes*. Academia.edu. Available online: https://www.academia.edu/5833043/The_Optimality_of_Naive_Bayes [Accessed 6/8/2022].
[54] 1.9. Naive Bayes. (n.d.) scikit-learn. Available online: https://scikit-learn.org/stable/modules/naive_bayes.html [Accessed 10/7/2022].

After tunning the hyperparameters, these were the results:

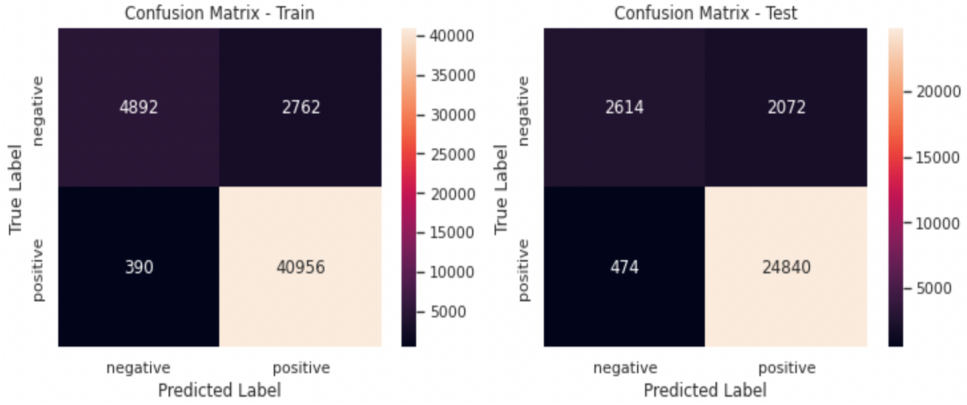
Vector	Algorithm	Hyperparameter-alpha	Train AUC	Test AUC
bow	naive-bayes	0.1	0.9737212081171733	0.9108361295403316
tfidf	naive-bayes	0.1	0.9737512385520164	0.9237328534738976

Support Vector Machine

Support Vector Machines(SMV) are supervised learning methods for classification and regression analysis of data.[55] I tried Linear SVM, which is a method used for linearly separable data.[56] I obtained the following results, with Linear SVM on TF-IDF resulting in the highest AUC on the test data[57] and Average Word2Vec resulting in the most generalised model.[58]



Area under train roc 0.9702775923844371
Area under test roc 0.9427696792393386

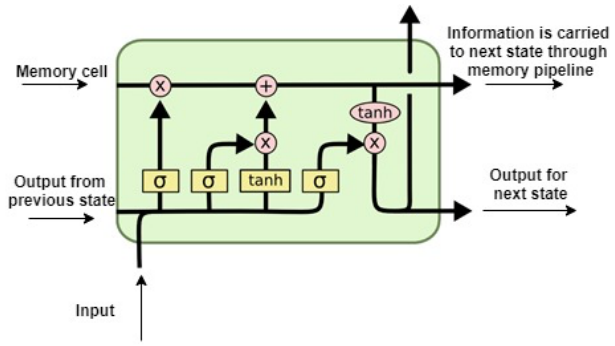


Vector	Algorithm	kernel	penalty	Hyperparam-alpha	Hyperparam-C	gamma	Train AUC	Test AUC
bow	SVM	linear	l2	0.001	-	-	0.9702775923844371	0.9427696792393386
tfidf	SVM	linear	l2	0.0001	-	-	0.9677053364122215	0.9500039132903875
avg-w2v	SVM	linear	l2	0.0001	-	-	0.9101625982736657	0.9054630562288741
tfidf-w2v	SVM	linear	l2	0.001	-	-	0.8882868065957238	0.9069258613732138

[55] 1.4. Support Vector Machines. (n.d.) scikit-learn. Available online: https://scikit-learn.org/stable/modules/svm.html [Accessed 6/8/2022].
[56] Support Vector Machine (SVM) Algorithm - Javatpoint. (n.d.) www.javatpoint.com. Available online: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm [Accessed 16/8/2022].
[57] Mohan, A. (2020b) Sentiment analysis on amazon food reviews: From EDA to deployment. Available online: https://towardsdatascience.com/sentiment-analysis-on-amazon-food-reviews-from-eda-to-deployment-f985c417b0c [Accessed May 18, 2022].
[58] Mohan, A. (2020b) Sentiment analysis on amazon food reviews: From EDA to deployment. Available online: https://towardsdatascience.com/sentiment-analysis-on-amazon-food-reviews-from-eda-to-deployment-f985c417b0c [Accessed May 18, 2022].

Deep Learning Approach

I used an LSTM model. Long Short Term Memory networks – LSTM are a type of Recurring Neural Network with the[59] capacity of avoid long-term dependency problems, thus remembering information for a longer period.[60]



(Understanding LSTM Networks -- colah's blog, 2015)

Firstly, the inputs were tokenized, and an index vocabulary was built for the training corpus, [61] which was then transformed to numeric vectors.[62] I used a sequential model with an Embedding that accepts input length, input and output dimensions, I added a SpatialDropout1D layer, LSTM layer, and Dense layer. Lastly, I added "Relu" as the activation unit as it tends to perform better than other activation units.[63]

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 embedding_1 (Embedding)      (None, 1217, 128)         12257152

 spatial_dropout1d_1 (Spatia  (None, 1217, 128)        0
 lDropout1D)

 lstm_1 (LSTM)                (None, 64)                49408

 dense_1 (Dense)              (None, 1)                 65

=================================================================
Total params: 12,306,625
Trainable params: 12,306,625
Non-trainable params: 0
_________________________________________________________________
None
```

InputLayer		Embedding		SpatialDropout1D		LSTM		Dense	
input:	output:	input:	output:	input:	output:	input:	output:	input:	output:
[(None, 1217)]		(None, 1217)		(None, 1217, 128)		(None, 1217, 128)		(None, 64)	
[(None, 1217)]		(None, 1217, 128)		(None, 1217, 128)		(None, 64)		(None, 1)	

[59] Understanding LSTM Networks -- colah's blog. (2015) Colah.github.io. Available online: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ [Accessed 7/8/2022].

[60] Understanding LSTM Networks -- colah's blog. (2015) Colah.github.io. Available online: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ [Accessed 7/8/2022].

[61] Sentiment Prediction Platform For Food Reviews. (2022) Medium. Available online: https://medium.com/@nbasatish/sentiment-prediction-platform-for-food-reviews-e77bf29c8a3a [Accessed 4/8/2022].

[62] Sentiment Prediction Platform For Food Reviews. (2022) Medium. Available online: https://medium.com/@nbasatish/sentiment-prediction-platform-for-food-reviews-e77bf29c8a3a [Accessed 4/8/2022].

[63] Sentiment Prediction Platform For Food Reviews. (2022) Medium. Available online: https://medium.com/@nbasatish/sentiment-prediction-platform-for-food-reviews-e77bf29c8a3a [Accessed 4/8/2022].

Results

The model was then trained and evaluated on 5 epochs, outputting the following results:

```
Epoch 1/5
191/191 [==============================] - 749s 4s/step - loss: 0.5220 - accuracy: 0.8120 - val_loss: 0.3520 - val_accuracy: 0.8791
Epoch 2/5
191/191 [==============================] - 743s 4s/step - loss: 0.3030 - accuracy: 0.8948 - val_loss: 0.2646 - val_accuracy: 0.9045
Epoch 3/5
191/191 [==============================] - 742s 4s/step - loss: 0.2472 - accuracy: 0.9155 - val_loss: 0.2424 - val_accuracy: 0.9155
Epoch 4/5
191/191 [==============================] - 741s 4s/step - loss: 0.2381 - accuracy: 0.9174 - val_loss: 0.4773 - val_accuracy: 0.8923
Epoch 5/5
191/191 [==============================] - 741s 4s/step - loss: 0.2679 - accuracy: 0.9104 - val_loss: 0.3499 - val_accuracy: 0.9036
```

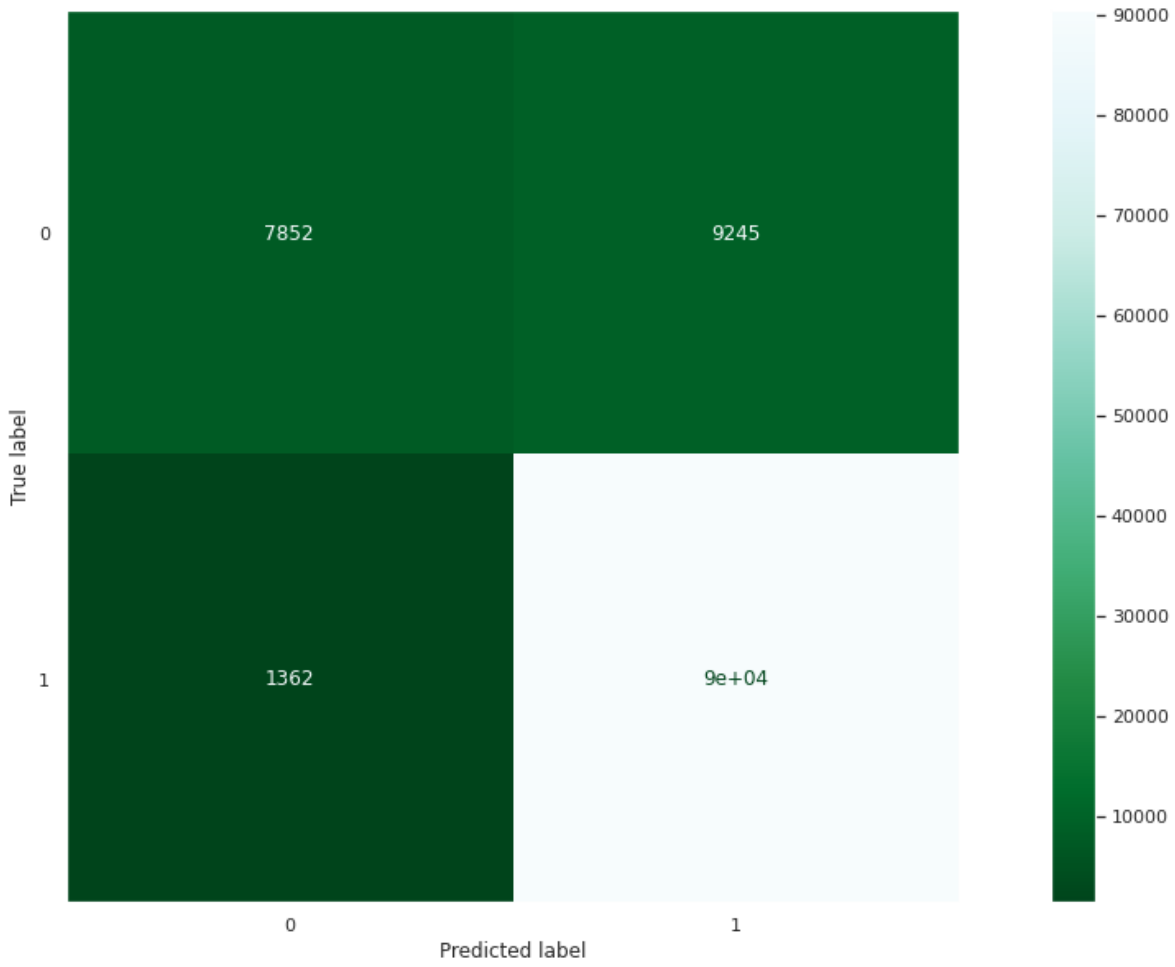
I note the training accuracy as 91%, validation accuracy as 90%, which is high and good for the model. The model prediction outputted the following results. The model's final accuracy score on prediction is 90%

```
The model accuracy score is: 0.902648775652557
The model precision score is: 0.8986601467328416
The model recall score is: 0.902648775652557
The model F1-score is: 0.890067797667348
              precision    recall  f1-score   support

           0       0.85      0.46      0.60     17097
           1       0.91      0.99      0.94     91859

    accuracy                           0.90    108956
   macro avg       0.88      0.72      0.77    108956
weighted avg       0.90      0.90      0.89    108956
```

	Pred. negative	Pred. positive
Act. negative	7852	9245
Act. positive	1362	90497



Conclusion

The model seems to have performed well with the given inputs, training and evaluation. I note that there is room for improvement, maybe with an increased data sample used for the training an evaluation, the outcome may be improved.

References:

Elli, M. & Wang, Y. (2015) Amazon Reviews, business analytics with sentiment analysis. Available online: https://www.semanticscholar.org/paper/%2C-business-analytics-with-sentiment-analysis-Elli-Wang/bbb4b549cae71fb74680764fd3fe4d72b705f4f4 [Accessed 6/6/2022].

Hamdallah, A. (2021) *Amazon Reviews using Sentiment Analysis*. RIT Scholar Works. Available online: https://scholarworks.rit.edu/theses/11060/ [Accessed 4/8/2022].

Tan, W., Wang, X. & Xu, X. (2018) *Sentiment Analysis for Amazon Reviews*. Cs229.stanford.edu. Available online: https://cs229.stanford.edu/proj2018/report/122.pdf [Accessed 8/7/2022].

Ummara Ahmed Chauhan, Muhammad Tanvir Afzal, Abdul Shahid, Moloud Abdar, Mohammad Ehsan Basiri & Xujuan Zhou (2020) A comprehensive analysis of adverb types for mining user sentiments on amazon product reviews. *World Wide Web,* (23, 1811-1829), .

Kosaka, M. (2020) Cleaning & preprocessing text data for sentiment analysis. Available online: https://towardsdatascience.com/cleaning-preprocessing-text-data-for-sentiment-analysis-382a41f150d6 [Accessed Aug 7, 2022].

Amazon Fine Food Reviews. (n.d.) Kaggle.com. Available online: https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews [Accessed 7/5/2022].

Sentiment analysis - Wikipedia. (n.d.) En.wikipedia.org. Available online: https://en.wikipedia.org/wiki/Sentiment_analysis [Accessed 7/5/2022].

Sentiment analysis - Wikipedia. (n.d.) En.wikipedia.org. Available online: https://en.wikipedia.org/w/index.php?title=Sentiment_analysis&oldid=1086472834 [Accessed 16/5/2022].

Bag-of-words model. (2022) Available online: https://en.wikipedia.org/w/index.php?title=Bag-of-words_model&oldid=1087243933 [Accessed Aug 8, 2022].

Mohan, A. (2020b) Sentiment analysis on amazon food reviews: From EDA to deployment. Available online: https://towardsdatascience.com/sentiment-analysis-on-amazon-food-reviews-from-eda-to-deployment-f985c417b0c [Accessed May 18, 2022].

SHUKOOR, N. (2021b) Amazon fine food reviews - EDA, data cleaning, data preprocessing, feature engineering 2/2. Available online: https://www.kaggle.com/code/naushads/1-2-amazon-fine-food-reviews-eda-data-cleaning-fe [Accessed Aug 9, 2022].

SHUKOOR, N. (2021a) Amazon fine food reviews - EDA, data cleaning, data preprocessing, feature engineering 1/2. Available online: https://www.kaggle.com/code/naushads/1-2-amazon-fine-food-reviews-eda-data-cleaning-fe [Accessed Aug 9, 2022].

Selvaraj, N. (2020) A beginner's guide to sentiment analysis with python. Available online: https://towardsdatascience.com/a-beginners-guide-to-sentiment-analysis-in-python-95e354ea84f6 [Accessed May 18, 2022].

Shrestha, N. (2020) A semantic analysis of amazon fine food reviews. Available online: https://ai.plainenglish.io/semantic-analysis-of-amazon-fine-food-reviews-8225fe11ddb1 [Accessed Aug 15, 2022].

Sentiment Prediction Platform For Food Reviews. (2022) Medium. Available online: https://medium.com/@nbasatish/sentiment-prediction-platform-for-food-reviews-e77bf29c8a3a [Accessed 4/8/2022].

Sinha, P. (2021) 140 python projects with source code. Available online: https://medium.datadriveninvestor.com/140-python-projects-with-source-code-fa12c9e2aeac [Accessed May 20, 2022].

Nguyen, K. (2020) *N-gram language model*. Medium. Available online: https://medium.com/mti-technology/n-gram-language-model-b7c2fc322799 [Accessed 8/8/2022].

Team, K. Keras documentation: Layer activation functions. Available online: https://keras.io/api/layers/activations/ [Accessed May 18, 2022].