Analyzing Amazon's Alexa Reviews Using Natural Language Processing

Session Spring 2021

Submitted By

Rishika Chaudhary

(1710110276)

Under Supervision

of

Dr. Sonal Singhal

(Assoc. Prof., Electrical Engineering Department)

Department Of Electrical Engineering
School Of Engineering
Shiv Nadar University
(Spring 2021)

ABSTRACT

In recent decades, as online marketplaces have become popular, the online vendors and retailers ask their customers to express their views about the items that they purchased. During their search for goods, these reviews were taken into account by other users. The industry has therefore found the source of providing the correct product searched by the consumer on the basis of user feedback using the principle of sentimental analysis. In a business setting, sentiment analysis is extremely helpful as it can help understand customer experiences, gauge public opinion, and monitor brand and product reputation. Analysing these opinions can be hard and time consuming for product manufacturers. This project considers the problem of classifying reviews by their overall semantic (positive or negative). The data contains five features (rating, date, variation, verified reviews, and feedback) and over three thousand samples (3,150 samples to be exact). To conduct this research, machine learning algorithms like Random Forest, have been utilized, and an accuracy of 94.28 percent was achieved. Also, advanced natural signal processing libraries like Spacy are used to gain deeper insights into the text.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1

Introduction

Around the globe, there are more than 6,500 daily newspapers selling close to 400 million copies every day. Additionally, there are blogs, micro blogs, periodicals, magazines, fanzines, etc. How can we make sense of all this information? How can we classify it and aggregate it so that we can perform quantitative analysis? Sentiment analysis (or opinion mining) is a technique of natural language processing, that is used to assess whether information is positive, negative or neutral. Sentiment analysis is often carried out on textual data to help marketers monitor customer input on brand and product sentiment and identify customer needs[1].

1.1 Problem Statement

Amazon is one of the giants of e-commerce that people use every day for online transactions where thousands of reviews can be read, dropped by other customers on their favorite items.These reviews provide useful views on a product, such as its property, quality and recommendations, which are valuable[2]. This is not only good for buyers, but also supporting sellers who produce their own items to better understand the customers and their needs.
Most e-commerce websites provide only an average rating (out of 5) for each product.Therefore, it is tedious to identify why people like or dislike a particular product. We aim to solve this problem.This project explores the issue of sentiment classification for online reviews using supervised methods to evaluate the overall customer reviews by categorizing them into positive and negative feelings.Apart from quantitative reviews (which are mostly skewed), amazon also records qualitative reviews. The objective is to assess those text reviews and determine whether they are negative or positive. Not only classification of sentiment but we also focus on determining how strong the sentiment is. Sentiment analysis research focuses on understanding the positive or negative tone of a sentence based on sentence syntax, structure, and content.
Sentiment classification aims to determine the overall intention of a written text which can be of admiration or criticism type. This can be achieved by using machine learning algorithms such as Decision Tree and Random Forest.

1.2 Motivation

As the number of reviews available for a product grows, it is becoming more difficult for a prospective customer to make a good decision as to whether buy the product or not. Different opinions about the same product on one hand and ambiguous reviews on the other hand makes customers more confused to get the right decision. Here the need for analyzing this contents seems crucial for all e-commerce businesses.
Sentiment analysis is a computational research which attempts to address this issue by extracting subjective information from the given texts in natural language, such as opinions and sentiments.

1.3 Dataset

The first step for conducting the research includes data collection for training and testing the classifiers. The dataset found is a tab-separated values (TSV) file. The TSV file can easily be converted into an excel format, to comprehend the data. The data contains 3150 samples and 5 features. The description of features are mentioned below :

- **Rating** - The number of stars given by the customer in their review, with 1 being the lowest and 5 being the highest.
- **Date** - The day on which the review was posted.
- **Variation** - The different model variations of the Amazon's Alexa product. We will dive deeper and analyze the differences between each model in the program
- **Verified Reviews** - Those are the typed reviews made by the Alexa consumers. They are the body of text that will be analyzed in the NLP model.
- **Feedback** - Feedback can be either 0 or 1. Negative feedback equals 0, and positive feedback equals 1. If the customers gave a rating of 2 stars or below, the feedback equals 0. If the customer gave a rating of 3 stars or above, the feedback equals 1.

Chapter 2

Literature Survey

Hui Zhang in May 2019, presented a paper on Sentiment Analysis on Amazon reviews, to explore the application of sentiment analysis and to understand its strength and limitations.This paper used dataset of the Amazon Alexa reviews, and then built a model to predict the sentiment of the comment given the comment declaration by using Python and machine leaning algorithm- Naïve Bayes and logistic regression. SMOTE is used to cope with the unbalanced dataset and AUC/ROC are used to evaluate which method is best.After applying the Naive Bayes and Logistic Regression, the accuracy of the Naive Bayes model is 87.1 per cent and that of Logistic Regression is 87.4 per cent.The limitation of the project is the accuracy of the sentiment analysis prediction is about 80 per cent [3].

Sentiment analysis, or opinion mining, is an active area of study in the field of natural language processing that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions via the computational treatment of subjectivity in text. It is not our intention to review the entire body of literature concerning sentiment analysis. Indeed, such an endeavor would not be possible within the limited space available (such treatments are available in Liu (2012) and Pang Lee (2008))[4]. With the help of sentiment analysis system, this unstructured information can be automatically translated into structured data about products, services, brands, politics, or other topics that people can express(Liu, 2012). This data is useful for marketing analysis, public relations, product reviews, customer service, etc. The classification of emotional polarity has three levels: document level, sentence level and entity and aspect level. The document level focuses on whether a document as a whole expresses negative or positive emotions; while the sentence level focuses on the emotional classification of each sentence; The entity level and the aspect level are the views that people like or don't like about them (Liu,2012; Fang and Zhan, 2015)[5].

Chapter 3

Work Done

3.1 Pre-Processing

To start with our python code, we will first import the necessary libraries required for our analysis and then perform a through exploratory data analysis to get a better understanding of the data provided to us.

3.1.1 Exploratory Data Analysis

An EDA is a thorough examination meant to uncover the underlying structure of a data set and is important for a company because it exposes trends, patterns, and relationships that are not readily apparent.

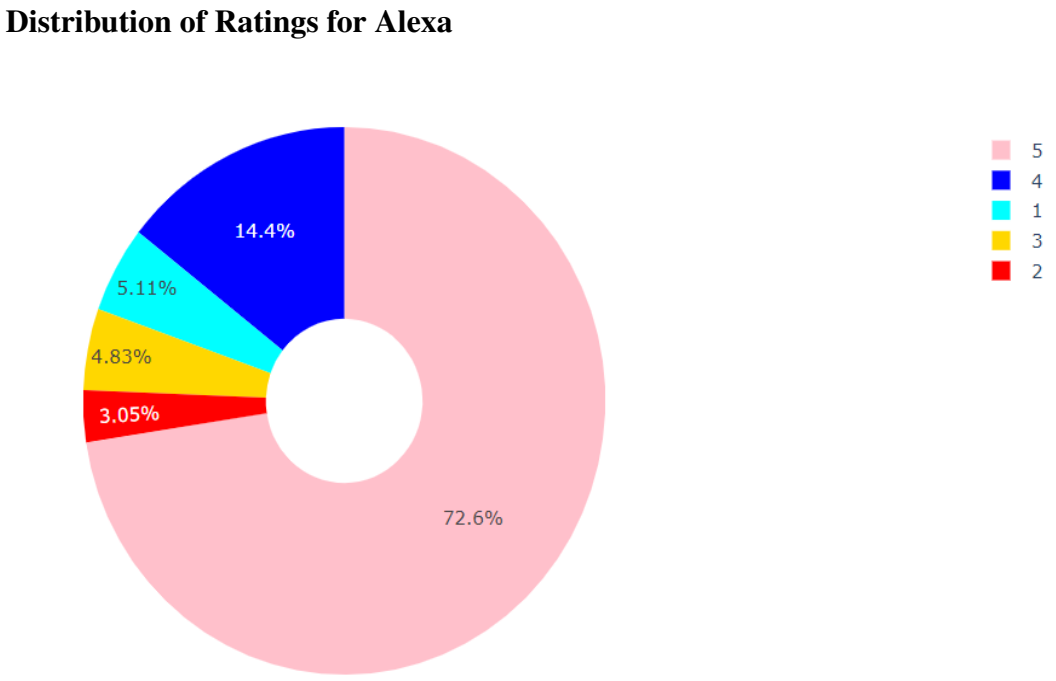

Figure 3.1: Distribution of Ratings for Alexa

By looking at the pie chart above, we can conclude that most of the ratings are positive for Alexa.

- 72.6 %customers have given Alexa a rating of 5 stars.
- 14.4% customers have given Alexa a rating of 4 stars.
- 4.38% of customers have given Alexa a rating of 3 stars.
- 3.05% of customers appear to not like Alexa as much as the other customers and chose to give only a 2-star rating to Alexa, whereas 5.11% people did not like Alexa and decided to give only 1-star rating.

This feedback shows a total of 8.16% of the customers were not satisfied with Alexa. Overall, the ratings feedback is very positive, showing almost 90% of the customers being very satisfied with the product.

Distribution Of Variation in Alexa

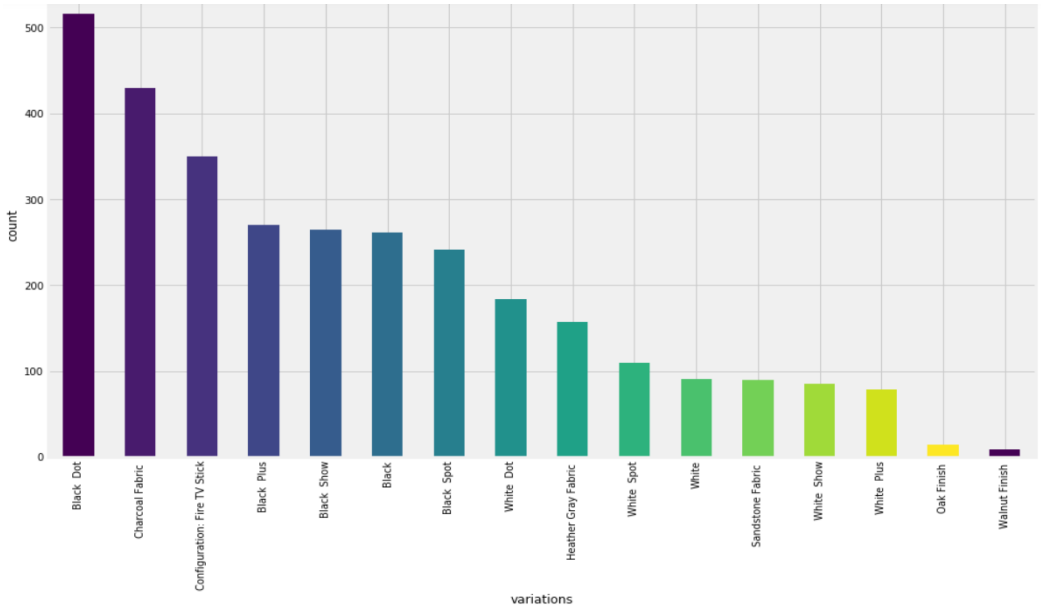

Figure 3.2: Distribution of Variation in Alexa

The bar chart shows all the different variations of Amazon's Alexa and their popularity as well. We see that there are 16 different variations of Alexa models. It goes from Black Dot all the way to Walnut Finish. It is clear that Black Dot is the most popular variation of Alexa with more than 500 units out of the 3150 samples in the dataset.

- Black Dot, Charcoal Fabric and Configuration: Fire TV Stick are the most popular models of Amazon's Alexa

- Oak Finish and Walnut Finish are the least popular variations

Distribution of feedback from Alexa

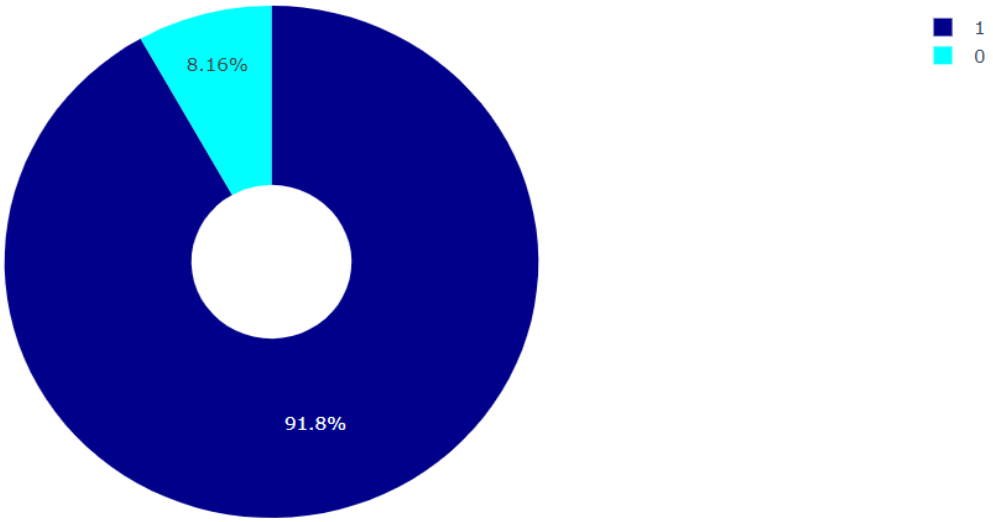

Figure 3.3: Distribution of feedback from Alexa

This pie chart represents the distribution feedback for Amazon's Alexa. 91.8% of customers have given a positive feedback for Alexa (3 stars or above), and only 8% of customers have given a negative feedback to Alexa (2 stars or below). This confirms that Alexa is has a very positive feedback from the majority of its customers, and only a small percentage did not like the product.

Distribution of length reviews

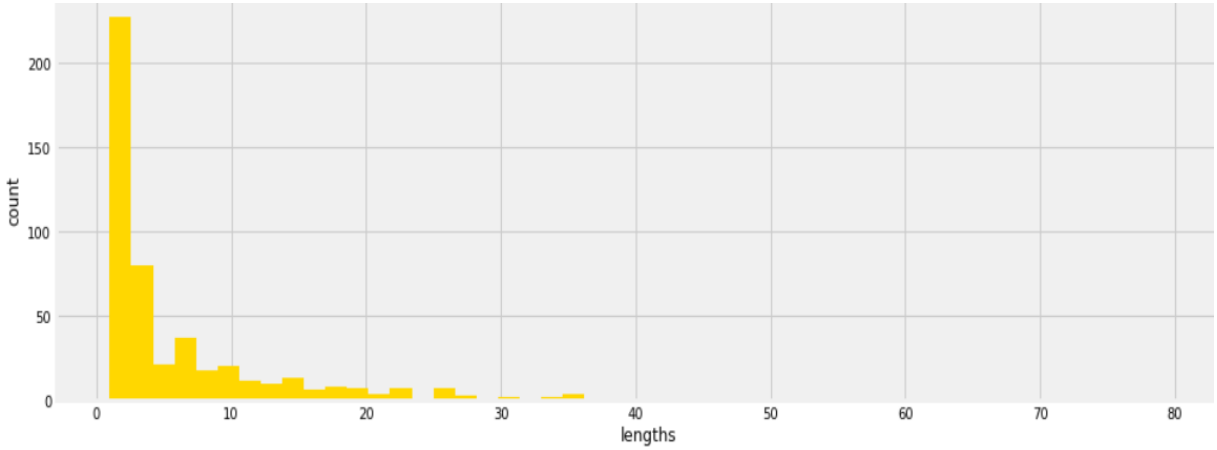

Figure 3.4: Distribution of length reviews

The above Distribution Plot shows the distribution of the length of the reviews written by the customers. This shows what is the average of the length of the reviews written by the customers of Amazon's Alexa. Most of the Reviews are very short, written in less than 3 words total. We can see that most customers write reviews that are between 5 to 20 words long. Very few customers write reviews that are longer than 30 words.

Variation VS Ratings

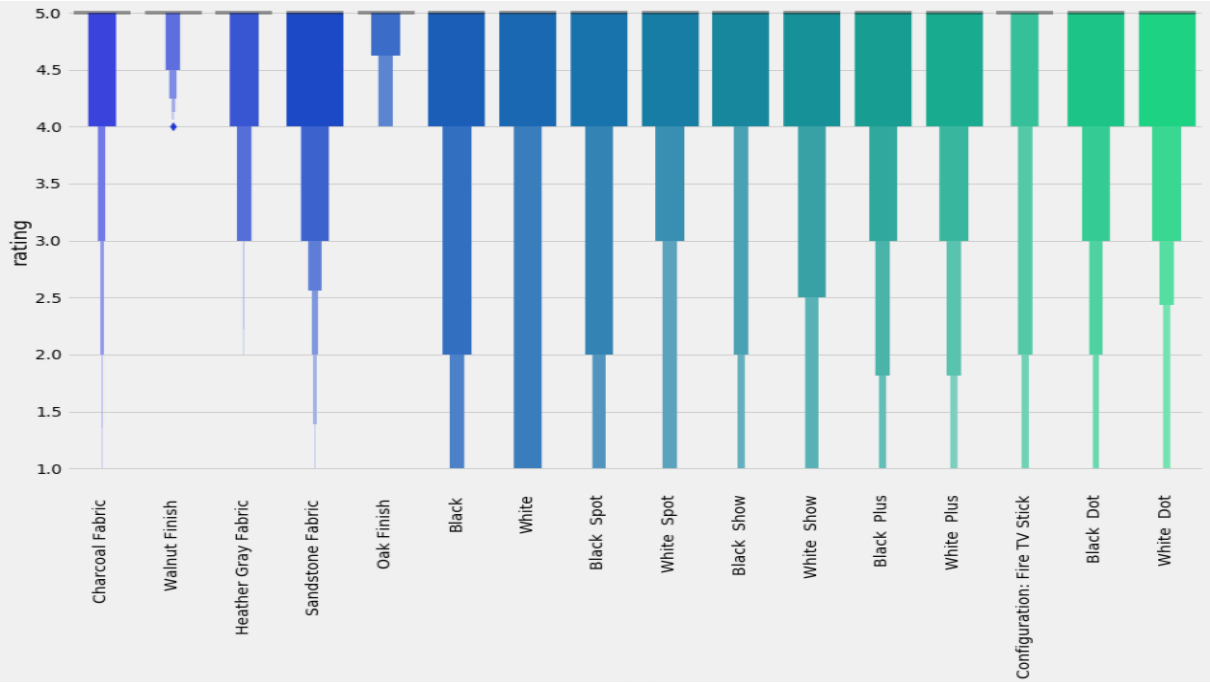

Figure 3.5: Variation VS Ratings

- Even though Walnut Finish and Oak Finish were the least popular variations of Alexa, they have very high ratings ranging from 4.5 to 5 stars. It shows that these Alexa variations are not very common among its customers, however, they have very positive ratings.
- The darker variations of Alexa, such as the Black variation model, have some negative ratings. We can infer this conclusion because of its popularity among the different Alexa model variations. Thus, due to the majority of customers owning this variation, it will have some negative reviews, even though most of them are positive.
- The White model besides being one of the least popular variation models, it is also one of the models that have the most negative ratings.We can see that there is a substantial number of ratings below 3 stars for the White model, especially below 2 stars

Variation vs Length of Ratings

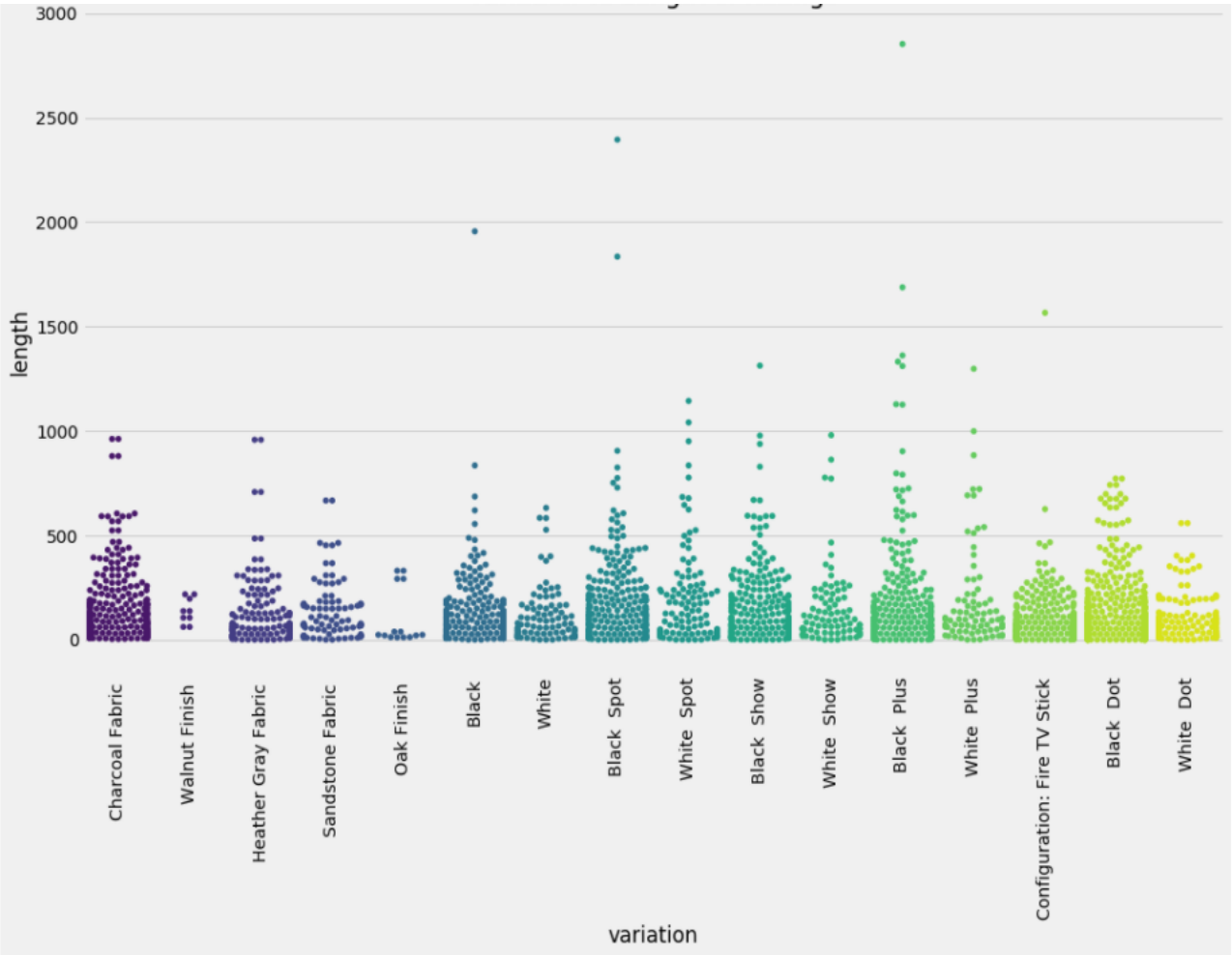

Figure 3.6: Variation VS Length of Ratings

- By looking at the graph, we can spot that the longest reviews were written by customers who own the Black Plus model, with the longest reviews with almost 3 thousand characters (including white space).
- Dissatisfied customers usually tend to leave long reviews explaining their reasons for not liking the product. Based on that, I was hoping to see long reviews for the White model since the ratings were low in this particular model accordingly with the previous graph. However, that was not the case. The reviews for the White model were relatively short when compared with other models.

Feedback vs Ratings

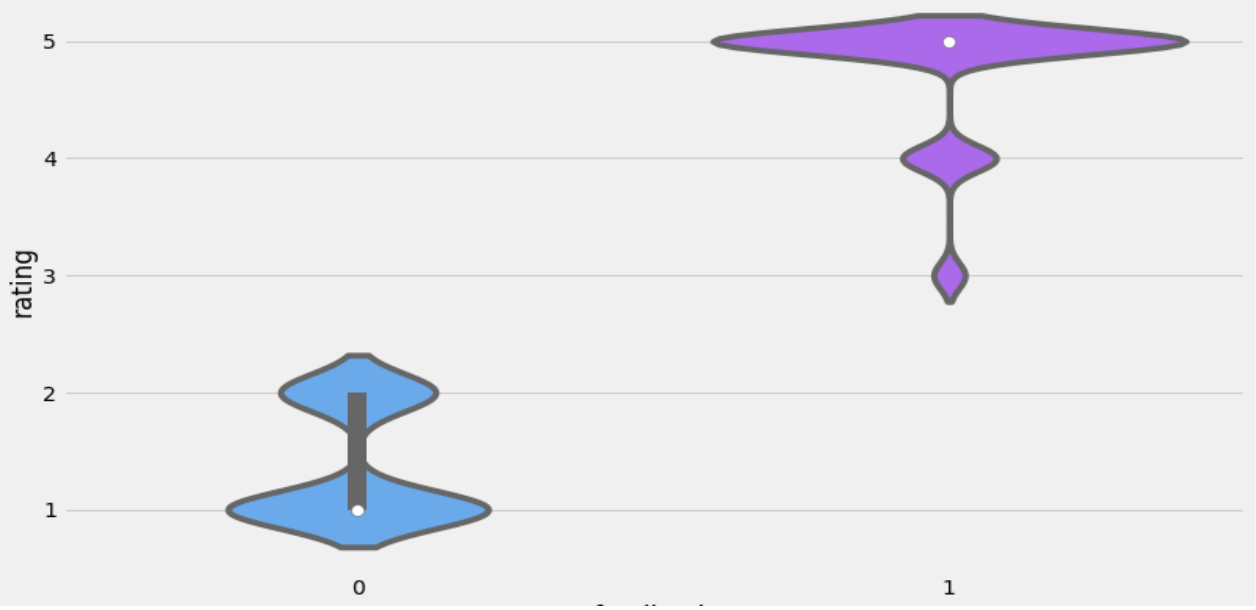

Figure 3.7: Feedback VS Ratings

We can conclude from the graph that the Alexa reviews that have 0 feedback have lower ratings ranging from 1 to 2 stars, but mostly 1 star only. Whereas the Alexa reviews having a feedback equal 1 have higher ratings between 3 to 5 stars. According to the graph, the majority of the feedback reviews has a 5-star rating.

Length vs Ratings

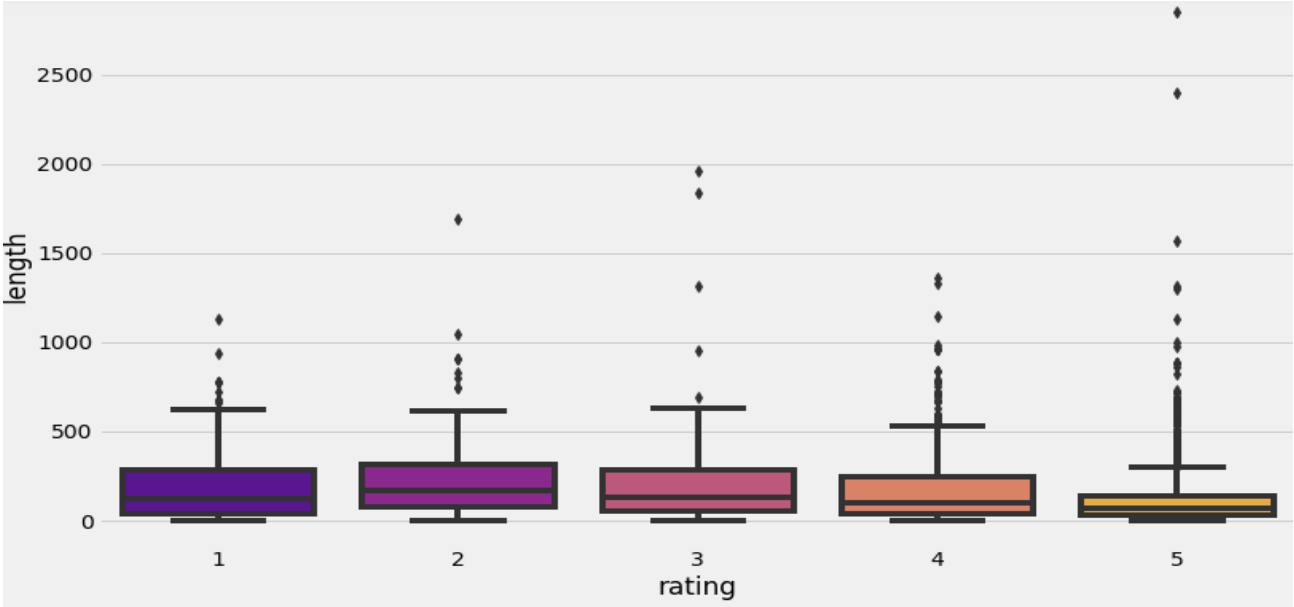

Figure 3.8: Length VS Ratings

The Bivariate plot shows the relationship between length and rating. Here we are looking how long the customer reviews are based on their rating. It is worth noting that all the reviews have pretty similar lengths regardless of their rating. However, there's a clear difference between the length of low rating reviews and high rating reviews.According to the graph and as previously mentioned, low rating reviews tend to be longer than high rating reviews. Most customers that gave Alexa 5-stars wrote shorter reviews than customers that gave 1 or 2-stars. That might be due to the fact that unsatisfied customers feel the need to explain the reasons for not liking the product, While satisfied customers feel happy, hence do not feel the same urgency to write long reviews.

Tokenization

In order to use textual data for predictive modeling, the text must be parsed to remove certain words – this process is called tokenization. These words need to then be encoded as integers, for use as inputs in machine learning algorithms. This process is called feature extraction (or vectorization). Scikit-learn's CountVectorizer is used to convert a collection of text documents to a vector of term/token counts. It also enables the pre-processing of text data prior to generating the vector representation. With the help of this feature, we can observe the chart of the most frequently used words in the reviews.

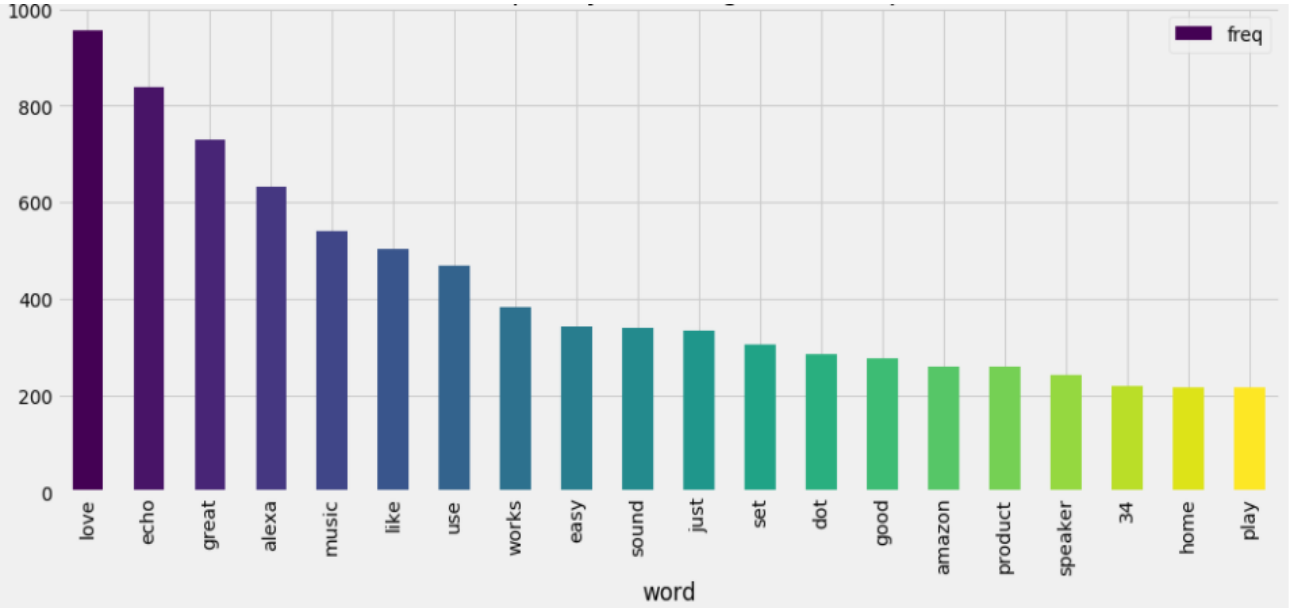

Figure 3.9: Most Frequently Occurring Words – Top 20

The bar plot represents the most frequent words among all of the reviews analyzed from the customers. By looking at the graph, we can have a good idea on how the customers think and feel regarding Amazon's Alexa. The words "love" and "great" are two of the most frequent words among all of the reviews which suggests that most customers had very positive feelings towards Alexa.

Vocabulary from Reviews

Figure 3.10: Vocabulary from Reviews

This Word Cloud visualization shows all the most frequently used and most relevant words analyzed from the customer reviews. The bigger the word, the higher is the frequency for that word been written by a customer. As seen in the previous result, “love”, “great”, “like”, are very frequent words written by Alexa customers. This reinforces the customer’s positive feedback towards Alexa.

Feedback vs Length

Figure 3.11: Feedback vs Length

The Bivariate graph above shows the relationship between feedback and length. We notice that there is a lot more positive feedback than negative feedback. This makes sense since the majority of customers had a positive rating towards Alexa. We also notice that the length of positive feedbacks is greater than the negatives one. This might be due to the total number of positive feedback reviews greatly surpassing the number of negative feedback reviews. Customers that are satisfied with Alexa wrote longer reviews than customers that did not like Alexa as much.

3.1.2 Natural Language Processing

After importing the necessary packages for NLP, we will move forward with our pre-processing.

Corpus

A corpus is a large and structured set of texts (nowadays usually electronically stored and processed). In corpus linguistics, they are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory. To create a corpus we need to clean the text, which entails the following steps [6] :

- Make text lowercase.

- Remove punctuation.
- Remove stopwords.
- Stemming.

Bag of Words

When we are interested in categorizing text, classifying it based on sentiment, or verifying whether it is a spam, we often do not want to look at the sequential pattern of words. Rather we would represent the text as a bag of words, as if it were an unordered set of words, while ignoring their original position in the text, keeping only their frequency.The bag-of-words is a representation of text that describes the occurrence of words within a document.t. It is a way of extracting features from the text for use in machine learning algorithms The count vectorizer converts a collection of text documents to a matrix of token counts. The max_features function creates a matrix of words when analyzing the corpus. In this case, we passed 2500 to max_features, that means creating a feature matrix out of the 2500 most frequent words across the text document. In summary, we are only analyzing the 2500 most frequent words among all of the Alexa reviews.

3.1.3 Model Training

This project comprises of three machine learning models, and we will read about them in detail.

- **Decision Tree** -In Machine Learning, Decision Trees is a predictive model that represents a mapping between object attributes and object values. A Decision Tree is a tree-like classifier that partitions every possible outcome of data recursively into classes. Decision Trees is similar to the flowchart, in which every non-leaf node indicates a test on a particular attribute, every branch represents an outcome of that test and every leaf node expresses a classification or decision. The node at the topmost label in the tree is called root node, which corresponds to the best predictor. By using Decision Trees, both numerical and categorical data can be processed. Based on maximum information gain, decision-makers can choose best alternative and traversal from root node to leaf nodes denoting unique class separation. The most popular Decision Trees methods are Iterative Dichotomiser 3 (ID3) , C4.5 , C5.0 and classification and regression tree (CART) which use entropy-based measures to grow the tree.
- **Random Forest** - Random forests are a very popular machine learning approach that addresses the shortcomings of decision trees using a clever idea. The goal is to improve prediction performance and reduce instability by averaging multiple decision trees (a forest of trees constructed with randomness). Random forest is another ensemble method based on decision trees. It split data into sub-samples, trains decision tree classifiers on each sub-sample and averages prediction of each classifier. Splitting dataset causes

higher bias but it is compensated by large decrease in variance. Random Forest is a supervised learning algorithm and it is flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and the fact that it can be used for both classification and regression tasks. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process. It does not suffer from the overfitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases. The algorithm can be used in both classification and regression problems. Random forests can also handle missing values. There are two ways to handle these: using median values to replace continuous variables, and computing the proximity-weighted average of missing values. You can get the relative feature importance, which helps in selecting the most contributing features for the classifier. Random forests is slow in generating predictions because it has multiple decision trees. Whenever it makes a prediction, all the trees in the forest have to make a prediction for the same given input and then perform voting on it. This whole process is time-consuming. The model is difficult to interpret compared to a decision tree, where you can easily make a decision by following the path in the tree.

- **Support Vector Machine** - Support Vector Machine (SVM) is a supervised machine learning algorithmic rule which might be used for each classification or regression challenges. However, it's principally utilized in classification issues. In this algorithmic rule, we plot each data item as a point in n-dimensional space where n is number of features one has with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyperplane that differentiates the two classes. Often researchers tend to plot every knowledge item as some extent in ndimensional area with the worth of every feature being the worth of a selected coordinate. Then, to perform classification by finding the hyper-plane that differentiate the 2 categories fine. It is a non-probabilistic binary linear classifier, however are often manipulated during a manner that it will perform non-linear and probabilistic classification also, creating it versatile algorithmic program. AN SVM model could be an illustration of the instances as points in area mapped, so they will be categorized and divided by a transparent gap. New instances are then mapped into the identical area and foreseen that within which class it would be supported which aspect of the gap they fall in. the most advantages of SVM is that the indisputable fact that it's effective in high dimensional areas. When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The clustering algorithm which provides an improvement to the

support vector machines is called support vector clustering. The main idea in SVM is to identify the optimal separating hyperplane which maximizes the margin of the training data.

3.1.4 Evaluation Metric

Summarization the performance of a classification algorithm is based on a technique which is known as confusion matrix. It is arguably the easiest way to regulate the performance of a classification model by comparing how many positive instances are correctly/incorrectly classified and how many negative instances are correctly/incorrectly classified. In a confusion matrix, as shown here, the rows represent the actual labels while the columns represent the predicted labels.

- **True Positives (TP)** : These are the occurrences where both the predictive and actual class is true (1), i.e., when the patient has complications (breast cancer in this case) and is also classified by the model to have complications.
- **True Negatives (TN)** : True negatives are the occurrences where both the predicted class and actual class is False (0), i.e., when a patient does not have complications and is also classified by the model as not having complications.
- **False Negative (FN)** : These are occurrences where the predicted class is False (0) but actual class is True (1), i.e., case of a patient being classified by the model as not having complications even though in reality, they do.
- **False Positive (FP)** : False positives are the occurrences where the predicted class is True (1) while the actual class is False (0), i.e., when a patient is classified by the model as having complications even though in reality, they do not.
- **Accuracy/Success Rate** : Evaluation of classification models is done by one of the metrics called accuracy. Accuracy is the fraction of prediction. It determines the number of correct predictions over the total number of predictions made by the model. The formula of accuracy is: Accuracy = (TP+TN)/(TP+TN +FP+FN).
- **Sensitivity** : Classifier's performance to spot positive results is related by Sensitivity. It is a measure of the number of patients who are classified as having complications among those who actually have the complications. Specificity is calculated as follows Precision = TP/(TP+FN).
- **Specificity** : Classifier's performance to spot negative results is related by Specificity. It is a measure of the number of patients who are classified as not having complications among those who actually did not have the complications. Specificity is calculated as follows: TN/(TN+FP).

- **F1 score** :The F1 score, also called the F score or F measure, is a measure of a test's accuracy. The F1 score is defined as the weighted harmonic mean of the test's precision and recall.

As the dataset is imbalanced, taking accuracy as an evaluation metric will not be sufficient, so we take F1 score to be our metric. F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall AND there is an uneven class distribution (large number of Actual Negatives).

Chapter 4

Results and Conclusions

The following table gives the comparison between the different algorithms performed. We observe that Random Forest and Decision Tree has outperformed SVM.

Algorithm	**Accuracy**	**F1 Score**
Decision Tree	94.12	93.00
Random Forest	94.00	93.00
SVM	93.10	90.00

Table 4.1: Results Achieved

In conclusion, the results show the vast majority of reviews written by Amazon's Alexa customers were highly positive. Overall, 87% of the customers gave Alexa at least 4- star ratings. When it comes to positive and negative feedback scores, 91.8customers have given a positive feedback, and only 8% of customers have given a negative feedback to Alexa. This shows that Alexa customers are very pleased with their purchase. Only a small percentage had some kind of complaint towards Alexa or did not like the product. The different model variations of Alexa that have darker color tones (e.g. Black Dot model) were ranked more popular against models with lighter color tones (e.g. White model). For the White model, besides being one of the least popular models, it is also among the models with the most negative ratings. In regard to the relationship between length of reviews and model variation, I was hoping to see longer reviews for the models who had very low rating scores. Thus, it was expected to see longer reviews for the White model since the ratings were low in this particular Alexa variation. However, that was not the case. The reviews for the White model were relatively short when compared with other models. The relationship between the length of reviews and their ratings, most customers that gave Alexa 5-star ratings wrote shorter reviews than customers that gave 1 or 2-stars. This result is expected since dissatisfied customers write longer reviews explaining their reasons for not liking the product. The NLP model was very effective in predicting the difference between positive and negative reviews. With 93.00% overall accuracy, I conclude that the random forest classifier algorithm is very effective and works really well for linguistic analysis.

Bibliography

[1] Akshit Arora Arush Nagpal. Amazon reviews sentiment analysis. *Thapar Institute of Engineering and Technology University*, 2018.

[2] SEPIDEH PAKNEJAD. Sentiment classification on amazon reviews using machine learning approaches. *KTH ROYAL INSTITUTE OF TECHNOLOGY SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE*, 2018.

[3] Hui Zhang. Sentiment analysis on amazon reviews. *Institute of Engineering and Technology University*, May 2019.

[4] Pang and Lee. Opinion mining and sentiment analysis. *Computer Science Department, Cornell University, Ithaca*, 2008.

[5] Liu J. Summary of english text mining preprocessing process [www document]. 2017.

[6] Muriel Kosaka. Cleaning preprocessing text data for sentiment analysis. *towards data science*, 2020.