5.1. A description of the dataset used.

This is a list of over 34,000 consumer reviews for Amazon products like the Kindle, Fire TV Stick, and more provided by Datafiniti's Product Database.

There are 24 features/variables in the data set, with 28,332 observations for each feature, except for the columns with missing values. There are 7 numerical variables : 'dateAdded', 'dateUpdated', 'manufacturerNumber', 'reviews.date', 'reviews.dateSeen', 'reviews.numHelpful', 'reviews.rating' and 14 categorical variables are: 'id', 'name', 'asins', 'brand', 'categories', 'primaryCategories', 'imageURLs', 'keys', 'manufacturer', 'reviews.didPurchase', 'reviews.doRecommend', 'reviews.id', 'reviews.numHelpful', 'reviews.sourceURLs', 'reviews.text', 'reviews.title', 'reviews.username', 'sourceURLs'.

Following is a short description of each feature:

- 'id' (identity of the review).
- 'dateAdded' (the date when the review was first added, for example, 2015-10-30).
- 'dateUpdated' (date of the updated review, for example, 2019-04-25 – approximately four years after the review was first posted).
- 'name' (name of the product that was reviewed by a customer, for example: AmazonBasics AAA Performance Alkaline Batteries, Fire Kids Edition Tablet, Fire HD 8 Tablet with Alexa, All-New Fire HD 8 Tablet with Alexa, etc).
- 'asins' (Amazon Standard Identification Number. It’s a unique identifier of 10 letters and/or numbers for a product that’s assigned by Amazon.com. It’s primarily used for product-identification within their product catalog of billions of items. For books, the ASIN is the same as the ISBN number. For all other products, a new ASIN is created when an item is uploaded to Amazon’s catalog. ASINs are only guaranteed to be unique within a marketplace. That means different national Amazon sites may use different ASINs for the same product.)
- 'brand' (for example, Amazon and Amazon Basics (often stylized as AmazonBasics) is Amazon's original private label brand that was launched in 2009. The brand offers many products, including home goods, electronics, travel, office supplies, and much more.)
- 'categories' (for example AA, AAA, Health, Electronics, Health & Household, etc.)
- 'primaryCategories' (for example, Electronics, Toys & Games, Health & Beauty).
- 'imageURLs' (shows the URL of the product’s image).
- 'keys' (provides delivery of packages and groceries into your garage, behind your gate and inside multi-unit buildings – discontinued).
- 'manufacturer' (for example Amazon and AmazonBasics).
- 'manufacturerNumber' (unique identifier to find specific products).
- 'reviews.date' (for example, 2016-12-21).
- 'reviews.dateSeen' (when the review was seen by the designated person, for example 2017-09-09).
- 'reviews.didPurchase' (NaN).
- 'reviews.doRecommend' (NaN, True).
- 'reviews.id' (NaN).
- 'reviews.numHelpful' (NaN, 0).
- 'reviews.rating' (for example: 2, 4, 5, where 1 is the lowest mark and 5 is the highest mark).

- 'reviews.sourceURLs'
- 'reviews.text' (the text of the product review, as entered by the consumer, for example: “Works as expected. Good value just as good as duracell”; “Good tablet. A lot of tablet for the money.”)
- 'reviews.title' (for example, Best Kindle Ever, Love My Echo, Cool little gadget, Seems like good budget batteries, Happy kids, Great small tablet, Great fun for the family).
- 'reviews.username' (for example, ByNannymac47, ByForced48, brettdavis4, ByIrish, Mynameisluist, KCRoyas).
- 'sourceURLs'

5.2. Details of the preprocessing steps.

- Made use of the lower() to convert upper case letters to lower cases.
- Made use of strip() to eliminate leading and trailing whitespaces.
- Lemmatization was done to reduce a word to its root form using lemma.
- Removed all missing values from the column.
- Removed stopwords, by utilizing the .is_stop attribute in spaCy.
- Removed punctuation, by utilizing .is_punct attribute in spaCy.

5.3. Evaluation of results.

Here are the results of the sentiment analysis for the Amazon products purchased or viewed by consumers:

Positive percentage: 82.98%
Negative percentage: 7.26%
Neutral percentage: 9.76%

It can be concluded that generally, consumers have a good feeling about Amazon products: 82.98% of sentiments are positive, 9.76 are neutral, and only 7.26% are negative.

Testing the model prediction for two values gives the following analysis of sentiment for two data points:
data[2] = well duracell price happy
polarity = 0.8, therefore the sentiment predicted is negative
data[10] = find amazon basics batteries equal superior name brand ones cant believe didnt start buying sooner packages large price great
polarity = 0.43, therefore the sentiment is negative.

5.4. Insights into the model's strengths and limitations.

The model gives a numerical measurement of the polarity and sentiments related to the consumers who posted reviews for the products they bought or viewed on Amazon. It is a good calculator of the polarity of the sentences in product reviews and in calculating consumer sentiment based on positive and negative values of polarity scores, where a polarity greater than 0 shows a positive sentiment, less than 0 – a negative sentiment and a polarity score neither greater than 0 nor less than 0 shows a neutral sentiment.

From the testing of the model, it can be deduced that the sentiment analysis is not quite accurate. A reading of data[1] shows positive–like words, such as “well” and “happy”, yet the model gave it 0.8 negative qualificatives. A reading of data[10] reveals that the consumer regrets not having bought the product sooner because Amazon batteries are equal to those of a superior brand name. So, the model is not infallible, it has its limitations and cannot predict with high precision the consumer sentiment inferred from product review.