Amazon Reviews Sentiment Analysis

Nourhan Ghanima

2023-12-02

Decoding Consumer Sentiments: Analyzing Amazon Reviews for Product Perception and User Satisfaction

Sentiment analysis is a technique that focuses on identifying the emotional tone or sentiment conveyed in a given piece of text.

It's particularly useful from a business perspective because it can provide important insight into customer opinions and preferences, which can then be used to make informed decisions about products or servies, after identifying potential areas of improvement. It also allows the monitoring of brand reputation, which is key to managing and enhancing brand perception.

The data for this project is taken from Kaggle, and includes user reviews left on a variety of products on Amazon.com.

After some basic data cleaning done on Excel and R with the tidyverse package, the descriptive statistics for the dataset are as follows:

```
summary(amazon_reviews)
```

```
##       id                name               asins               brand
##  Length:67992       Length:67992       Length:67992       Length:67992
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   categories         manufacturer       reviews.numHelpful reviews.rating
##  Length:67992       Length:67992       Min.   :  0.000     Min.   :1.000
##  Class :character   Class :character   1st Qu.:  0.000     1st Qu.:4.000
##  Mode  :character   Mode  :character   Median :  0.000     Median :5.000
##                                        Mean   :  0.572     Mean   :4.556
##                                        3rd Qu.:  0.000     3rd Qu.:5.000
##                                        Max.   :814.000     Max.   :5.000
##                                        NA's   :12746       NA's   :33
##  reviews.text       reviews.title      reviews.username
##  Length:67992       Length:67992       Length:67992
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

The data includes 67992 reviews, for a total of 7 brands, all owned by Amazon.

Which brand has the highest number of reviews?

```
amazon_reviews |>
  group_by(brand) |>
  summarise(Count = n())
```

```
## # A tibble: 7 x 2
##   brand                    Count
##   <chr>                    <int>
## 1 Amazon                   49854
## 2 Amazon Coco T                1
## 3 Amazon Digital Services Inc.  10
## 4 Amazon Echo                636
## 5 Amazon Fire                256
## 6 Amazon Fire Tv             5056
## 7 AmazonBasics              12179
```

Which brand has the highest average rating?

From this plot showing average review ratings per brand, it's obvious that Amazon Digital Services has the lowest average rating.

```
average_rating |>
  ggplot(aes(x = brand, y=average_rating)) +
  geom_col(fill= "blue") +
  xlab("Brand") +
  ylab ("Average Rating") +
  theme_bw()
```
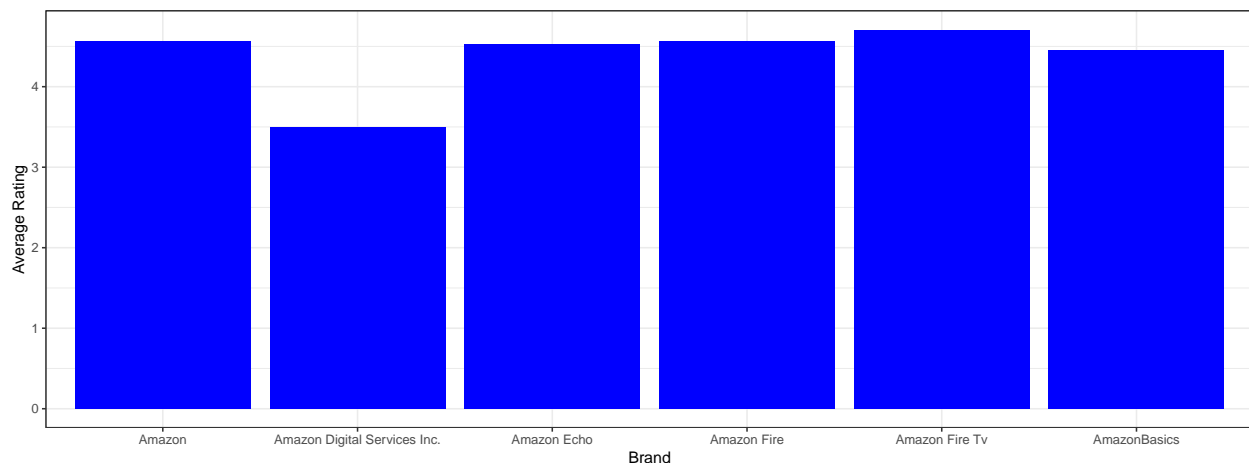


```
amazon_reviews |>
  filter(brand != "Amazon Coco T") |>
  group_by(brand) |>
  summarise(average_rating = mean((reviews.rating), na.rm = TRUE))
```

```
## # A tibble: 6 x 2
##   brand                    average_rating
##   <chr>                            <dbl>
## 1 Amazon                            4.57
## 2 Amazon Digital Services Inc.      3.5
## 3 Amazon Echo                       4.53
## 4 Amazon Fire                       4.56
```

```
## 5 Amazon Fire Tv                     4.71
## 6 AmazonBasics                       4.45
```

The average star rating for Amazon Fire TV is the highest.

Setting up sentiment analysis

The first step is to create a corpus.

```
corpus <- iconv(amazon_reviews$reviews.text)
corpus <- Corpus(VectorSource(corpus))
```

I then transform the corpus to lowercase to avoid duplicates.

```
corpus <- tm_map(corpus, tolower)
```

I then clean up the corpus by removing punctuation, numbers, and stopwords.

```
corpus <- tm_map(corpus, removeNumbers)
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, removeWords, stopwords("english"))
amazon_reviews_final <- corpus
```

Creating the term frequency matrix, which is a matrix of the frequency of each word in the reviews:

```
tfm <- TermDocumentMatrix(amazon_reviews_final)
tfm <- as.matrix(tfm)
```

The first ten rows and five columns of the frequency matrix are displayed below.

```
tfm[1:10, 1:5]
```

```
##               Docs
## Terms          1 2 3 4 5
##   ability      1 0 0 0 0
##   children     1 0 0 0 0
##   content      1 0 0 1 1
##   control      1 0 0 0 0
##   disappointed 1 0 0 0 0
##   ease         1 0 0 0 0
##   far          1 0 0 0 0
##   like         1 0 0 1 0
##   love         1 0 0 2 0
##   monitor      1 0 0 0 0
```

Plot of word frequency

```
w <- rowSums(tfm)
w <- subset(w, w >= 25)
barplot(w, las = 2, col = "blue")
```
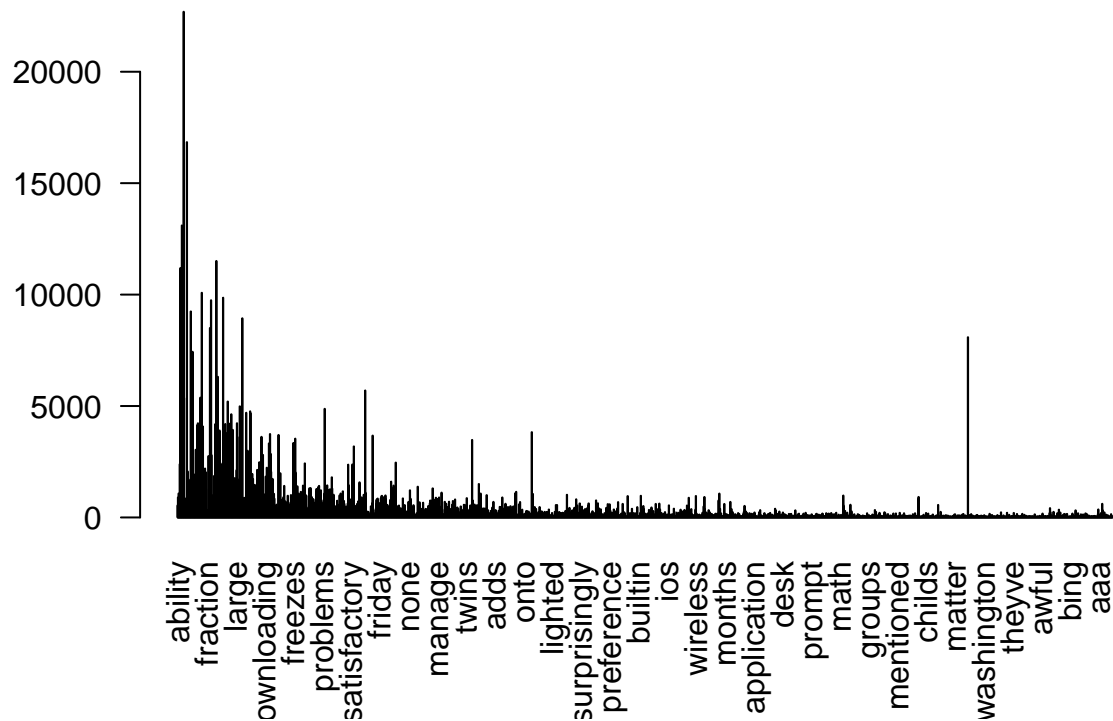
I limit the words to words that have a frequency of 25 or higher.

The word cloud

Assigning sentiment scores for Amazon brand

```
amazon_reviews_amazonbrand <- amazon_reviews |>
  filter(brand == "Amazon")
sentiment_data <- iconv(amazon_reviews_amazonbrand$reviews.text)
sentiment <- get_nrc_sentiment(sentiment_data)
sentiment[1:10, ]
```

```
##   anger anticipation disgust fear joy sadness surprise trust negative positive
## 1     1            0       1    0   2       1        0     1        1        4
## 2     0            1       0    0   1       0        1     1        0        2
## 3     0            0       0    0   0       0        0     0        0        2
```

4

```
## 4        0               1        0     2   2       0          0      3        1        5
## 5        0               0        0     0   2       0          0      1        0        3
## 6        0               2        0     1   2       0          1      2        0        2
## 7        0               1        1     1   1       1          1      1        1        3
## 8        0               1        0     0   1       0          1      0        0        1
## 9        0               3        1     0   2       0          1      2        0        5
## 10       0               0        0     0   1       0          0      0        0        3
```

calculating the review wise score

```
sentiment$score <- sentiment$positive - sentiment$negative
sentiment[1:30, ]
```

```
##      anger anticipation disgust fear joy sadness surprise trust negative positive
## 1        1               0        1     0   2       1          0      1        1        4
## 2        0               1        0     0   1       0          1      1        0        2
## 3        0               0        0     0   0       0          0      0        0        2
## 4        0               1        0     2   2       0          0      3        1        5
## 5        0               0        0     0   2       0          0      1        0        3
## 6        0               2        0     1   2       0          1      2        0        2
## 7        0               1        1     1   1       1          1      1        1        3
## 8        0               1        0     0   1       0          1      0        0        1
## 9        0               3        1     0   2       0          1      2        0        5
## 10       0               0        0     0   1       0          0      0        0        3
## 11       0               0        0     0   0       0          1      0        1        0
## 12       0               0        0     0   1       0          0      1        0        2
## 13       0               2        0     1   1       0          1      2        0        1
## 14       0               0        0     0   0       0          0      0        0        1
## 15       0               1        0     0   2       0          1      1        0        2
## 16       0               1        0     0   1       0          0      2        1        1
## 17       0               2        0     0   2       0          1      3        0        4
## 18       1               1        0     0   0       0          0      0        1        0
## 19       0               1        0     0   1       0          0      0        0        1
## 20       0               0        0     1   2       1          0      1        0        2
## 21       0               3        0     0   2       0          0      2        0        3
## 22       0               1        0     0   1       0          0      1        1        4
## 23       0               1        0     0   1       0          0      1        0        1
## 24       0               2        0     0   3       0          1      2        0        4
## 25       0               1        0     0   1       1          0      1        0        4
## 26       0               0        0     0   0       0          0      0        0        0
## 27       0               1        0     0   1       0          0      1        0        2
## 28       0               0        0     0   0       0          0      1        0        0
## 29       0               1        0     0   1       0          0      2        0        2
## 30       0               0        0     1   0       0          0      1        0        1
##      score
## 1        3
## 2        2
## 3        2
## 4        4
## 5        3
## 6        2
## 7        2
## 8        1
## 9        5
```

```
## 10     3
## 11    -1
## 12     2
## 13     1
## 14     1
## 15     2
## 16     0
## 17     4
## 18    -1
## 19     1
## 20     2
## 21     3
## 22     3
## 23     1
## 24     4
## 25     4
## 26     0
## 27     2
## 28     0
## 29     2
## 30     1
```

Most of the sentiment scores for the Amazon brand are positive.