Sentiment Analysis of Amazon Reviews

Using Product Reviews

Submission Details

16 December 2022

Course Title: ISMG 6470: Text Data Analysis E01

Prepared and submitted by Lakshmi Brahmani Suravajjala, 110377044

# Executive Summary

**Background**

Analyzing customer reviews help businesses in identifying their flaws and helps them to make necessary changes in their future products. Amazon contains a big collection of reviews and feedbacks that these businesses can grab insights from for their analysis. This includes comments on services, and product reviews, etc. Thus, there is a need for a sentiment analysis system that can extract sentiments about a particular product or a service, which will boost their brand and improve customer satisfaction.

**Methodology**

In this project, The Amazon Data collected is in the text format, and the corpus that has been built contains 40 text files. The Text files include the customer reviews collected from a product on amazon called the Dyson supersonic dryer and performed **Sentiment Analysis.**

**Key Findings**

* Using the insights from the word clouds after performing sentiment analysis, we can identify the most impactful positive and negative words. Positive words being “great”, “style”,”good”, etc. And the negative words being “counterfit”, “rather”, “die”, etc.
* “Trust” and “anticipation” were the major emotions shown in the reviews according to the analysis.
* There are more positive words compared to the negative words in the corpus.

**Key Recommendations**

* The analysis shows that the customers trust the brand and anticipate better products.
* Dyson should work on removing counterfeits which was a major concern of the customers writing negative reviews.

# Introduction

## **Background**

Customer reviews are particularly important for product-based companies. Ratings and reviews create a great avenue for these businesses to get current content on their site and bring attention to their product. Also, Analysing the sentiment of these reviews, they can work and improve upon their future products for better customer satisfaction.

In this project Amazon data has been collected in the form of 40 text files which include the reviews from customers on a hair styling product called dyson supersonic hair dryer. Using the data collected, Sentiment analysis has been performed to analyse the sentiment of each customer review and analyze key words used in the reviews.

## **Objectives**

The objectives of this sentiment analysis are to:

* Identifying and obtaining the sentiment of each review.
* Getting the most positive impact words and negative words.
* Identifying and getting various emotions that were displayed in the reviews.
* Analysing the sentiment distribution across the reviews

Methodology

## **Research Questions**

1. What are the main concerns of the customers who bought The Dyson supersonic hairdryer?
2. How many customers enjoyed this product, and how many disapproved?
3. What are the major emotions people showed in the reviews?
4. How are the reviews spread across the sentiments?

## **Research Design**

This project will use sentiment analyses method to answer the research questions. The following table summarises the methods used to answer each question:

**Table 1 Methods used to answer research questions**

|  |  |
| --- | --- |
| **Research Question** | **Method Used to Answer Question** |
| What are the main concerns of the customers who bought The Dyson supersonic hairdryer? | Using the word clouds of positive and negative words after performing the sentiment analysis. |
| How many customers enjoyed this product, and how many disapproved? | By analysing the positive customer reviews and negative customer reviews. |
| What are the major emotions people showed in the reviews? | By using NRC sentimental analysis, we get a distribution of Words and their corresponding emotions. |
| How are the reviews spread across the sentiments? | By plotting the density distribution of reviews across the polarities. |

## **Data Collection**

The data that is used for the sentiment analysis is collected from Amazon. The corpus that was built in the project consists of 40 text files and the text files include the customer reviews on a hair styling product called The Dyson supersonic hair dryer. The reviews were carefully chosen to include both high star rated and low star rated customer reviews. These reviews have been chosen for this project because they include multiple paragraphs discussing various aspects of the product.

The reviews were collected from this page:

[Amazon.com: Dyson Supersonic Hair Dryer, Iron/Fuchsia : Beauty & Personal Care](https://www.amazon.com/Dyson-Supersonic-Hair-Dryer-Fuchsia/dp/B01FIG1JIM/ref=sr_1_2_sspa?keywords=dyson%2Bsupersonic%2Bhair%2Bdryer&qid=1671252412&sprefix=dyson%2Bsuper%2Caps%2C202&sr=8-2-spons&ufe=app_do%3Aamzn1.fos.2b70bf2b-6730-4ccf-ab97-eb60747b8daf&spLa=ZW5jcnlwdGVkUXVhbGlmaWVyPUFQM05WSVhIVUVKODYmZW5jcnlwdGVkSWQ9QTA1NTAzNDdESkdMN0EyQVJNSjkmZW5jcnlwdGVkQWRJZD1BMDg3MjM4MTNUMDdFMFhIQjZGUFImd2lkZ2V0TmFtZT1zcF9hdGYmYWN0aW9uPWNsaWNrUmVkaXJlY3QmZG9Ob3RMb2dDbGljaz10cnVl&th=1)

## **Data Cleaning Procedure**

The number of unique words before pre-processing were 1678

1. There were emoticons in the reviews collected, these were replaced by their meanings. Some of the emoticons present in the dataset were not available in the emoticon package so, these meanings have been appended.
2. Since the data has reviews which were multi lined, line breaks have been removed.
3. The data has been converted to **lowercase** because it helps to maintain the consistency flow during sentiment analysis.
4. After carefully analysing the words in the corpus, there were some unique and irregular words which had to be replaced for e.g., “110v” was replaced with “110 volts”
5. The **punctuations** have been removed as they are not informative and they just add up noise that makes the corpus complicated to analyze and by removing them, we can analyze the data efficiently.
6. While dealing with any English sentences there are many words that do not bring any new information and they just create redundancy which just makes everything complicated so, The **stop words** have been removed, removing these words will help clean the data and we can focus on important words which carry information about the sentiment.
7. Removed all the **Whitespaces**, all the extra whitespaces between two words will be stripped off making the data easy to work with.
8. **Stemming** has been performed to break the words into their root forms.

The number of unique words after pre-processing are 991

## **Data Analysis**

Sentiment analysis was employed in this project since it has various business benefits and is an efficient and reliable tool for assessing any product’s performance and we can identify the brand experience insights and also the customer reviews and improve the customer service accordingly. It also helps manage company’s brand reputation by enabling them to make timely decisions on how to respond to negative brand mentions and thus avert risks.

Three methods were employed for sentiment analysis in the project:

**Method 1**: sentiments were analysed using “**qdab”** library. By using this library, the polarities of each of the pre-processed reviews were calculated. Document term matrix was created using tf-idf algorithm for both positive and negative reviews. Using this Document term matrix Word clouds were plotted to show the most impactful positive and negative words in the reviews.

**Method 2**: Sentiments were analysed using “**Bing”** lexicon analysis method.before applying this method, The document term matrix was created using the pre-processed corpus. The document term matrix was converted into dataframe using “**Tidy”** library.and this dataframe is joined with”**Bing”** library to get the sentiment of their respective words.The total number of positive words and negative words have been identified for each review. And the polarity was calculated by getting the difference between the positive and negitive words. After that, the distribution of polarities across the reviews been plotted, The overall sentiments of the customer reviews have been identified.

**Method 3**: Sentiments were analysed using “**NRC”** lexicon method, this method outputs the number of words describing various emotions like “anger”, “anticipation”, “disgust”, “fear”, “joy”, “surprise”, “sadness”, and “trust”, along with the positive and negative counts for each review in the corpus. The count for each emotion is aggregated across the corpus to plot the distribution of emotions across the reviews.

## **Limitations**

1. Here the project focusses only on reviews in English but amazon has a large amount of international audience. This approach should be extended to classify sentiment with a language specific positive/negative keyword list in other languages as well.
2. The data collected is handpicked for this project, It does not represent the whole population sentiment who purchased this product. To address this a more holistic sample has to be analysed.
3. The sentiments derived from the lexicon approach is not robust to slang terms and technical terms that were used in the reviews. This can be addressed by using state of art machine learning language models.

# Results

**Question 1:** What are the main concerns of the customers who bought The Dyson supersonic hairdryer?

This question can be addressed by analysing the word clouds of most impacted positive and negative words.

Figure 1: The word cloud generated after pre-processing and before sentiment analysis.



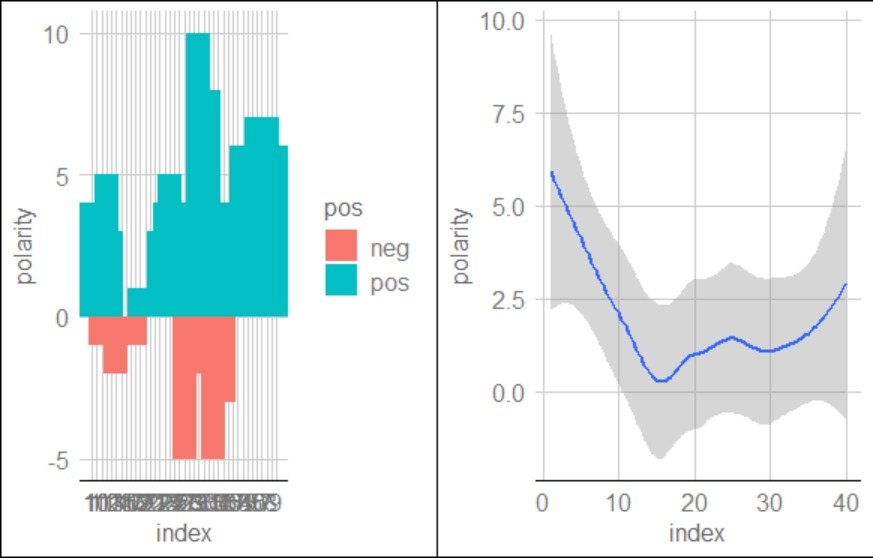
Figure 2: The word cloud which shows most impactful words after sentiment analysis



**Question 2:** How many customers enjoyed this product, and how many disapproved?

This can be analysed by plotting the reviews and and their polarities.

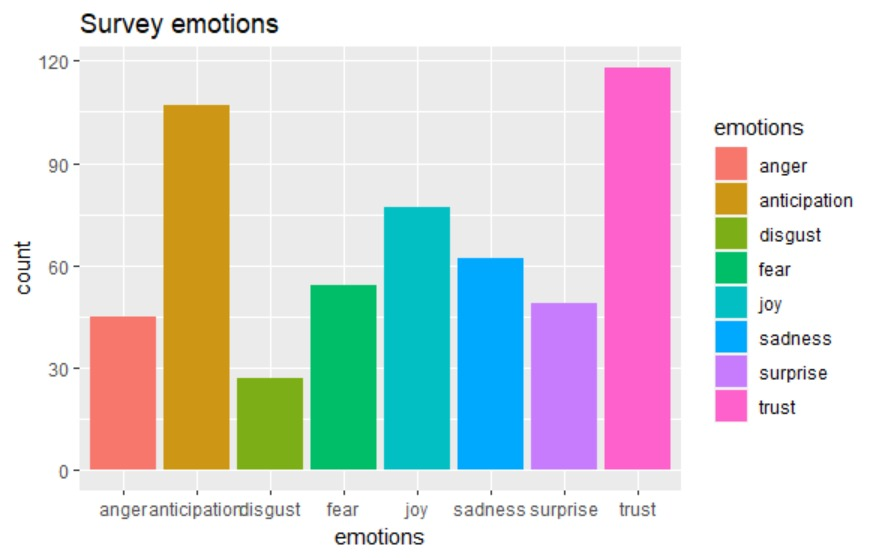
Figure 3:



**Question 3:** What are the major emotions people showed in the reviews?

This can be analysed by observing the distribution of words across various emotions present in the corpus.

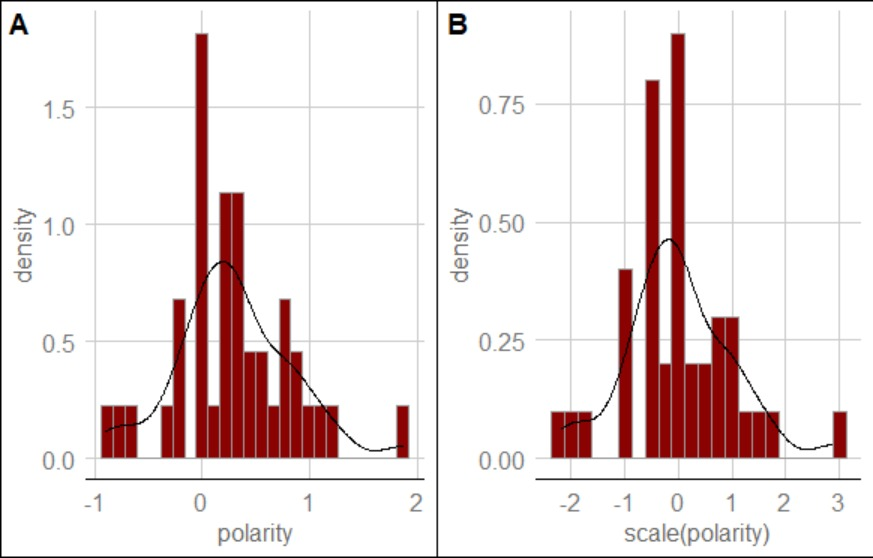
Figure 4:



**Question 4:** How are the reviews spread across the sentiments?

This can be analysed by studying the distribution of reviews across the polarities

Figure 5:



Discussion

* To find the most impactful positive and negative words Word clouds were plotted after the sentiment analysis of the corpus. By observing the word clouds in figures 1 and 2 we can find the distinctive difference how the most frequent words have been changed before and after the sentiment analysis. From figure 1, the most frequent words are “dryer’, “Dyson”, “product”, “hair”, which were very predictable because the product is a hair styling tool. From figure 2 we can identify the most impactful positive and negative words. Positive words being “great”, “style”, “good”. And the negative words being “counterfeit”, “rather”, “die”
* By observing Figure 3 Which shows many of the reviews in the corpus are classified as positive from this observation we can infer that there are more customers who purchased this product are satisfied than the customers who are not satisfied with the purchase
* “Trust” and “anticipation” were the major emotions shown in the reviews according to the Figure 4 From this we can infer that customers have huge trust in the brand Dyson and anticipate their new products
* Figure 5 shows that the polarity density is more skewed towards positive before scaling and there are few reviews with high positive and negative polarity but most of the reviews are centred around 0 which implied, they are tending towards neutral rather than positive or negative.

# Recommendations

Based on the results from the sentiment analysis the following recommendations are made:

* Dyson should work on removing counterfeits which was a major concern of the customers writing negative reviews.
* Dyson should work on improving longevity of the products because the second most impactful negative word is “Repair”. They also should work on better customer service for their damaged products.
* The third most impactful negative word is “die” which might be caused due to the batteries and power issues. Dyson should make their products compatible across various voltage ranges.

References

Sanil Mhatre, 13th May 2020, Text Mining and Sentiment Analysis with R, <https://www.red-gate.com/simple-talk/databases/sql-server/bi-sql-server/text-mining-and-sentiment-analysis-with-r/>