Business Insight Report

Individual Assignment

Prepared by Pedro Nascimento

Prepared for: Professor Thomas Kurnicki

Text Analytics - DAT-5317 - FMSBA3

Due to: February 14th, 2020

Table of Contents

1. MOTIVATION

This document is a business insight report, based on the writer's passion about tourism and travel, comparing Airbnb and Booking.com reviews' from Trustpilot website, and analyzing these companies' customer service and customer experience, aiming to generate business insights to Airbnb, as it is the writer preferred company.

2. ANALYSIS

2.1. PROCESS

The rationale behind the analysis was to compare both companies using tokenization, sentiment analysis and a correlogram. On Trustpilot website, both companies are not doing well and have ratings below 2 stars (See Figure 1). The data is gathered by scraping the website's html, using the "rvest" package. Then, the first step is a preliminary analysis using the "sentiment" function, from the "sentimentr" package. This function allows analyzing the sentiment not only token by token, but also by sentence. R analyzes the whole text on the review's object, breaking into sentences and assigning positive and negative values for them.

This strategy allowed finding the overall sentiment both companies, showing that both Airbnb and Booking.com customer service have a negative perception by their customers (-103.6474 for Airbnb and --111.8753 for Booking.com, on the date this report was written). Applying the same process on "refund" token, it showed that "refund" is a recurrent problem (See Figure 2). On the "bing" library, this token's sentiment is flagged as "positive", but in this case, the writer changed it to "negative", to reflect the bad experiences most of the customers are having – also, there is a token "refunded", with a positive sentiment, that remained with that sentiment to reflect the customers who had a better experience.

After this initial changes, the analysis was based on two libraries: "afinn" and "bing", and is summarized in a table (See Figure 3), with the minimum and maximum sentiment values for tokens in the datasets, and the overall sentiment value for both companies - using the "tidytext" package -, charts with the most used tokens opened by sentiment, and a correlogram to compare Airbnb and Booking.com.

2.2. FINDINGS AND SUGGESTIONS

The analysis supports the idea that both companies needs to improve their customer support area. The overall sentiment of the sample – 595 reviews for Airbnb and 592 reviews for Booking.com, on the date this report was written – is negative, accounting for "-344" and "-280" using "afinn" library for Airbnb and Booking.com, respectively (See Figure 3).

The problems between the companies vary, which can be inferred by the difference among the tokens, especially looking at the negative sentiment. The company that not only attacks its problems – considering its weakness, but also work on the competitor's ones – considering opportunities, will have a competitive advantage in the short-term and improve its financial performances – especially considering that switching costs are virtually inexistent for customers, thus the experience is an advantage in favor of the company.

As the goal is to improve Airbnb's experience, the focus will be on Airbnb and Booking.com weaknesses. Reviewing the refund policy is essential. "Refund" was used 229 times in Airbnb case, almost 30% - in 154 of the total 595 reviews, numbers founded by filtering "refund" using the "grepl" function, in the original Airbnb reviews' table. That's the issue that should be attacked immediately, as both companies are having the same issue, and solving it would be a way of getting a comparative advantage.

Second, some Airbnb customers claimed about dirty places (See Figure 4) and is interesting to find out that customers associate housing problems, as "dirty" (32 tokens), with the brand, and not the host only. Therefore, it's crucial to provide training and onboarding to the hosts, showing the importance of good reviews, and as rewards, reduce the service fee of Airbnb, decreasing the overall price of the stay, making that residence more accessible.

Last, "misleading" and "disappointed" are within the top 15 most repeated tokens in the Booking.com dataset. As this company is most used to book hotels, these are probably customers complaining about wrong information and the hotels not fulfilling the expectations. A business insight here is to have onboarding with the "superhosts" (See Figure 7), providing useful information, local tips about the region and the importance of setting the right expectation on the customer, to improve customer experience, especially with the information on the "Experiences" section from the website. Partnerships with other websites, as TripAdvisor, are also recommended.

2.3. CONCLUSION

Airbnb's revenue, according to Craft.Co, was $3.6 billion in 2018 [4], with a market valuation of $35 billion on March 2019. By improving the customer experience, the company can not only improve its financial performances but also the market share and recurrence of customers.

There four tokens that based this report are "refund", "dirty", "misleading" and "disappointed". The first token is the most recurring problem of both companies, and not getting a refund is a negative experience, especially considering that those customers are having issues probably during their vacation time, thus either review the refund policy or increase the free cancellation period, as today is around 14 days, are good ideas[7].

"Dirty" arises the question of how customers associate bad experiences not with hosts, but with companies' brands, therefore training with hosts, and giving financial benefits to the ones who have high reviews. Even though financial benefits to superhosts is already being done[6], as stated on Airbnb's website (See Figure 7), having periodic onboarding and trainings will increase hosts' consciousness over the importance of having good reviews and impress the customers, and potentially increase customer loyalty and brand value, and in the end, customer associate the hosts with Airbnb, which is something that doesn't happen with hotels and Booking.com.

Last, "misleading" and "disappointed" suggest that customers look for information about where they are and there is an opportunity to increase their experience, as it reflects in good reviews for the host – which, as explained on the "dirty" case, the experience are reflected not only on the host, but also on the company.

3. APPENDIX – R SCRIPT

library(rvest)

library(magrittr)

```r
library(tidyverse)

library(tidytext)

library(dplyr)

library(scales)

library(sentimentr)

# This R Script will compare the information of the first 30 pages from Airbnb and Booking.com, from Trustpilot reviews' website

# First, we get the two companies' url

list_airbnb_pages <- str_c('https://www.trustpilot.com/review/www.airbnb.com?page=', 1:30)    # Getting the first 20 pages of Airbnb

list_booking_pages <- str_c('https://www.trustpilot.com/review/www.booking.com?page=', 1:30)  # Getting the first 20 pages of Booking.com

# Creating a function to get the reviews' text

geting_table <- function(html, company_name){

    # Extract the Basic information from the HTML

  review <- read_html(html) %>% html_nodes('.review-content__text') %>% html_text()

    # Append into a tibble

  append_table <- tibble(review = review)

   # Tag the individual data with the company name

  append_table %>%

    mutate(company = company_name) %>%

    select(company, review)

}

scrape_write_table <- function(list_company_url, company_name){

  # Apply the extraction and bind the individual results back into one table,

  list_company_url %>%

    # Apply to all URLs

    map(geting_table, company_name) %>%

    # Combine the tibbles into one tibble

    bind_rows()
```

```
}

airbnb <- scrape_write_table(list_airbnb_pages, 'airbnb')

booking <- scrape_write_table(list_booking_pages, 'booking')

# Now that we have 2 dataframes, one for each company. Thus, it's important to do two processes: tokenization and perform an

# anti join with the stop_words dataset.

airbnb_tokens_nostop <- airbnb %>%

  unnest_tokens(word, review) %>%

  anti_join(stop_words) %>%

  count(company,word, sort = TRUE)

booking_tokens_nostop <- booking %>%

  unnest_tokens(word, review) %>%

  anti_join(stop_words) %>%

  count(company,word, sort = TRUE)

airbnb_reviews  <- airbnb[,2]

 booking_reviews <- booking[,2]

 airbnb_reviews  <- sentiment(airbnb_reviews$review)

 booking_reviews <- sentiment(booking_reviews$review)

 airbnb_reviews %>% summarise(company_name = 'airbnb', overall_sentiment = sum(sentiment))

 booking_reviews%>%  summarise(company_name = 'booking.com', overall_sentiment = sum(sentiment))

 airbnb_refund <- filter(airbnb, grepl('refund', review))

 booking_refund <- filter(booking, grepl('refund', review))

 airbnb_refund  <- sentiment(airbnb_refund$review)

 booking_refund <- sentiment(booking_refund$review)

 airbnb_refund %>% summarise(company_name = 'airbnb', overall_sentiment = sum(sentiment))

 booking_refund%>%  summarise(company_name = 'booking.com', overall_sentiment = sum(sentiment))

# A preliminary analysis showed that most of experiences with 'refund' are bad. Basically, customers

# are not getting their money refunded. Therefore, the sentiment is going to be changed to 'negative',
```

```
# and consider only the token 'refunded' as 'positive'.

filter(airbnb, grepl('refund', review)) %>% count()        #number of times “refund” was used in Airbnb

filter(booking, grepl('refund', review)) %>% count()       #number of times “refund” was used in Booking

# Finally, the main goals are:

# 1) Count the frequency of the remaining tokens.

# 2) Perform an inner join with 2 sentiment libraries: afinn and bing.

# 3) Perform some calculations and plot some graphics to generate business insights. The calculations are summing the

# overall sentiment over the two companies', according to Trustpilot's reviews, and a count of the tokens per feeling. Then,

# plotting the top 10/20 tokens, by positive and negative feeling and a correlogram to generate business insight. Last,

airbnb_afinn <- airbnb_tokens_nostop %>% inner_join(get_sentiments('afinn'))

airbnb_bing <- airbnb_tokens_nostop %>% inner_join(get_sentiments('bing'))

booking_afinn <- booking_tokens_nostop %>% inner_join(get_sentiments('afinn'))

booking_bing <- booking_tokens_nostop %>% inner_join(get_sentiments('bing'))

# 1) Performing the sum, using the afinn library

airbnb_afinn %>%

  summarise(min_value = min(value), max_value = max(value), overall_value = sum(airbnb_afinn$value))

booking_afinn %>%

  summarise(min_value = min(value), max_value = max(value), overall_value = sum(booking_afinn$value))

# 2) Working with the bing library

airbnb_bing %>% count(sentiment) %>% arrange(desc(n))

airbnb_bing$sentiment[airbnb_bing$word == 'refund'] <-  'negative'

booking_bing %>% count(sentiment) %>% arrange(desc(n))

booking_bing$sentiment[booking_bing$word == 'refund'] <-  'negative'

bind_rows(airbnb_bing %>% filter(sentiment == 'positive') %>% top_n(15, n), airbnb_bing %>%

        filter(sentiment == 'negative') %>% top_n(15, n))%>%mutate(word2 = fct_reorder(word, n)) %>%
```

```
      ggplot(aes(x = word2, y = n, fill = sentiment)) + geom_col(show.legend = F) + theme_classic() +
      facet_wrap(~ sentiment, scales = "free") + coord_flip() + theme(plot.title = element_text(hjust =
0.5)) +
      labs(x = 'Words', y = 'Frequency', title = 'Airbnb - Top 15 Tokens - Bing Library')+
      geom_text(aes(label = n), hjust = .6, size = 3)
bind_rows(booking_bing %>% filter(sentiment == 'positive') %>% top_n(15, n), booking_bing %>%
      filter(sentiment == 'negative') %>% top_n(15, n))%>%mutate(word2 = fct_reorder(word, n))
%>%
      ggplot(aes(x = word2, y = n, fill = sentiment)) + geom_col(show.legend = F) + theme_classic() +
      facet_wrap(~ sentiment, scales = "free") + coord_flip() + theme(plot.title = element_text(hjust =
0.5)) +
      labs(x = 'Words', y = 'Frequency', title = 'Booking - Top 15 Tokens - Bing Library')+
      geom_text(aes(label = n), hjust = .6, size = 3)
 # 3) Doing a correlogram between Airbnb x Booking.com
correlogram_table <- bind_rows(airbnb_bing, booking_bing) %>%
  mutate(proportion = n/sum(n))%>%
  select(-n) %>%
  spread(company, proportion) %>%
  gather(company, proportion, `booking`)
ggplot(correlogram_table, aes(x=proportion, y=`airbnb`,
            color = sentiment))+
  geom_abline(color="grey40", lty=2)+
  geom_jitter(alpha=.1, size=2.5, width=0.3, height=0.3)+
  geom_text(aes(label=word), check_overlap = TRUE, vjust=1.5) +
  scale_x_log10(labels = percent_format())+
  scale_y_log10(labels= percent_format())+
  #scale_color_gradient(limits = c(0,0.001), low = "darkslategray4", high = "gray75")+
  facet_wrap(~company, ncol=1)+
  #theme(legend.position = "none")+
  labs(y= "Airbnb", x=NULL) + theme_classic()
```

4. OUTPUTS

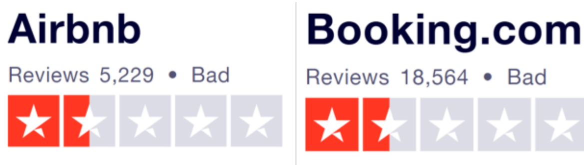

Figure 1: Airbnb and Booking.com Trustpilot's Ratings

Airbnb	Booking.com
-38.88519	-46.58112

Figure 2: "Refund" Token Sentiment of the Companies' Reviews

Airbnb	min_value	max_value	overall_value
	<dbl>	<dbl>	<dbl>
	-5	5	-344

Booking.com	min_value	max_value	overall_value
	<dbl>	<dbl>	<dbl>
	-4	5	-280

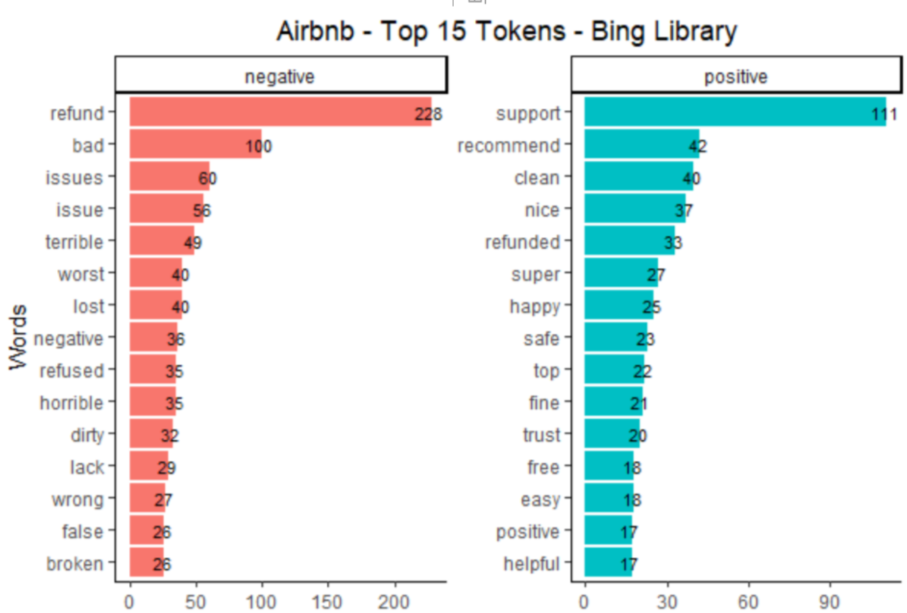

Figure 4: Airbnb Top 15 Tokens by Sentiment

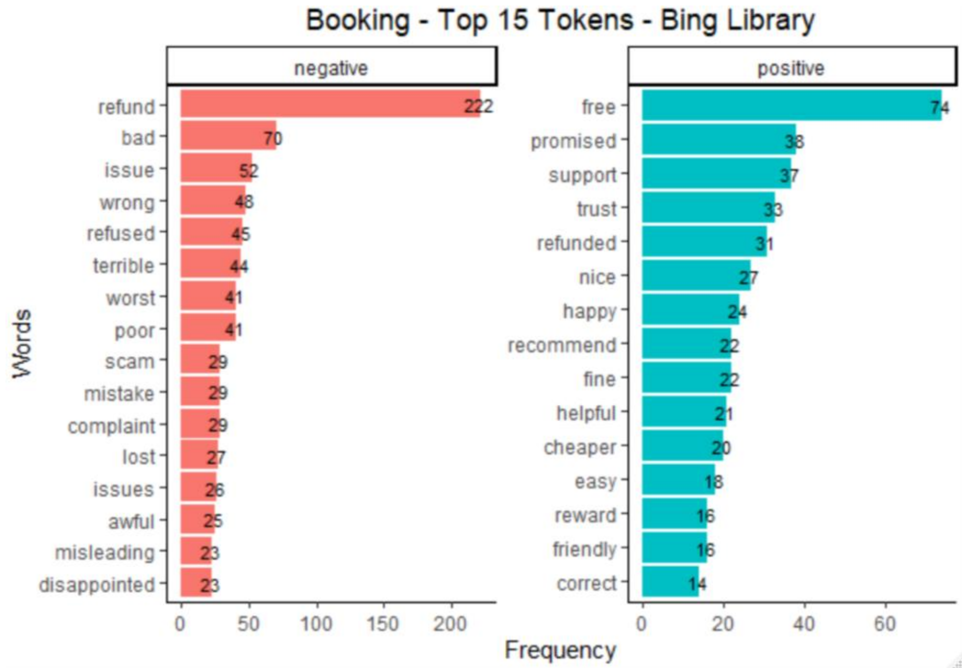

Figure 5: Booking.com Top 15 Tokens by Sentiment

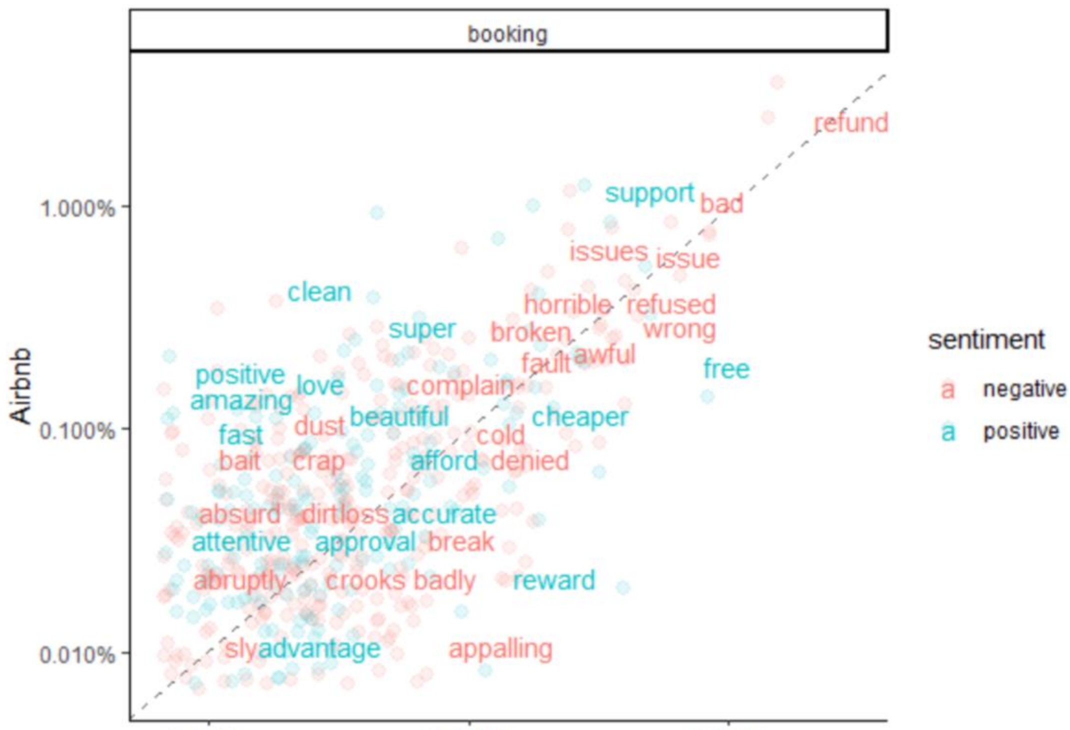

Figure 6: Correlogram Airbnb vs Booking.com by Sentiment

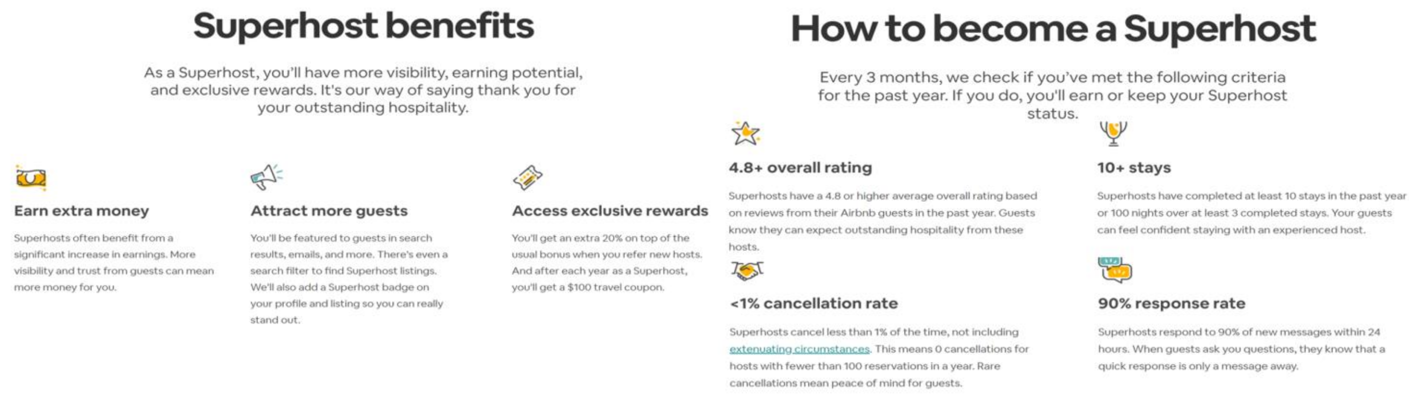

Figure 7: Superhost Benefits - Airbnb 2020

5. SOURCES

[1] Kingl, A. DataCamp. (February 27, 2018). *Web Scraping in R: rvest Tutorial*. DataCamp. Retrieved February 8, 2020 from https://www.datacamp.com/community/tutorials/r-web-scraping-rvest

[2] Kaushik, S. Analytics Vidhya. (March, 27, 2017). *Beginner's Guide on Web Scraping in R (using rvest) with hands-on example*. Analytics Vidhya. Retrieved February 8, 2020 from https://www.analyticsvidhya.com/blog/2017/03/beginners-guide-on-web-scraping-in-r-using-rvest-with-hands-on-knowledge/

[3] Lindblad, B. Towards Data Science. (April 29, 2018). *Scrape Data and Build a Webapp in R Using Rvest and Shiny*. Towards Data Science. Retrieved February 8, 2020 from https://towardsdatascience.com/scrape-data-and-build-a-webapp-in-r-using-rvest-and-shiny-f20d84dd1b74

[4] Nishida, K. Blog Explanatory. (July 11, 2016). Introduction to Text Sentiment Analysis in Exploratory. Blog Explanatory. Retrieved February 8, 2020 from https://blog.exploratory.io/twitter-sentiment-analysis-scoring-by-sentence-b4d455de3560

[5] Airbnb. CraftCo. (February, 2020). Airbnb. Retrieved February 8, 2020 from https://craft.co/airbnb/metrics

[6] Airbnb. (February, 2020). *Airbnb Superhost*. Airbnb. Retrieved February 8, 2020 from https://www.airbnb.com/superhost

[7] Airbnb. (February, 2020). *What if I need to cancel because of an emergency or unavoidable circumstance?*. Airbnb. Retrieved February 8, 2020 https://www.airbnb.com/help/article/1320/what-if-i-need-to-cancel-because-of-an-emergency-or-unavoidable-circumstance