Text Analytics on Reviews of Indian IT Companies

Sadaat Tameem

2023-06-03

Introduction

I conducted text analytics on customer reviews for Indian IT companies as part of my preparation to open a new business in the industry. The dataset consisted of reviews obtained from various sources, providing insights into the pain points experienced by customers.

By analyzing the reviews, I aimed to understand the common issues and challenges faced by customers when dealing with Indian IT companies. This information was crucial in developing a customer-centric business strategy that addresses these concerns effectively.

Through text analytics, I gained valuable insights into recurring themes and specific pain points highlighted by customers. By addressing these pain points, I aimed to create a business model that caters to the specific needs and expectations of Indian IT customers.

Understanding and resolving customer pain points would enable me to design the right products, services, and customer experiences that mitigate the challenges faced by customers in the Indian IT industry. This customer-centric approach would lay a strong foundation for my business,trust, loyalty, and success in the market.

Research Questions

The research questions of interest are as follows:

1. How does sentiment analysis correlate with star ratings across companies?
2. What are the dominant topics in customer reviews across different companies?
3. Which aspects or features are most frequently mentioned in customer reviews?
4. How do bigram analysis and the relationships between words in customer reviews contribute to understanding the context and sentiment of the feedback?

Unveiling Intriguing Patterns: How These Research Questions Shed Light on Key Insights

- The correlation between sentiment analysis and star ratings is crucial for understanding the drivers of positive or negative customer experiences. By examining this relationship, companies can gain deep insights into the specific aspects that contribute to customer satisfaction or dissatisfaction. Armed with this knowledge, companies can make data-informed decisions to refine their products, fine-tune their services, and ultimately enhance overall customer satisfaction, leading to increased loyalty and positive brand perception.

- Identifying the dominant topics in customer reviews across different companies provides a valuable opportunity to uncover recurring themes and patterns. This knowledge empowers companies to prioritize their efforts, addressing areas that customers find most noteworthy or problematic. By proactively addressing these key topics, companies can optimize their offerings, provide targeted solutions, and improve the overall customer experience. This strategic focus can result in increased customer loyalty, positive word-of-mouth, and a competitive advantage in the market.

- Uncovering the most frequently mentioned aspects or features in customer reviews is instrumental in understanding customer preferences and pain points. By analyzing these insights, companies can align their resources and strategies to address the areas that customers value the most or are dissatisfied with. This customer-centric approach enables companies to allocate resources effectively, make informed business decisions, and develop products or services that meet or exceed customer expectations. The result is enhanced customer satisfaction, increased brand loyalty, and improved business performance.

- Exploring bigrams in customer reviews and analyzing the relationships between words offer a deeper understanding of the context and sentiment expressed by customers. By identifying significant word combinations and the impact they have on the overall meaning of reviews, companies can uncover nuanced insights. This understanding of word relationships allows companies to extract valuable information, identify emerging trends, and derive actionable insights to refine their offerings. By leveraging these insights, companies can effectively communicate with their customers, address specific pain points, and enhance their products or services to align with customer preferences.

Overall, finding answers to these research questions provides tangible benefits to companies by enabling them to make data-driven decisions, enhance the customer experience, and ultimately improve their business performance. By strategically addressing each question, companies can cultivate stronger customer relationships, drive customer satisfaction, and achieve long-term success in their respective markets.

Data Description

Data Source:

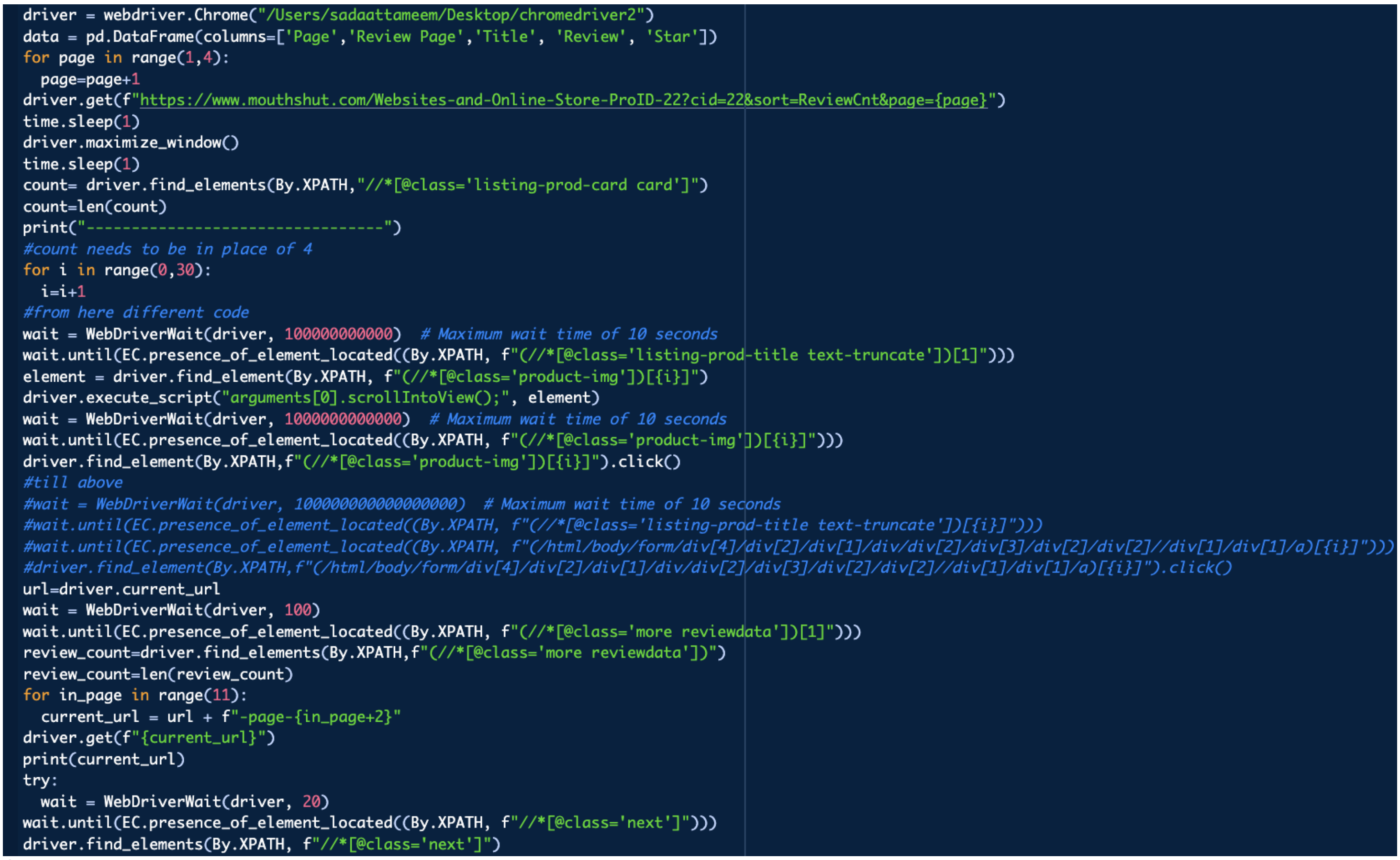
The most challenging and intriguing part of this project was the web scraping process in Python conducted on the website "https://www.mouthshut.com/websites". This involved understanding the webpage structure and meticulously investigating each page's XPath. The task required dealing with various loops, as the main page contained multiple sections, each with its own set of pages, and within those pages were numerous reviews.

One of the major obstacles encountered during web scraping was handling the numerous ad pop-ups. I implemented logic to wait for elements to become visible before executing the code. Despite these challenges, the web scraping endeavor proved to be highly rewarding. It emerged as a project within the project, enabling me to successfully extract three crucial data points: company name, review title, and average star rating.

To ensure robustness and continuity, I incorporated a loop for both review page numbers and the page number from which the information was scraped. This approach allowed the scraping process to restart from a specific webpage in case of any failures or interruptions in the script.

Overall, the web scraping phase presented a complex yet fascinating aspect of the project. It demanded a deep understanding of webpage structures, effective handling of ad pop-ups, and meticulous implementation of loops. Through this process, I successfully gathered valuable information, laying the foundation for further analysis and insights in the project.

Attaching the screen shot of Python code of webscraping



```
driver = webdriver.Chrome("/Users/sadaattameem/Desktop/chromedriver2")
data = pd.DataFrame(columns=['Page','Review Page','Title', 'Review', 'Star'])
for page in range(1,4):
  page=page+1
driver.get(f"https://www.mouthshut.com/Websites-and-Online-Store-ProID-22?cid=22&sort=ReviewCnt&page={page}")
time.sleep(1)
driver.maximize_window()
time.sleep(1)
count= driver.find_elements(By.XPATH,"//*[@class='listing-prod-card card']")
count=len(count)
print("--------------------------------")
#count needs to be in place of 4
for i in range(0,30):
  i=i+1
#from here different code
wait = WebDriverWait(driver, 100000000000)  # Maximum wait time of 10 seconds
wait.until(EC.presence_of_element_located((By.XPATH, f"(//*[@class='listing-prod-title text-truncate'])[1]")))
element = driver.find_element(By.XPATH, f"(//*[@class='product-img'])[{i}]")
driver.execute_script("arguments[0].scrollIntoView();", element)
wait = WebDriverWait(driver, 1000000000000)  # Maximum wait time of 10 seconds
wait.until(EC.presence_of_element_located((By.XPATH, f"(//*[@class='product-img'])[{i}]")))
driver.find_element(By.XPATH,f"(//*[@class='product-img'])[{i}]").click()
#till above
#wait = WebDriverWait(driver, 100000000000000000)  # Maximum wait time of 10 seconds
#wait.until(EC.presence_of_element_located((By.XPATH, f"(//*[@class='listing-prod-title text-truncate'])[{i}]")))
#wait.until(EC.presence_of_element_located((By.XPATH, f"(/html/body/form/div[4]/div[2]/div[1]/div/div[2]/div[3]/div[2]/div[2]//div[1]/div[1]/a)[{i}]")))
#driver.find_element(By.XPATH,f"(/html/body/form/div[4]/div[2]/div[1]/div/div[2]/div[3]/div[2]/div[2]//div[1]/div[1]/a)[{i}]").click()
url=driver.current_url
wait = WebDriverWait(driver, 100)
wait.until(EC.presence_of_element_located((By.XPATH, f"(//*[@class='more reviewdata'])[1]")))
review_count=driver.find_elements(By.XPATH,f"(//*[@class='more reviewdata'])")
review_count=len(review_count)
for in_page in range(11):
  current_url = url + f"-page-{in_page+2}"
driver.get(f"{current_url}")
print(current_url)
try:
  wait = WebDriverWait(driver, 20)
wait.until(EC.presence_of_element_located((By.XPATH, f"//*[@class='next']")))
driver.find_elements(By.XPATH, f"//*[@class='next']")
```

Screenshot Description

Data dictionary:

1. Page: This is a numerical value representing the main page number from which the record is obtained.
2. Review Page: This is a numerical value indicating the review page number from which the review information is scraped.
3. Title: This refers to the company name associated with the review.
4. Review: This contains the user's review about the company.
5. Star: This represents the average star rating given to the company.

The data dictionary provides a clear description of each data point extracted through the scraping process, enabling better understanding and interpretation of the collected information.

Data set observations: Total Columns = 5 Total Rows = 42,331

Reading and Preprocessing: (After webscraping in Python)

Excel File Reading and Saving into Proj Variable

```
Proj <- read_excel("/Users/sadaattameem/Desktop/R_text_analysis_project_data.xlsx")

#Viewing the first 5 rows
head(Proj,n=5)
```

```
## # A tibble: 5 × 5
##    Page `Review Page` Title        Review                                    Star
##   <dbl>         <dbl> <chr>        <chr>                                    <dbl>
## 1     1             0 Flipkart.com "Flipkart has the worst refund and ret…   3.1
## 2     1             0 Flipkart.com "I placed an order for Sony speaker wo…   3.1
## 3     1             0 Flipkart.com "I have order aarya shape ware not rec…   3.1
## 4     1             0 Flipkart.com "I bought a dress on your site. But th…   3.1
## 5     1             0 Flipkart.com "I have purchased a Samsung 5g phone o…   3.1
```

```
Proj
```

```
## # A tibble: 46,188 × 5
##     Page `Review Page` Title       Review                               Star
##    <dbl>         <dbl> <chr>       <chr>                               <dbl>
##  1     1             0 Flipkart.com "Flipkart has the worst refund and re…   3.1
##  2     1             0 Flipkart.com "I placed an order for Sony speaker w…   3.1
##  3     1             0 Flipkart.com "I have order aarya shape ware not re…   3.1
##  4     1             0 Flipkart.com "I bought a dress on your site. But t…   3.1
##  5     1             0 Flipkart.com "I have purchased a Samsung 5g phone …   3.1
##  6     1             0 Flipkart.com "Flipkart is doing fraud with their c…   3.1
##  7     1             0 Flipkart.com "I have a order but I don't receive m…   3.1
##  8     1             0 Flipkart.com "Extremely Dissatisfied ..\n\nThe ser…   3.1
##  9     1             0 Flipkart.com "Total west my money 7128 cash on del…   3.1
## 10     1             0 Flipkart.com "Flipkart is the worst application I …   3.1
## # ℹ 46,178 more rows
```

Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA): Data Structure, Handling Null Values, and Histogram Analysis of the DataFrame

```
#Filtering the data for 11 pages to maintain uniformity for further text analytics
Proj <- Proj %>%
  filter(`Review Page` <= 11)

#Viewing the structure of the Dataframe
str(Proj)
```

```
## tibble [42,388 × 5] (S3: tbl_df/tbl/data.frame)
##  $ Page       : num [1:42388] 1 1 1 1 1 1 1 1 1 1 ...
##  $ Review Page: num [1:42388] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Title      : chr [1:42388] "Flipkart.com" "Flipkart.com" "Flipkart.com" "Flipkart.com" ...
##  $ Review     : chr [1:42388] "Flipkart has the worst refund and returns policy.\n\nI have been trying to return my product since last week bu"| __truncated__ "I placed an order for Sony speaker worth Rs. 25000 with Flipkart however when I received the product I was shoc"| __truncated__ "I have order aarya shape ware not received but marked as delivered worst customer service while without resolut"| __truncated__ "I bought a dress on your site. But this is a damaged product.It is money that everyone earns hard. Don't lose y"| __truncated__ ...
##  $ Star       : num [1:42388] 3.1 3.1 3.1 3.1 3.1 3.1 3.1 3.1 3.1 3.1 ...
```

```
#finding null data
null_counts <- colSums(is.na(Proj))

#viewing null data
null_counts
```

```
##        Page Review Page       Title      Review        Star
##           0           0           0          57          54
```

```
#Dropping null values
Proj <- na.omit(Proj)

#Again viewing null values to confirm
null_counts <- colSums(is.na(Proj))
null_counts
```

```
##        Page Review Page       Title      Review        Star
##           0           0           0           0           0
```

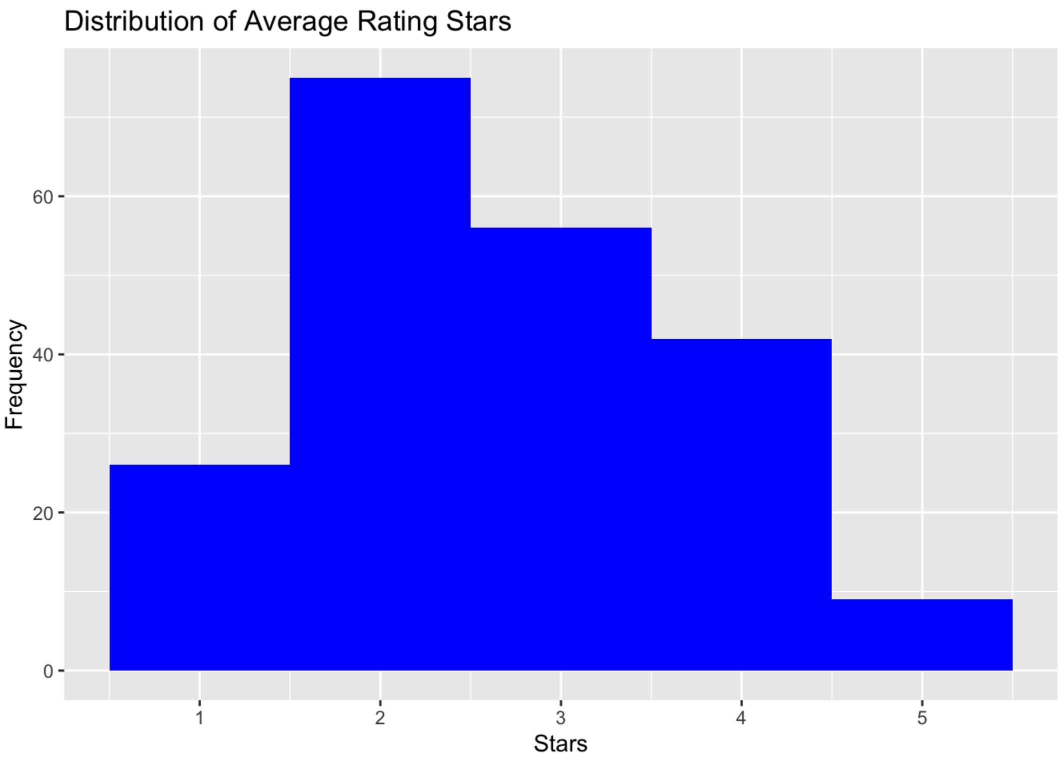
```
#Group the data by the "Title" column and calculate the maximum value for each title, storing the result in a new data frame.
Proj_unique_by_company <- Proj %>%
  group_by(Title) %>%
  summarize(Star = max(Star))

#view the new data frame
Proj_unique_by_company
```

```
## # A tibble: 208 × 2
##    Title                  Star
##    <chr>                 <dbl>
##  1 1mg.com                2.07
##  2 2captcha.com           3.25
##  3 5paisa.com             1.46
##  4 70trades.com           2.11
##  5 99acres.com            2.78
##  6 Abof.com               3
##  7 Agoda.com              1.14
##  8 Akbartravelsonline.com 3.14
##  9 AliExpress.com         2.25
## 10 Amazon.com             3.5
## # ℹ 198 more rows
```

```
#Summary statistics of numerical variables
summary(Proj$Star)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.770   2.500   2.655   3.490   4.740
```

```
#Histogram of stars
ggplot(Proj_unique_by_company, aes(Star)) +
  geom_histogram(binwidth = 1, fill='blue') +
  labs(x = "Stars", y = "Frequency") +
  ggtitle("Distribution of Average Rating Stars")
```

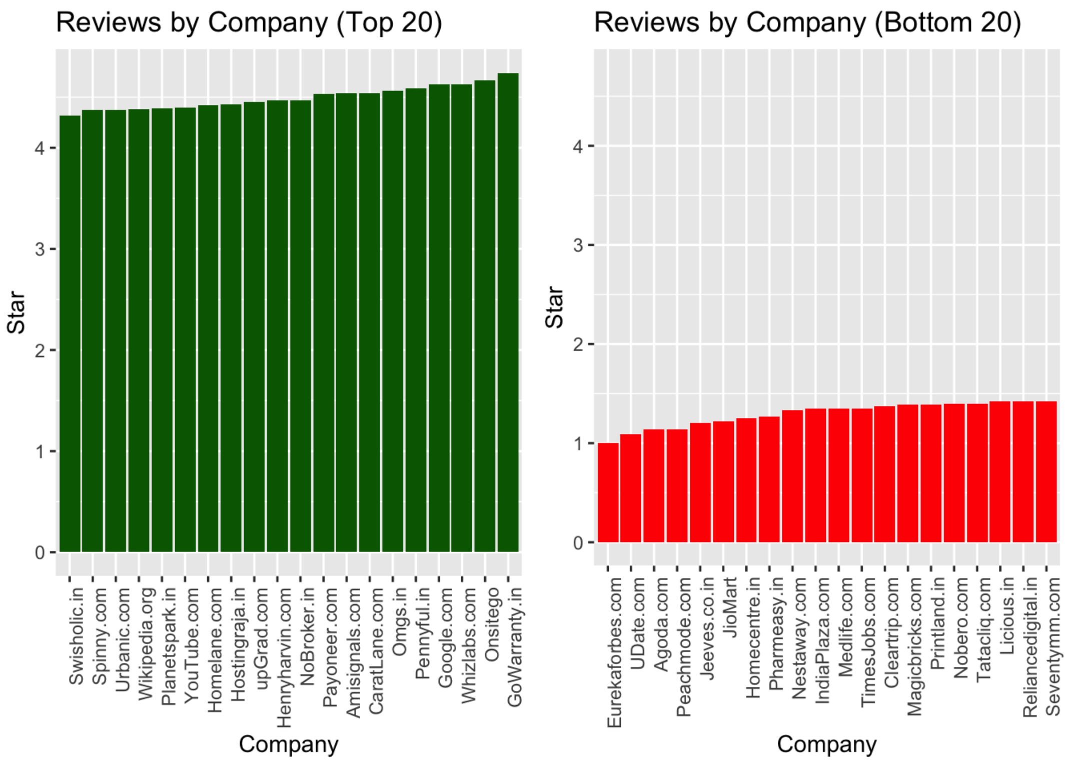


When observing the Histogram, it becomes apparent that the average star rating is predominantly concentrated between 2 and 4. This indicates that there is room for improvement by promptly addressing the pressing issues that users are experiencing, focusing on resolving their pain points

Combining the plots

Generating a plot to perform analysis on the Average Star Ratings of the Top 20 and Bottom 20 Companies

```
#Top 20 companies
company_counts_top <- Proj %>%
  group_by(Title) %>%
  summarize(Star = max(Star)) %>%
  arrange(desc(Star)) %>%
  head(20)

plot_top <- ggplot(company_counts_top, aes(reorder(Title, Star), Star)) +
  geom_bar(stat = "identity", fill = "darkgreen") +
  labs(x = "Company", y = "Star") +
  ggtitle("Reviews by Company (Top 20)") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))


#Bottom 20 companies
company_counts_bottom <- Proj %>%
  group_by(Title) %>%
  summarize(Star = max(Star)) %>%
  arrange(Star) %>%
  head(20)

plot_bottom <- ggplot(company_counts_bottom, aes(reorder(Title, Star), Star)) +
  geom_bar(stat = "identity", fill = "red") +
  labs(x = "Company", y = "Star") +
  ggtitle("Reviews by Company (Bottom 20)") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

#Get the maximum y-axis value from both plots
y_max <- max(c(company_counts_top$Star, company_counts_bottom$Star))

#Combine the plots with same y-axis limits
combined_plot <- grid.arrange(
  plot_top + ylim(0, y_max),
  plot_bottom + ylim(0, y_max),
  ncol = 2
)
```



The plot above illustrates the extreme ends of the spectrum, showcasing the Top 20 companies with the highest star ratings and the Bottom 20 companies with the lowest star ratings. This visualization serves as a valuable tool for conducting in-depth analysis.

Text Analysis

Tokenization

The text data was tokenized by splitting it into individual words, phrases, or tokens based on whitespace and punctuation marks. This process enables further analysis and extraction of valuable insights from the text data.

```
text <- Proj %>%
  unnest_tokens(output = word,
               input = Review)

#Viewing the Data file
text
```

```
## # A tibble: 3,705,875 × 5
##     Page `Review Page` Title         Star word
##    <dbl>         <dbl> <chr>        <dbl> <chr>
##  1     1             0 Flipkart.com   3.1 flipkart
##  2     1             0 Flipkart.com   3.1 has
##  3     1             0 Flipkart.com   3.1 the
##  4     1             0 Flipkart.com   3.1 worst
##  5     1             0 Flipkart.com   3.1 refund
##  6     1             0 Flipkart.com   3.1 and
##  7     1             0 Flipkart.com   3.1 returns
##  8     1             0 Flipkart.com   3.1 policy
##  9     1             0 Flipkart.com   3.1 i
## 10     1             0 Flipkart.com   3.1 have
## # ℹ 3,705,865 more rows
```

Stop Word Removal and Custom Stop Word Creation

In this step, stop words are removed from the text data, and a custom stop word dictionary is created and used to eliminate additional words. By removing these common and non-informative words, we can improve the accuracy and relevance of our text analysis.

During the analysis, it was observed that the stop words included some words present in the title. To address this, a customized stop word list was created by incorporating those title words, and they were subsequently removed from the text data. This step ensures that the title words do not interfere with the analysis and allow for more accurate results.

```
word_counts <- text %>%
  anti_join(stop_words) %>% #remove stop words
  count(Title,word,sort=T)
```

```
## Joining with `by = join_by(word)`
```

```
# Print the updated data frame

Proj$Title_2 <- sapply(strsplit(Proj$Title, ".", fixed = TRUE), "[", 1)

stop_company_names<-unique(Proj$Title_2)

stop_company_names <- tibble(word = stop_company_names)
stop_company_names<-tibble(tolower(stop_company_names$word))
colnames(stop_company_names)
```

```
## [1] "tolower(stop_company_names$word)"
```

```
my_stop <- rename(stop_company_names, "word" = "tolower(stop_company_names$word)")

#now removing the company names:
company_word <- text %>%
  anti_join(stop_words) %>% #remove stop words
  anti_join(my_stop) %>%
  count(Title,word,sort=T)
```

```
## Joining with `by = join_by(word)`
## Joining with `by = join_by(word)`
```

```
#viewing the data after removing custom stop words
company_word
```

```
## # A tibble: 398,631 × 3
##    Title           word        n
##    <chr>           <chr>   <int>
##  1 Redbus.in       bus      1587
##  2 TravelGuru.com  hotel     835
##  3 Cricbuzz.com    cricket   642
##  4 Veenaworld.com  tour      577
##  5 Google.com      search    559
##  6 Dominos.co.in   pizza     550
##  7 OyoRooms.com    oyo       543
##  8 Paytmmall.com   product   539
##  9 Bikedekho.com   bike      525
## 10 TimesJobs.com   job       511
## # ℹ 398,621 more rows
```

Sentiment Anlaysis

Sentiment Analysis #1 using get_sentiments()

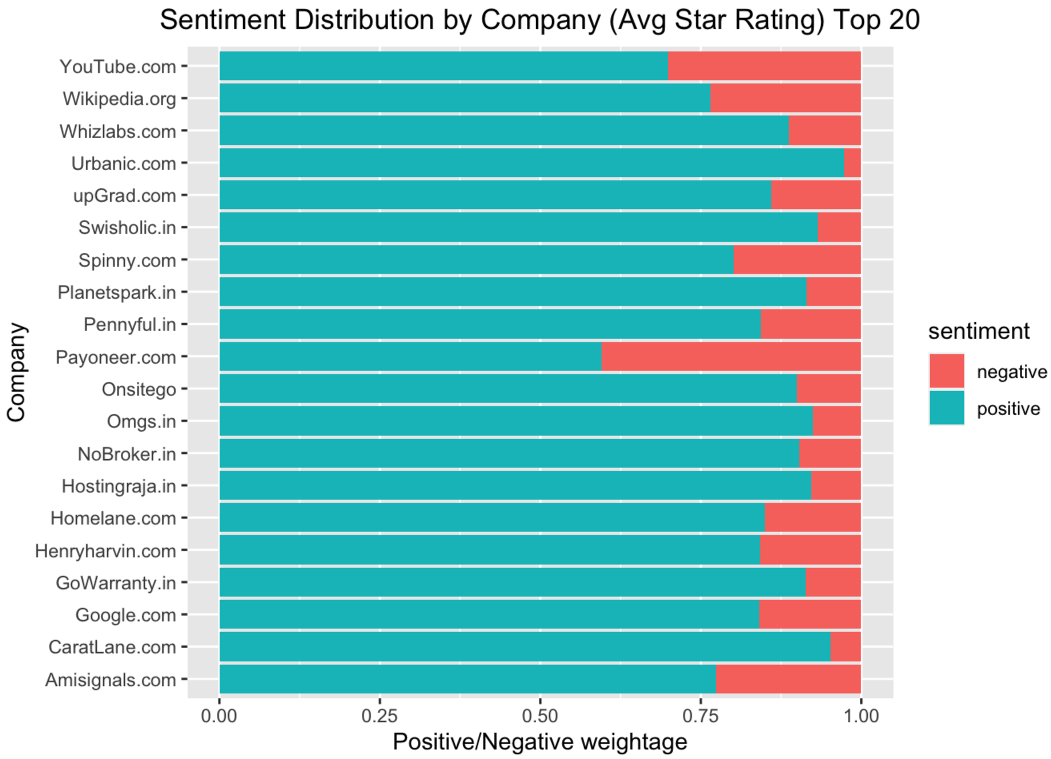
Correlation between Average Star Rating and Sentiment for Top 20 and Bottom 20 IT Companies in India

```
afin<-get_sentiments("afinn")

company_word<- text %>%
  anti_join(stop_words) %>%
  left_join(afin,relationship = "many-to-many") %>%
  left_join(get_sentiments(),relationship = "many-to-many") %>%
  #left_join(get_sentiments("nrc"),relationship = "many-to-many") %>%
  filter(!is.na(value)) %>%
  filter(!is.na(sentiment)) %>%
  count(Title,sentiment)
```

```
## Joining with `by = join_by(word)`
## Joining with `by = join_by(word)`
## Joining with `by = join_by(word)`
```

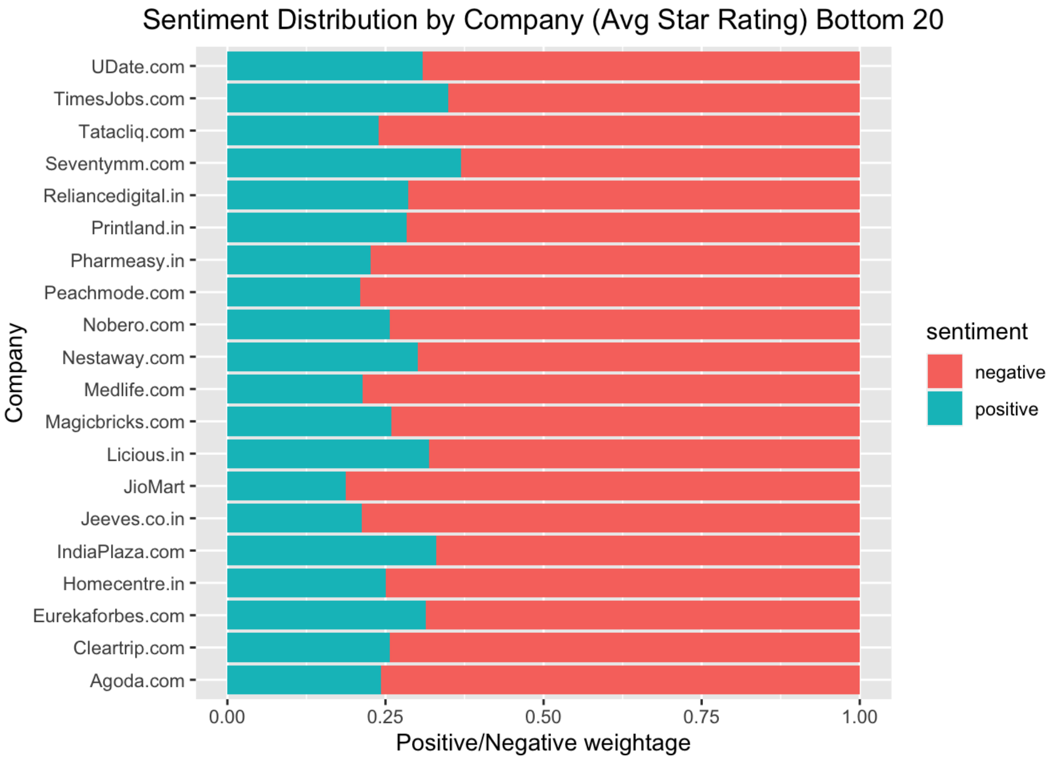
```
#Top 20 Companies sentiment plot
select(company_word,Title,sentiment,n)%>%
  inner_join(select(company_counts_top,Title)) %>%
  filter(!is.na(Title))  %>%
  ggplot() +
  geom_col(aes(x = n, y = Title, fill = sentiment), position = "fill") +
  labs(title = "Sentiment Distribution by Company (Avg Star Rating) Top 20",
       x = "Positive/Negative weightage",
       y = "Company")+ theme(plot.title = element_text(hjust = 0.5))
```

```
## Joining with `by = join_by(Title)`
```



```
#Bottom 20 Companies sentiment plot

select(company_word,Title,sentiment,n)%>%
  inner_join(select(company_counts_bottom,Title)) %>%
  filter(!is.na(Title))%>%
  ggplot() +
  geom_col(aes(x = n, y = Title, fill = sentiment), position = "fill") +
  labs(title = "Sentiment Distribution by Company (Avg Star Rating) Bottom 20",
       x = "Positive/Negative weightage",
       y = "Company")+ theme(plot.title = element_text(hjust = 0.5))
```

```
## Joining with `by = join_by(Title)`
```



Correlation between Average Star Rating and Sentiment for Top 20 IT Companies in India

After plotting the top 20 companies, we can observe a significant correlation between Average Star Rating and sentiment. The analysis reveals that approximately 75% of the sentiments expressed were predominantly positive. This finding indicates a strong positive sentiment trend among the top-ranked companies, reflecting high customer satisfaction and positive experiences.

Correlation between Average Star Rating and Sentiment for Bottom 20 IT Companies in India

After plotting the bottom 20 sentiments, we can observe a notable correlation between Average Star Rating and sentiment. The analysis reveals that approximately 25% of the sentiments expressed were exclusively positive. This finding indicates a positive sentiment trend among the bottom-ranked companies.

Sentiment Analysis #2 using Afinn sentiments

Correlation between Average Star Rating and Sentiment for Top 20 IT Companies in India

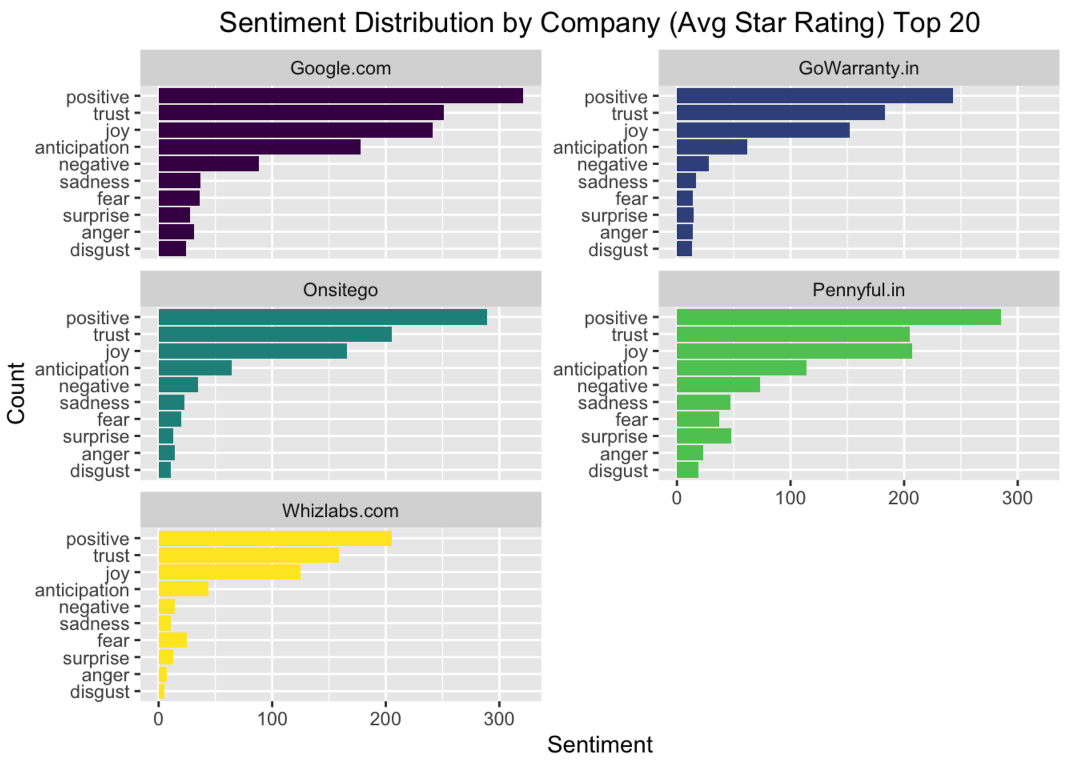
```
company_word_sentiments <- text %>%
  anti_join(stop_words) %>%
  left_join(afin, relationship = "many-to-many") %>%
  left_join(get_sentiments("nrc"), relationship = "many-to-many") %>%
  filter(!is.na(value)) %>%
  filter(!is.na(sentiment)) %>%
  count(Title, sentiment)
```

```
## Joining with `by = join_by(word)`
## Joining with `by = join_by(word)`
## Joining with `by = join_by(word)`
```

```
# Sentiment analysis of top 5 companies
company_counts_top <- company_counts_top %>% slice(1:5)

sent_top_f <- select(company_word_sentiments, Title, sentiment, n) %>%
  inner_join(select(company_counts_top, Title)) %>%
  filter(!is.na(Title))
```

```
## Joining with `by = join_by(Title)`
```

```
suppressWarnings({
  sent_top_f %>%
    group_by(Title, sentiment) %>%
    summarize(n = sum(n), .groups = "keep") %>%
    arrange(desc(n), Title) %>%
    ggplot() +
    geom_bar(aes(x = n, y = reorder(sentiment, n), fill = Title), stat = "identity") +
    scale_fill_viridis_d() +
    labs(title = "Sentiment Distribution by Company (Avg Star Rating) Top 20",
         x = "Sentiment",        y = "Count") +
    theme(plot.title = element_text(hjust = 0.5)) +
    facet_wrap(~ Title, ncol = 2, scales = "free_y") +
    guides(fill = "none")
})
```



Based on the provided data, the interpretations for the top 3 sentiments in each company are as follows:

1.GoWarranty.in:

Positive: GoWarranty.in has received a significant number of positive sentiments from customers, indicating a favorable perception of their services or products. This suggests that customers are satisfied and happy with their experience. Joy: The presence of joy as one of the top sentiments suggests that customers have expressed feelings of happiness, delight, or contentment while interacting with GoWarranty.in. This indicates a positive emotional connection with the company. Trust: The high count of trust indicates that customers have confidence and faith in GoWarranty.in. Trust is an essential sentiment that suggests customers believe in the company's reliability, credibility, and ability to deliver on promises.

2.Google.com:

Positive: Google.com has received a large number of positive sentiments from customers, indicating a strong positive perception. This suggests that customers have had positive experiences, interactions, or outcomes related to Google.com. Joy: The presence of joy as a top sentiment suggests that customers have expressed feelings of happiness, excitement, or satisfaction while using Google.com. This indicates a positive emotional connection and enjoyable experiences. Trust: The high count of trust indicates that customers have a strong belief in the reliability, credibility, and trustworthiness of Google.com. Trust is crucial for customers when using services or products provided by Google.com.

3.Onsitego:

Positive: Onsitego has received a significant number of positive sentiments from customers, indicating a favorable perception of their services. This suggests that customers have had positive experiences and outcomes with Onsitego. Joy: The presence of joy as a top sentiment suggests that customers have expressed feelings of happiness, pleasure, or satisfaction while interacting with Onsitego. This indicates a positive emotional connection and enjoyable experiences. Trust: The high count of trust indicates that customers have confidence and faith in the reliability and credibility of Onsitego. This suggests that customers trust Onsitego to provide reliable and trustworthy services.

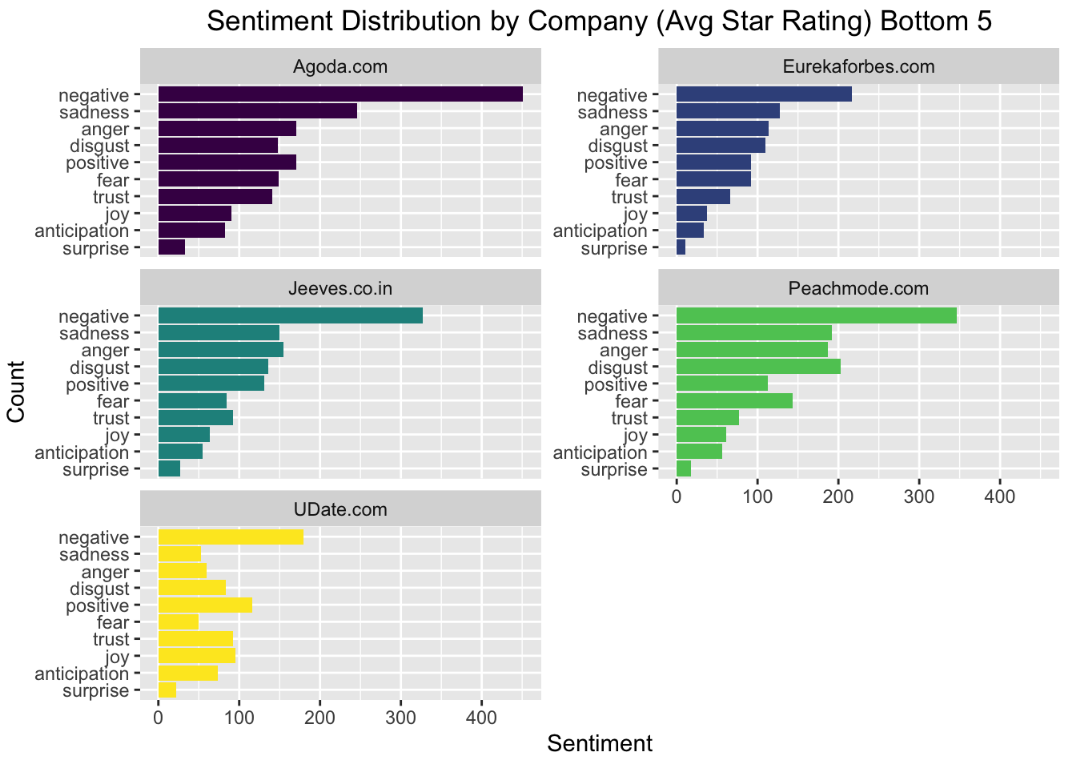
The interpretations for Pennyful.in and Whizlabs.com follow a similar pattern, where positive sentiment, joy, and trust are the top sentiments expressed by customers. These sentiments reflect the overall positive perception and emotional connection customers have with these companies.

Understanding these top sentiments can provide valuable insights into customer satisfaction, emotional experiences, and the level of trust customers have in the companies. This information can help companies identify areas of strength and focus on maintaining positive customer experiences and building trust.

```
#Performing sentiment for bottom 5 companies
company_counts_bottom<-company_counts_bottom %>%  slice(1:5)
#filtering for those 5 companies
sent_bottom_f<-select(company_word_sentiments,Title,sentiment,n)%>%
  inner_join(select(company_counts_bottom,Title)) %>%
  filter(!is.na(Title))
```

```
## Joining with `by = join_by(Title)`
```

```
#plotting the bar graph
suppressWarnings({
  sent_bottom_f %>%
    group_by(Title, sentiment) %>%
    summarize(n = sum(n), .groups = "drop") %>%
    arrange(desc(n)) %>%
    ggplot() +
    geom_bar(aes(x = n, y = reorder(sentiment, n), fill = Title), stat = "identity") +
    scale_fill_viridis_d() +
    labs(
      title = "Sentiment Distribution by Company (Avg Star Rating) Bottom 5",
      x = "Sentiment",
      y = "Count"
    ) +
    theme(plot.title = element_text(hjust = 0.5)) +
    facet_wrap(~ Title, ncol = 2, scales = "free_y") +
    guides(fill = "none")
})
```



1.Agoda.com: Negative: Customers expressed dissatisfaction and disappointment with Agoda.com. Sadness: Customers experienced unhappiness and sorrow while dealing with Agoda.com. Disgust: Customers felt aversion and revulsion towards Agoda.com.

2.Eurekaforbes.com: Negative: Customers expressed dissatisfaction and disappointment with Eurekaforbes.com. Sadness: Customers experienced unhappiness and sorrow while dealing with Eurekaforbes.com. Disgust: Customers felt aversion and revulsion towards Eurekaforbes.com.

3.Jeeves.co.in: Negative: Customers expressed dissatisfaction and disappointment with Jeeves.co.in. Sadness: Customers experienced unhappiness and sorrow while dealing with Jeeves.co.in. Anger: Customers felt frustration and irritation towards Jeeves.co.in.

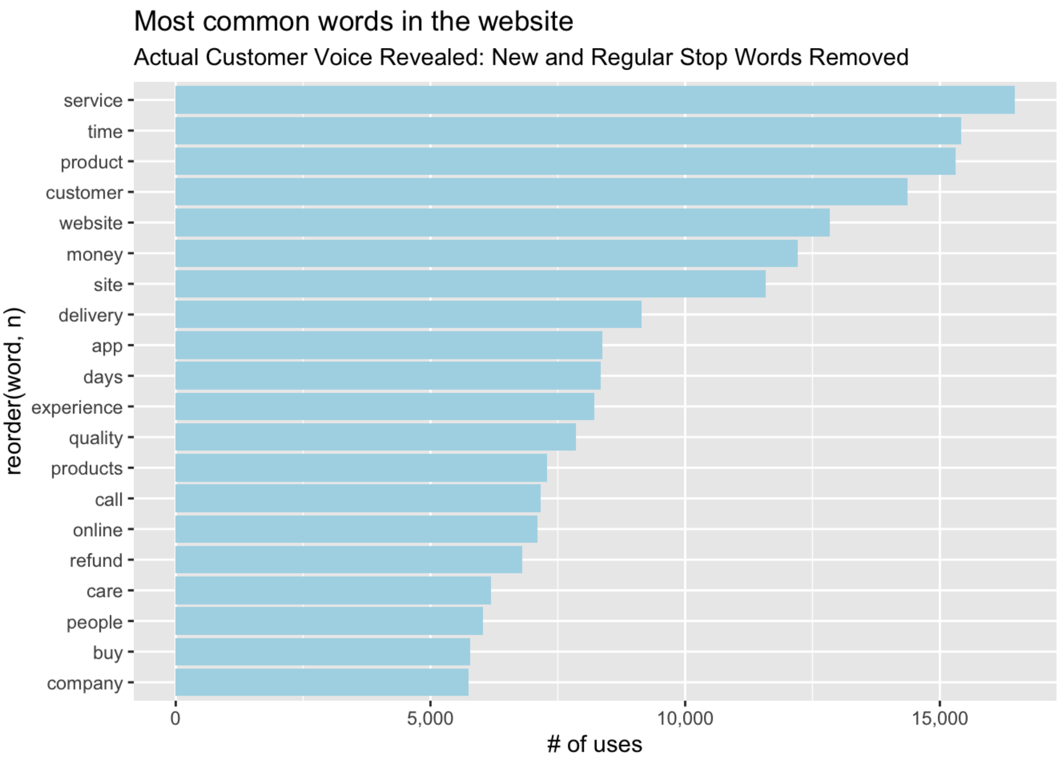
4.Peachmode.com: Negative: Customers expressed dissatisfaction and disappointment with Peachmode.com. Disgust: Customers felt aversion and revulsion towards Peachmode.com. Sadness: Customers experienced unhappiness and sorrow while dealing with Peachmode.com.

5.UDate.com: Negative: Customers expressed dissatisfaction and disappointment with UDate.com. Positive: Despite the negative sentiments, there were customers who had positive experiences with UDate.com. These summaries provide a brief overview of the top sentiments expressed by customers for each company.

Word Frequency

```
#creating a plot for word frequency overall
text %>%
  anti_join(stop_words) %>% #remove stop words
  anti_join(my_stop) %>%
  count(word,sort=T) %>%
  head(20) %>%
  ggplot(aes(reorder(word, n), n)) +  # Reorder the word variable by count (-n)
  geom_col(fill = "lightblue") +
  scale_y_continuous(labels = comma_format()) +
  coord_flip() +
  labs(
    title = "Most common words in the website",
    subtitle = "Actual Customer Voice Revealed: New and Regular Stop Words Removed",
    y = "# of uses"
  )
```

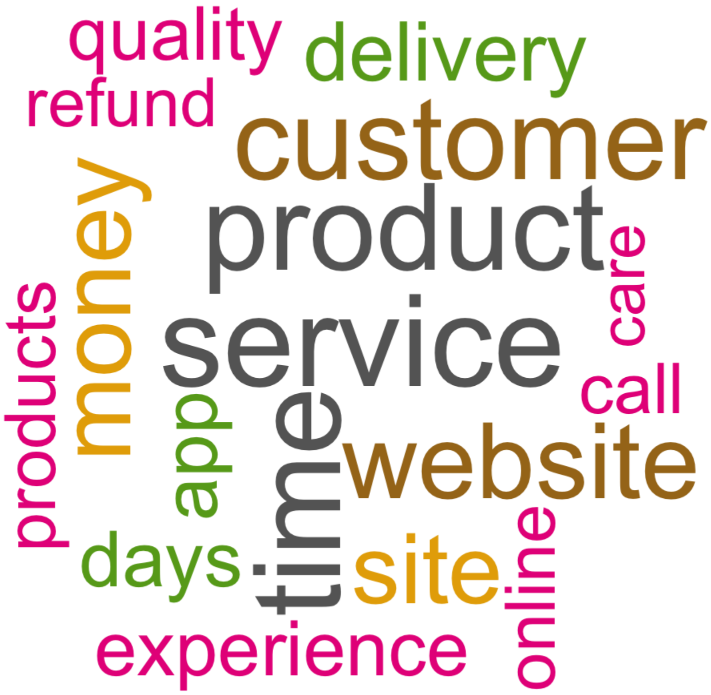
```
## Joining with `by = join_by(word)`
## Joining with `by = join_by(word)`
```



```
#creating a world cloud
w<-text %>%
  anti_join(stop_words) %>% #remove stop words
  anti_join(my_stop) %>%
  count(word,sort=T) %>% head(17)
```

```
## Joining with `by = join_by(word)`
## Joining with `by = join_by(word)`
```

```
w_filtered <- w[!grepl("\\d", w$word), ]

# Adjust plot margins to avoid warning
par(mar = c(0, 0, 0, 0))

# Create a word cloud with the filtered data
  suppressWarnings({wordcloud(words = w_filtered$word, freq = w_filtered$n, scale = c(5, 1),
          random.order = FALSE, rot.per = 0.35, colors = brewer.pal(8, "Dark2"),
          main = "Most common words in the website",
          sub = "Actual Customer Voice Revealed: New and Regular Stop Words Removed",
          ylab = "# of uses", max.words = 50)
  })
```



The Overall Word frequency analysis reveals the following insights:

Service: The word “service” appears most frequently in the analysis, indicating that customers frequently discuss their experiences with the service provided by the company. This suggests that service quality is an important aspect for customers.

Time: The word “time” is another frequently mentioned word, implying that customers often express their opinions or concerns regarding the time it takes for various processes, such as delivery, response, or resolution. This suggests that time efficiency is an important factor for customers.

Product: The word “product” is highly mentioned, indicating that customers frequently discuss the company’s offerings and express their opinions about the quality, features, or performance of the products. This suggests that product satisfaction is a significant aspect for customers.

Customer: The word “customer” is prominently present, suggesting that customers often refer to themselves or express their perspectives as consumers. This indicates that customer-centricity and understanding customer needs are important for the company’s success.

Website: The word “website” is frequently mentioned, indicating that customers discuss their experiences or interactions with the company’s website. This suggests that website usability, navigation, and overall user experience play a vital role in customer satisfaction.

Money: The word “money” appears prominently, suggesting that customers often express their concerns or feedback regarding the cost, pricing, or value for money aspects of the company’s products or services. This implies that pricing strategies and perceived value are significant factors influencing customer perceptions.

Site: The word “site” is frequently mentioned, implying that customers refer to the company’s website or online platform. This suggests that customers discuss their experiences, usability, or issues related to the company’s online presence.

Delivery: The word “delivery” is highly mentioned, indicating that customers frequently discuss their experiences or satisfaction with the company’s delivery service. This suggests that efficient and reliable delivery is an essential aspect of customer satisfaction.

App: The word “app” is prominently present, indicating that customers refer to the company’s mobile application and discuss their experiences or opinions about its usability, features, or performance. This suggests that a well-designed and user-friendly app is important for customer satisfaction.

Days: The word “days” appears frequently, suggesting that customers often mention specific time frames, such as delivery days or processing time, and express their satisfaction or dissatisfaction related to these time aspects. This indicates that promptness and meeting delivery timelines are important for customer satisfaction.

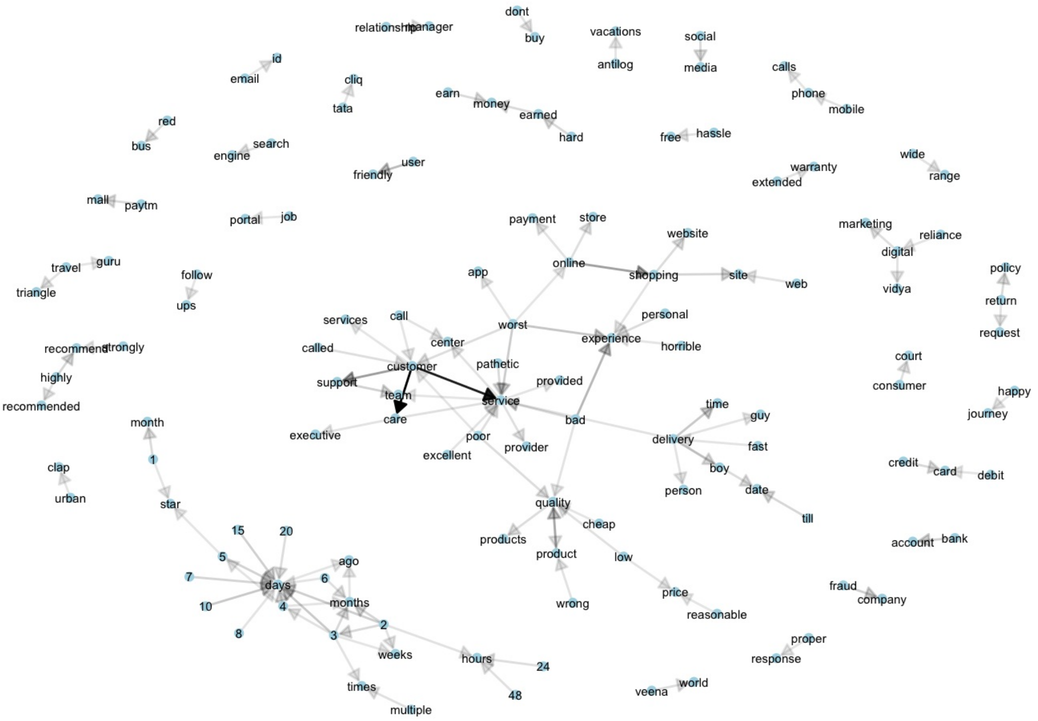
Overall, this word frequency analysis provides insights into the common topics and concerns expressed by customers. Understanding these key words and their frequencies can help the company identify areas of improvement, prioritize customer satisfaction initiatives, and enhance the overall customer experience.

Bi Gram Analysis

After observing the high frequency of certain words such as “customer” and “services” separately, as well as “delivery,” “days,” “website,” and “apps,” it became apparent that conducting a bi-gram analysis would provide valuable insights into their relationship and overall importance. By examining these words in pairs, we can uncover any patterns or associations that may not be evident when considering them individually.

```
# Perform bi-gram analysis on the project dataset
company_bigram <- Proj %>%
  unnest_tokens(bigram, Review, token = "ngrams", n = 2)

# Separate the bi-gram into individual words
company_bigram <- company_bigram %>%
  separate(bigram, c("word1", "word2"), sep = " ")

# Filter out stop words from the bi-grams
company_bigram <- company_bigram %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word)

# Remove rows with missing values in word1 or word2
company_bigrams <- company_bigram %>%
  filter(!is.na(word1)) %>%
  filter(!is.na(word2))

# Unite word1 and word2 to create the final bi-gram column
company_bigrams <- company_bigrams %>%
  unite(bigram, word1, word2, sep = " ")

# Count the occurrences of each bi-gram by company
company_bigrams_count <- company_bigrams %>% count(Title, bigram, sort = TRUE)

# Perform bi-gram analysis on the project dataset for overall word frequency
bigram_counts <- Proj %>%
  unnest_tokens(bigram, Review, token = "ngrams", n = 2) %>%
  separate(bigram, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  filter(!is.na(word1)) %>%
  filter(!is.na(word2)) %>%
  count(word1, word2, sort = TRUE)

# Create a graph from the bi-gram counts
bigram_graph <- bigram_counts %>%
  filter(n >= 150) %>%
  graph_from_data_frame()

# Visualize the bi-gram graph using the Fruchterman-Reingold layout
suppressWarnings({
  ggraph(bigram_graph, layout = "fr") +
    geom_edge_link(aes(alpha = n), show.legend = FALSE,
                   arrow = grid::arrow(type = "closed", length = unit(2, "mm")),
                   end_cap = circle(1, "mm")) +
    geom_node_point(color = "lightblue", linewidth = 2) +
    geom_node_text(aes(label = name), size = 2) +
    theme_void() +
    theme(legend.key.size = unit(0.5, "points")) +
    guides(linewidth = FALSE)

})
```

```
## Warning: Using the `size` aesthetic in this geom was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` in the `default_aes` field and elsewhere instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Here are some notable bigrams and their interpretations:

customer->care, customer->service, customer->support: These bigrams suggest a strong association between the word “customer” and different aspects of service and support. It indicates that customer care, service, and support are important considerations in the data.

online->shopping, shopping->site, shopping->website: These bigrams indicate a relationship between the words “online” and “shopping” as well as their association with websites and sites. It suggests that the data may contain feedback or experiences related to online shopping platforms.

bad->experience, worst->experience, pathetic->service: These bigrams highlight negative experiences, indicating dissatisfaction or disappointment expressed by the customers. It suggests that there are instances of poor service, bad experiences, and dissatisfaction in the data.

delivery->time, delivery->date, delivery->boy: These bigrams suggest a connection between the word “delivery” and aspects related to time, date, and delivery personnel (boy). It indicates that the data may include feedback or comments related to delivery experiences.

product->quality, earn->money, return->policy: These bigrams indicate a relationship between the words “product” and “quality” as well as associations between earning money and return policies. It suggests that the data may involve discussions about product quality, earning money, and return policies of certain companies.

fraud->company, pathetic->service, bad->service: These bigrams highlight negative experiences related to fraud, pathetic service, and bad service. It indicates that there are instances where customers encountered fraudulent activities or experienced unsatisfactory service from specific companies.

Overall, the bigram analysis helps identify frequently co-occurring word pairs, shedding light on the relationships and associations between the words. It provides valuable insights into customer experiences, service quality, online shopping, delivery, and other aspects discussed in the data.

Topic Modelling

Through topic modeling, we can gain a comprehensive understanding of the customer feedback landscape. It provides a powerful tool for extracting actionable insights, enabling us to make informed decisions and prioritize areas of focus based on customer needs and preferences.

Now, let’s dive into the process of topic modeling and explore the fascinating world of customer feedback analysis.

After conducting topic modeling analysis, I removed common words like “product,” “delivery,” “service,” and “job” to gain more meaningful insights. I started with 3 topics and gradually increased the number to find the best model. The optimal number of topics was determined to be 5, which successfully captured the diverse themes in the customer feedback

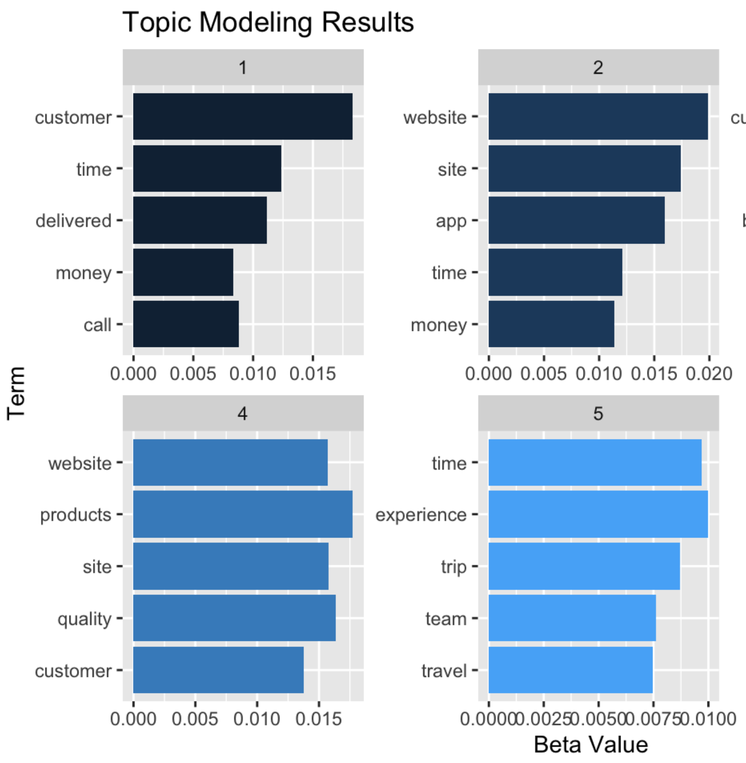
```
# Find document-word counts
company_word <- text %>%
  anti_join(stop_words) %>% # Remove stop words
  anti_join(my_stop) %>%
  count(Title, word, sort = TRUE)
```

```
## Joining with `by = join_by(word)`
## Joining with `by = join_by(word)`
```

```r
unwanted_words <- c("product", "delivery", "service", "job")  # List of words to remove

filtered_word_counts <- subset(company_word, !word %in% unwanted_words)

word_counts <- filtered_word_counts

# Make into DTM (Document-Term Matrix)
title_dtm <- word_counts %>%
  cast_dtm(document = Title,
           term = word,
           value = n)

# Display word counts
word_counts
```

```
## # A tibble: 397,925 × 3
##    Title           word        n
##    <chr>           <chr>   <int>
##  1 Redbus.in       bus      1587
##  2 TravelGuru.com  hotel     835
##  3 Cricbuzz.com    cricket   642
##  4 Veenaworld.com  tour      577
##  5 Google.com      search    559
##  6 Dominos.co.in   pizza     550
##  7 OyoRooms.com    oyo       543
##  8 Bikedekho.com   bike      525
##  9 World4ufree     movies    496
## 10 Ideacellular.com idea     495
## # ℹ 397,915 more rows
```

```r
# Perform Topic Modeling with 5 topics
title_lda <- LDA(title_dtm, k = 5, control = list(seed = 1234))

# Extract the topic-word probabilities
title_topics <- tidy(title_lda, matrix = "beta")

# Find the most common words in each topic
top_terms <- title_topics %>%
  group_by(topic) %>%
  slice_max(beta, n = 5) %>%
  ungroup() %>%
  arrange(topic, -beta)



# Plot the most common words in each topic
top_terms %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(x = beta, y = term, fill = topic)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  scale_y_reordered() +
  labs(title = "Topic Modeling Results",
       x = "Beta Value",
       y = "Term")
```



Labeling Topics

After analyzing the words in each topic, several insights emerged:

1.Customer Service: This topic captures discussions related to customer service interactions, response time, and phone support. The presence of terms like “customer,” “time,” and “call” suggests a focus on addressing customer queries and providing efficient assistance.

2.Website and App Experience: This topic revolves around users’ experiences with the company’s website and mobile application. Keywords such as “website,” “site,” and “app” indicate conversations related to user interface, navigation, and overall satisfaction with the digital platforms.

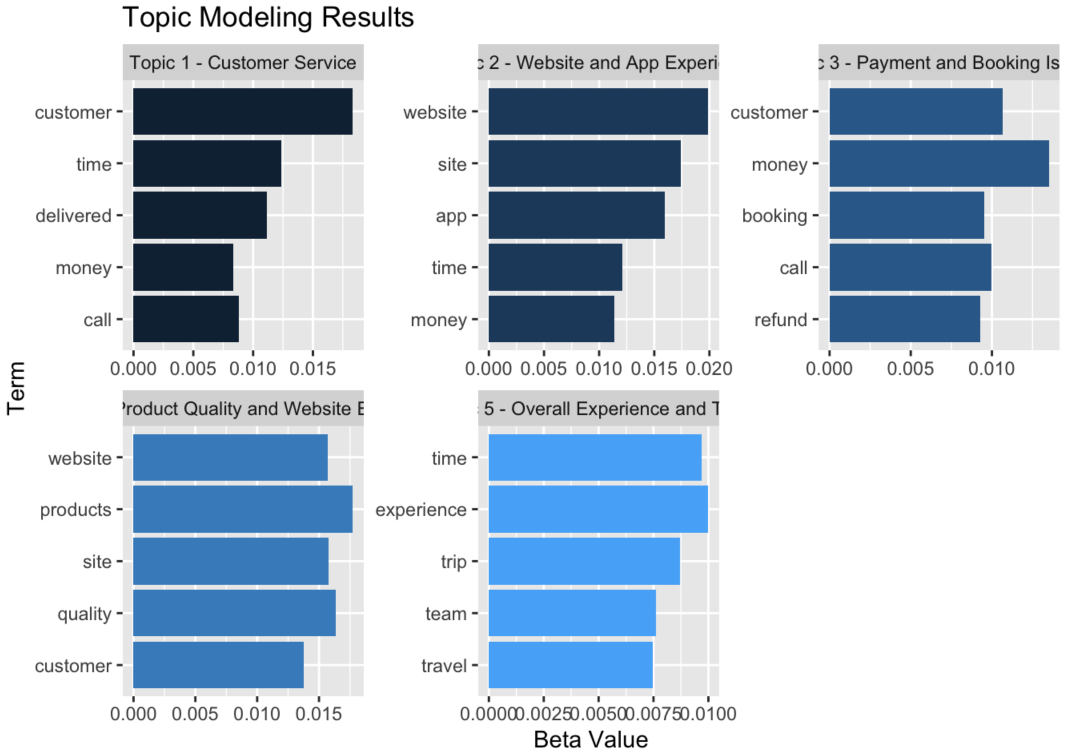
3.Payment and Booking Issues: This topic highlights concerns regarding payments, customer reimbursements, and booking problems. Terms like “money,” “customer,” and “refund” suggest discussions centered around payment failures, refund requests, and difficulties in the booking process.

4.Product Quality and Website Experience: This topic delves into the quality of products offered by the company and user experiences on their website. The presence of terms like “products,” “quality,” “site,” and “website” indicates conversations about product reliability, satisfaction, and ease of online browsing and shopping.

5.Overall Experience and Travel: This topic encompasses broader discussions about customers’ overall experiences and interactions with the company during their travel. Terms such as “experience,” “time,” “trip,” “team,” and “travel” signify a focus on travel-related aspects like customer experiences, trip arrangements, and interactions with the company’s team.

By identifying these distinct topics, the company can gain valuable insights into specific areas that require attention and improvement. Addressing customer service concerns, enhancing the website and app experience, resolving payment and booking issues, ensuring product quality, and optimizing the overall travel experience will contribute to enhanced customer satisfaction and loyalty.

```r
# Define topic labels
topic_labels <- c("Topic 1 - Customer Service", "Topic 2 - Website and App Experience", "Topic 3 - Payment and Booking Issues", "Topic 4 - Product Quality and Website Experience", "Topic 5 - Overall Experience and Travel")
# Plot the most common words in each topic with reordered topics and original topic colors
top_terms %>%
  mutate(term = reorder(term, beta),
         topic_original = topic) %>%  # Add a new column to preserve the original topic values
  mutate(topic = recode(factor(topic), "1" = topic_labels[1], "2" = topic_labels[2], "3" = topic_labels[3],
                        "4" = topic_labels[4], "5" = topic_labels[5])) %>%
  ggplot(aes(x = beta, y = term, fill = topic_original)) +  # Use the original topic values for fill
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  scale_y_reordered() +
  labs(title = "Topic Modeling Results",
       x = "Beta Value",
       y = "Term")
```



Conclusion

Conclusion:

The text analytics performed on the data revealed valuable insights into customer sentiments, topic modeling, word frequency, and sentiment analysis for top and bottom companies. Here is a summary of the key findings:

1.Bi-gram Analysis: The bi-gram analysis highlighted frequently occurring word pairs such as “customer care,” “online shopping,” “bad experience,” and “user-friendly.” These pairs provide insights into common phrases and concepts expressed by customers.

2.Topic Modeling: The topic modeling analysis identified five main topics discussed by customers. These topics were related to customer service, website and app experience, payment and booking issues, product quality and website experience, and overall experience and travel. Each topic was associated with specific terms that provided an understanding of the key themes in customer conversations.

3.Word Frequency: The word frequency analysis revealed the most frequently mentioned words in the data. The top words included “service,” “time,” “product,” “customer,” “website,” and “money.” These words give an indication of the most common concerns and interests expressed by customers.

4.Sentiment Analysis for Bottom Companies: The sentiment analysis for the bottom 20 companies highlighted negative sentiments such as dissatisfaction, disappointment, sadness, and disgust expressed by customers towards these companies. This analysis provides an overview of the negative experiences reported by customers.

5.Sentiment Analysis for Top Companies: The sentiment analysis for the top 20 companies indicated positive sentiments such as satisfaction, joy, and trust expressed by customers towards these companies. This suggests that customers had positive experiences, emotional connections, and trust in these companies.

In conclusion, the text analytics performed on the data provided valuable insights into customer sentiments, topics of discussion, frequently mentioned words, and sentiments towards different companies. These findings can help companies gain a better understanding of customer feedback, identify areas for improvement, and enhance their overall customer experience.

Next Steps

Based on the analysis I conducted, here are some steps I can take to improve the customer experience and grow our business, leveraging the insights gained and their potential impact:

1.Customer Service Enhancement: Focus on improving customer service interactions by addressing common concerns related to customer care, support, and response time. Analyze customer feedback to identify pain points and implement measures to enhance customer satisfaction in these areas.

2.Website and App Experience Optimization: Pay attention to users’ experiences with the website and mobile application. Identify areas for improvement based on discussions around website and app issues. Make necessary updates to enhance user-friendliness, navigation, and overall experience.

3.Payment and Booking Issue Resolution: Address concerns related to payments, refunds, and booking problems. Streamline the payment processes, improve transparency, and provide clear policies to build customer trust. Resolve issues promptly and offer timely customer reimbursements where applicable. Product Quality and Website Experience:

4.Evaluate customer discussions regarding product quality and experiences on the website. Take necessary steps to improve the quality of products offered and address any issues reported by customers. Enhance the user experience on website, ensuring ease of navigation, product information, and a smooth purchasing process.

5.Overall Customer Experience and Travel: Analyze broader discussions about overall customer experiences and interactions during travel. Identify areas for improvement in terms of communication, support, and service delivery. Work on enhancing customer satisfaction at various touchpoints throughout their travel journey.

6.Address Negative Sentiments: Take note of the negative sentiments expressed by customers towards business. Investigate the underlying issues and take corrective actions to address their concerns. Implement measures to improve customer satisfaction, resolve complaints, and rebuild trust.

7.Capitalize on Positive Sentiments: Leverage the positive sentiments expressed by customers towards business. Continue delivering excellent service, joy, and building trust. Encourage satisfied customers to share their positive experiences through testimonials, reviews, or referrals.

8.Continuous Monitoring and Feedback Analysis: Regularly monitor customer feedback, sentiments, and trends. Implement a system to capture and analyze customer feedback in real-time. Use the insights gained to make data-driven decisions and prioritize improvements.

I strongly believe that key is to listen to the customers, proactively address their concerns, and continuously strive to enhance their experience. By taking these steps, I can work towards improving customer satisfaction, loyalty, and the overall success of the business.