CNBC Articles Analysis with Subtopic: Consumers

Aysen Akpinar

2023-08-14

Introduction

The world is changing toward a data-driven environment. Data is now one of the most important concepts that drive innovation and decision-making and it also helps to understand what is happening around us. Not only businesses or institutions benefit from data, but also individuals. Before we buy a product as a consumer, we are more likely to check reviews or articles about the product we are interested in. Even if it is an article or product review, they certainly impact our decision-making process in a positive way or negative way. There are also studies in the literature that suggest marketing professionals encourage to show more positive product reviews for their products (Jang et al, 2012).

In this project, using text as data, we aim to provide insights from consumer articles that reflect recent trends, preferences, and the most common brands among customers. The intention is to grasp the categories (within text analysis) these brands are associated with. For this purpose, we decided to choose CNCB.com as the data source and "consumer" as a category.

Note:

This project refers to a comprehensive book called "Text as Data (Justin Grimmer, Margaret E. Roberts, Brandon M. Stewart)". The book is cited as "Grimmer, Roberts, & Stewart" throughout the project.

CNBC as Data Source

CNBC is one of the reputable business news channels in the United States. It also has a broad range of audiences on its websites and cable television broadcasts. It covers a wide range of topics including financial markets, consumer, and business events. Analyzing CNBC news or articles can provide important insights into recent trends since it is considered a trustworthy source. And compared to other mainstream news channels, the CNBC website has a well-organized HTML structure that provides an easier web-scraping process.

Consumer as a Subtopic

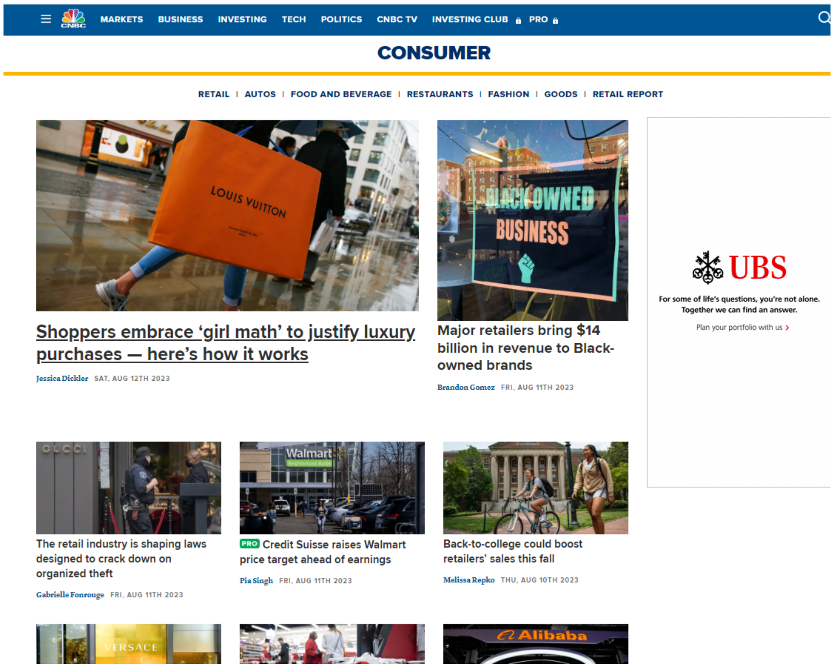
Consumers are key factors in society, the economy, and businesses. Nowadays, many companies are not only interested in selling their products, but also, they follow consumer feedback or trends to drive innovation through their organizations. Consumer topics also pique the attention of social researchers. Once we see the text as source data, it gives new opportunities to study (Grimmer, Roberts, & Stewart, pg 34). Besides, consumer behaviors are impacted by recent trends and can reshape the next change.

Research Question

In this project we aim to answer the following research question:

“Which companies appear most in Consumer news? What are the primary categories for these brands/customers?”

Corpus



Document selection is a central part of the analysis process which might be limited by constraints such as the availability to gather a larger collection.

Looking through our target website’s interface, the chosen category consists of pictures, titles, author names, categories (defined by the website), and date information for each article. To get full article information, the user needs to click on each article. Since we are going to analyze the textual part of it, we are not interested in images for this project.

Scrolling down to the bottom of the page, we see the ‘Load More’ button to see previous articles published on the website. The ‘load more’ button itself, does not provide functional information about the volume of articles that the category has. This led researchers to think about ‘quantities of interest’. Related to the research question, researchers ask questions about the quantity of data they need to gather (or in some cases, the population of interest). In this website, we overcome this constraint with the code we use in the scraping process which determines the number of pages in total and the number of articles. Even if the data is non-representative, a careful analysis can yield meaningful results (Grimmer, Roberts, & Stewart, pg. 47-48).

Corpus Gathering

For this project, we used Selenium and related Google drivers to scrape data from the website. The inspect function of Google aids to target elements we are interested in. For the web-scraping part, we used Jupyter notebook, for the analysis part we preferred R markdown.

Data and Pre-Processing

First, we start with the necessary libraries for this analysis.

```
library(qdap) # quantitative discourse analysis of transcripts
library(ggplot2) # plotting discourse data
library(data.table) # for easier data manipulation
library(scales) # to help us plot
library(tidyverse) # to help import data files
library(viridis) # inclusive color palates
library(tm)
library(wordcloud)
rm(list=ls())
```

Next, we set the working directory.

```
working_directory = "C:/Users/hodor/Desktop/TextasDataSummer/labs"

setwd(working_directory)
```

In this section, we gather our .txt files together to create R data frame so we can work on it.

(In this model, Prof. Posch's code has been used with some adjustments such as adding extra column, changing date format)

```
# Set the path to the folder containing the text files
folder_path <- "cnbc_v2"

# Get the list of text files in the folder
file_list <- list.files(folder_path, pattern = "*.txt", full.names = TRUE)

# Initialize an empty dataframe
articles_df <- data.frame(Title = character(),
                 Author = character(),
                 Date = character(),
                 URL = character(),
                 Category = character(),
                 Full_Text = character(),
                 stringsAsFactors = FALSE)

# 1.1 Loop through each file and read its content into the dataframe
for (file in file_list) {
  # Read the contents of the file
  file_content <- readLines(file)

  # Initialize variables
  title <- ""
  author <- ""
  date <- ""
  category <- ""
  url <- ""
  full_text <- ""

  # 1.2 Extract variables from the file content
  for (i in 1:length(file_content)) {
    line <- file_content[i]
    if (grepl("^Title:", line)) {
```

```r
      title <- trimws(sub("^Title:", "", line))
    } else if (grepl("^Author:", line)) {
      author <- trimws(sub("^Author:", "", line))
    } else if (grepl("^Date:", line)) {
      date <- trimws(sub("^Date:", "", line))
    } else if (grepl("^Category:", line)) {
      category <- trimws(sub("^Category:", "", line))
    } else if (grepl("^URL:", line)) {
      url <- trimws(sub("^URL:", "", line))
    } else if (grepl("^Full Text:", line)) {
      full_text <- trimws(sub("^Full Text:", "", line))

      # Extract the full text that spans multiple lines
      j <- i + 1
      while (j <= length(file_content) && !grepl("^\\s*$", file_content[j])) {
        full_text <- paste(full_text, file_content[j], sep = "\n")
        j <- j + 1
      }

      # Remove leading and trailing whitespace from the full text
      full_text <- trimws(full_text)
    }
  }

  # 1.3 Create a dataframe with the extracted variables
  file_df <- data.frame(Title = title,
                        Author = author,
                        Date = date,
                        Category = category,
                        URL = url,
                        Full_Text = full_text,
                        stringsAsFactors = FALSE)

  # 1.3 (cont.) Add the file dataframe to the main dataframe
  articles_df <- bind_rows(articles_df, file_df)
}

# see a truncated version of the data
head(truncdf(articles_df),10)
```

```
##         Title     Author       Date        URL   Category   Full_Text
## 1  Mediterran Amelia Luc 2023-06-15 https://ww RESTAURANT  In this ar
## 2  Michelin G Audrey Wan 2023-06-16 https://ww CNBC TRAVE  Singapore'
## 3  How restau Kate Roger 2023-06-17 https://ww RESTAURANT WATCH NOW\n
## 4  Chipotle w Ian Thomas 2023-06-18 https://ww     EVOLVE  In this ar
## 5  Domino's r Yuheng Zha 2023-06-20 https://ww RESTAURANT  In this ar
## 6  U.S. regul Amelia Luc 2023-06-21 https://ww FOOD & BEV  Chicken pr
## 7  Olive Gard Amelia Luc 2023-06-22 https://ww RESTAURANT  In this ar
## 8  Burger Kin Amelia Luc 2023-06-23 https://ww RESTAURANT  In this ar
## 9  Starbucks  Kate Roger 2023-06-23 https://ww RESTAURANT  In this ar
## 10 Stocks fal Kevin Stan 2023-06-23 https://ww
```

Our data frame comes with 198 articles and 6 variables such as "Title", "Author", "Date", "URL", "Category", and "Full_Text".

Over time, it was realized that the total number of articles on the website was always equal to 200, even though new articles were being published every day. This may be related to recently popular SEO practices, such as removing outdated content*.

As a pre-process, we check how many articles are missing:

```
x <- articles_df %>%
  filter(Full_Text == "") %>%
  select(Full_Text)

print(x)
```

```
##    Full_Text
## 1
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16
## 17
```

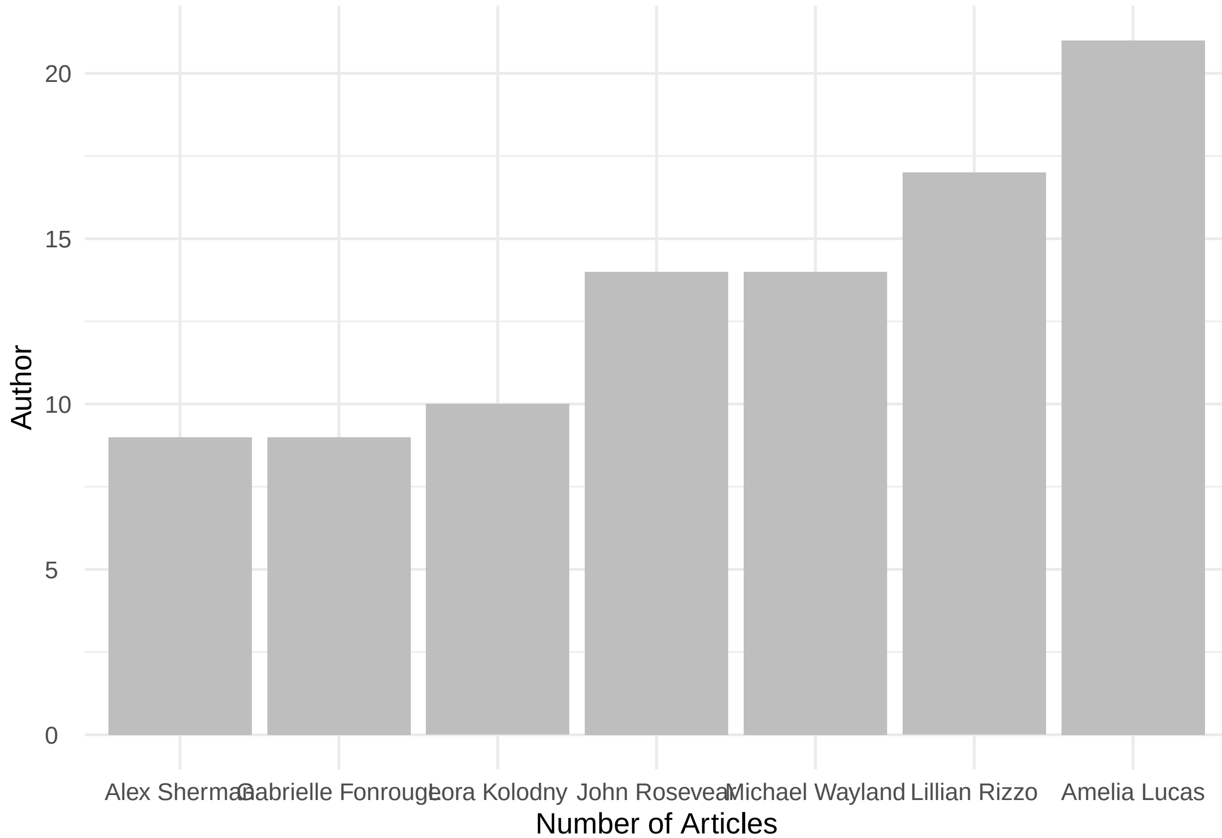
As we see, there are 17 articles missing full_text information. At first, it seemed like something was wrong with the scraping process. So, we checked the scraping process from the beginning. However, once we go to the website and check the missing articles, we realized that these articles were under Pro category, which requires CNBC Pro subscriptions.

The bar chart below shows the top 7 authors with the most articles. The rest of the authors have contributed to this category with less than 5 articles.

```
author_counts = articles_df %>%
  group_by(Author) %>%
  summarize(ArticleCount = n()) %>%
  arrange(ArticleCount)
top_authors_num = 7

top_authors = author_counts %>% top_n(top_authors_num)

ggplot(top_authors, aes(x = reorder(Author, ArticleCount), y = ArticleCount)) +
  geom_bar(stat = "identity", fill = "gray") +
  labs(x = "Number of Articles", y = "Author") +
  theme_minimal() +
  theme(axis.text.y = element_text(hjust = 0))
```



Models and Discovery

Word Cloud

In the context of text analysis, Bag of Word is the most popular way to represent text data. The logic behind it comes from counting the words that most appear in a text. For this method, there are some important tasks that the researcher needs to consider. Firstly, we need to choose the unit of analysis. In our case, we are going to use all articles since we have a limited number of articles. To reduce complexity, we should pay attention to removing punctuation and stop words, applying lowercase, and creating equivalence classes.

Another important thing is to decide on the number of frequencies. (Grimmer, Roberts, & Stewart, pg.49) In this project, the word cloud library is used to visualize the words.

To start with, min. frequency is set to 300 words. This frequency yielded only six words such as “new”, “company”, “said”, “also”, “will” and “year”. Considering the consumer category, these words have not provided meaningful results, since they could be the outcome of any other category, such as finance or technology. So, they are removed from the corpus.

Min. frequency is set to 100 words. This time, company names have started to appear on the word cloud, which is interested in this project related to the research question (most appear brands). These companies are Disney, Tesla, and Ford.

Min. frequency is set to 70 words. This frequency had led to see more companies such as Netflix and Starbucks.

Lastly, when it is set to 40 words, Amazon, Twitter, and Walmart have also appeared. After this point, reducing the frequency of words has not brought different company names.

(In this model, Prof. Posch's code has been used with some adjustments such as removing extra words)

```
# Load necessary packages
library(tidyverse)
library(tm)
library(wordcloud)

# Assuming you have your data frame "articles_df" with a "Full_Text" column

# Combine the Full_Text column into a single string
combined_text <- paste(articles_df$Full_Text, collapse = " ")

# Define custom stop words to remove
custom_stopwords <- c("new", "company", "said", "also", "will", "year", "i", "_")

# Define additional characters to remove
custom_characters <- c("|")

# Define additional patterns to remove
custom_patterns <- c("\"", "'s\\b")

# Remove custom stop words and characters
remove_custom <- function(x) {
  # Remove characters
  for (char in custom_characters) {
    x <- gsub(char, "", x)
  }
  # Remove patterns
  for (pattern in custom_patterns) {
    x <- gsub(pattern, "", x)
```

```r
  }
  return(x)
}

# Apply custom function to remove characters and patterns
combined_text_cleaned <- remove_custom(combined_text)

# Create a corpus from the cleaned text
corpus <- Corpus(VectorSource(combined_text_cleaned))

# Preprocess the corpus
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, removeNumbers)
corpus <- tm_map(corpus, removePunctuation)

# Remove custom stop words
corpus <- tm_map(corpus, removeWords, c(stopwords("en"), custom_stopwords))

corpus <- tm_map(corpus, stripWhitespace)

# Create a term document matrix
tdm <- TermDocumentMatrix(corpus)
matrix <- as.matrix(tdm)

# Get word frequencies
word_freq <- sort(rowSums(matrix), decreasing = TRUE)

# Set the color palette
color_palette <- viridis(length(word_freq))

# Set the seed for reproducible results
set.seed(1)

# Create the word cloud
wordcloud(words = names(word_freq), freq = word_freq,
          min.freq = 70, random.order = FALSE,
          colors = color_palette)
```

In addition to company or brand names, this word cloud represents relevant keywords to the customer category. Such as "customer", "price", "retailers", "media", "restaurant" etc. And it also shows how consumers are connected to businesses and the economy with words such as "stocks", "sales", "earnings", "market" etc.

Topic Modeling

Topic modeling can be defined as one of the clustering algorithms that allow researchers to discover underlying topics in a corpus. The difference is that topic models assign each document to all categories. In this way, topic models give more insight than clustering algorithms. (Grimmer, Roberts, & Stewart, pg.147)

There are different types of algorithms for topic modeling such as LSA (Latent Semantic Analysis (LSA), and NMF (Non-Negative Matrix Factorization), but in this project, Latent Dirichlet Allocation (LDA) is used as a topic model.

The corpus is prepared for topic modeling and then we adjust the number of topics and words to get meaningful insight.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	To 19
cnbc	year	year	million	million	also	quarter	company	sales	now	year	also	“	million	according	can	sales	year	’s
“	also	also	quarter	free	million	last	’s	year	images	’s	can	year	billion	company	quarter	quarter	according	sa
market	million	first	also	sales	can	now	according	company	customers	last	’s	cnbc	also	ceo	“	year	last	pr
companies	billion	free	billion	share	company	business	can	free	share	billion	quarter	market	year	can	sales	customers	—	la
getty	cnbc	can	cnbc	billion	billion	“	—	last	free	—	free	also	can	electric	—	ceo	cnbc	no
sales	images	revenue	year	images	share	first	million	per	year	according	—	business	business	people	according	business	customers	cr
share	according	“	market	ceo	companies	year	ceo	market	according	free	last	tesla	revenue	’s	million	first	including	bu
tesla	last	vehicles	electric	according	quarter	also	share	growth	business	customers	getty	—	—	per	electric	time	’s	el
first	quarter	may	revenue	electric	cnbc	ceo	including	“	disney	ceo	now	can	images	workers	last	companies	million	bi
time	first	according	people	getty	growth	according	per	can	sales	first	ceo	sales	per	business	time	billion	years	re
years	”	reported	companies	quarter	last	cnbc	consumers	reported	last	tesla	”	images	customers	second	stocks	growth	favorite	ac
quarter	like	market	—	told	like	market	price	may	growth	disney	first	per	now	just	revenue	images	follow	fi
billion	can	still	according	business	electric	sales	follow	revenue	can	now	still	companies	reported	stores	like	free	also	ve
last	vehicle	theft	growth	favorite	—	like	sports	earlier	years	revenue	customers	share	years	ford	also	including	per	st
told	including	growth	ceo	reported	vehicles	tesla	also	expected	first	favorite	people	electric	vehicles	“	ceo	stocks	one	ye
ford	production	vehicle	share	may	reported	revenue	retail	ceo	stocks	told	year	retail	sales	last	years	prices	business	co
according	ceo	ford	reported	can	time	may	expected	cnbc	create	sales	including	’re	disney	deal	business	market	price	ea
including	industry	told	disney	disney	still	customers	industry	disney	article	company’s	companies	create	tesla	watch	video	second	can	ju

(In this model, Prof. Posch’s code has been used with some adjustments such as removing extra words - and also help of AI) https://chat.openai.com/share/605ccaed-f685-4f05-b3e6-480b094aba9e

First, the model is run with 40 topics and 20 top words. The topic data frame shows that each topic is similar to each other, and it is hard to differentiate them. Then the number of topics was reduced to 20 and the number of words as well. However, this adjustment also doesn’t change the outcome, still similar.

In this case, the model ends with a situation in which a document can be represented by multiple topics. This is considered a limitation in that the model is unable to predict or explain a document that is more distinct from topics (Grimmer, Roberts, & Stewart, pg. 161).

##Time Period

58 days (2 months) shows a narrow period to use in this analysis. The research question aims to find general brands and categories that shape the Consumer category.

This limitation prevents the application of time analysis. Otherwise, the results cause a bias toward the recent activities of mentioned brands in this project.

Conclusion

For the research question, the word cloud method provides more meaningful results compared to Topic modeling. Word cloud model shows which brands were most present over the last two months in the consumer category. These brands and categories are:

1. Amazon: E-commerce, Retail
2. Starbucks: Food and Beverage, Retail
3. Disney: Entertainment, Media, Tourism
4. Tesla: Electric Vehicles, Technology, Energy
5. Ford: Automotive, Manufacturing
6. Walmart: Retail, E-commerce

In addition to brand names, the word cloud used is also able to catch these categories, which answers the categorical part of the research question. And they are parallel with CNBC’s existing categories. (The blank row represents Pro section, which requires a subscription.

```
x = articles_df %>%
  group_by(Category) %>%
  summarize(count = n()) %>%
  arrange(desc(count))
head(x, 10)
```

```
## # A tibble: 10 x 2
##    Category            count
##    <chr>               <int>
##  1 "AUTOS"                33
##  2 "MEDIA"                33
##  3 ""                     25
##  4 "RETAIL"               22
##  5 "TECH"                 22
##  6 "RESTAURANTS"          19
##  7 "PERSONAL FINANCE"      9
##  8 "FOOD & BEVERAGE"       7
##  9 "SPORTS"                3
## 10 "CNBC DISRUPTOR 50"     2
```

As expected, these companies are leaders in their categories, which leads them to be present in the news coverage every day. There are many reasons that might explain their presence, such as new product launches, quarterly profit releases, new store openings, etc.

However, observing only 6 companies in 200 articles shows their power in the media coverage as well. This power not only shows their successes or failures among the news but also their ability to take the attention of society.

References

https://link.springer.com/article/10.1007/s11002-012-9191-4

Grimmer, J., Roberts, M. E., & Stewart, B. (2022). Text as data: A new framework for Machine Learning and the Social Sciences. Princeton University Press.

- https://www.theverge.com/2023/8/9/23826342/cnet-content-pruning-deleting-articles-google-seo