Riley Finley

Falmouth, Maine

**Email**: Finley.r@northeastern.edu

**Phone:** (207)807-5529

**Social scraping to measure brand engagement**

For Garrand Moehlenkamp

TABLE OF CONTENTS

TABLE OF CONTENTS……………………………………………………………………....….2 STATEMENT OF PURPOSE…………………………………………………………………….3

BACKGROUND………………………………………………………………………………….3

PROJECT SCOPE………………………………………………………………………………...3

OBJECTIVES…………………………………………………………………………………..3

METHODOLOGY………………………………………………………………………………..5

RESEARCH AND LITERATURE……………………………………………………………….6

IMPLEMENTATION……………………………………………………………………….…….8

SAMPLE PROJECT………………………………………………………………………………9

CONCLUSION…………………………………………………………………………………..10

REFERENCES…………………………………………………………………………………..12

APPENDIX……..………………………………………………………………………………..13

WEB SCRAPING STEPS.………………………………………………………………....….13

TEXT MINING STEPS…………………………………………………………………...…..14

R CODE WORDLCOUD……………………………………………………………………..16

STATEMENT OF PURPOSE

While market research is typically carried out through surveys and interviews, bias and limited responses may prevent accurate findings. Web scraping is a powerful tool in analytics to learn about a market in the most raw and unfiltered form. Scraping social media may be a great resource to increase consumer data. With the increase in social media usage and its advancements, there is a great opportunity on these platforms to learn more about consumers as they interact with brands, and other users, and share personal information with thousands to millions of people. Through research and analysis, consumer interactivity on social media will be explored and utilized to make informed business decisions, as well as to measure brand awareness and brand success. By exploring the various ways social media users interact with each other and brands, products, and services, there is a great potential for new and exceptional success using social media to create brand awareness, consumer-brand relationships, and greater purchase decisions. Providing replicable web scraping and text mining techniques will allow Garrand Moehlenkamp to use social media as a source of consumer insight. Applying A/B testing to the analysis will also be useful to compare engagement with brands and products/services before and after applying different marketing tactics to measure its effectiveness on consumers.

BACKGROUND

Social media provides a space for users to share information and thoughts in online communities. (Tufts, 2021). With its extensive growth over the past few years, there are now close to 4.5 billion users. There are many opportunities within social media to reach thousands of people instantly, making it an excellent tool for not online advertisement but to also learning about different markets. To make decisions regarding social media marketing content, one must first understand what consumers are saying. As consumers are now the center of marketing, they have the power to drive brand decisions and success through how they communicate and interact with each other and brands online. A major advantage of social media is its ability to generate more data, allowing for more market research that can be used to understand consumers. Learning about their engagement with brands, products and services provides actionable information to change or improve upon these conversations online, which may subsequently impact brand outcomes.

PROJECT SCOPE

OBJECTIVES

* ***Understand what people are saying about brands online***
* ***Analyze the interactions on social media sites***
* ***Track online activity after creating new brand content***

The proposed project will analyze consumer behavior and conversations on social media through web scraping and text mining. Social media data will be gathered through the web scraping process performed in RStudio. Data will need to be cleaned in R as well before text mining is conducted. Text mining will also be performed in R, where word clouds and/or frequency visuals can be created. A/B testing may also be conducted to analyze how the implemented marketing tactics impacted consumer thoughts and attitudes about a brand and product/service on social media. This will entail two periods of web scraping to compare and contrast the different conversations and interactions taking place online before and after. Client data may be collected through surveys and used in conjunction with web scraping and text mining. Survey data should be cleaned through any preferred method and can be analyzed through methods of choice. Using survey data along with social media data, GM can further understand the consumer narrative that at times is not fully investigated in market research.

Shape, arrow

Description automatically generated

METHODOLOGY

The following techniques will be used together to gather and understand consumer insights from a social media view and will be used in conjunction with literature that reflects upon current brand awareness, interactivity, and trends, with a focus on this all within the social media realm. Together these methods will explore the use of social media in marketing and how to understand consumers through a new channel.

**Web scraping**

Web scraping extracts data from a website automatically by browsing sites and copying the information that is programmed to be collected. The user can program the code to collect and store certain features of the website, such as only comments, or only posts. Web scraping saves time as this work would otherwise be done manually and can collect a greater volume of data very quickly. Web scraping does require the use of coding, however, there may be some non-coder tools and sites that can help. Another issue is that website structure may change frequently, which may mess up one’s data.

**Text mining**

Text mining consists of transforming unstructured text into a structured format where patterns within the text can be found and visualized, providing insight into what is being said on a site. Text mining is also done through coding in R and may require additional cleaning to provide the most sensible outcomes.

**R**

RStudio will be a necessary software program to use R, a programming language for statistical computing and analysis, which is necessary for both web scraping and text mining. The code to extract and present the data will be run here. R does tend to have a steep learning curve; thus, prior experience will aid in this analysis. R supports various data types and can perform data cleaning, wrangling, and web scraping all in the same location.

**A/B testing**

A/B testing is incredibly useful in comparing two versions of a variable to track progress and changes. This is achieved by collecting data in two different segments and comparing the findings. Regarding social media marketing, A/B testing can be utilized by comparing social media marketing tactics and social media interactivity now to new social media marketing tactics and the interactions that occur online after implementation. A/B testing may take longer to get results, so adequate planning is necessary to incorporate it.

**Icon

Description automatically generated**

**Shape

Description automatically generated with medium confidence**

RESEARCH & LITERATURE

Social media consists of a significant amount of content, users, and interactions. Having an appropriate alignment of marketing content with user needs can be incredibly powerful to improve brand outcomes. However, one must understand how individuals use and perceive social media in general, and more specifically, the role brands play within this realm (Zhu & Chen, 2015).

Multiple frameworks exist that address consumer behavior, several of which are summarized below:

|  |  |
| --- | --- |
| **The hierarchy of effects model** | Suggests an order in which consumers perceive, process, and use marketing communication.   * Cognitive Phase: awareness and knowledge of a product * Affective stage: develops feelings towards the product (positive or negative) * Cognitive stage: accepts or rejects the product   (Dennhardt, Füller, Hautz & Hutter, 2013) |
| **Brand awareness** | Consumers tend to purchase based on how familiar and well-known the brand is to them. This can influence the strength of the brand association with the brand image, which impacts the strength of how one attaches information to a brand.   * Social media allows increase exposure to a brand with active use, such as through high brand page commitment. * Falls within the cognitive phase of the HOE model   (Dennhardt, Füller, Hautz & Hutter, 2013) |
| **Word of mouth** | Includes any kind of communication about a company, brand, or product/service. This increases familiarity and makes the information more reliable, credible, and trustworthy. WOM has a strong influence on the cognitive and affective stages of the HOE model.   * Social media presents an opportunity to conversate about and expose brands and products quickly and reach millions instantly.   (Dennhardt, Füller, Hautz & Hutter, 2013) |
| **Purchase intention** | Refers to the decision to purchase, which is done in the affective stage of the HOE model.  (Dennhardt, Füller, Hautz & Hutter, 2013) |
| **Listening and discovery** | Seeking information about what is going on in the social media pace around one’s market and then drawing valuable insights from that |
| **Social search** | filtering information from online sources to pull out brand mentions or content that pertains to the brand, to gain insights that will help with major business decisions   * Access to opinions of their target audience * Examine the competitive landscape and identify market opportunities * Discover trends, customer needs, opportunities, positive or negative information about brands, and the health of your relationships with customers * Can reflect what is learned from listening in to content and communications   (Perkins, 2015) |
| **Category trends** | Patterns happening within a given market |
| **Consumer trends** | Patterns of behavior   * 53% of people use social media multiple times per month to compliment and give credit to brands that do a great job. * 46% of people use social media multiple times per month to express concerns or complain about brands or products.   (Perkins, 2015) |
| **Brand experience** | Describes the interactions between a customer and an organization in a personal and memorable way. Some of the attributes that make up this experience include sensory, affective, intellectual, and behavioral.  (Japutra & Molinillo, 2019) |
| **Brand personality** | Refers to human personality traits and characteristics that are associated with a brand, which include five dimensions: sincerity, excitement, competence, sophistication, and ruggedness.  (Japutra & Molinillo, 2019) |
| **Consumer brand relationships** | Formed between a consumer and a brand and the most important aspects of this are satisfaction, trust, and loyalty.   * Both responsible and active brand personalities positively affect satisfaction, trust, and brand loyalty * Active brand personality is a better predictor of satisfaction and brand loyalty, while responsible brand personality is a better predictor of trust.   (Japutra & Molinillo, 2019) |

IMPLEMENTATION

The first step to learning about the audiences on social media is to gather data from these various platforms. Web scraping is a great tool to gather data from these online sites, such as Facebook or Twitter. This can be done in R Studio using multiple packages, which include Swirl, Pacman, rvest, Stringr, Httr, and Pander. Once all packages are installed and loaded, page reading times must be generated which will determine how long the data will be collected from the site. Following that, variables must be created to store the data that will be gathered. Next, an HTML session will need to be opened, which allows for a user to simulate a page interacting with the website. Chrome will need to be utilized as the selector gadget tool will be useful. Opening the site that is intended to be scraped in chrome, will allow the selection of specific page attributes through the selector gadget. For example, this will easily allow the selection of all comments, or all posts exclusively. Each component can be stored in its own vector in R. Once all is completed, each vector created can be combined to create a final data frame of all the data, which can then be stored as a csv file.

Once web scraping is completed, text mining can begin. Text mining will also be executed in R Studio using the packages, tm, snowball, wordcloud, and RColorBrewer. Once packages and installed and loaded, the data must be read into R and then changed to a corpus, which will help with normalizing text to search for occurrences. Next, the text will need to be transformed which will clean up the data removing any special characters such as !, @, or / symbols. Additional cleaning may take place such as changing all text to lower case and removing numbers or specific words. Following clean-up, the data will be transformed into a matrix that describes the frequency of terms that occur in a collection of documents. Finally, a word cloud can be generated, as well as additional exploration such as frequency charts and additional visuals

(STHDA).

Based on the findings of the interactions online from web scraping and text mining, social media content can be created to change, improve, and/or increase the conversations surrounding specific brands and overall improve business outcomes. After some time has passed, the web scraping and text mining process may be repeated periodically to track the progress of the implemented content to measure if social media conversations and interactions are changing and if so, how. This may be done every couple of weeks or a couple of months depending on the project timeline.

SAMPLE PROJECT

Web scraping was performed on the top 100 Facebook posts on MaidPro residential cleaning service. The social media feature looked at in this specific web scrape included the post text. In the cleaning process, symbols (%, ?) were removed, the text was changed to lowercase, only English words were included, and all numbers and punctuation were removed as well. This allowed for a clear and accurate output of word frequencies.

Text

Description automatically generatedThe final word cloud output is provided below:

Figure 1.

From this word cloud it is evident that “maidpro” and “cleaning” are the most frequently used words in the 100 Facebook posts. “Clean”, “home”, “time”, “work”, “happy”, “service”, and “great” were also popular word choices as well. The Facebook accounts in this particular file were all MaidPro accounts and thus this would not tell much about the consumers, but rather the MaidPro company itself. This may be useful for learning about competition in the residential cleaning service industry. Knowing the buzz words and posts that others are using in their marketing tactics, can provide some insight into how to follow popular trends but also how to differentiate a company as well.

Using this information along with survey data, as shown below in Figure 2, may be the most advantageous. Figure 2 is pulled from MaidPro’s interior sure data, which highlights that many respondents use Facebook either frequently each day or once a day, and thus Facebook may be a great platform to market MaidPro’s residential cleaning services.

Chart, bar chart

Description automatically generated

Figure 2.

Additional web scraping and survey data analysis may need to be performed to determine the ideal messages and marketing tactics to implement. Web scraping posts not posted by MaidPro accounts will also be essential to understand the consumer's perspective and interactions surrounding this brand and its services on this platform.

CONCLUSION

With social media increasingly infiltrating the business space, Garrand Moehlenkamp has the opportunity to explore and utilize a new channel of information that can highlight brand awareness, engagement, attitudes, and purchasing decisions. Social media provides unfiltered, raw insights with limited restrictions that multiple-choice and ranking questions tend to have. Having a consumer narrative goes beyond simple answers and can be greatly useful for implementing social media marketing in the most optimal ways and receiving favorable outcomes. Execution of this through web scraping and texting mining tools provides a simple way to measure social media conversations and interactions, as well as how these may change in response to advertising.

Web scraping is one technique to carry out a social search where information online is filtered to pull out specific content. Not only does this provide insight into opinions and customer needs, but also highlights potential competition and opportunities within the market. One specific aspect of social media content that surveys and interviews can’t explore as well as brand awareness and conversation through word of mouth. Under a social search, word-of-mouth content can be examined by looking at what social media users converse about regarding certain brands and products/services. Word of mouth includes honest opinions and feelings which can provide more accurate, detailed data. Through this, brand experience may be revealed as intellectual and behavioral experiences with a brand may be portrayed on social media sites through posts and comments. Based on brand awareness, experience, and all information explored through a social search, consumer-brand relationships can be evaluated, which mainly consist of three important aspects, satisfaction, trust, and loyalty. Using the data from social media, informed business decisions can be made on how to proceed with the use of social media within marketing to improve relationships, awareness, and ultimately consumer satisfaction and business outcomes.

REFERENCES

Bowditch, W. (2017, January 27). *Social Media Research using R*. Retrieved June 15, 2022, from https://willbowditch.github.io/scraping\_with\_R/

*Business templates, tools, and calculators*. Wise. (n.d.). Retrieved May 21, 2022, from https://wise.com/us/business-templates/

Hutter, K., Hautz, J., Dennhardt, S., & Füller, J. (2013, August 19). *The impact of user interactions in social media on brand awareness and purchase intention: The case of mini on Facebook*. Journal of Product & Brand Management. Retrieved May 19, 2022, from https://www.emerald.com/insight/content/doi/10.1108/JPBM-05-2013-0299/full/html

Japutra, & Molinillo, S. (2019). Responsible and active brand personality: On the relationships

with brand experience and key relationship constructs. *Journal of Business Research*, *99*,

464–471. https://doi.org/10.1016/j.jbusres.2017.08.027

Louth, J. D. (2021, March 31). *The changing face of marketing*. McKinsey & Company. Retrieved May 21, 2022, from https://www.mckinsey.com/business-functions/growth-marketing-and-sales/our-insights/the-changing-face-of-marketing

Panel®, E. (2022, April 14). *Council post: 15 top marketing trends to keep an eye on in 2021*. Forbes. Retrieved May 21, 2022, from https://www.forbes.com/sites/forbesagencycouncil/2021/01/27/15-top-marketing-trends-to-keep-an-eye-on-in-2021/?sh=5503af312a5e

Perkins, L. (2015). Listening and Discovery. In: The Community Manager’s Playbook. Apress,

Berkeley, CA. <https://doi-org.ezproxy.neu.edu/10.1007/978-1-4302-4996-2_7>

STHDA . (n.d.). *Text mining and word cloud fundamentals in R : 5 simple steps you should know*. STHDA. Retrieved June 15, 2022, from http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know/

*The evolution of social media: How did it begin and where could it go next?* Maryville Online. (2021, March 3). Retrieved May 21, 2022, from https://online.maryville.edu/blog/evolution-social-media/

Tufts University Relations. (2021, March 5). *Social Media Overview*. Communications. Retrieved June 18, 2022, from https://communications.tufts.edu/marketing-and-branding/social-media-overview/

UMaine. (2021). *Social Media Statistics Details - undiscovered Maine - University of Maine*. Undiscovered Maine. Retrieved May 21, 2022, from https://umaine.edu/undiscoveredmaine/small-business/resources/marketing-for-small-business/social-media-tools/social-media-statistics-details/

Zhu, Y.-Q., & Chen, H.-G. (2015). *WWW-sciencedirect-com.ezproxy.neu.edu*. Retrieved June 2, 2022, from https://www-sciencedirect-com.ezproxy.neu.edu/science/article/pii/S0007681316300866

APPENDIX

The following includes step-by-step instructions to conduct web scraping and text mining in RStudio as well as a real example of R code that was used to create the word cloud in Figure 1.

**Web scraping**

**Step 1:** **Install and load all necessary packages**. Packages rvest, stringr, httr, and pander will need to be loaded under the pacman library

Pacman: : p\_load (rvest, stirngr, httr, pander)

**Step 2: Generate page reading times** using rnorm. rnorm generates random numbers using a normal distribution.

rnorm(n, mean, sd)

Where n= the sample size

Mean: mean value of sample data (default is 0)

Sd: standard deviation (default is 1)

**Step 3: Create variables** to store data

A<- data.frame()

B<- data.frame()

Ostring<-

**Step 4: Open html session** using html\_session function

html.session<-html\_session (“URL”, user\_agent(Ostring))

\*user\_agent: this will make the session appear like an actual user and helps the site operate normally when scraping

**Step 5:** **Use the selector gadget** in chrome on the site being scraped

This will provide a tag to gather all comments, for example, that will be used in the html nodes function in step 6

**Step 6: Grab the information desired using the** html\_nodes function

New\_vector<-read\_html(html.session) %>%

Html\_nodes(“tag from selector gadget”) %>%

Html\_text()

**Step 7: Create a data frame** with all information gathered in step 6

Final\_data<-(New\_vector….)

**Step 8: Save as a csv file** using write.csv

Write.csv (Final\_data, “what you want to name the csv file”)

(Bowditch, 2017)

**Text mining**

**Required packages**

* tm
* snowball
* **word cloud**
* RColorBrewer

**Step 1: Install and load all necessary packages**

**Step 2: Load the data file into R** using readLines()

Text<-readLines(file.choose(“File name”)

**Step 3: Load the data as a corpus** using corpus()

Text<-Corpus(VectorSource(text))

**Step 4: Transform text** to get rid of any special characters using content\_transformer and tm\_map()

to\_space<- content\_transformer(function (x, pattern) gsub(pattern, “ “, x)

text<-tm\_map(text, to\_space, “special charcter”)

where the exact symbol is used in the place of a special character

\*This will change special characters to a space

\

**Step 5: Clean the text** using tm\_map

Tm\_map (text, cleaning function)

\*Where cleaning function could be

Content\_Transformer(tolower))= all lower case

removeNumbers= removes numbers

removeWords, c(“xxx”, “yyy”)= remove specific words

removePunctuation= removes punctuations

stripWhitespace= eliminates extra white space

**step 6: Build a matrix** using Term\_documentmatrix() and as.matrix()

text<-TermDocumentMatrix(text)

text\_matrix<-as.matrix(text)

text\_sort<-sort(rowSum(text\_matric), decreasing =TRUE)

text\_data<-data.frame(word=names(text\_sort), freq= text\_sort)

**Step 7: Generate word cloud** using word cloud

Set.seed(1234)

Wordcloud(words=text\_data$word, freq=text\_data$freq, min.freq=q, max.words=200, random.order=FALSE, rot.per=.35, colors=brewer.pal(8, “Dark2”))

**Step 8: Further exploration**

* Words that appear x number of times
  + findFreqTerms(text\_data, lowfreq=x)
* associated words
  + findAssocs(text\_data, terms=”Word”, corlimit=.4)

**R Code to create a word cloud:**

#Read the file into R

```{r}

facebook<-read.csv("MaidPro\_FB.csv")

#only interested in the text of the posts

facebook<-as.data.frame(facebook$post\_text)

#saving text posts as its own file

write.csv(facebook, 'facebook')

```

```{r}

#install.packages("tm")

#update.packages("tm", checkBuilt = TRUE)

#install.packages("SnowballC")

#install.packages("wordcloud")

#install.packages("RColorBrewer")

library(tm)

library(wordcloud)

library(RColorBrewer)

```

```{r}

#read new file into R

text<-readLines(file.choose("facebook"))

#change to a unicode

text2 <- iconv(text, to ="utf-8")

```

```{r}

#normalize text

text2<-Corpus(VectorSource(text2))

```

```{r}

#clean up the post text

to\_space<- content\_transformer(function (x, pattern ) gsub(pattern, "", x))

text3<-tm\_map(text2, to\_space, "!")

text3<-tm\_map(text3, to\_space, "@")

text3<-tm\_map(text3, to\_space, "-")

text3<-tm\_map(text3, to\_space, "?")

```

```{r}

#additional cleaning of post text

text4<-tm\_map(text3, content\_transformer(tolower))

text4<-tm\_map(text4, removePunctuation)

text4<-tm\_map(text4, removeNumbers)

text4 <- tm\_map(text4, removeWords, stopwords("english"))

text4

```

```{r}

#creating a dataframe of the posts

text4<-TermDocumentMatrix(text4)

text\_matrix<-as.matrix(text4)

text\_sort<-sort(rowSums(text\_matrix))

text\_data<-data.frame(word=names(text\_sort), freq= text\_sort)

text\_data

```

```{r}

#install.packages("RCurl")

#install.packages("XML")

library(RCurl)

library(XML)

library(wordcloud)

```

```{r}

#create a word cloud

set.seed(1234)

wordcloud(words=text\_data$word, freq=text\_data$freq, max.words = 100, random.order = FALSE, colors = brewer.pal(5,"Greens"))

```