

Aqila Habib (#47162)

ITC 336 Data Mining

Project:

Text Mining

Dr.Asadullah Jawid

Date: 12.12.2024

Semester: Fall/2024

Text Mining Project

1. Select a company with Online customer reviews (it can also be a single product on Amazon, for instance)
2. Define your problem
3. copy the text reviews and the rating of at least 30 customers
4. pre-process your data
5. Extract features
6. Perform a sentiment analysis
7. Make conclusions.

Introduction

Product and Company: CeraVe is a well-known skincare brand that offers a range of products formulated with essential ceramides and other beneficial ingredients. One of its flagship products is the CeraVe Hydrating Facial Cleanser, designed to cleanse the skin without disrupting its natural protective barrier. This product is particularly popular among

individuals with dry or sensitive skin, as it aims to provide hydration while effectively removing dirt and makeup. I have collected reviews from Amazon.

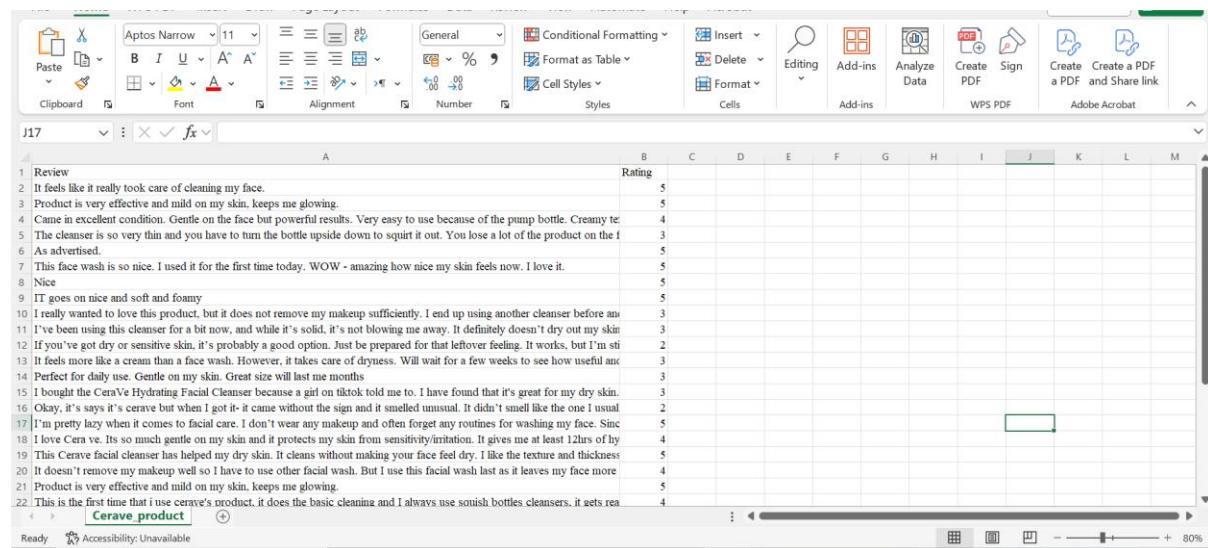
Objectives:

The objective of this project is to analyze customer reviews of the CeraVe Hydrating Facial Cleanser to determine overall sentiment and identify common themes in customer feedback. Specifically, we aim to:

1. Assess the overall sentiment (positive or negative) towards the product.
2. Identify frequently mentioned themes and concerns in the reviews.
3. Provide actionable insights for potential product improvements.

Data:

The dataset consists of 30 customer reviews collected from an online platform, along with corresponding ratings.



A screenshot of a Microsoft Excel spreadsheet titled "Cerave_product". The spreadsheet contains two columns: "Review" (A1:A30) and "Rating" (B1:B30). The reviews are listed in column A, and the corresponding rating for each review is listed in column B. The reviews are numbered from 1 to 22, followed by a blank row, and then numbered again from 1 to 18. The ratings range from 2 to 5. The Excel ribbon at the top shows various tabs like Home, Insert, Page Layout, etc., and the formula bar shows "J17". The status bar at the bottom indicates "Ready" and "Accessibility: Unavailable".

Review	Rating
1 Review	
2 It feels like it really took care of cleaning my face.	5
3 Product is very effective and mild on my skin. keeps me glowing.	5
4 Came in excellent condition. Gentle on the face but powerful results. Very easy to use because of the pump bottle. Creamy texture.	4
5 The cleanser is so very thin and you have to turn the bottle upside down to squirt it out. You lose a lot of the product on the tip.	3
6 As advertised.	5
7 This face wash is so nice. I used it for the first time today. WOW - amazing how nice my skin feels now. I love it.	5
8 Nice	5
9 It goes on nice and soft and foamy	5
10 I really wanted to love this product, but it does not remove my makeup sufficiently. I end up using another cleanser before and after.	3
11 I've been using this cleanser for a bit now, and while it's solid, it's not blowing me away. It definitely doesn't dry out my skin.	3
12 If you've got dry or sensitive skin, it's probably a good option. Just be prepared for that leftover feeling. It works, but I'm still not a fan.	2
13 It feels more like a cream than a face wash. However, it takes care of dryness. Will wait for a few weeks to see how useful and effective it is.	3
14 Perfect for daily use. Gentle on my skin. Great size will last me months!	3
15 I bought the Cerave Hydrating Facial Cleanser because a girl on tiktok told me to. I have found that it's great for my dry skin.	3
16 Okay, it's says it's cerave but when I got it, it came without a sign and it smelled unusual. Didn't smell like the one I usually buy.	2
17 I'm pretty lazy when it comes to facial care. I don't wear any makeup and often forget any routines for washing my face. Since I've been using this, I've noticed a difference in my skin.	5
18 I love Cerave. Its so much gentle on my skin and it protects my skin from sensitivity/irritation. It gives me at least 12hrs of hydration.	4
19 This Cerave facial cleanser has helped my dry skin. It cleans without making your face feel dry. I like the texture and thickness.	5
20 It doesn't remove my makeup well so I have to use other facial wash. But I use this facial wash last as it leaves my face more hydrated.	4
21 Product is very effective and mild on my skin, keeps me glowing.	5
22 This is the first time that I use cerave's product. it does the basic cleaning and I always use snush bottles cleansers. it gets the job done.	4

Approach:

1. **Data Preprocessing:** The first step is to clean the dataset by the preprocessing method. In this method, we removed stop words, punctuations, and numbers, and converted texts to lowercase, and stemming was applied to standardize word forms.
2. **Feature Extraction:** A Document-Term Matrix (DTM) was created to convert the text data into a numerical format suitable for analysis.
3. **Sentiment Analysis:** Customer ratings were used to classify reviews as positive or negative. The analysis focused on identifying sentiment trends and common themes.
4. **Visualization:** Graphical representations of the data were created using the ggplot2 package to illustrate key findings.

Results:

Basic statistics:

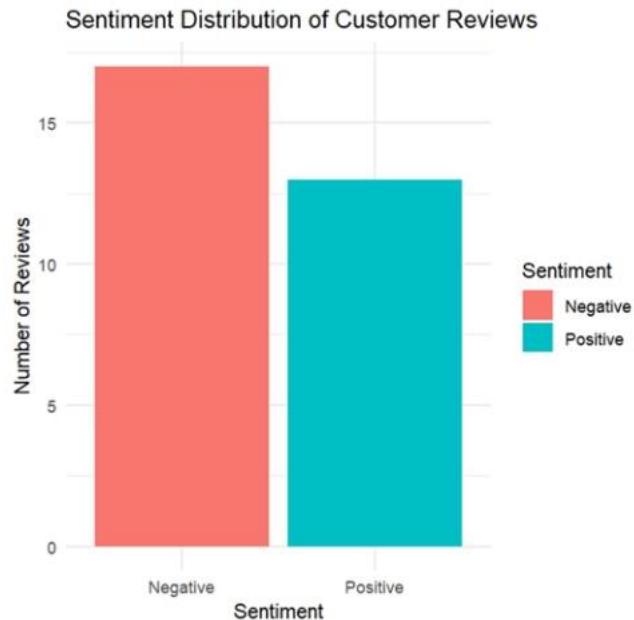
Total Review: 30

Positive Reviews: 13 (43%)

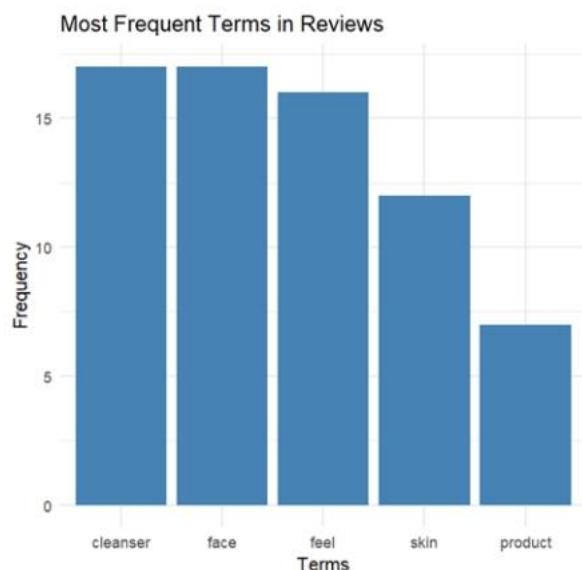
Negative Reviews: 17 (57%)

Graphs with comments (try to plot graphs in ggplot2)

I have used a bar plot from ggplot2 to illustrate the sentiment analysis of reviews based on the negative and positive reviews of customers.



Interpret: The bar plot illustrates the distribution of sentiments among customer reviews, highlighting negative feedback.



Interpret: This bar plot visualizes the frequency of specific terms mentioned in customer reviews for a product. The x-axis displays the terms ("face", "feel", "product", "skin", "cleanser"), while the y-axis shows the corresponding frequency of each term. The bars are colored in steel blue, and the plot is titled "Most Frequent Terms in Reviews". This

visualization allows for a quick assessment of which terms are most commonly referenced by customers, highlighting key aspects of their feedback and potentially guiding further analysis or product improvements.

pre-process your data

```
> setwd("C:\\\\Users\\\\Habibullah\\\\Desktop\\\\Data Mining")
> x = read.csv("C:\\\\Users\\\\Habibullah\\\\Desktop\\\\Data
Mining\\\\cerave_product.csv", stringsAsFactors = FALSE)
> library(qdap)
> fr=freq_terms(x)
> fr
> plot(fr)
```

Word	Count
it	55
i	50
the	48
and	45
my	32
a	28
is	26
of	24
this	23
to	21
face	18
on	17
skin	17
use	16
for	14
but	12
product	11
that	11
not	10
so	10

```
> ##Create documents
> ibiscr=Corpus(VectorSource(x$Review))
> ibiscr[[11]]$content
[1] "If you<d5>ve got dry or sensitive skin, it<d5>s probably a good
option. Just be prepared for that leftover feeling. It works, but I<d5>m
still not 100% sold on the finish it leaves behind. Not bad overall, just
not amazing for me."
> #Preprocessing
> #1. trun all words to lower case
> ibiscr= ibiscr %>%
+   tm_map(tolower)
> ibiscr[[11]]$content
[1] "if you<d5>ve got dry or sensitive skin, it<d5>s probably a good
option. just be prepared for that leftover feeling. it works, but i<d5>m
still not 100% sold on the finish it leaves behind. not bad overall, just
not amazing for me."
```

Note: This part of the code applies a transformation to the entire corpus, converting all text to lowercase. The <d5> in the text appears to be a placeholder or encoding issue.

```
> #2. Remove punctuations
> ibiscr = ibiscr %>%
+   tm_map(removePunctuation)
> ibiscr[[11]]$content
[1] "if youd5ve got dry or sensitive skin itd5s probably a good option
just be prepared for that leftover feeling it works but id5m still not 100
sold on the finish it leaves behind not bad overall just not amazing for
me"

> stopwords("english")
[1] "i"          "me"         "my"        "myself"      "we"
[6] "our"        "ours"       "ourselves" "you"        "your"
[11] "yours"      "yourself"    "yourselves" "he"         "him"
[16] "his"        "himself"    "she"       "her"        "hers"
[21] "herself"    "it"         "its"       "itself"     "they"
[26] "them"       "their"     "theirs"    "themselves" "what"
[31] "which"      "who"        "whom"     "this"       "that"
```

```

[36] "these"      "those"       "am"          "is"          "are"
[41] "was"        "were"        "be"          "been"        "being"
[46] "have"        "has"         "had"         "having"      "do"
[51] "does"        "did"         "doing"       "would"       "should"
[56] "could"       "ought"       "i'm"         "you're"      "he's"
[61] "she's"       "it's"        "we're"       "they're"     "i've"
[66] "you've"      "we've"       "we've"       "they've"     "you'd"
[71] "he'd"        "she'd"       "he'd"        "she'd"       "i'll"
[76] "you'll"      "he'll"       "she'll"      "we'll"       "they'll"
[81] "isn't"        "aren't"      "wasn't"      "weren't"     "hasn't"
[86] "haven't"     "hadn't"     "doesn't"    "don't"       "didn't"
[91] "won't"        "wouldn't"   "shan't"      "shouldn't"   "can't"
[96] "cannot"      "couldn't"   "mustn't"     "let's"       "that's"
[101] "who's"        "what's"      "here's"      "there's"     "when's"
[106] "where's"     "why's"       "how's"       "a"           "an"
[111] "the"          "and"         "but"         "if"          "or"
[116] "because"     "as"          "until"       "while"       "of"
[121] "at"           "by"          "for"         "with"        "about"
[126] "against"     "between"    "into"        "through"     "during"
[131] "before"       "after"       "above"       "below"       "to"
[136] "from"         "up"          "down"        "in"          "out"
[141] "on"           "off"         "over"        "under"       "again"
[146] "further"      "then"        "once"        "here"        "there"
[151] "when"         "where"       "why"         "how"         "all"
[156] "any"          "both"        "each"        "few"          "more"
[161] "most"         "other"       "some"        "such"        "no"
[166] "nor"          "not"         "only"        "own"         "same"
[171] "so"           "than"        "too"         "very"        ""

> ibiscr = ibiscr %>%
+   tm_map(removewords, c("for", "use", "not", stopwords("english")))
> ibiscr[[11]]$content
[1] " youd5ve got dry sensitive skin itds probably good option just
prepared leftover feeling works idm still 100 sold finish leaves
behind bad overall just amazing "

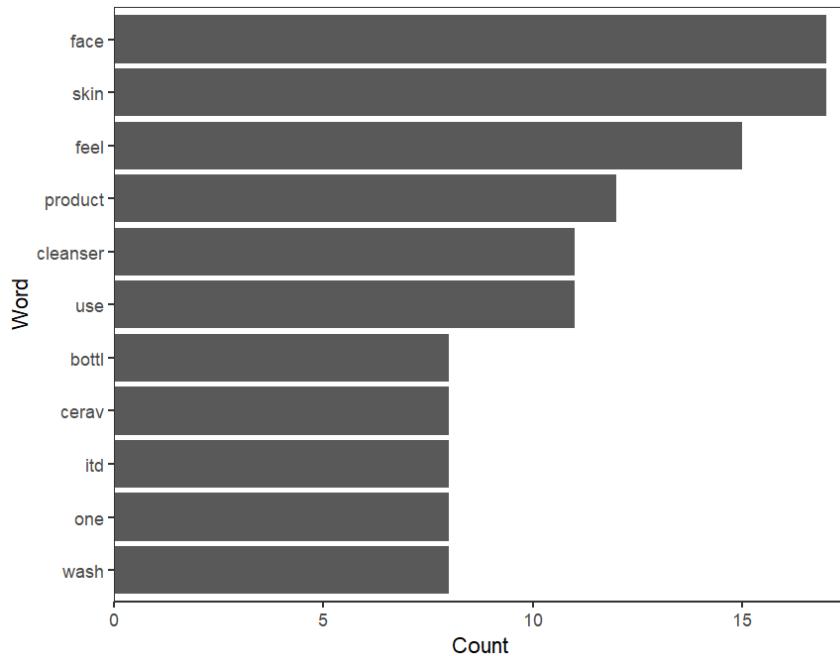
> #4. Remove numbers
> ibiscr = ibiscr %>%
+   tm_map(removeNumbers)
> ibiscr[[11]]$content
[1] " youdve got dry sensitive skin itds probably good option just
prepared leftover feeling works idm still sold finish leaves
behind bad overall just amazing "

> #5. stripping the white space
> ibiscr= ibiscr %>%
+   tm_map(stripwhitespace)
> ibiscr[[11]]$content
[1] " youdve got dry sensitive skin itds probably good option just
prepared leftover feeling works idm still sold finish leaves behind bad
overall just amazing "

> #6. Stemming
> ibiscr = ibiscr %>%
+   tm_map(stemDocument)
> ibiscr[[11]]$content
[1] "youdv got dri sensit skin itd probabl good option just prepar leftov
feel work idm still sold finish leav behind bad overal just amaz"

> fr1=freq_terms(ibiscr, 10)  #5 most frequent words
> plot(fr1)

```



1. Extract features

Most frequent terms mentioned in the reviews are "face," "feel," "product," "skin," "cleanser," and "wash." These terms highlight key aspects of the product that customers focus on, such as its effectiveness and texture.

```
> ibisfreq=DocumentTermMatrix(ibiscr)
>
> dim(ibisfreq)      #39 rows and 289 variables
[1] 30 293
>
> inspect(ibisfreq)
<<DocumentTermMatrix (documents: 30, terms: 293)>>
Non-/sparse entries: 542/8248
Sparsity           : 94%
Maximal term length: 16
Weighting          : term frequency (tf)
Sample             :
Terms
Docs bottl cleanser face feel itd one product skin use wash
10    0       1   0   1   2   0       0   1   1   0
11    0       0   0   1   1   0       0   1   0   0
14    0       1   1   3   1   0       0   3   0   0
15    0       0   0   0   2   1       0   0   0   0
16    0       0   3   3   0   0       2   1   1   1
17    0       0   1   0   0   0       0   2   1   0
18    0       1   1   2   0   0       0   1   0   0
22    0       0   2   1   2   2       0   3   0   2
26    2       2   0   0   0   2       0   0   0   0
28    2       1   2   0   0   1       4   0   3   0
>
> inspect(ibisfreq)[1:4, 1:3]  #document 1:4 columns 1:3
<<DocumentTermMatrix (documents: 30, terms: 293)>>
Non-/sparse entries: 542/8248
Sparsity           : 94%
Maximal term length: 16
Weighting          : term frequency (tf)
Sample             :
Terms
Docs bottl cleanser face feel itd one product skin use wash
10    0       1   0   1   2   0       0   1   1   0
11    0       0   0   1   1   0       0   1   0   0
```

14	0	1	1	3	1	0	0	3	0	0
15	0	0	0	0	2	1	0	0	0	0
16	0	0	3	3	0	0	2	1	1	1
17	0	0	1	0	0	0	0	2	1	0
18	0	1	1	2	0	0	0	1	0	0
22	0	0	2	1	2	2	0	3	0	2
26	2	2	0	0	0	2	0	0	0	0
28	2	1	2	0	0	1	4	0	3	0

Docs	bottl	cleanser	face
10	0	1	0
11	0	0	0
14	0	1	1
15	0	0	0

```
>
> findFreqTerms(ibisfreq) #list of words that are present in the corpus
[1] "care"           "clean"          "face"           "feel"
"like"
[6] "reallii"
"keep"
[11] "mild"
"came"
[16] "colour"
"excel"
[21] "gentl"
"textur"
[26] "cleanser"
"squirt"
[31] "thin"
"amaz"
[36] "first"
"time"
[41] "today"
"foami"
[46] "goe"
"makeup"
[51] "one"
"away"
[56] "bit"
"doesndt"
[61] "dri"
"leav"
[66] "lotionlik"
"rins"
[71] "solid"
"bad"
[76] "behind"
"idm"
[81] "just"
"prepar"
[86] "probabl"
"work"
[91] "youdv"
"see"
[96] "take"
"daili"
[101] "great"
"size"
[106] "advers"
"bought"
[111] "can"
"color"
[116] "contain"
"fan"
[121] "found"
"ingredi"
[126] "moistur"
"repeat"

Terms
```

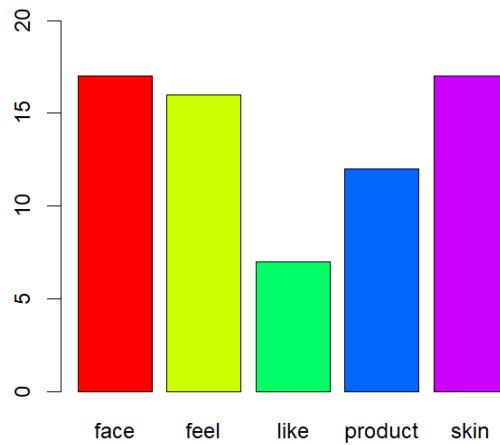
[131] "smooth"	"strip"	"tight"	"tiktok"
"told"			
[136] "toxic"	"white"	"without"	"didndt"
"guess"			
[141] "idll"	"okay"	"say"	"sign"
"smell"			
[146] "unusu"	"usual"	"alon"	"also"
"alway"			
[151] "base"	"becom"	"better"	"chang"
"come"			
[156] "day"	"dondt"	"everyday"	"exfoli"
"fact"			
[161] "fix"	"forget"	"get"	"harsh"
"instant"			
[166] "lazi"	"look"	"money"	"much"
"notic"			
[171] "often"	"pretti"	"prone"	"purchas"
"routin"			
[176] "safe"	"shower"	"sinc"	"sponsor"
"twice"			
[181] "unlik"	"wear"	"well"	"worth"
"best"			
[186] "cera"	"fragranc"	"free"	"give"
"hrs"			
[191] "least"	"part"	"protect"	"scent"
"sensitivityirrit"			
[196] "help"	"pleasant"	"thick"	"valu"
"amount"			
[201] "basic"	"design"	"difficult"	"easili"
"left"			
[206] "press"	"squish"	"back"	"certain"
"cheaper"			
[211] "combin"	"new"	"oil"	"tad"
"thing"			
[216] "tri"	"wish"	"amazon"	"bigger"
"consist"			
[221] "older"	"possibl"	"qualiti"	"right"
"smaller"			
[226] "someth"	"target"	"video"	"wateri"
"wrong"			
[231] "youdll"	"crack"	"disappoint"	"empti"
"noth"			
[236] "put"	"send"	"anyon"	"awar"
"beauti"			
[241] "burn"	"chemic"	"common"	"contact"
"cool"			
[246] "cosmet"	"dark"	"decad"	"fair"
"ferment"			
[251] "find"	"formula"	"furious"	"googl"
"grab"			
[256] "immedi"	"includ"	"incred"	"item"
"kick"			
[261] "less"	"longer"	"lotion"	"never"
"note"			
[266] "paraben"	"phenoxyethanol"	"pictur"	"previous"
"redden"			
[271] "reformul"	"regular"	"respons"	"rest"
"scentless"			
[276] "ship"	"somehow"	"sour"	"space"
"store"			
[281] "sunburn"	"tend"	"two"	"unaccept"
"uncertain"			
[286] "unscent"	"way"	"worst"	"box"
"broken"			
[291] "port"	"brand"	"start"	

> l=findFreqTerms(ibisfreq, lowfreq = 8) #List of words that appear at least 8 times in the corpus
> l #Display the frequent terms found

```

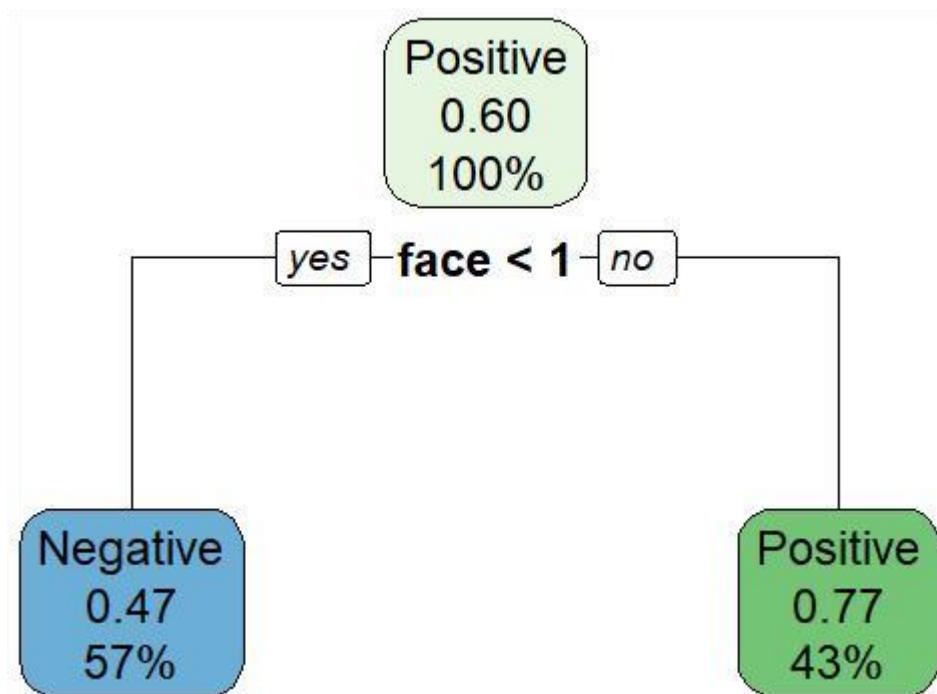
[1] "face"      "feel"      "product"   "skin"       "bottl"      "cleanser"
"use"        "wash"      "one"       "itd"        "cerav"
[10] "length(1) #Get the count of the frequent terms identified
[1] 11
> ##we have many features with too many zeros, high sparsity
> ibissparse=removeSparseTerms(ibisfreq, 0.80)
> ##keep only the terms that appears in 20% or more of the
feedback/documents/columns
> dim(ibissparse)
[1] 30 13
> inspect(ibissparse)
<<DocumentTermMatrix (documents: 30, terms: 13)>>
Non-/sparse entries: 108/282
Sparsity            : 72%
Maximal term length: 8
Weighting           : term frequency (tf)
Sample              :
Terms
Docs bottl cerav cleanser face feel one product skin use wash
 14     0    1      1   1   3   0      0    3   0    0
 16     0    1      0   3   3   0      2    1   1    1
 17     0    0      0   1   0   0      0    2   1    0
 18     0    1      1   1   2   0      0    1   0    0
 22     0    0      0   2   1   2      0    3   0    2
 23     0    0      0   1   1   0      0    1   1    1
 26     2    0      2   0   0   2      0    0   0    0
 28     2    2      1   2   0   1      4    0   3    0
 6      0    0      0   1   1   0      0    1   1    1
 9      0    0      1   0   0   1      1    0   1    0
> ##convert it to dataframe
> ibis_review=as.data.frame(as.matrix(ibissparse))
> dim(ibis_review)
[1] 30 13
> ibis_names=colnames(ibis_review)
>
>
> ibis_freq=c() # Initialize an empty vector to store the frequency
counts
>
> for (i in 1:5){
+   ibis_freq[i]=sum(ibis_review[,i]) #calculte the total freq for each
of the first 5 terms(col)
+ }
> ibis_freq # print the 5 terms
[1] 17 16 7 12 17
> barplot(ibis_freq,
+           col=rainbow(5),
+           names.arg = ibis_names[1:5],
+           ylim=c(0,20))

```



2. Perform a sentiment analysis

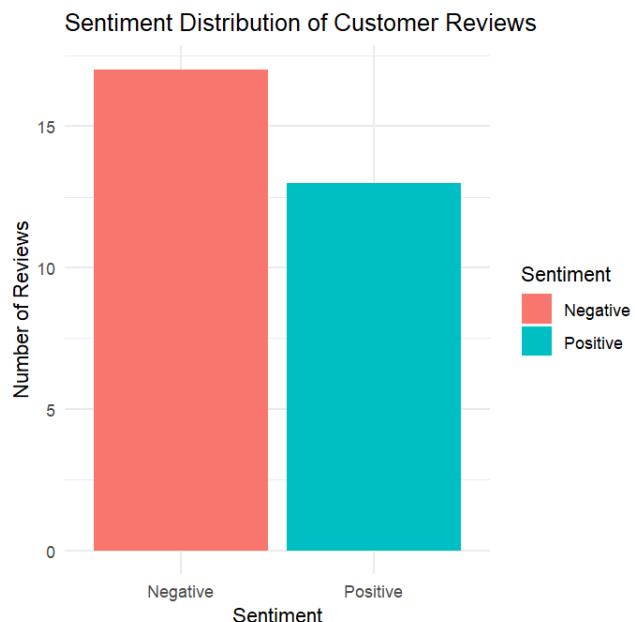
```
> #Supervised
> #####Sentiment analysis
> #create your sentiment variable
> View(x)
> rate=x$Rating
> ibis_review= ibis_review %>%
+   mutate(y=ifelse(rate>3, "Positive", "Negative"))
> View(ibis_review)
Intrepretation of the DT model:
```



Data Representation: 57% of the observations are negative if $\text{face} < 1$ while 43% of the observations are positive if $\text{face} \geq 1$. However, the probability of negative is 47% and the likelihood of positive is 77%. This indicates a slightly more negative sentiment overall among the customers. This also suggests that mentions of the term "face" may correlate with more favorable reviews.

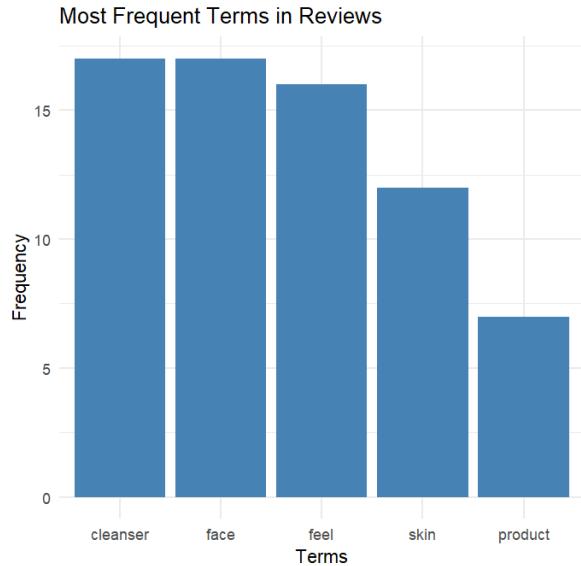
Ggplot2 for better analysis.

```
> # Created a data frame for plotting
> review_counts <- data.frame(
+   Sentiment = c("Positive", "Negative"),
+   Count = c(13, 17)
+ )
>
> ggplot(review_counts, aes(x = Sentiment, y = Count, fill = Sentiment)) +
+   geom_bar(stat = "identity") +
+   labs(title = "Sentiment Distribution of Customer Reviews",
+        x = "Sentiment",
+        y = "Number of Reviews") +
+   theme_minimal()
```



Comment: This bar plot illustrates the distribution of sentiments among customer reviews, highlighting a strong of negative feedback.

```
> term_freq <- data.frame(
+   Term = c("face", "feel", "product", "skin", "cleanser"),
+   Frequency = c(17, 16, 7, 12, 17)
+ )
>
> ggplot(term_freq, aes(x = reorder(Term, -Frequency), y = Frequency)) +
+   geom_bar(stat = "identity", fill = "steelblue") +
+   labs(title = "Most Frequent Terms in Reviews",
+        x = "Terms",
+        y = "Frequency") +
+   theme_minimal()
```



This bar plot visualizes the frequency of specific terms mentioned in customer reviews for a product. The x-axis displays the terms ("face", "feel", "product", "skin", "cleanser"), while the y-axis shows the corresponding frequency of each term. This visualization allows for a quick assessment of which terms are most commonly referenced by customers, highlighting key aspects of their feedback and potentially guiding further analysis or product improvements.

Conclusions

The analysis of customer reviews for the CeraVe Hydrating Facial Cleanser showed a slightly negative sentiment, with 57% of reviews being negative. Key themes included the product's effectiveness and texture; while many users appreciated its cleansing ability, concerns about a residual feeling and dryness were common. To improve customer satisfaction, CeraVe [must think. Overall, these insights can help guide product improvements and better meet customer needs.