BIG DATA ANALYTICS REPORT

Text Analysis - Ben & Jerry's IceCream

CONTENT

All the sections of the report

1 - DISCOVERY

Learn the business domain :

The global ice cream and sorbet market was estimated at USD 57.9 billion in 2018. It is expected to grow at a compound annual growth rate of 4.9% between 2019 and 2025 to reach USD 84.9 billion in 2025. The global market is highly fragmented, but some of the major players in the market include Unilever, Mars, Nestle, General Mills, Lotte and Dunkin' Brands.

Ben & Jerry's is an American brand founded in 1978 by Ben Cohen and Jerry Greenfield and based in South Burlington (United States of America). The company is specialized in the market ice cream and sorbets. Since 2000, Ben & Jerry's is part of the Unilever group.

Assess the resources available to support the project :

For this project, two datasets are used, downloaded from Kaggle :
- Product.csv which is composed of 57 rows and 7 columns (key, name, subhead, description, rating, rating_count and ingredients)
- Reviews.csv which is composed of 7943 rows and 8 columns (key, author, date, stars, title, helpful_yes, helpful_no and text)

Frame the business problem as an analytics challenge :

As Data Scientist working for Ben & Jerry's, we can applicate big data solutions to improve our marketing strategy. In fact, based on market sentiment, we want to incorporate data from customer's reviews in order to create a fuller picture of the customers' preferences and potential needs about all our products, through a product review sentiment analysis.

Formulate initial hypotheses to test :

Our initial hypotheses are the following :

- Ice cream being a product strongly associated with happiness, the sentiment analysis should mainly bring out the emotion of joy.
- Sentimental analysis scores and costumer's scores out of 5 should be close.
- Consumer feedback could help us identify the exact problem with a product.

2 - DATA PREPARATION

We explain here how we did the extract, transform and load steps to get our data.

1

Select only data about Ben & Jerry's IceCreams

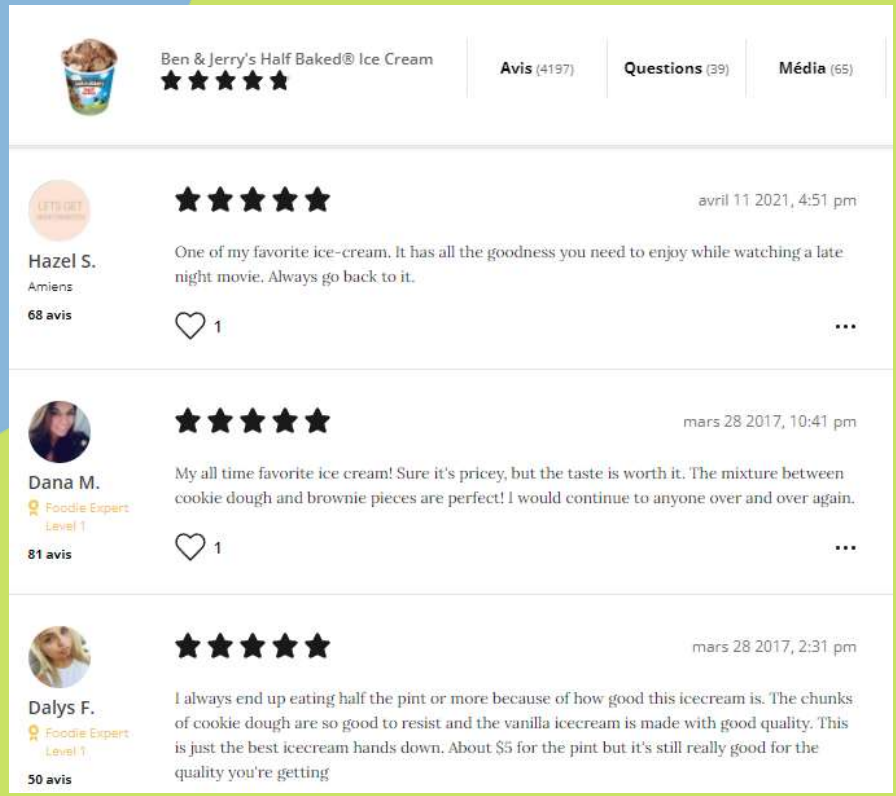
2

Import the two datasets modifying the encoding format

3

Merge both datasets to create only one corpus

3 - MODEL PLANNING



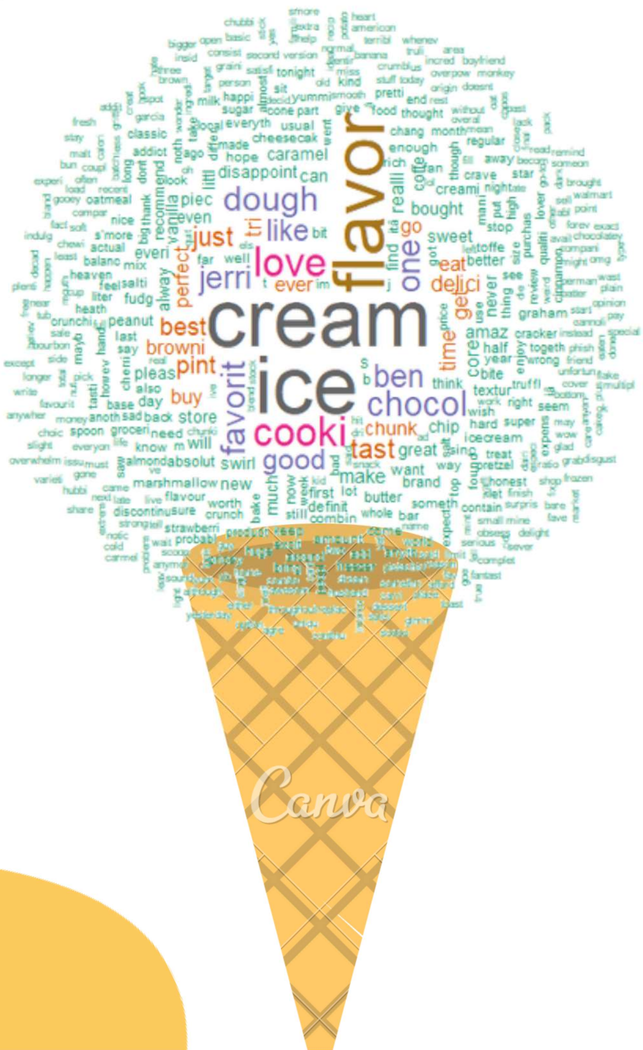
We decided to follow all these steps :

- Word Frequencies & Wordcloud
- Tokens
- Sentiment Analysis
- Sentiment Barplots of Ice Cream Flavors
- Investigating the worst IceCream flavor

4 - MODEL BUILDING / ESTIMATION

We explain in the following pages our estimation strategy and results.

WORDCLOUD



cream
ice
flavor
love
cooki
chocol
dough
favorit
like
jerri
ben
good
tast
just
best
one
delici
chunk
pint
browni
buy
Canva

TOKENS

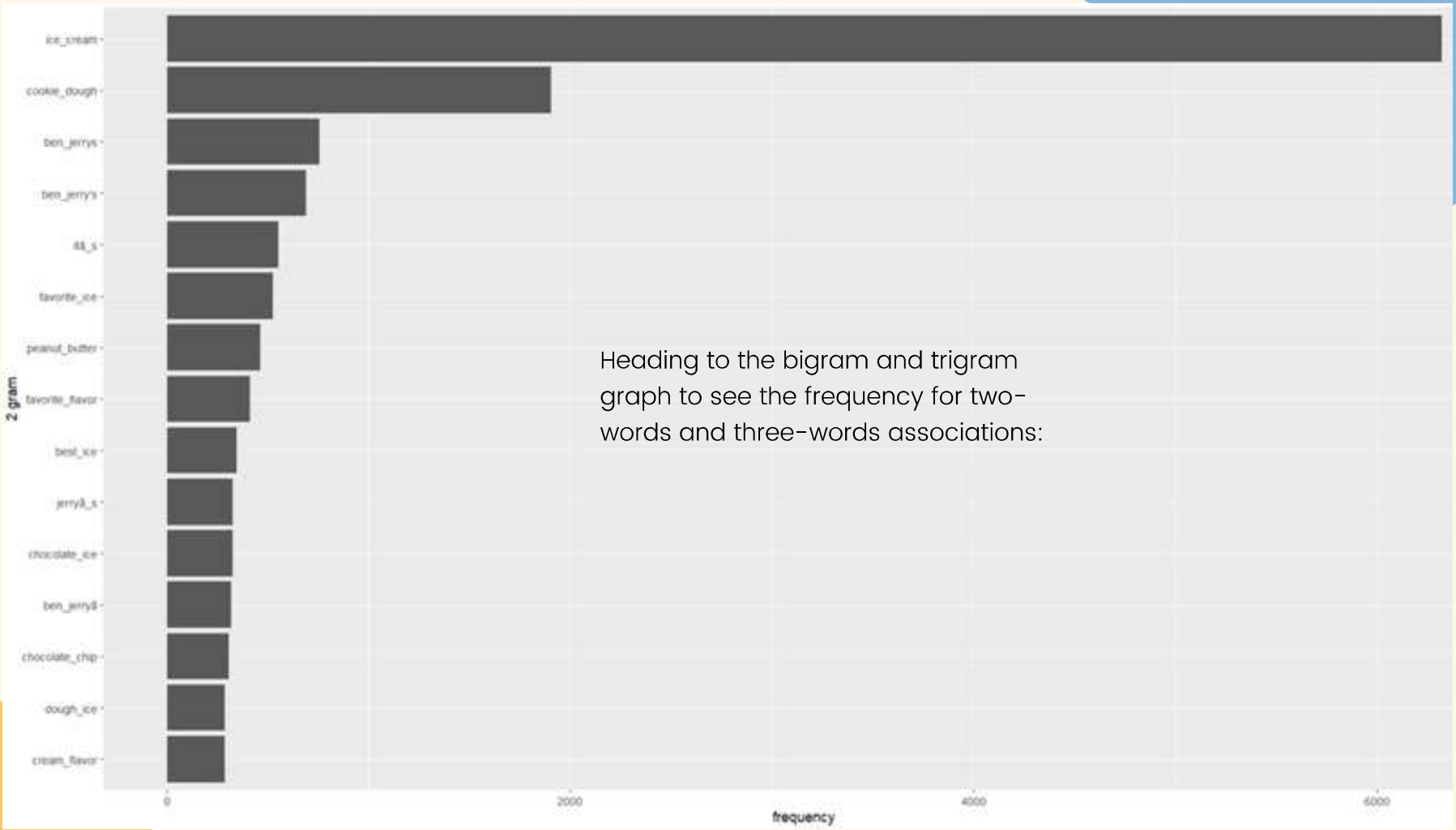
TOKENIZING IS A SIMPLE WAY TO SEARCH TEXT, PROCESS OF SPLITTING TEXT INTO INDIVIDUAL TERMS.

```
tok <- dfm$text %>%
  gsub("#","", . ) %>%
  corpus %>%
  tokens(what="word",
         remove_numbers=TRUE,
         remove_punct=TRUE,
         remove_symbols=TRUE,
         remove_separators=TRUE,
         remove_url=TRUE)
tok <- tokens_remove(tok,stopwords("english"))
```

As every occurrence of a term means a token, we're creating a bag of words

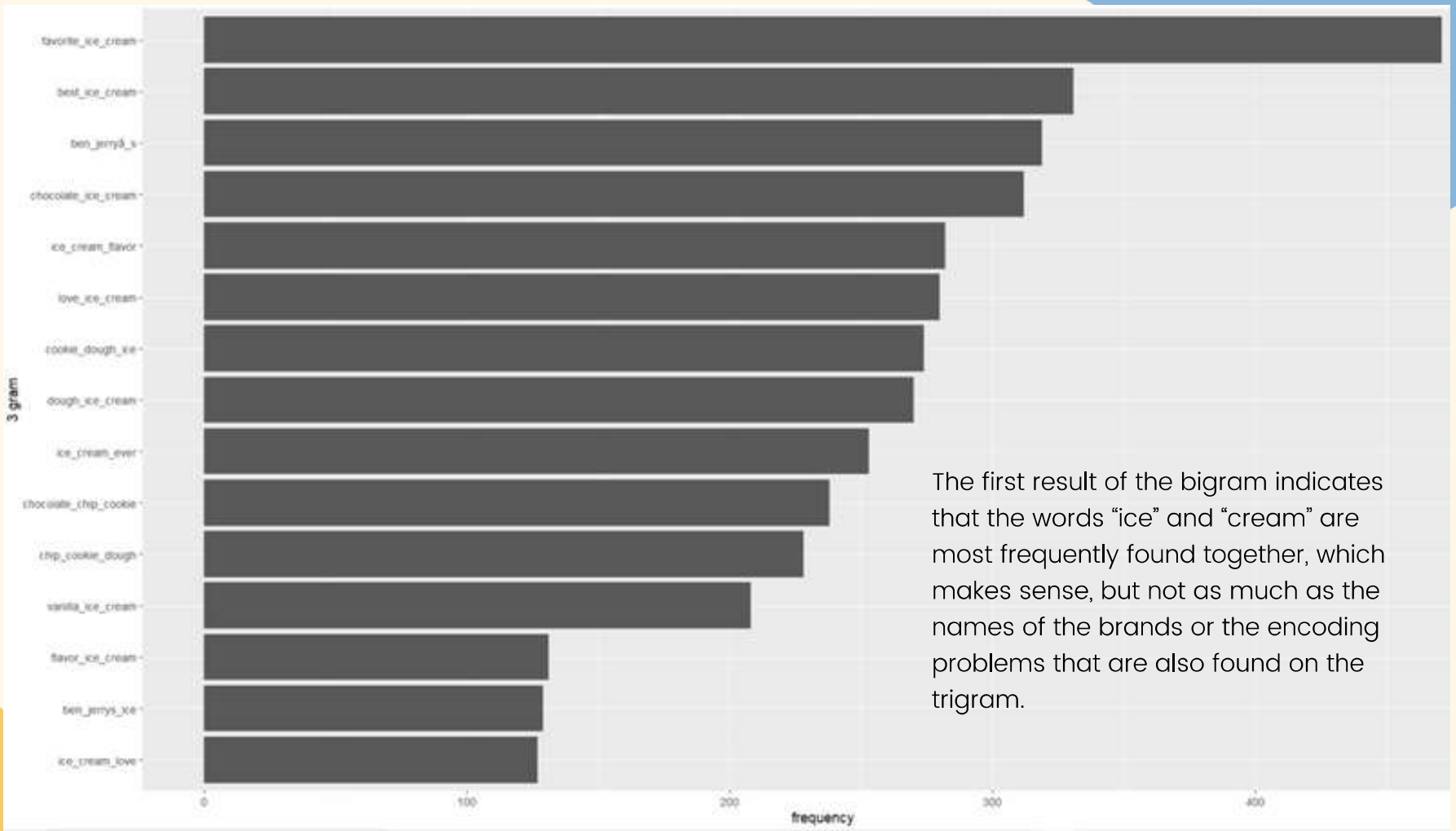
The most common types of tokens are characters, words, sentences, documents, and regular expressions, but we'll focus only for words, as we got reviews on products. We're looking for word associations, therefore here we're putting "what = 'word':"
Here we wanted to be sure to remove all the elements that would not be "pure text".

Tokens : bigram graph 1



Heading to the bigram and trigram graph to see the frequency for two-words and three-words associations:

Tokens : trigram graph 1

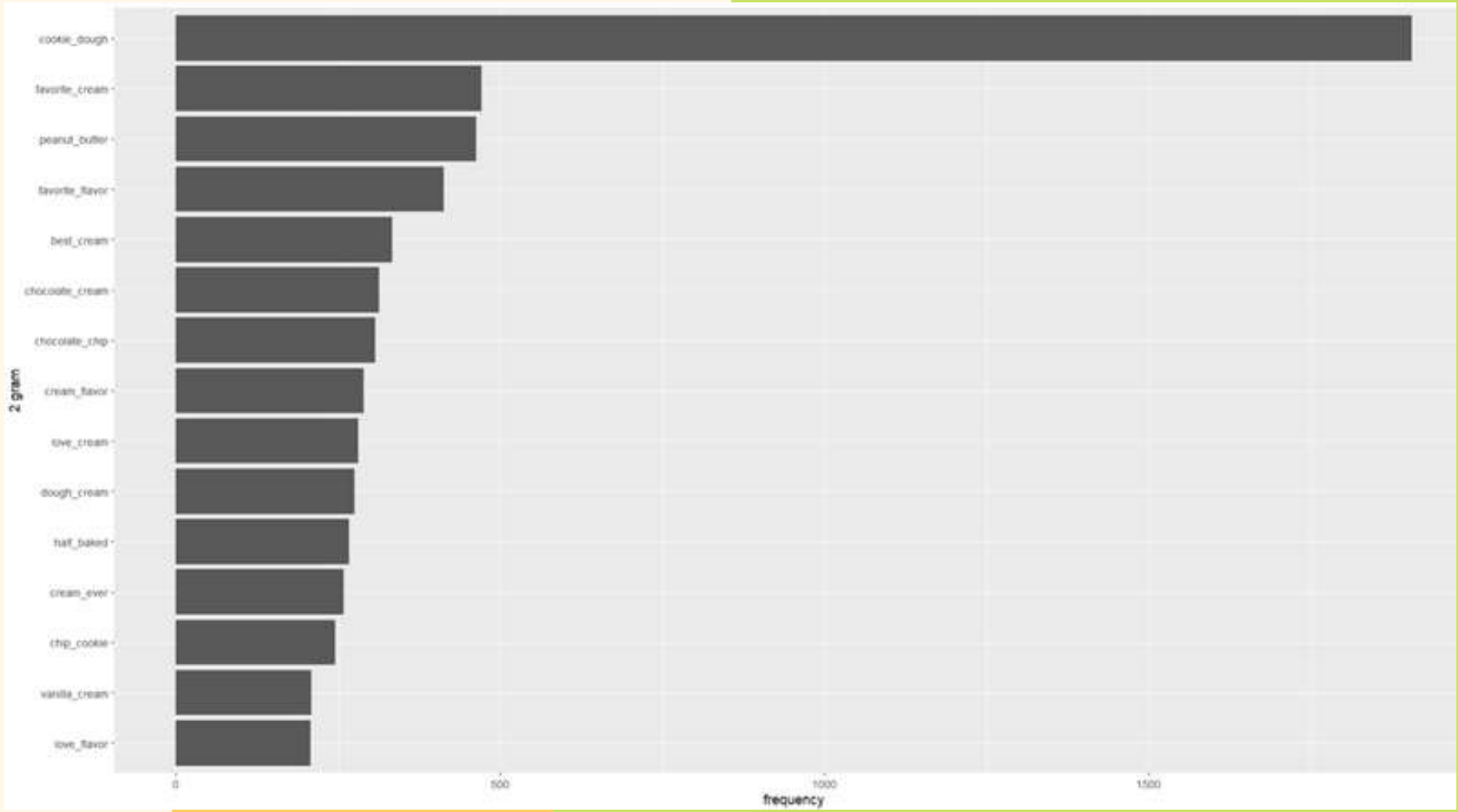


The first result of the bigram indicates that the words “ice” and “cream” are most frequently found together, which makes sense, but not as much as the names of the brands or the encoding problems that are also found on the trigram.

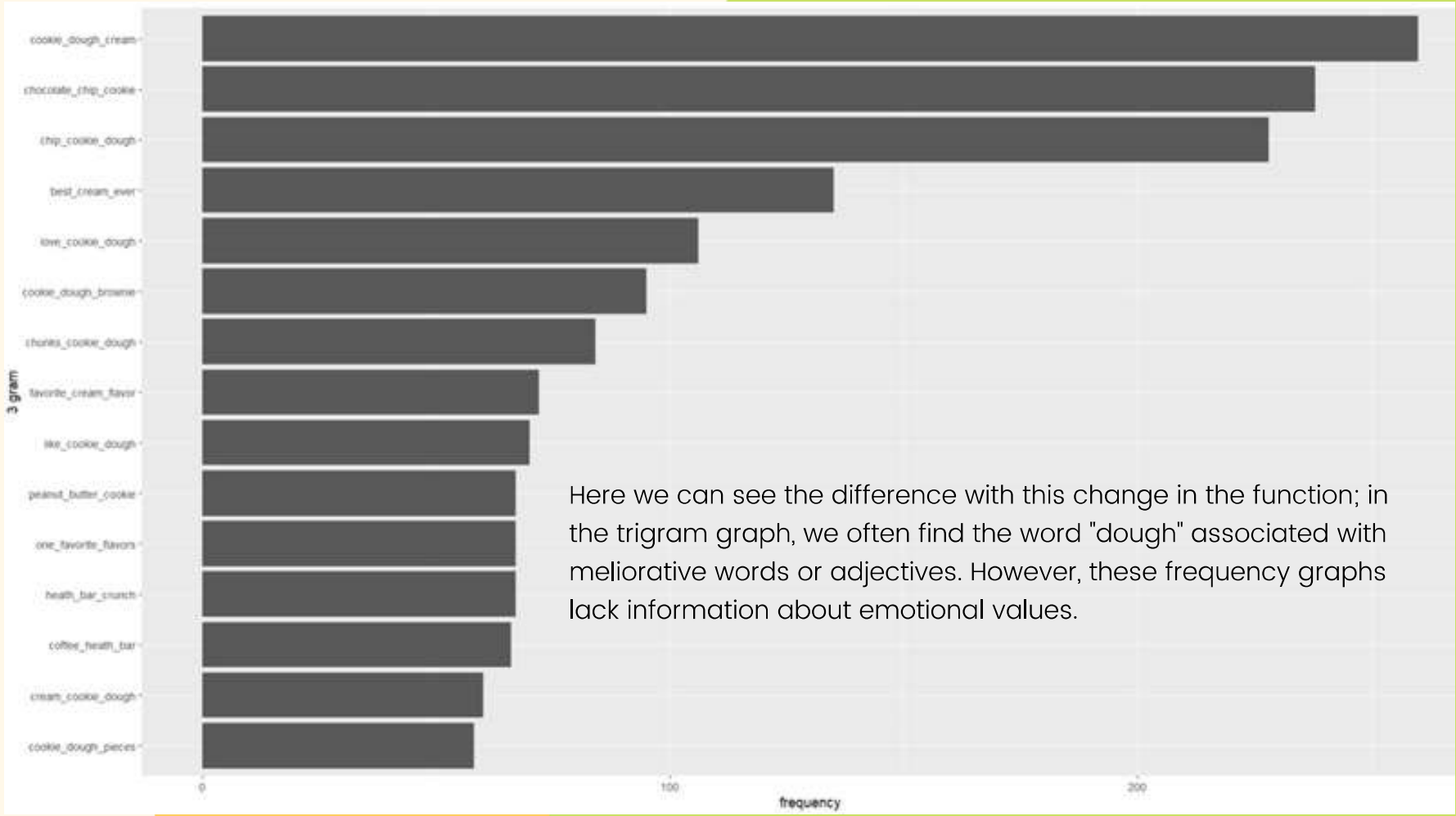
By removing certain terms from the tokens, we obtain more relevant associations, for example by specifying in the code below to remove some stemmed words

```
TokensStemmed <- tokens_remove(tok,c(stopwords("English"),"itâ", "iâ", "ben",
                                     "didnâ", "core", "ice", "jerryâ", "b" ))

dfm2 <- dfm(tokens_ngrams(TokensStemmed,n=2))

dfFreq2 <- textstat_frequency(dfm2)

#Top 15 2-grams tokens
ggplot(dfFreq2[1:15,], aes(x=reorder(feature, frequency), y=frequency)) +
  geom_col() +
  coord_flip() +
  scale_x_discrete(name = "2 gram") +
  theme(text=element_text(size=12))
```

Tokens : bigram graph 2

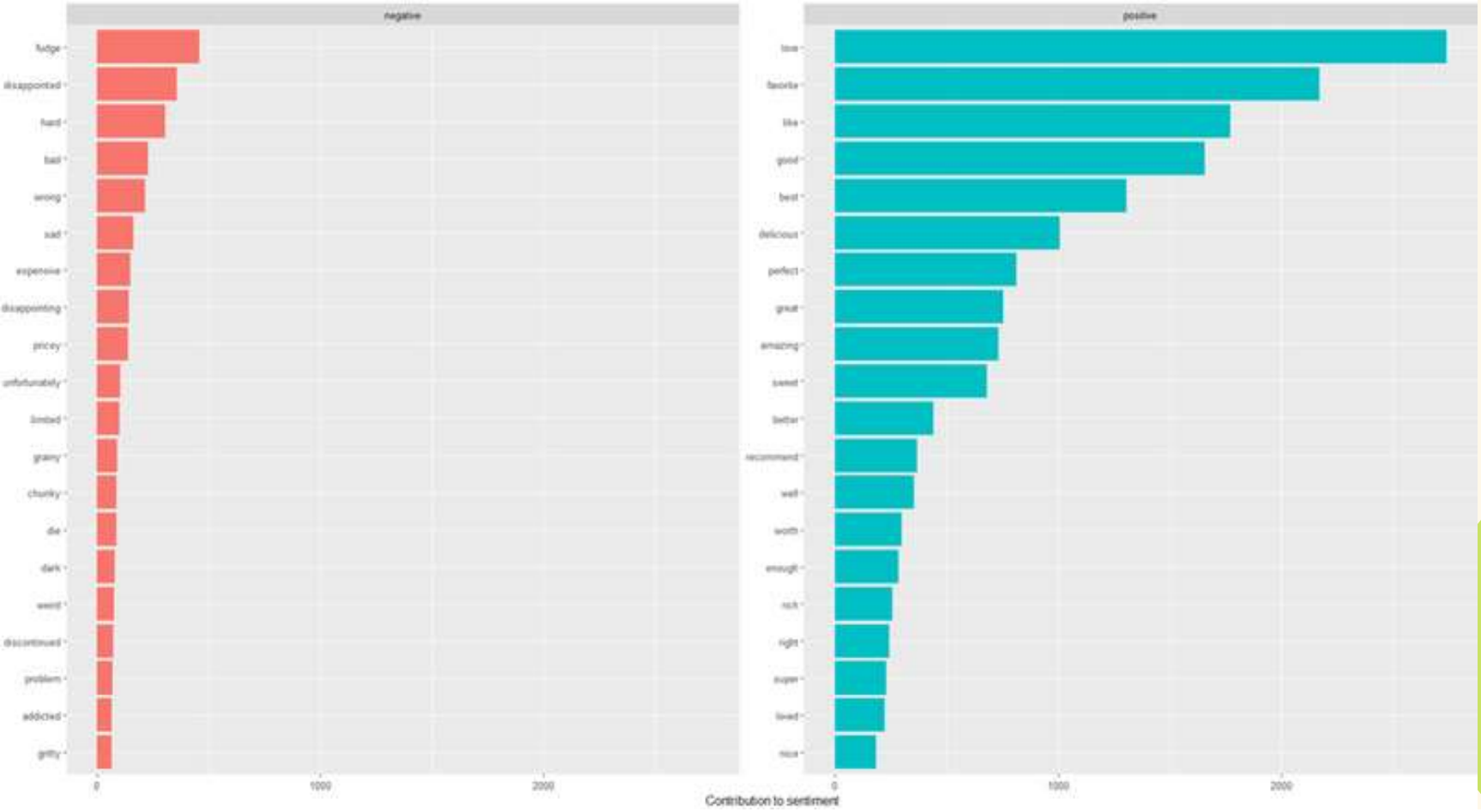


Tokens : trigram graph 2



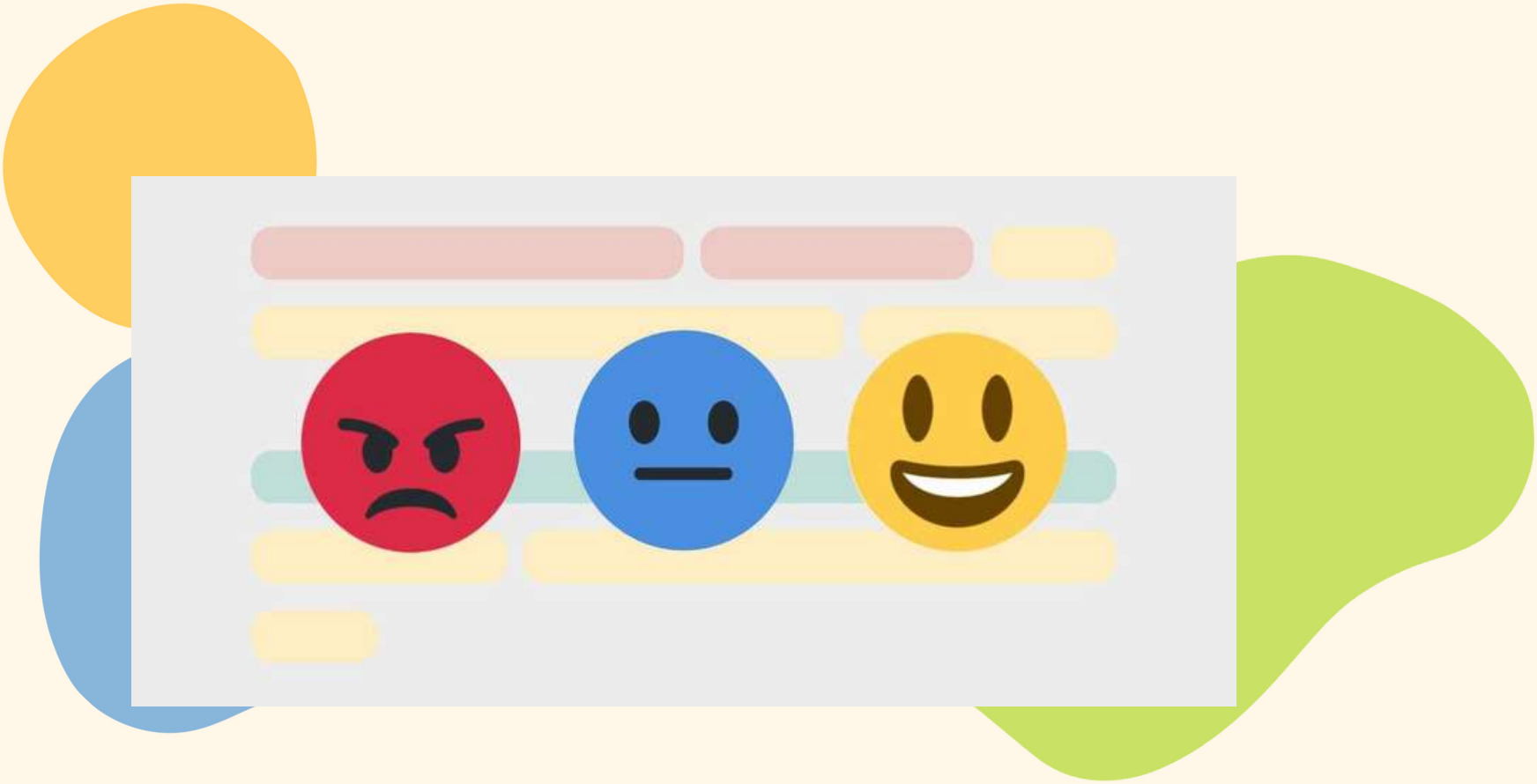
Here we can see the difference with this change in the function; in the trigram graph, we often find the word "dough" associated with meliorative words or adjectives. However, these frequency graphs lack information about emotional values.

Having removed words that don't have emotional values, the contribution to sentiment graph displays the words categorized positively and negatively:
Quite frequently, the words and adjectives used are positive, although the context is missing as an example, the word fudge is used in several ways, describing both a pastry and an escape (in a pejorative sense).

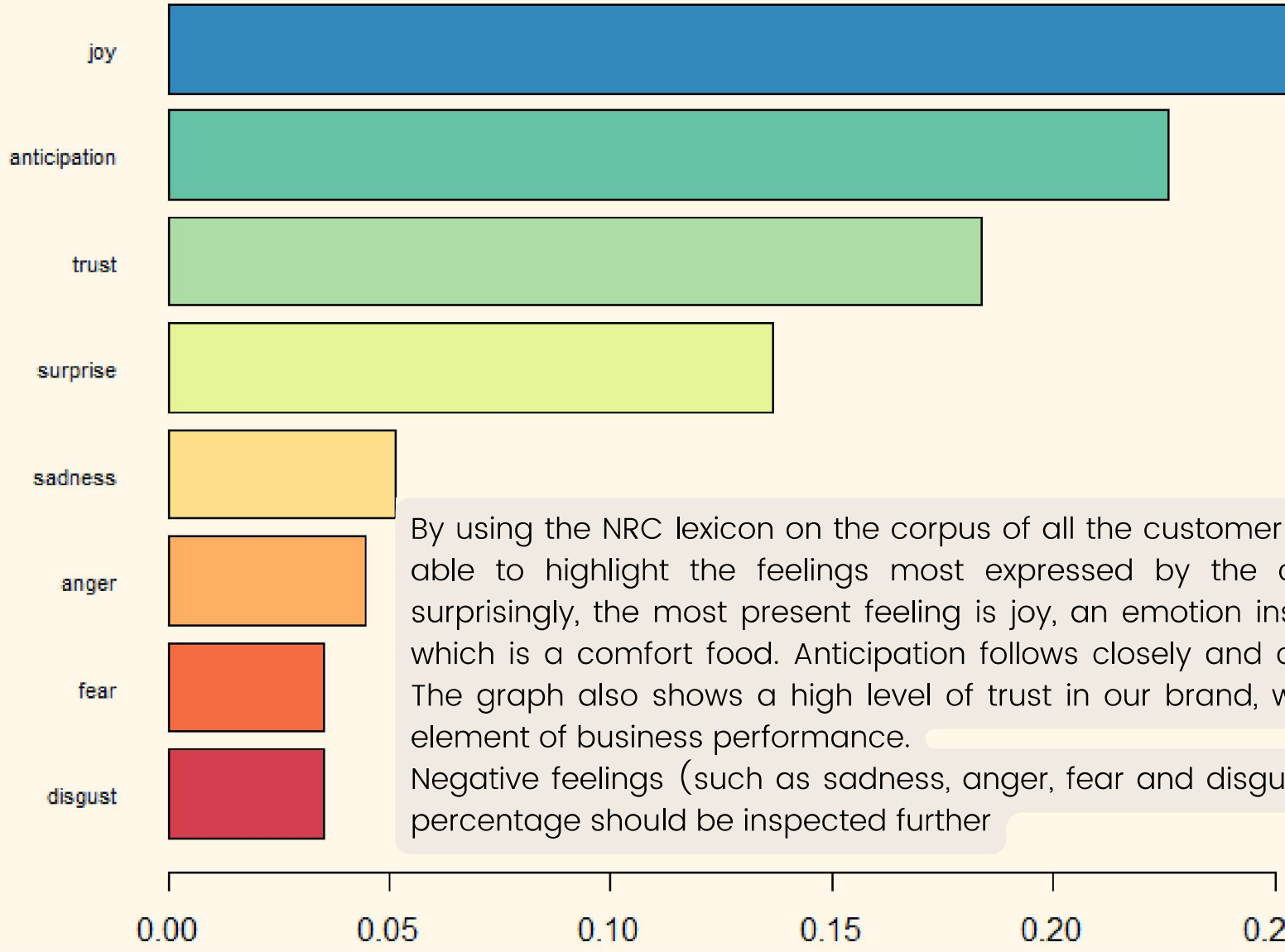


We can see here the limit of the model used with dictionaries when the context has not been specified. Moreover, we find several adjectives concerning the dissatisfaction of prices (expensive, pricey), an element to be taken into account in the improvement of the product, especially the offer.

SENTIMENT ANALYSIS



Emotions Expressed in IceCream product reviews

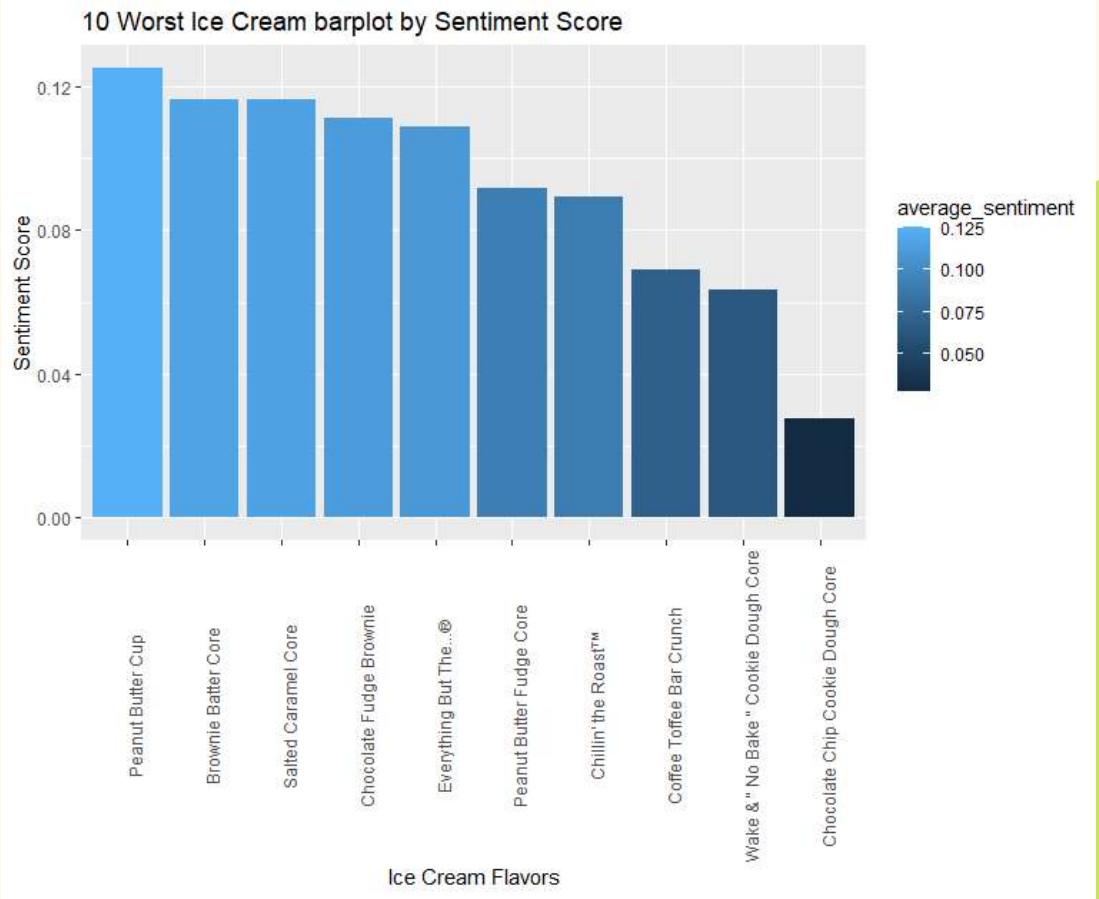


By using the NRC lexicon on the corpus of all the customer feedbacks, we were able to highlight the feelings most expressed by the consumers. Not too surprisingly, the most present feeling is joy, an emotion inspired by ice cream which is a comfort food. Anticipation follows closely and could echo appetite. The graph also shows a high level of trust in our brand, which is an essential element of business performance.

Negative feelings (such as sadness, anger, fear and disgust) despite their low percentage should be inspected further

The flavors ranking allows us to have a more precise idea of the reception of our offer. By grouping the dataset by flavor we can more easily get an idea of the ice creams that demand our attention.
Whether it's the Sentimental Score or the Score out of 5, the worst ice cream remains the same, which reassures us in the validity of our sentimental score.

The final part is dedicated to the research of the reasons of this low score.
In the absolute it would be necessary to examine all the ice creams in order to know the reasons why the best are the most appreciated and vice versa for the worst.
However within the framework of this exercise it is more relevant and concise to concentrate on one product.



COMMUNICATE RESULTS

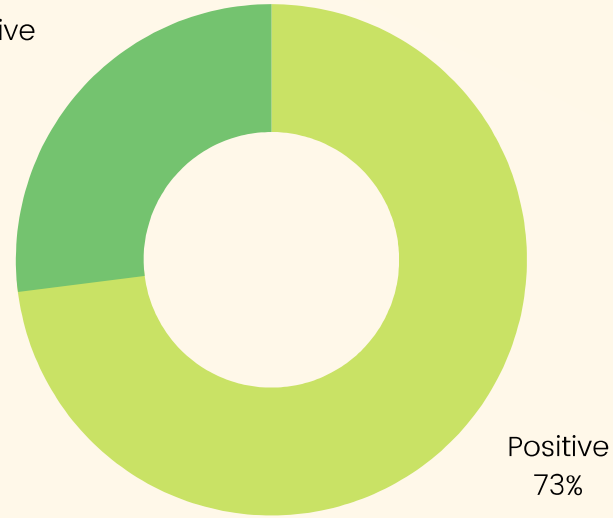
The Chocolate Chip Cookie Dough Core is indeed the most unappreciated flavor. A product that doesn't carry its weight will taint the customer's experience with the brand and by extension the image that consumers have of the company.

The resources used to produce this flavor can be put to better use, so it's important to look at it and decide whether or not to discontinue the product.

There seems to be a much stronger positive feeling in the comments than negative ones.

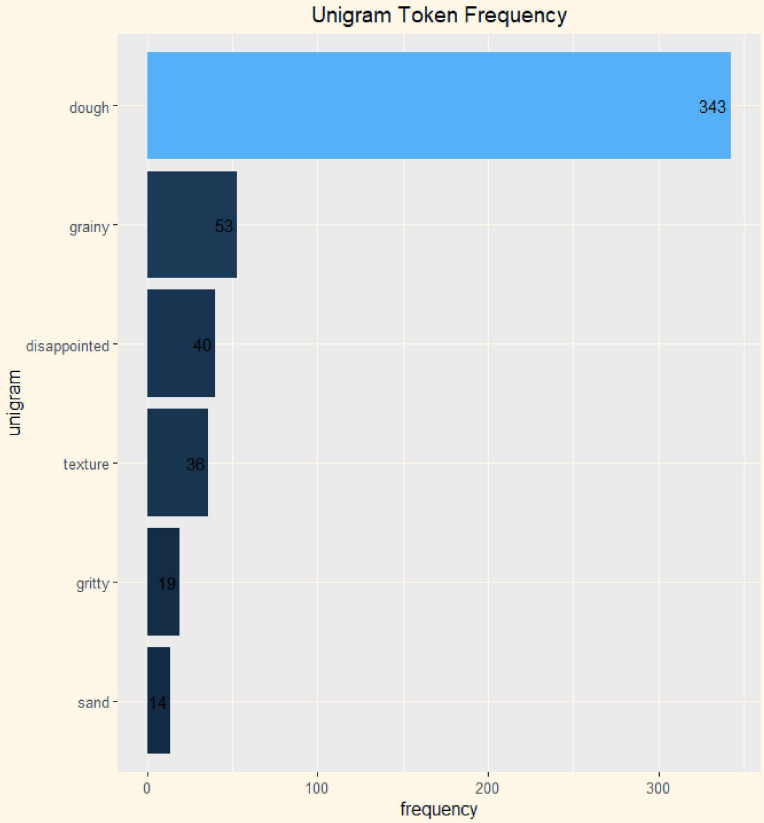
It leads us to believe that the product is not a hopeless case and that it can be improved. One specific element seems to lower the overall rating and it would be interesting to identify it.

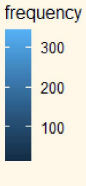
Percentage of Positive and Negative Sentiments of the Chocolate Chip Cookie Dough Core Flavor



Barplot of Negative Words

In the frequency table of negative words, we notice two adjectives related to texture ("grainy" and "gritty"). An element of the Ice cream seems be rough to the tongue because of a substance made of relatively large particles.





Unigram Token Frequency

This lead becomes concrete when we trace the frequency of unigrams tokens that reveal words of the same lexical field, 14 consumers had even compared the texture to sand. The most frequent word being "dough" it is possible that it is our culprit

PRODUCT REVIEW	SENTIMENT SCORE
I bought the cookie dough core and the **cookie dough** was crumbly when I tried to scoop it out. It was hard, had a grainy **texture** and didn't taste very good at all,nothing like the cookie dough in the half baked. The **icecream part** was good though	0.0528
Bought this last night and was incredibly disappointed. Never write reviews but don't buy this. The **cookie dough** was literally like **sweet** sand. It was gritty and grainy. Threw us away after about 5 bites. **I'll stick to the original Cookie Dough.**	0.0353
... I was so excited to dig in but while the **ice cream** is delicious, the **cookie dough core** is terrible. Rock hard, gritty, sandy, and very **inconsistent in size** from the top of the pint to the bottom. I hate having to write this review, especially because the **other 2 new cookie cores** were pretty good, but this one is REALLY dreadful! I'm really bummed because I wanted badly to love it.	0.030

Extracting low score reviews gives us new insights

When we linger on the comments containing these terms, we can underline that indeed the element which comes to spoil the product is especially the texture of the dough which is rough what is not the case of the pastes in the other flavors of ice cream. In addition, the rest of the ice cream seems to please which explains the presence of a majority of relatively positive comments.

CONCLUSION

Based on these results there is no need to discontinue the production of the Chocolate Chip Cookie Dough Core. This flavor is at the bottom of the ranking mainly because of the dough that can be replaced by the ones used in the other products. The Ben &Jerry's cookie dough is so popular that it is also sold separately so it would not be too expensive to replace this element of the ice cream.
The rest of the ice cream was mostly associated with positive adjectives, making changes to it would therefore not be justified.

Lastly our results seem to confirm the hypotheses. Joy (along with anticipation) are indeed the emotions that our consumers express the most which is a good sign regarding the customer experience the lowest scoring ice cream can be improved as we have managed to identify the exact disruptive element