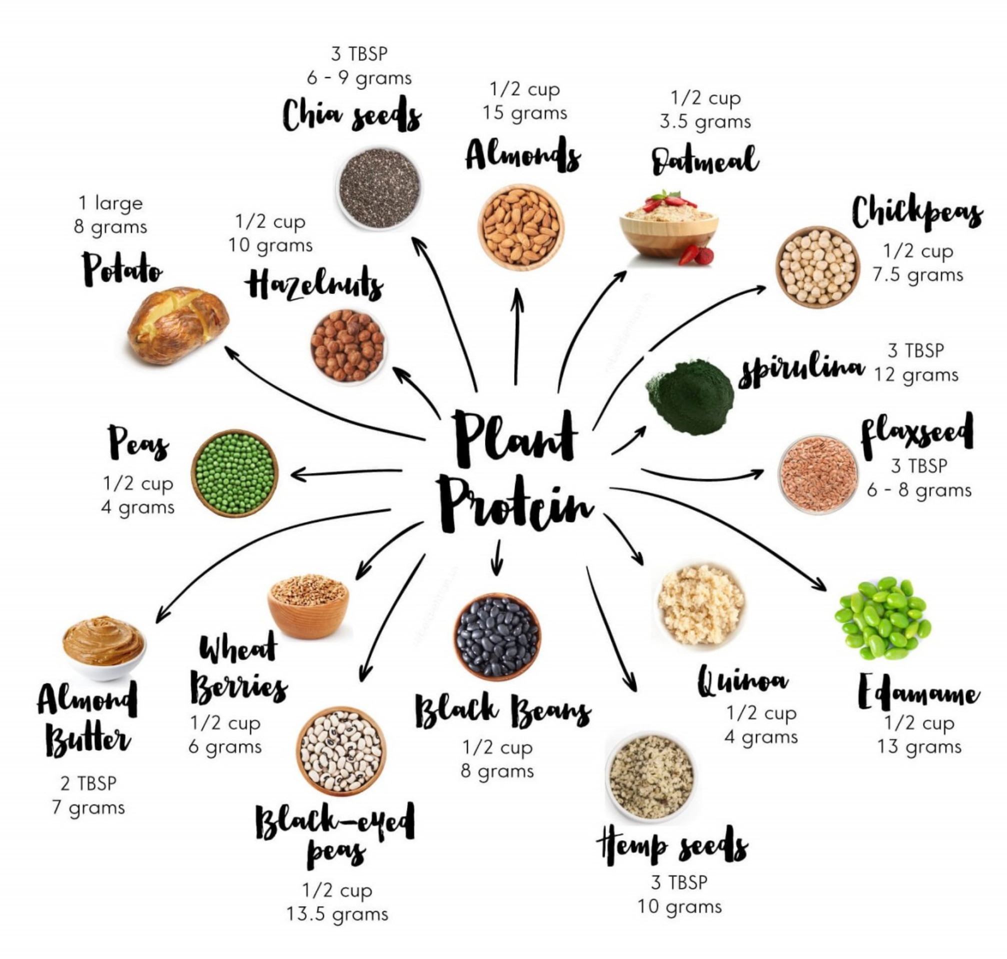


Business Insights on Online Discussions around Plant-based Diet

How do people Tweet about #plantbased, #vegetarian, and #protein?

Yi Yuan
Feb 10, 2020
Text Analytics
Word count: 1453

Introduction

The existing food system is unsustainable for the climate as well as global health. Having global markets adopt a plant-based diet offers a viable solution to avert an antibiotic crisis, reduces global climate change, feeds the global population with high quality protein and addresses animal suffering (Lagally, 2017). Last year was a breaking point for the plant-based industry to become highly in-demand by consumers. Media has increased coverage of new plant-based meat development in the business world. Impossible, Beyond, Ikea, Burger King are just a few of the names that support, adopt or push forward the plant-based meat market.

However, the market for plant-based products is still in an early phase without a mainstream customer base. As innovative plant-based products aim to drive up market share, it is critical to accurately understand consumer sentiments towards plant-based products and deliver the right message to consumers (Lagally, 2017). Because legumes and grains as a pair offers the complete protein profile comparable to animal protein, plant-based meat stands as a promising alternative source of protein (Lagally, 2017). Therefore, the goal of this project is to provide a picture of how online communities perceive topics around plant-based diet and products. The analysis explores Twitter discussions around hashtags #plantbased, #vegetarian, and #protein, identifies high frequency words in each hashtag dataset, and compare and contrast the three datasets, thereby generating business insights on the plant-based meat market.

Text Analysis

First, the analysis used Twitter API to pull three different datasets, each containing 1000 recent Twitter posts containing #plantbased, #vegetarian, or #protein based (Yuan, 2020). Excluding stop words such as “me”, “you”, “http”, “and”, the analysis identified the most common words in each dataset, assigning sentiments to each word in the Tweets by “positive” or “negative”:

Most common words in #plantbased dataset

	word	sentiment	n
1	love	positive	44
2	support	positive	32
3	plight	negative	28
4	gain	positive	27
5	wholesome	positive	26
6	easy	positive	25
7	healthy	positive	24
8	free	positive	23
9	lose	negative	23
10	damage	negative	22
11	delicious	positive	18
12	joker	negative	17
13	suffering	negative	16
14	fresh	positive	15
15	lovely	positive	14
16	win	positive	13

#vegetarian dataset

	word	sentiment	n
1	healthy	positive	91
2	delicious	positive	40
3	free	positive	38
4	love	positive	36
5	sweet	positive	34
6	perfect	positive	31
7	pan	negative	24
8	cold	negative	21
9	easy	positive	21
10	warm	positive	17
11	cheesy	negative	16
12	fantastic	positive	16
13	celebrate	positive	15
14	pure	positive	15
15	super	positive	15
16	fresh	positive	13

#protein dataset

	word	sentiment	n
1	fat	negative	117
2	benefit	positive	86
3	healthy	positive	82
4	delicious	positive	28
5	cold	negative	24
6	breaking	negative	22
7	free	positive	20
8	love	positive	20
9	shake	negative	19
10	celebrate	positive	15
11	easy	positive	15
12	happy	positive	12
13	parody	negative	10
14	lean	positive	9
15	amazing	positive	8
16	clean	positive	7

As shown in the table, the general Twitter user sentiments toward plantbased, vegetarian, and protein are positive. The analysis reveals that certain words are shared across discussions around plant-based, vegetarian, and protein: “healthy”, “love”, “delicious”, “easy”. Since the sentiment package that assigns sentiment to words are not customized to the context of the discussion, certain words that are in fact positive are labeled as negative, i.e. “plight” which most likely refers to a pledge to be herbivorous, “fat” which most likely refers to healthy fat. The analysis of common words in our datasets uncovers the business insight that plant-based, vegetarian, and protein are overwhelmingly positively perceived in Tweeter discussions.

The high frequency words that are not shared across datasets shares interesting revelations. For example, in the table on the left, “joker” appears as a high frequency word associated with #plantbased. Rather than a negative word, “joker” is certainly positive after Joaquin Phoenix as “Joker” in the eponymous film won the Best Actor in Oscars Academy Awards 2020 (Piper, 2020). Phoenix has been an active proponent of plant-based diet to address global warming and animal “suffering”, which also appears in the most common words in this table. “Suffering” in this case would also be perceived as positive towards #plantbased, since suffering most likely refers to the animal sufferings resulted from factory farming, which can be eliminated by the global adoption of plant-based diet. Phoenix’s massive influence as a celebrity person has driven up social awareness of the benefits of plant-based diet.

Moreover, in line 11 in the middle table, “cheesy” appears as a common word in discussions revolving around vegetarian. For plant-based companies, this reveals that cheese has a popular appeal to vegetarians who do not consume meat nonetheless tend to consume a large amount of cheese. If the plant-based cheese product achieves price parity, flavor and functionality parity with dairy cheese, positions to vegetarians with the right message, there is a promising business opportunity for plant-based cheese products.

On the table on the right, “fat” appears as the top frequency word. This can be interpreted as consumers who care about protein also care about healthy fat intake. For plant-based companies, it is valuable to consider the protein passionate consumers, formulate protein and healthy fat-rich products with a plant-based formula to increase market share.

Correlation between top common words in #plant-based vs #protein   #plant-based vs #vegetarian

Taking a closer look at the scatterplot above to compare and contrast #protein vs #plantbased, and #vegetarian vs #plantbased, our analysis extracts more granular nuances in the three overlapping customer bases. In the above correlograms, words closer to the diagonal are shared both customer bases of similar frequency of appearance; words there are further from the diagonal are peculiar to one of the customer bases. The analysis focuses on top common words that are not shared by the two customer bases to extract business insight.

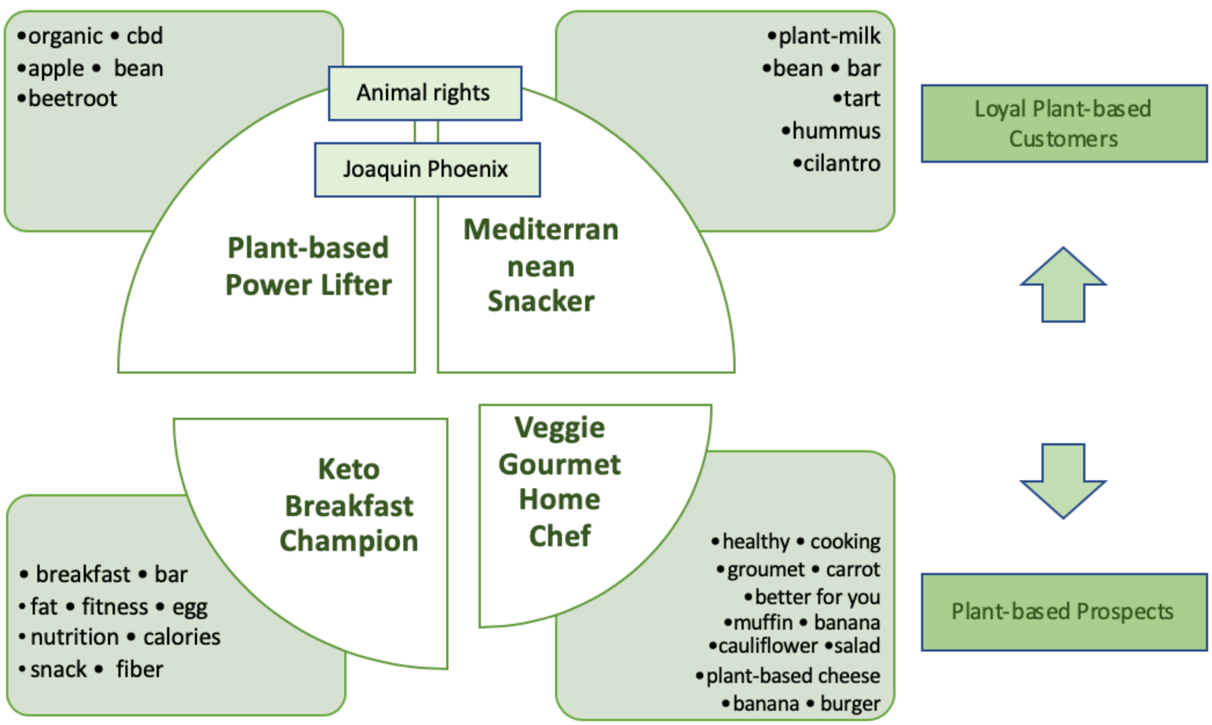
In both correlograms, again we identify “oscars”, confirming the impact from Phoenix’s winning speech on plant-based diet adoption. However, we observe that customers who talk about plant-based are more interested in certain foods and topics. In terms of topics, plant-based consumers predominantly care about animal rights; translating to actionable business terms, plant-based companies should consider incorporating animal suffering relief and animal rights in the messaging of plant-based products in marketing materials.

Plant-based businesses can also investigate top mentioned words tweeted by consumers who are plant-based but have nuanced focuses, i.e. protein or being vegetarian, to upsell or cross sell to the existing customer base. In terms of foods, plant-based consumers who also care about protein but to a lesser degree focus significantly more on “organic”, “avocado”, “cbd”, “apple”, “bean” and “beetroot” in descending order. On the other hand, plant-based consumers who also care about vegetarian but to a lesser degree focus significantly more on “dairy”, “bar” (as in snack bars), “bean”, “tart”, “hummus” and “cilantro” in descending order. Particularly, plant-based consumers who also care about vegetarian also focus on calories.

We can also investigate top mentioned words tweeted by consumers who care more about protein or being vegetarian than plant-based to expand the market share of plant-based meat. In terms of topics, protein loving consumers who care about plant-based but to a lesser degree protein focus significantly more on “fat”, “nutrition”, “fitness”, “breakfast”, “body” and “calories” in descending order. In terms of food, they focus significantly more on “snacks”, “bars”, “eggs”, and “fiber” in descending order. On the other hand, in terms of topics, vegetarian consumers who also care about plant-based but to a lesser degree focus significantly more on “healthy”, “cooking”, “foodporn”, “benefits” in descending order. In terms of food, they focus significantly more on “salad”, “veggie”, “cauliflower”, “cheese”, “muffins”, “carrot”, “banana” and “burger”.

These keywords provide valuable insights on how to improve product offering or highlight food ingredients in marketing materials to plant-based consumers by catering to specific needs, i.e. protein or vegetarianism. For example, according to our analysis, consumers who care about plant-based and protein in their diet are likely to be interested in products that highlight “organic”, contains “cbd”, “beans”, “avocado”. A matrix of four customer segments is presented below, including recommended product varieties and value propositions.

Plant-based Customer Segments – Taste and Preferences



The correlation between top common words in plant-based discussion and protein discussion is 68%, lower than the correlation between top common words in plant-based discussion and vegetarian discussion which is 80%. This follows our assumption since vegetarians have a plant-based diet except for dairy and egg consumptions. However, the correlation score between #protein and #plantbased reveals protein-driven customer segment as an opportunity for plant-based companies to expand their market share if plant-based alternative protein becomes strategically accessible and offers functionality parity and price parity with animal protein.

Conclusion

Adopting a plant-based diet on a global scale provides critical solutions to relieve climate change, antibiotic crisis and food shortage, and there has been accelerated growth in the business of plant-based products. However, the market is extremely immature compared to factory meat. Plant-based meat must achieve price, taste, and accessibility parity with animal protein, while marketed to the right consumers with the right message.

Thanks to celebrity activists such as Joaquin Phoenix, there has been growing media attention to plant-based diets, driving up online discussions and public awareness of the benefits of plant-based diet, including eliminating animal suffering. Overall, public sentiments on Twitter are very positive towards plant-based diet. This text analysis project uncovers valuable business insights by creating four customer segments, including consumers who care about plant-based diet, vegetarianism, and protein on various degrees. The report has identified the essential tastes and preferences of each segment and offered specific recommendations on product variety and market messaging for each segment. The result is visualized in a two by two matrix, easy to digest in the hope of helping plant-based companies and organizations to drive up the market share of plant-based products.

References

Lagally, C. (2017). *Plant-Based Meat Mind Maps: An Exploration Of Options, Ideas, And Industry*. The Good Food Institute. Retrieved from gfi.org/files/PBMap.pdf

(n.d.). photograph. Retrieved from https://trisalexandranutrition.com/blog/plant-based-protein-sources

Piper, K. (2020, February 10). We don't talk enough about animal suffering. That's why Joaquin Phoenix's Oscars speech matters. Retrieved from https://www.vox.com/future-perfect/2020/2/10/21131025/joaquin-phoenix-speech-animal-rights-oscars-2020

Yuan, Y. (2020, February 10). TwitteR Library: #plantbased, #vegetarian, #protein.

Appendix – R Code and Outputs

Appendix I – R Code

```
library(tidyverse)
library(tidytext)
library(textdata)
library(dplyr)
library(widyr)
library(tidyr)
library(stringr)
library(scales)
library(twitteR)
library(tm)
library(ggplot2)
library(igraph)
library(ggraph)
library(reshape2)
library(wordcloud)

#to get your consumerKey and consumerSecret see the twitteR documentation for instructions
consumer_key <- 'fXZKq3cBuzf0SxNF3HtBhS1QP'
consumer_secret <- 'WGU570efXI1mqEewnO11ayK1VoiAcUUIXEOWGGtHq5FHid5xOi'
access_token <- '1217533548321619968-WS9uPTcaPEmusN7DUi3tlXqo9UREfG'
access_secret <- 'gtpnH4Ea9fC0DC8A4GgZ0BlrbqgHKb1dPQBOa43X4O9Fl'

setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)

####################################
# pull 3 Twitter datasets, #########
# associated with one common theme : plant-based diet #####
plantbased <- twitteR::searchTwitter('#plantbased', n = 1000, since = '2015-01-
01',retryOnRateLimit = 1e3)
pb = twitteR::twListToDF(plantbased)

vegetarian <-twitteR::searchTwitter('#vegetarian', n = 1000, since = '2015-01-01',
retryOnRateLimit = 1e3)
veg = twitteR::twListToDF(vegetarian)

protein <- twitteR::searchTwitter('#protein', n = 1000, since = '2015-01-01', retryOnRateLimit =
1e3)
prtn = twitteR::twListToDF(protein)
```

```
# create my own stop word library
cust_stop <- data_frame(word = c("http", "https", "rt", "t.co", "amp" ,"h", "a", "q", "b", "c", "n",
"w", "o", "f", "g", "i", "m", "d", "u", "th", "aber", "it", "t", "al", "el"), lexicon = rep("cust", each
=25))

# tokenize, rmv stop words
# protein df
tidy_prtn <- prtn %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(cust_stop)
View(tidy_prtn)

# plantbased df
tidy_pb <- pb %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(cust_stop)
View(tidy_pb)

# veg df
tidy_veg <- veg %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(cust_stop)
View(tidy_veg)

#### We want to combine all the datasets and calculate frequencies
# correllation is the best framework to compare
frequency_twitter <- bind_rows(mutate(tidy_veg, author = "vegetarian"),
                               mutate(tidy_pb, author = "plantbased"),
                               mutate(tidy_prtn, author = "protein")
) %>%
  mutate(word = str_extract(word, "[a-z']+")) %>%
  count(author, word) %>%
  group_by(author) %>%
  mutate(proportion = n/sum(n))%>%
  select(-n) %>%
  spread(author, proportion) %>%
  gather(author, proportion, `protein`, `vegetarian`)

#correlograms
ggplot(frequency_twitter, aes(x=proportion, y=`plantbased`,
                              color = abs(`plantbased`- proportion)))+
```

```
  geom_abline(color="grey40", lty=2)+
  geom_jitter(alpha=.1, size=2.5, width=0.3, height=0.3)+
  geom_text(aes(label=word), check_overlap = TRUE, vjust=1.5) +
  scale_x_log10(labels = percent_format())+
  scale_y_log10(labels= percent_format())+
  scale_color_gradient(limits = c(0,0.001), low = "darkslategray4", high = "gray75")+
  facet_wrap(~author, ncol=2)+
  theme(legend.position = "none")+
  labs(y= "plantbased", x=NULL)

# compare how multiple groups are talking
cor.test(data=frequency_twitter[frequency_twitter$author == "vegetarian",],
         ~proportion + `plantbased`)

cor.test(data=frequency_twitter[frequency_twitter$author == "protein",],
         ~proportion + `plantbased`)

######## Most common positive and negative words #############

bing <- get_sentiments('bing')

# we want to find bing sentiments, ranked by contribution per token
bing_tidy_pb <- pb %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(cust_stop) %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()

bing_tidy_pb %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(y="Contribution to sentiment", x=NULL)+
  coord_flip()

bing_tidy_veg <- veg %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
```

```
  anti_join(cust_stop) %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()

bing_tidy_veg %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(y="Contribution to sentiment", x=NULL)+
  coord_flip()

bing_tidy_prtn <- prtn %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  anti_join(cust_stop) %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort=T) %>%
  ungroup()

bing_tidy_prtn %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word=reorder(word, n)) %>%
  ggplot(aes(word, n, fill=sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y")+
  labs(y="Contribution to sentiment", x=NULL)+
  coord_flip()
```

Appendix II – Outputs

>View(tidy_pb)

	word	sentiment	n
1	love	positive	44
2	support	positive	32
3	plight	negative	28
4	gain	positive	27
5	wholesome	positive	26
6	easy	positive	25
7	healthy	positive	24
8	free	positive	23
9	lose	negative	23
10	damage	negative	22
11	delicious	positive	18
12	joker	negative	17
13	suffering	negative	16
14	fresh	positive	15
15	lovely	positive	14
16	win	positive	13

>View(tidy_veg)

	word	sentiment	n
1	healthy	positive	91
2	delicious	positive	40
3	free	positive	38
4	love	positive	36
5	sweet	positive	34
6	perfect	positive	31
7	pan	negative	24
8	cold	negative	21
9	easy	positive	21
10	warm	positive	17
11	cheesy	negative	16
12	fantastic	positive	16
13	celebrate	positive	15
14	pure	positive	15
15	super	positive	15
16	fresh	positive	13

>View(tidy_prtn)

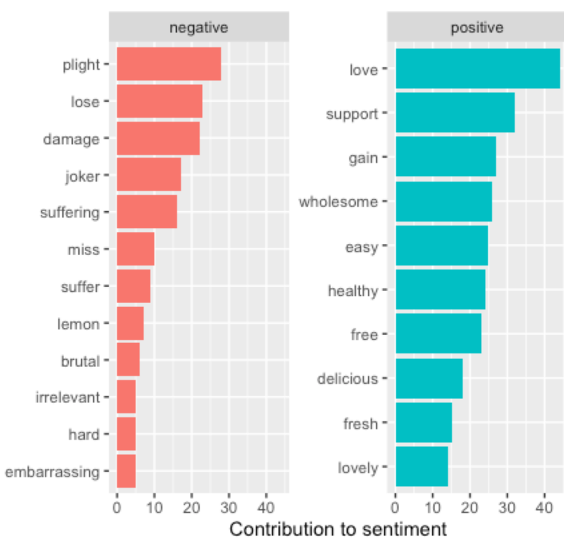
	word	sentiment	n
1	fat	negative	117
2	benefit	positive	86
3	healthy	positive	82
4	delicious	positive	28
5	cold	negative	24
6	breaking	negative	22
7	free	positive	20
8	love	positive	20
9	shake	negative	19
10	celebrate	positive	15
11	easy	positive	15
12	happy	positive	12
13	parody	negative	10
14	lean	positive	9
15	amazing	positive	8
16	clean	positive	7

#correlogram

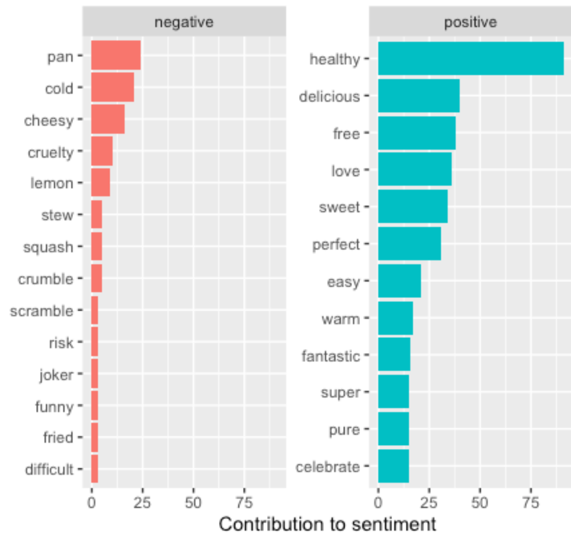

```
> cor.test(data=frequency_twitter[frequency_twitter$author == "vegetarian",],
+         ~proportion + `plantbased`)

        Pearson's product-moment correlation

data:  proportion and plantbased
```

```
t = 41.775, df = 981, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7764102 0.8215235
sample estimates:
      cor
0.8000953

>
> cor.test(data=frequency_twitter[frequency_twitter$author == "protein",],
+         ~proportion + `plantbased`)

	Pearson's product-moment correlation

data:  proportion and plantbased
t = 25.377, df = 745, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6404598 0.7175928
sample estimates:
    cor
0.68091
```

#bing counts in pb dataset



#bing counts in veg dataset



#bing counts in protein dataset