Presentation

Trend Prediction Based on Social Media Data: A Case Study on Veganism

Runxi Wang

01 Introduction

02 Exploratory Analysis

03 Model Building

04 Results Review

05 Conclusion and Discussion

01

Introduction

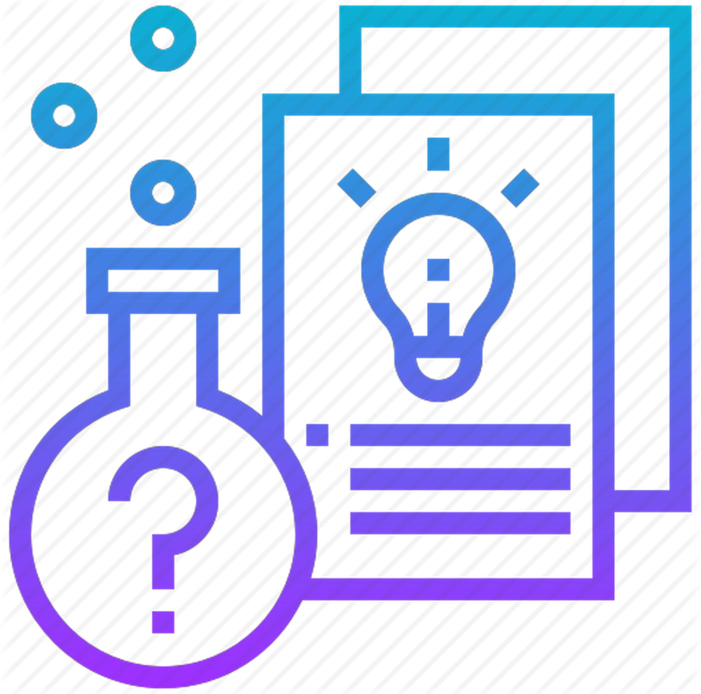
Introduction

- Background：
  - Growing Acceptance of Veganism
  - Boosting Vegan Market
  - Social Media is used more and more often as source of marketing decision process

- Research purpose：
  - Investigate how the conversation of vegan looks like on social media (Twitter and Instagram) in the US in 2018
  - Provide suggestions to marketers in the vegan market on how to reach out to social media users better

Measure and Procedure

- **Hypothesis:**
  - Different media type of the post, posting day of week, post time, location, and demographic factors, such as gender and profession are significantly related to the engagement number of social media posts related to vegan.



Measure and Procedure

- **Research Methods:**
  - Descriptive analysis
  - Regression analysis: multiple linear regressions

Data Description

- Datasource: The dataset was pulled by NetBase from the Vegan topic in this April, 2019. I was doing an analysis about the changing trend of vegan conversation. Netbase captured about 4 Million mentions of Vegan in 2018.
- Dataset format: The dataset was saved in a csv file, which contains 1,000 randomly selected tweets and Instagram posts, authors’ gender, followers of the authors, posting day and time, posting locations, format of posts, and engagements (Likes and Comments).
- Analytics Tool: I’m using R as my analysis tool and plan to analyze the data.
- The data was split to train (80%) and test (20%) randomly.

Sound Bite Text	Source	Post T	Media	URL	Publis	Author Gen
I made 3 dishes for our Geek Elite Holida	Twitter	Original	Link	http://twi	Dec 23, 2(	Unknown
I liked a @YouTube video youtu.be/GO68	Twitter	Original	Link	http://twi	Jun 15, 20	Unknown
Is eating ass vegan? Cause I might just tu	Twitter	Original	No Media	http://twi	Mar 19, 2(	Unknown
Hollup @Starbucks! What is this picture	Twitter	Original	Image	http://twi	Apr 5, 201	Female
Okay, I'm making vegan creamy broccoli	Twitter	Original	No Media	http://twi	Jul 3, 2018	Female
I liked a @YouTube video youtu.be/lfIl3x	Twitter	Original	Link	http://twi	May 21, 2	Male
I told my therapist I was giving up drinkir	Twitter	Original	No Media	http://twi	Mar 19, 2(	Unknown
I thought getting vegan soul food would	Twitter	Original	No Media	http://twi	May 13, 2	Female
#whatveganseat #vegan Can't Quit This:	Twitter	Original	Link	http://twi	Feb 11, 2(	Female
Is she still a vegan if she eats my meat?	Twitter	Original	No Media	http://twi	Dec 19, 2(	Unknown
Kiss Me I'm Raw-ish! Sort of Raw? Raw ti	Twitter	Original	Link	http://twi	Sep 24, 2(	Female
I found a recipe for vegan pumpkin sugar	Twitter	Original	No Media	http://twi	Oct 28, 2(	Female
My gf suggested that we cut palm oil out	Twitter	Original	No Media	http://twi	Nov 13, 2(	Unknown
I can’t eat that, I’m vegan! *does a line o	Twitter	Original	No Media	http://twi	Jul 19, 20:	Unknown
This is going in the slow cooker for 🏈 n	Twitter	Original	Link	http://twi	Jan 13, 20	Female
I followed a vegan diet for 3 months, it w	Twitter	Original	No Media	http://twi	Jan 21, 20	Unknown
This meal was so good last night, I'm hav	Twitter	Original	Image	http://twi	Apr 19, 2(	Male
I thought Californication libs/DemocRATS	Twitter	Original	Link	http://twi	Sep 1, 201	Unknown
I love that Halo Top put out a few vegan	Twitter	Original	No Media	http://twi	Jun 13, 20	Female
Peanuts? No, thanks. Pretzels? I'll pass. A	Twitter	Original	Link	http://twi	Aug 13, 2(	Unknown
I could never date a vegan. if you think I’	Twitter	Original	No Media	http://twi	Sep 12, 2(	Female

Data Clean Process

- **Remove Non-Plain Text Elements:**
  - Social media data contains a lot of elements other than plain text
  - Remove punctuations and Remove stopwords
- **Sentiment Calculation:**
  - Using NLP process to calculate the sentiment of the post
- **Update Post Time:**
  - The time is based on EST time zone. Convert the time to author's local time base on their location
  - Split the time to Date and Hour
  - Change the Date to Day of Week
- **Encode Categorical Data:**
  - Encode Hour into day parts, then encode to Weekday (1) and Weekend (0)
  - Encode State (West Coast ->1, Others ->0) into binary code
  - Encode post type: Video (1) and No Video (0)

02

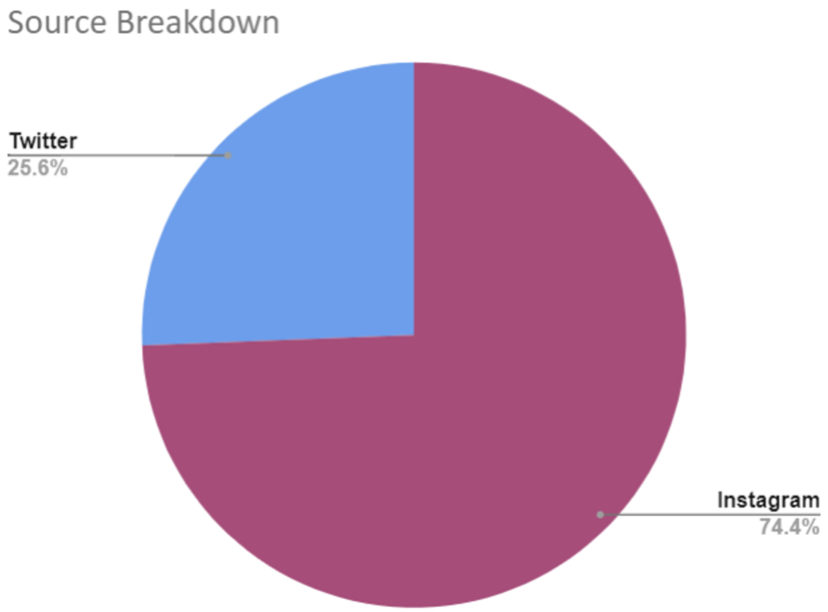
Exploratory Analysis

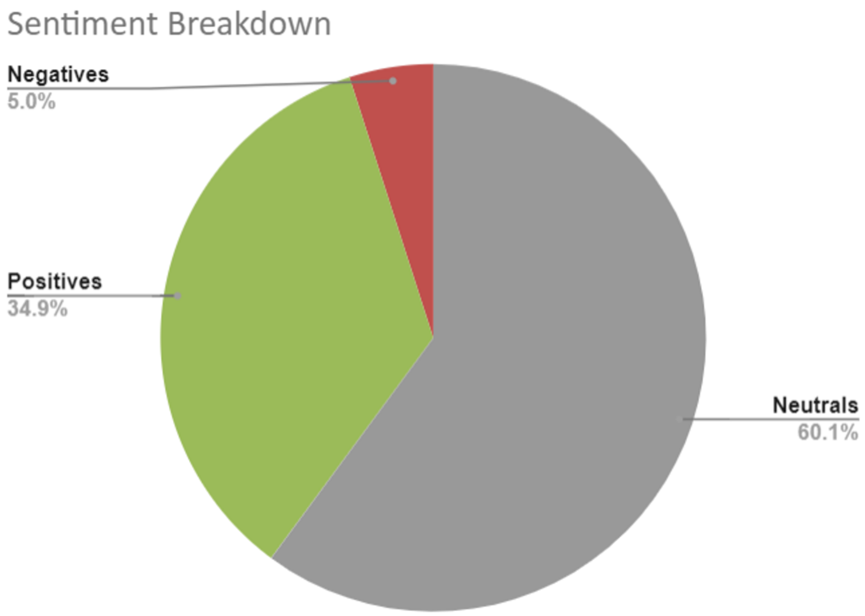
Check the frequency distribution

- **Word cloud** is a good way to show how many times a word is used in the conversation.
- The word cloud above indicates that people are using word like “healthy”, “love”, “delicious” to describe their vegan dining experience.
- “Chocolate”, “Salad” and “Protein” are also mentioned frequently together with vegan.

Source and Sentiment

- In this random sample, nearly 75% of the social media posts were published on Instagram, which indicates that social media users are more likely to share their experience and opinions about vegan on Instagram.
- About 35% of the sample are leaning towards positive. Generally, those who talk about vegan on social are holding a positive emotion toward the topic.

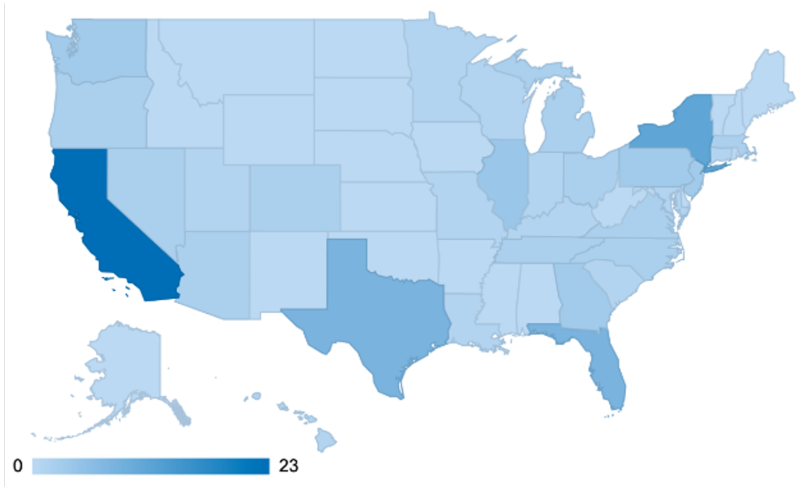




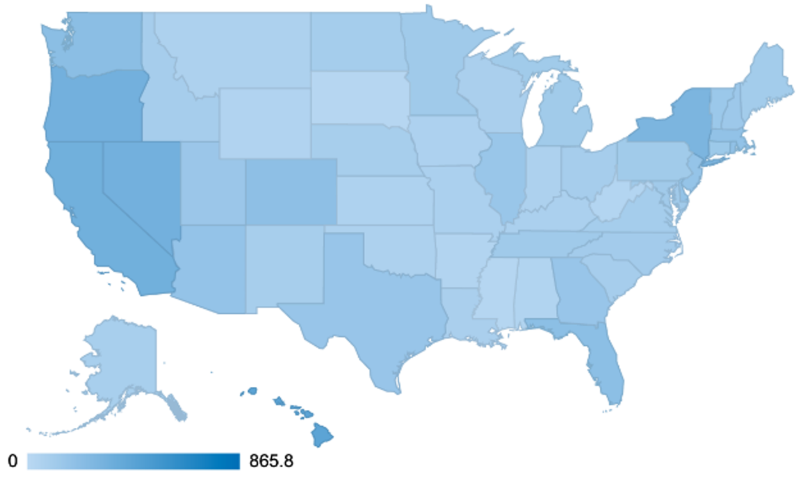
Author’s Location: State Level

- In this random sample, more than 20% of the posts are from California, followed by New York and Texas.
- However, California is the state with largest population, which makes it always the top state in any social media conversation. Therefore, I calculated an indexing suggests the social media posts compared the population and the post, which indicates that vegan social media conversation are more likely to happen in the states on west coast.

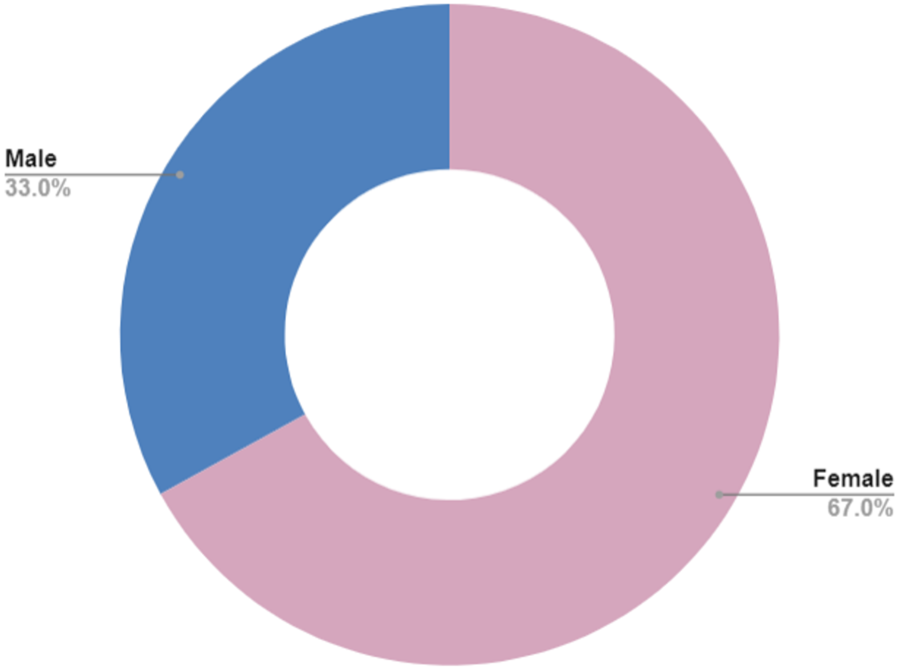
Share



Indexing Score



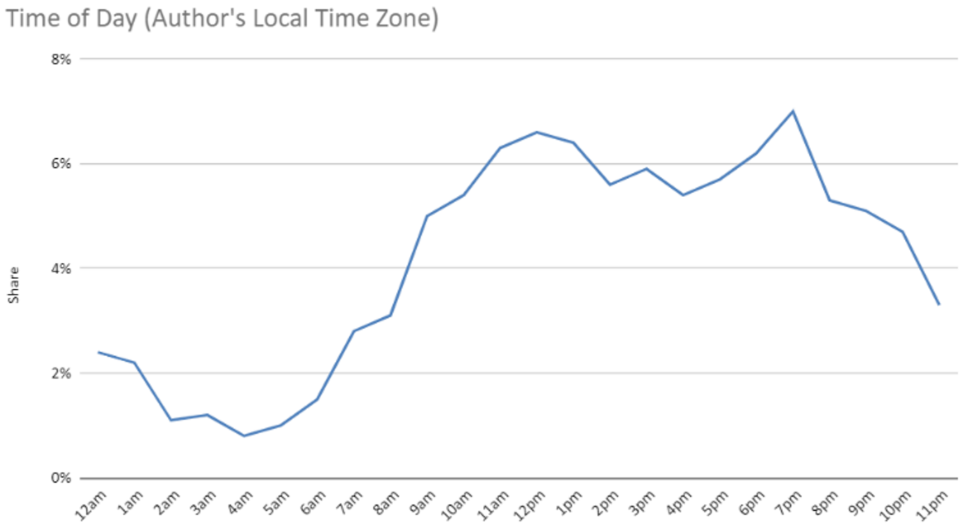
Gender Overview

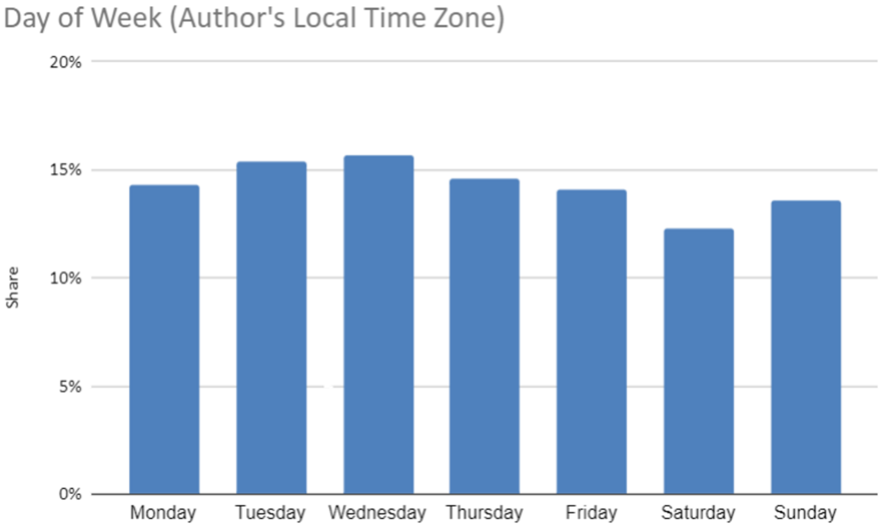


- In this random sample, more than 2/3 of the conversation are from female audience, which indicates that women conversation contributors are more likely to share and post about vegan on social media.

Posting Time

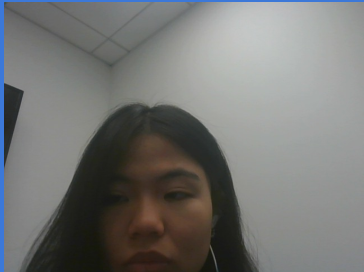
- In this random sample, most of the conversation about vegan is published at 7pm and 12pm, the lunch and dinner hours.
- Tuesday and Wednesday see the largest share of vegan social media posts.





03

Model Building



Statistical Analysis - Regression Model

- **Independent Variables:**
  - Sentiment
  - Author's gender
  - Author's profession
  - Author's geo
  - Author's number of followers
  - Post time: Day
  - Post time: Day part
  - Post Type

- **Dependent Variables:**
  - Number of Engagements

Model Test

The data was split into train and test data sets. The train dataset contains 80% of the total data and the test data set contains 20%.

```
model1<-
lm(Engagements~westcoast+sentimentNegatives+sentimentPositives+sentimentNeutrals+weekend+Author_Gender+daypart1+daypart2+daypart3+daypart4+video+followers, data=train)
```

Model Summary:

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 100.4 on 885 degrees of freedom
  (29 observations deleted due to missingness)
Multiple R-squared:  0.4618,	Adjusted R-squared:  0.415
F-statistic: 9.862 on 77 and 885 DF,  p-value: < 2.2e-16
```

Test Result

- **Independent variables that are significant related to dependent variables (p-value < .05)**
  - Daypart
  - Followers
  - Professions

- **Independent variables that are not significant related to dependent variables (p-value > .05)**
  - Weekday
  - Author_Gender
  - Westcoast
  - Sentiment
  - Media_Type

Model Update

```
> model2<-lm(Engagements~Author_Gender+daypart1+daypart2+daypart3+daypart4+followers+Professions,
data=train)
```

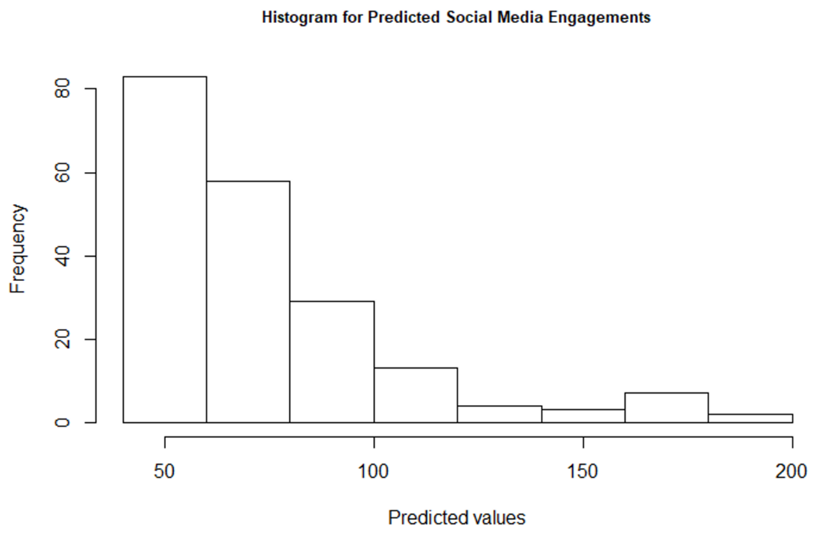
```
Residuals:
    Min      1Q  Median      3Q     Max
-162.08  -32.99  -11.97   14.21 1830.28

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
daypart1                         57.316282  10.392256   5.515 4.83e-08
daypart2                         47.220880   6.547399   7.212 1.38e-12
daypart3                         35.105509   9.350910   3.754 0.000188
daypart4                         45.604868   8.632313   5.283 1.68e-07
Followers                         0.028866   0.003514   8.215 9.59e-16

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 97.26 on 731 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared:  0.4507,    Adjusted R-squared:  0.4222
F-statistic: 15.79 on 38 and 731 DF,  p-value: < 2.2e-16
```

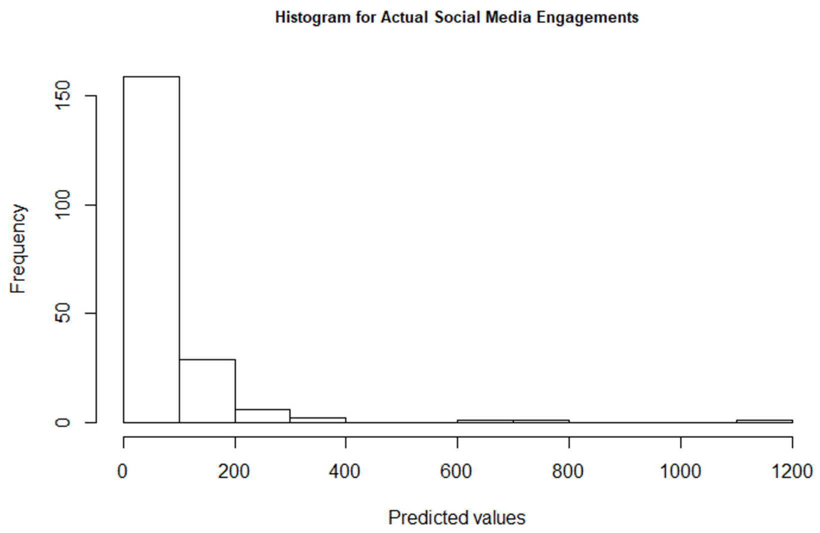
04

Results Review

Make Prediction on Test Dataset

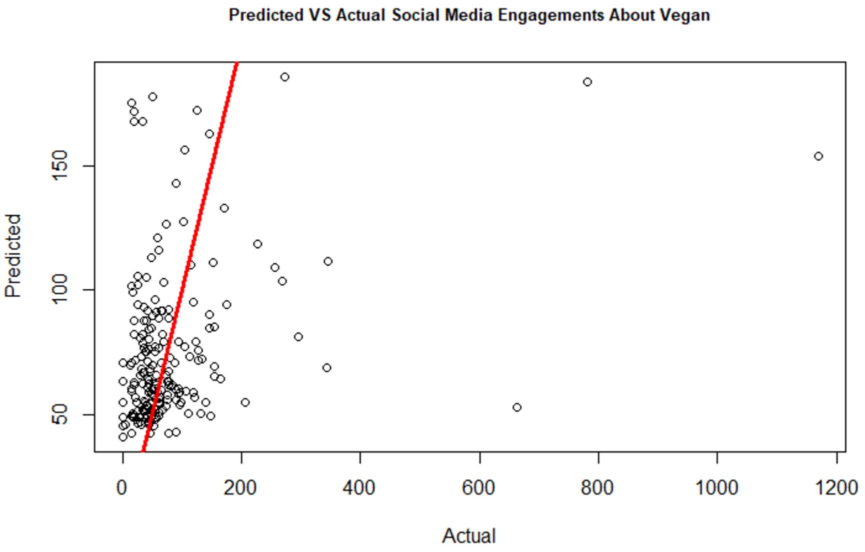
- The left chart shows the predicted results on the test dataset
- The right chart shows the actual engagement numbers of the social posts around vegan

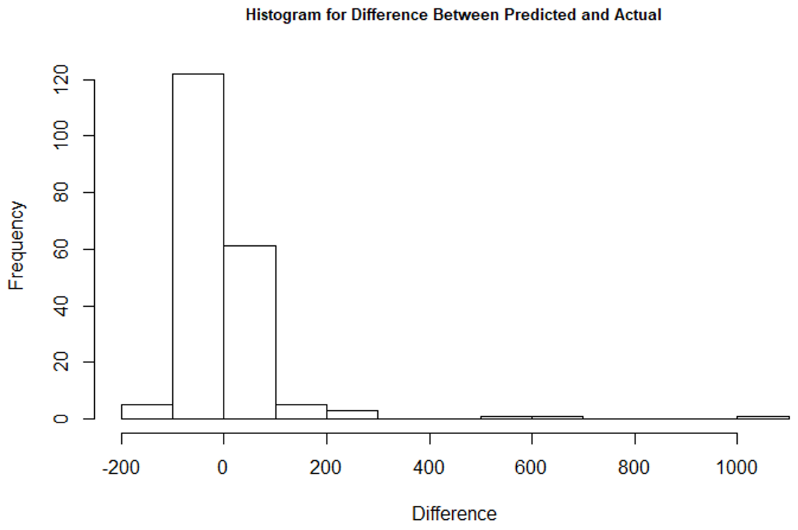




Make Prediction on Test Dataset

- From the chart we can tell that in most of the cases, the predicted numbers of engagements are seeing a difference between -100 to +100.





05

Conclusion and Discussion

Conclusion

- As a social media analyst, my daily job is using social listening tools, mainly NetBase, to collect and analyze social media data. However, this is the first time I tried applying the quantitative method in social media analysis.
- According to validation, the linear regression model supports part of the hypothesis that the demographic factors (professions in creative arts), follower numbers and posting time are significantly related to the engagements of the social media posts.
- Overall, follower numbers have the most significant impact on the number of engagements.
- The posts related to vegan see the highest average engagements in the morning hours (87).
- Suggestions to marketers:
  - They could work together closely with social media influencers
  - Boost their marketing social media posts during the morning hours will see a less crowded post poll but could be more engaging among their target audience.
- Limitations:
  - For the regression model, the R square is lower than 65%, which is a little bit blow the industry benchmark.
  - The sample size is only 1000, which is slightly too small for social media analysis.
  - The sentiment score algorithm still expects the improvements.

06

Appendix

Appendix

Data Source:

Date Range: 1/1/2018 12:00 AM - 12/31/2018 11:59 PM (GMT-04:00) New York

Sources: Twitter, Instagram

Post Types: Original

Followers/Visitors: 10 - 5K

Reference:

- Andy Bromberg: http://andybromberg.com/sentiment-analysis/
- Tidy Text Mining: http://tidytextmining.com/sentiment.html
- Julian Hillebrand: Create Twitter Sentiment Word Cloud in R
- Veera Raghava Reddy: Sentiment Analysis Using R Language
- Bing: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
- AFINN: http://www2.imm.dtu.dk/pubdb/views/publication_details.php
- Luc Gendrot: Sentiment Analysis in R

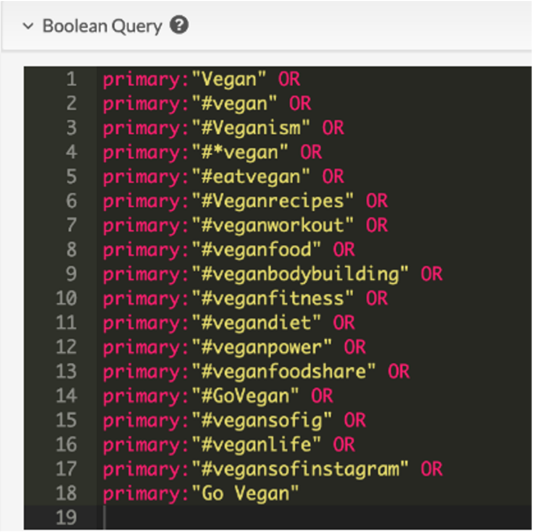
Tool used to collect data: NETBASE

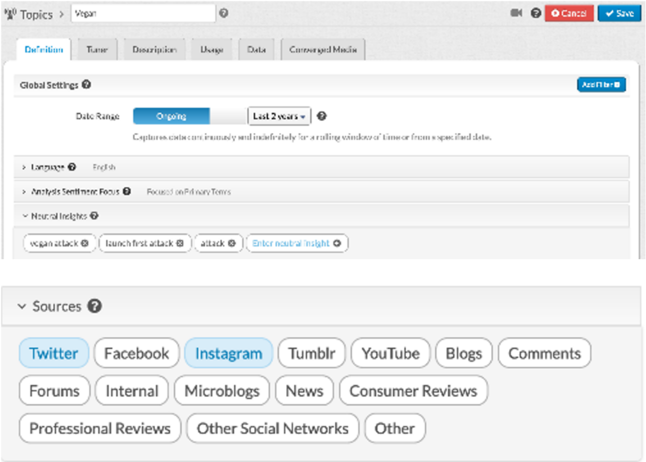
[NetBase] is the social analytics platform that global companies use to run brands, build businesses, and connect with consumers every second. Its platform processes millions of social media posts daily for actionable business insights for marketing, research, customer service, sales, PR and product innovation. NetBase is recognized by analysts and customers as the leader in Social Analytics. It was also named by Forrester as an Enterprise Leader.

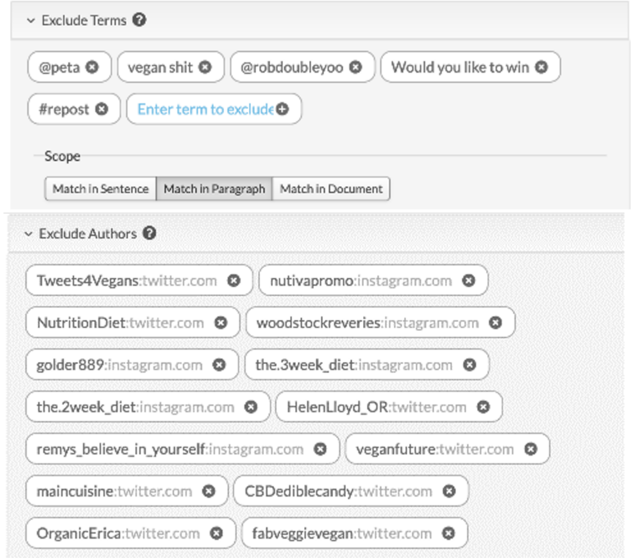
＊ The brief introduction is from the offical website of NetBase

Build a topic to capture social media posts related to vegan

Create the Boolean Query with keywords → Define the Time Range and Channel → Exclude some spamming posts and authors





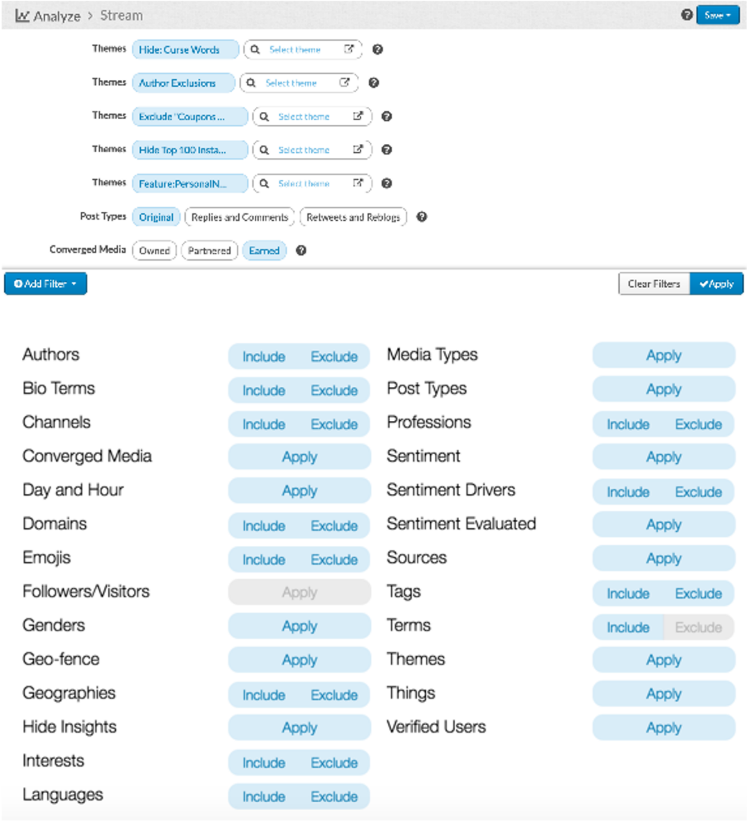


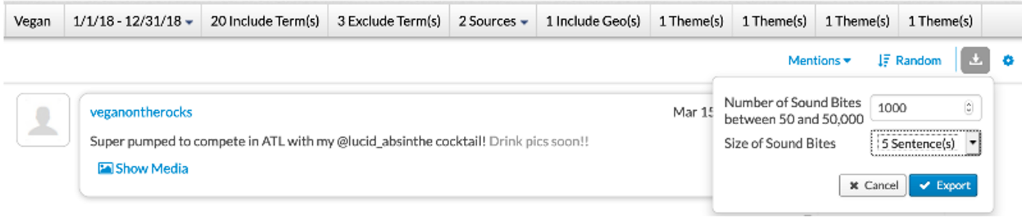
- Netbase provide direct access to social posts / tweets.
- To capture conversation, a topic is created. A topic contains: a boolean query with keywords; the date range of the conversation; the social channel; the spamming accounts and tweets need to be excluded.
- After the topic is created, NetBase will pull the public social media posts which match all the conditions.

* The brief introduction is from the [offical website of NetBase]

Clean the topic and pull a random sample of 1,000 tweets

Clean the top with filters → Select and download a random sample of 1,000 tweets





- NetBase utilizes natural language processing (NLP) to track and analyze sentiment, passion, behavior and more around the topic keywords.
  - It assigns each tweet a score, which shows whether it is positive, negative or neutral.
- By using the “analyze” function, we can keep adding more filters to clean up the social media conversation. In this case, I removed the retweets and reposts, and look at only original tweets and Instagram posts.
- How to create a dataset for further analysis:
  - Go to the Stream section → change the way of ranking to “random” → click ⬇ button → change Number of Sound Bites to “1000” → click “Export” → save the csv file in the laptop.

Tokenize the data and build corpus for frequency distribution

```
> #Create corpus
> corpus = Corpus(VectorSource(tweets$Tweet))
> #Convert to lower-case
> corpus = tm_map(corpus, tolower)
> corpus = tm_map(corpus, PlainTextDocument)
> #Remove punctuation
> corpus = tm_map(corpus, removePunctuation)
> #Remove stopwords and apple
> corpus = tm_map(corpus,
removeWords,stopwords("english"))
```

- As we all know, social media data contains a lot of elements other than plain text.
- These elements, such as @, # and other punctuations, won't contribute to the sentiment analysis. Therefore, before analyzing the tweets, we need to remove them.
- Stopwords is also causing a lot of problems, especially in frequency distribution analysis.
- Emojis are contributing to the sentiment. However process of analyzing emojis is very complicated. As a result, they were removed.
- By using the tm package, I remove both the punctuations and stopwords. The tm_map() function is predefined transformations (mappings).

Figure out Sentiment: Create word polarity list

Load up word polarity list and format it

```
> afinn_list <- read.delim(file='AFINN-111.txt', header=FALSE, stringsAsFactors=FALSE)
> names(afinn_list) <- c('word', 'score')
> afinn_list$word <- tolower(afinn_list$word)
```

categorize words and add some additional words

```
> vNegTerms <- afinn_list$word[afinn_list$score==-5 | afinn_list$score==-4]
> negTerms <- c(afinn_list$word[afinn_list$score==-3 | afinn_list$score==-2 |
afinn_list$score==-1], "second-rate", "third-rate", "boring", "disgusting", "senseless",
"confused", "disappointing", "not surprising", "silly", "tired", "predictable", "stupid",
"uninteresting", "trite",  "outdated", "dreadful", "bland","break","leak","died-battery","not
work","stop
working","short","risky","unsafe","problem","messup","hacked","struggle","unremarkable","un
amazing","overrated","unnecessary","unremarkable","pointless","unnecessary","groupies")
> posTerms <- c(afinn_list$word[afinn_list$score==3 | afinn_list$score==2 |
afinn_list$score==1], "first-rate", "insightful", "clever", "charming",  "enjoyable", "absorbing",
"sensitive", "powerful", "pleasant", "surprising","high-quality","long battery life","working","
safer"
,"easier","cool","effective","fast","trendy","durable","clever","deluxe","testament","light","spe
edier","excited","sleek")
> vPosTerms <- c(afinn_list$word[afinn_list$score==5 | afinn_list$score==4], "uproarious",
"riveting", "fascinating", "dazzling", "legendary","best","highest quality","revolutionary")
```

- Referring to Andy Bromberg's blog post, I found the AFINN word list, which has 2477 words and phrases rated in a scale from -5 [very negative] to +5 [very positive].
- Andy Bromberg reclassified the AFINN words into four categories (3):
    - Very Negative (rating -5 or -4)
    - Negative (rating -3, -2, or -1)
    - Positive (rating 1, 2, or 3)
    - Very Positive (rating 4 or 5)
- I applied his classifying way, and also added in a few more words specific to Apple and iPhone (from NetBase and here) to round out my wordlist.
- Andy Bromberg chose to ignore neutral words. I agree with him. In my daily, we also focus more on the tweets with sentiments.
- The number of words in each tweet that fit one of those four categories: sentence | #vNeg | #neg | #pos | #vPos | sentiment

Create the function to calculate number of words

- Before we proceed with sentiment analysis, a function needs to be defined that will calculate the sentiment score.
- I used [Veera Raghava Reddy](#)’s and [Andy Bromberg](#)’s blog posts as reference in creating the code.
- The code on the right showcases how sentiment analysis is written and executed. The code will assign each posts a score.

```
> sentimentScore <- function(sentences, vNegTerms, negTerms, posTerms, vPosTerms){
+   final_scores <- matrix('', 0, 5)
+   scores <- laply(sentences, function(sentence, vNegTerms, negTerms, posTerms, vPosTerms){
+     initial_sentence <- sentence
+     #remove unnecessary characters and split up by word
+     sentence <- gsub('[[:punct:]]', '', sentence)
+     sentence <- gsub('[[:cntrl:]]', '', sentence)
+     sentence <- gsub('\\d+', '', sentence)
+     sentence <- tolower(sentence)
+     wordList <- str_split(sentence, '\\s+')
+     words <- unlist(wordList)
+     #build vector with matches between sentence and each category
+     vPosMatches <- match(words, vPosTerms)
+     posMatches <- match(words, posTerms)
+     vNegMatches <- match(words, vNegTerms)
+     negMatches <- match(words, negTerms)
+     #sum up number of words in each category
+     vPosMatches <- sum(!is.na(vPosMatches))
+     posMatches <- sum(!is.na(posMatches))
+     vNegMatches <- sum(!is.na(vNegMatches))
+     negMatches <- sum(!is.na(negMatches))
+     score <- c(vNegMatches, negMatches, posMatches, vPosMatches)
+     #add row to scores table
+     newrow <- c(initial_sentence, score)
+     final_scores <- rbind(final_scores, newrow)
+     return(final_scores)
+   }, vNegTerms, negTerms, posTerms, vPosTerms)
+   return(scores)
+ }
```

Build tables of positive and negative sentences with scores

Build tables of positive and negative sentences with scores

```
> posResult <- as.data.frame(sentimentScore(posTweet, vNegTerms, negTerms, posTerms,
vPosTerms))
> negResult <- as.data.frame(sentimentScore(negTweet, vNegTerms, negTerms, posTerms,
vPosTerms))
> posResult <- cbind(posResult, 'positive')
> colnames(posResult) <- c('sentence', 'vNeg', 'neg', 'pos', 'vPos', 'sentiment')
> negResult <- cbind(negResult, 'negative')
> colnames(negResult) <- c('sentence', 'vNeg', 'neg', 'pos', 'vPos', 'sentiment')
> results <- rbind(posResult, negResult)
```

Combine the positive and negative tables and check the display

```
> str(results)
'data.frame':   844 obs. of  6 variables:
 $ sentence : Factor w/ 833 levels "??????? @apple come back for new twit's",..: 179 205 220
256 2 25 234 263 108 209 ...
 $ vNeg     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ neg      : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ pos      : Factor w/ 5 levels "0","1","2","3",..: 3 2 2 3 2 2 2 2 2 1 ...
 $ vPos     : Factor w/ 3 levels "0","1","2": 2 1 1 1 2 3 1 1 1 2 ...
 $ sentiment: Factor w/ 2 levels "positive","negative": 1 1 1 1 1 1 1 1 1 1 ...
```

- The code on the left shows how each tweet gets their new sentiment score, after analyzed by AFINN word list.
- The new positive and negative results are stored in a new dataset called “results.”
- By using the str() function, we can see what and how many variables entries are there in the new dataset.