Trend Prediction Based on Social Media Data:

A Case Study on Veganism

Runxi Wang

Applied Project in Analytics

Table of Contents

Abstract

Social media is currently considered as an important resource to extract customer and industry intelligence. Analyzing conversation online helps companies and business to respond fast to the trends in the industry. However, the data captured on social channels is unstructured, while no standardized social media analysis and prediction making process has been presented. This paper discussed extracting food general trend from social media, with the case study on social media data of Veganism. Veganism is identified as one of the top food trends and the most popular food diet by marketing experts, this project used the content analysis frame work to first organize and structure a sample of 1,000 posts collected from two major social media platforms Instagram and Twitter. After the content cleaning, the sample size was reduced to 714. K-mean Clustering and Multiple Linear Regressions were applied quantitative analysis methods. The research indicated that the author's professions, posting time and the number of followers were significantly related to the number of engagements. Result also suggests that the posts in the morning got more engagements than those posts published in other time. Breakfast was the most frequently mentioned meal among the vegan social media conversation. This research project provides marketers directional suggestions on their communication with customers. It also developed a standardized method of doing analysis on social media.

Keywords: Food Trend, Social Media, Cluster Analysis, Linear Regressions, Content Analysis

Introduction

Social media is generally playing an increasing important role in marketing and advertising industry (Brandwatch, 2018). Currently, it is considered as an important resource to extract customer and industry intelligence (Brandwatch, Pi Datametrics, 2018). Analyzing conversation online helps companies and business to respond fast to the trends in the industry, and even help them to make decisions in not only marketing, but also customer service and even product research and development (R&D) (Brandwatch, 2018). However, not as search engine marketing, there is no direct way to measure how the social media data is connected to consumer behavior. Researchers are still exploring how people talk about a specific trend and what they actually purchase. In other words, the correlation between social media trends and product sales still needs to be tested and evaluated.

When it comes to food industry, food, diet and fitness-related content has gained significant popularity among social media users (Gemma J. , 2018). Veganism, the practice of abstaining from eating animal in diet and using or containing any other animal products, is identified as one of the top food trends and the most popular food diet by marketing experts from Brandwatch, a social media monitoring company (Gemma J. , 2018). The people who followed this type of diet or life style, is known as vegan (Gemma J. , 2018). In this case study, I hypothesize there is a correlation between conversation on social media (Twitter, Instagram, Forums, and Blogs) about Veganism and the actual sales of vegan food products.

This case study has three major contributions: 1) a method of gather, clean and process large unstructured social media data, 2) a method of integrating the social media data into an improved prediction model for marketing and product R&D decision making, 3) a method of evaluation social media prediction model, using first- and second-party sales data. These methods and models can also be applied in making prediction of other consumer / human behaviors using social media data, and influencing not only commercial decisions, but also public or political decisions.

The data used in this research is collected from different sources: 1) user-generated social media conversation is collected by NetBase, a social media analytical tool, using key word based Boolean query from Twitter, Facebook, Instagram, YouTube, Forums and Blogs: dataset will include textual content, image and video content, posting time (if available), geo-location (if available) and demographic information (if available). The date frame will be the year of 2017, from January to December 2) Vegan food sales data: Amazon, Instagram Buying and other e-commerce sites ranking, as well as Mintel and other industry sales reports.

This research is facing several challenges and problems, especially in collecting representative data. Firstly, the social media is of fluid and rapidly changing nature (Pila E. , Mond, Griffiths, Mitchison, & Murray, 2017). The size of conversation and access of posts I need to collect data will get affected by the algorithm updating and social media platforms (for example, Twitter and Instagram) authority. Secondly, the social media data collected for analysis is not necessarily representative of all vegan consumers. Nowadays, consumers are having serious concerns about their privacy. As a result, some of them have changed account setting to avoid posting anything publicly. Visualization elements are important on social media. The content on vegan without mentioning term “vegan” or any relevant verbal terms will be difficult to collect and examined in the study. Thirdly, sales data is usually considered as confidential information by some companies and business. Alternatively, I’ll use the e-commerce sites ranking data and sales data from Mintel and similar type of industry reports as data sources.

However, these data probably will not be as accurate as the first-hand data directly collected from food retails and companies.

To overcome the problem, I plan to test several Boolean queries, to make sure the queries I use are as inclusive as possible. In addition, I will also reach out to food experts in Allrecipe.com for advice. Further, the social media data collecting tool used in the research, NetBase, provides data from various sources, which is considered as one of the best in the industry. Basic search data from Google Trends will also be utilized as reference. I'll also evaluation the sales data I collect first, before syndicating them. All these efforts will be contributing to a representative dataset of actual consumer behavior.

Data processing tasks will also be a challenging task in this research project. First of all, social media data always come in large volume and unclear structure. A clear and straightforward coding procedure will be necessary to categorize social media posts and "identify tags or labels (i.e., hashtags)." In addition, "the popularity of online information exhibits rich temporal variation and follows various evolution patterns (Hu, Hu, Fu, Fang, & Xu, 2017)." Not many standardized methods provided by previous research are available to identify the trending items and topics in the vegan conversation and measure "how trendy" they are. Lastly, to approach the online and offline consumer behavior, this research will use data from various sources. They come in different structures, both categorical and numerical. How to syndicate data together and link them, will be another problem.

Literature Review

The previous research and papers on social media data analysis can be classified into three major categories: 1) reviews of tools and theories; 2) qualitative analysis and 3) quantitively modeling and prediction.

Research Tools and Theoretical Review

Recently, analyzing and predicting social media has become an increasing important research topic. The tremendous size of users' demographic and behavior data are available on social media, which provide a new source and opportunities of information people can use to do research and predict the future events (Schoen, et al., 2013). People in both academic and industry are exploring an optimized methodology of analyzing social media conversation and figuring out consumer behavioral insights. The first step of doing any kind of analysis is choosing tools. However, in each step from data scraping, collecting to presentation, there are a large number of software tools. The social media analysis and prediction also need theoretical support and foundation. The research performed by Batrinca and Treleaven (2014) presented the available sources of social media data, how to collect and store the social media data, how to clean, sentiment analyze and visualize the data. They The researchers reviewed a large number of software tools available in the market. These tools were grouped into various number of sets: Social media data (platforms and formats); Social media programmatic access; Text cleaning and storage tools; Text analysis tools; Social media platforms (Batrinca & Treleaven, 2014).

Another problem all the researchers who try to analyze social media data faced was the lack of widely accepted evaluation process and research fragmented (Schoen, et al., 2013). Batrinca and Treleaven's work also provided a comprehensive process and methodology of analyzing social media conversation. Schoen, Gayo-Avello and their team (2013) explored the current models, including Prediction market models, Survey models and Statistical models, and their adaptation to the special circumstances of social media data analysis (influenza incidence, product sales, stock market movement, and electoral results). While in the process of looking forward to a more optimized social media analytical model, statistical models were considered the most reliable models for now (Schoen, et al., 2013) after the researchers review the relative advantages and disadvantages of the current existing models. Model predicting social media's competitive advantages include: it can lower the cost; models can overcome the human error of overvaluing small probabilities and undervaluing high probabilities; models can avoid making decision based on human bias; it also has the ability of processing huge size of data and responding fast (Yu & Kak, 2012). The social media predicting models can be used in human related and public events or topics, such as Marketing, Movie box-office, Information dissemination, Elections and Macroeconomic (Yu & Kak, 2012). By mining the attributes and contents of social media, Yu and Kak (2012) provided both theoretical support and practical technical suggestions for social media trend prediction. They advised to use sentiment metrics, time series metrics, terminology, density, degree, centrality and structural hole as predictors and regressions, Bayes classifier, K-mean clustering, Decision Tree and Artificial Neural network as model bases.

Since social media launched in marketing playground, marketers started to debate that whether social media had an impact on consumers behavior, because they do not directly lead to sales increase. The effectiveness of social media marketing needs to be aligned with people's various needs of social media, because of the unique characteristics of social media: "nature of connection (profile-based versus content-based) and level of customization of messages (broadcast versus customized) (Zhu & Chen, 2015).” In their research, they summarized several categories of needs of social media users: the first need was building and developing relationship; the second one is self-media, which let marketers to better leverage the power of celebrities; the third one was collaboration, which provide marketers and business a possibility to target their customers better. Their work provides a theoretic support for the social media predicting research.

Qualitative Analysis

The investigation on peer communication on social media and its pact on purchasing behavior (Wang, Yu, & Wei, 2012) provided theoretical bases for the consumer behavior prediction and social media marketing and adverting effort. They applied existing measures for peer communication, product involvement, product attitudes, purchase intentions, and consumer’s need for uniqueness, the moderating variable (Wang, Yu, & Wei, 2012) as research methodology. The research outcome supported the hypnosis that peer communication on social media had a significant impact on consumers’ attitude and purchasing behaviors, in both direct and indirect ways.

Content analysis is a commonly applied qualitative method in social media analysis. In 2016, Pila, Mond and their team’s research provided a standardized way to approach the qualitative analysis, especially content categorization and sentiment analysis on social media. They did a case study on #cheatmeal images on Instagram. The team applied a thematic content analysis method, “extracted more than 1.6 million images marked with the #cheatmeal hashtag on Instagram, coded the photographic and textual elements of a sample (n5600) (Pila E. , Mond, Griffiths, Mitchison, & Murray, 2017).” In the categorized content, the researchers found that the

Instagram content with #cheatmeal exemplified the idealization of overconsumption, a strict commitment to fitness, and a reward-based framework around diet and fitness. The research is also a good combination of qualitative and quantitative research. A quantitative test was also provided as an additional evaluation. The analysis explored the correlation between the demographic variable and people's eating behavior presented on Instagram.

Vaterlaus, Patten and their team explored the influence of social media on people's health behaviors in 2015. Their research was mainly focused on young adults (18-25 years old) and the impact of social usage and information on the health behaviors, such as diet and exercise, of young adults. Exploratory qualitative study was the major research method. The team recruited 34 young adults as participants for eight focus groups and interviewees for individual interviews from five courses at an American midwestern university (Vaterlaus et al., 2015).

From the qualitative analysis, Vaterlaus, Patten and their team summarized three themes describing the connection between social media and young adults' diet and exercise. Firstly, young adult took social media as a source to explore food choices. Secondly, young adults considered social media as a platform of self-presenting, by showing the food they ate or prepared on social, as well as how they exercised in daily life. However, social media also impacted young adults on their eating and exercising habits negatively. The participants acknowledged that social media interrupted their face-to-face communication while they are eating or practice. The inaccurate information on social media might also lead to poor food/diet choices (Vaterlaus et al., 2015).

This research showed the influence of social media on young people's health habit, such indicates that social media is an important way to learn people, especially youth's behavior. It also provides evidence of the effectiveness of food marketing effort. The qualitative research method used in this paper will also be applied in my research, to provide suggestions for the research direction.

Holmberg, Chaplin and their team (2016) did a study aiming to explore how adolescents (teenagers) communicate food images on Instagram. The research team examined how and in what context food was presented and the type of food items that were portrayed by following the hashtag #14år (“14 years”) on Instagram. The research team created a categorizing frame of food items that were shared by youth. They also explored how food was displayed, and the context of food were shared. Instagram is a platform allows user to share both visual and written content. Therefore, the researchers also used the text (captions and hashtags) as a context of the shared pictures (Holmberg, et al., 2016).

The paper indicated the type of pictures teenagers presented on Instagram most frequently: 1. the “aesthetical features or home-made qualities of the food”; 2. food as a part of a lifestyle (Holmberg, et al., 2016). About 50% of the images showed the brand name clearly, which indicated that the way teens posted food on social media were highly impacted by advertising or food marketing campaigns.

The result of this research proved the influence and effective of the marketing effort of the food companies. They also indicated that the images were an important part of a lifestyle that the young people and is a powerful to communicate to youth on social media. The method to practice the content analysis on social media topics, from data collection to categorizing, was applied in my research project.

Quantitative Modeling and Prediction

Statistical models were considered the most reliable models for now (Schoen, et al., 2013).

Sentiment analysis can also be approached through quantitative methodologies. Kang, Yong and their team (2016) did was focused on the methodology of opining mining and sentiment analysis of social media conversation. In their case study of cosmetic industry, they developed a new framework to cluster the brand names and performance on social media, especially Twitter. “The distance between two brand names is in inverse proportion to the frequency of co-mention (Kang, Yong, & Hwang, 2016).” Two brands were considered similar or alternative as a “pair” by customer if their pair appeared frequently. MDS (Multi-Dimensional Scaling) was used to project brand names on to 2-dimensional and a 3-dimentional space (Kang, Yong, & Hwang, 2016). This study provides a practical approach to clustering social media brands and can be applied in the real-life marketing strategy development and customer relationship management.

Statistical models commonly used in other fields are also applied and tested in the social media analysis and prediction. The model developed by Altshuler, Pan, and Pentland (2011) was based on information diffusion models. The researchers stated that their model provided “a capability of predicting future trends based on the analysis of past social interactions between the community’s members (Altshuler, Pan, & Pentland, 2011).” They used G (graph), V (community’s members), E (social links among community members) and pattern (dependent variable) as variables (Altshuler, Pan, & Pentland, 2011). This model report focused on the developing theoretical work on the prediction of the likelihood of how a topic or small pattern from some specific social media community spread out of their network and become a trend.

They found out that the maximal rate of global adoption of a “trend,” or so called the “maximal outreach,” is dominated by the topology of the network, and the local adoption features (Altshuler, Pan, & Pentland, 2011). They explored the variables which impacted the trending process and built an analytics model of social diffusion dynamics to predict the trends.

In 2016, Ying Hu, Changjun Hu and their team developed a model in predicting and evaluation popularity of online information. To identify the key events in popularity evolution, research team also developed universal method for different patterns. “Two directions of popularity evolution: popularity evolution patterns and popularity evolution prediction” were related to their research (Hu, Hu, Fu, Fang, & Xu, 2017). By analyzing a large sample of popular and trending hashtags on Twitter, they designed a new evaluation metric to evaluation the “burst, peak and fade” social conversation circle and found most popular topics “bursts suddenly, peaks very soon, and then fade quickly (Hu, Hu, Fu, Fang, & Xu, 2017).” Figuring out when the three key events (burst, peak and fade) occur, will contribute to a more effective marketing system and strategy.

The work by Mihuandayani, Ramandita, Setyanto, and Sumafta (2018) is a recent example of applying both qualitative and quantitative methods, using social media data to explore the social media’s impact on real life purchasing and make predictions. The hypothesis they offered was associations between social media food conversation and offline restaurant sales and orders exist. K-mean clustering and Simple Additive Weighting (SAW) were the analyzing methods they used. Data were collected from two sources: Twitter’s Application Programming Interface (API) and company (restaurant) sales (Mihuandayani, Ramandita, Setyanto, & Sumafta, 2018).

The whole research process was following six steps. After the data were collected, the researchers “cleaned, featured extraction, did clustering analysis, ranked and evaluated the results (Mihuandayani et al., 2018).” The sales data collected from observation from randomly sample restaurants were used to evaluate the accuracy of the predicted trending food. After the comparison, the researchers got the accuracy for four weeks prediction as 0.81, 0.72, 0.75 and 0.63, which arrived in an average accuracy of 0.7275. In sum, the result of data analysis supported the hypothesis, which suggested there was an association between the food trend prediction from social media and popular dishes sold in restaurants. Such result indicates that the algorithms used in this paper is practical in defining food trends, and the food trends on social media are consistent with the food trends offline.

The research provides a method of how to use social media data to predict the actual food trends in local restaurants and companies. In addition, the analyzing and evaluating methodology would also be applied in social media data analysis in other categories.

Fried and his team (2015), stated that it also reflected people’s identity and personality. The researchers explored the predictive power of food conversation on social and investigated the method to use such conversation to predict people’s “location, likelihood of diabetes, and political preferences (Fried, Surdeanu, Kobourov, Hingle, & Bell, 2015). ”

The team gathered over three million food-related tweets as the original dataset, then developed a Natural Language Processing (NLP) model to explore the connections between the language people used to describe food, their geographic locale, and author’s political views and community characteristics. They also claimed that their complex NLP model language-based models performed better than the major research baselines. The research also provided a practical solution to collect and visualize real time tweets and social media data. “Geo-referenced heatmaps, semantics-preserving word clouds and temporal histograms (Fried et al., 2015)" were implemented to investigate the language people used in food related tweets.

The researchers, Fried and his team, suggested that the information can be used for a variety of purposes, from public health communication to brand marketing and customer targeting. They provided a practical method and complete process of doing data analysis using social media conversation, from data collection, cleaning, visualization, model building to prediction making (Fried et al., 2015).

Yoo, Song and Jeong proposed their system, Polaris, for analyzing the predicting sentiment of media messages about real time events. The data was collected from Twitter API for a month-long period. Then Yoo and the research team utilized the method of weighted values to analyze social media content and explored the sentiment of the expression. "Sentimental paths were predicted by analyzing the sentiment of the contents for keywords for a certain event (Yoo et al., 2018)." The analysis also took geo location into consideration. By analyzing the sentimental path, "the place where a certain event is expected to occur, and the sentiment are predicted and shown in advance (Yoo et al., 2018)." The team proposed the sentiment path prediction model, could be widely used in different situations of prediction, from disaster notice service to social events organizing and marketing programs.

Xu and his team (2015) presented Social-Forecast, a model to make prediction based on timeline data to a forecast the popularity of videos promoted by social media. "Dynamically changing and evolving propagation patterns of videos in social media (Xu, Van Der Schaar, Liu, & Li, 2015)" are two major difficulties in making predictions using social media data. The Social-Forecast model took these two problems into consideration and provided a solution for making popularity forecasts, by exploring the correlation using context information of the video.

The model's accuracy was tested with the video data collected from a large Chinese social media website. The test result showed that the method with context of video performed better in popularity prediction than the existing view-based approaches by 30% (Xu et al., 2015).

The team suggested that their methodology and algorithm can be easily adapted to predict other trends in social media. The food marketer can utilize their model to identify popular or relevant food content on social.

Linear Regression is also commonly used as a modeling method to explain performance and make predictions on social media data. Coyne and his team (2017), used the social media posts collected from StockTwits, a social media platform for investors, to build model and predict individual stock prices. The multiple linear regression model was one of the three models they built. It was chosen because of linear regression's simple and low risk of over fitting. The model was tested on Apple ($AAPL) with an average accuracy came out to 52.45% (Scott Coyne, 2017). They also stated that many stocks did perform well, over large periods of test data. Their method of handling user information (number of likes, follower count and how often the user is right about a stock) provides me a solution of processing demographic (Scott Coyne, 2017).

Çizmeci and his team (2018) explored the use of Factorization Machines approach to predict movie success by predicting IMDb ratings for newly released movies using social media data. They also created linear regression as a baseline model (Beyza Çizmeci, 2018). Their linear regression model was built base on the genre, country, runtime, director and actor metadata features to predict IMDb movie rating.

Methodology

As being depicted in Fig. 1, the general process of this research consists of seven steps. Data collection, data cleaning, coding and text analysis, clustering, ranking, evaluation, data visualization and strategy suggestion. The deliverable includes both output (the accuracy of the food trend prediction compared to the sales data collected from industry report, newspaper and e-commerce sites) and outcome (the veganism food trends based on social media).

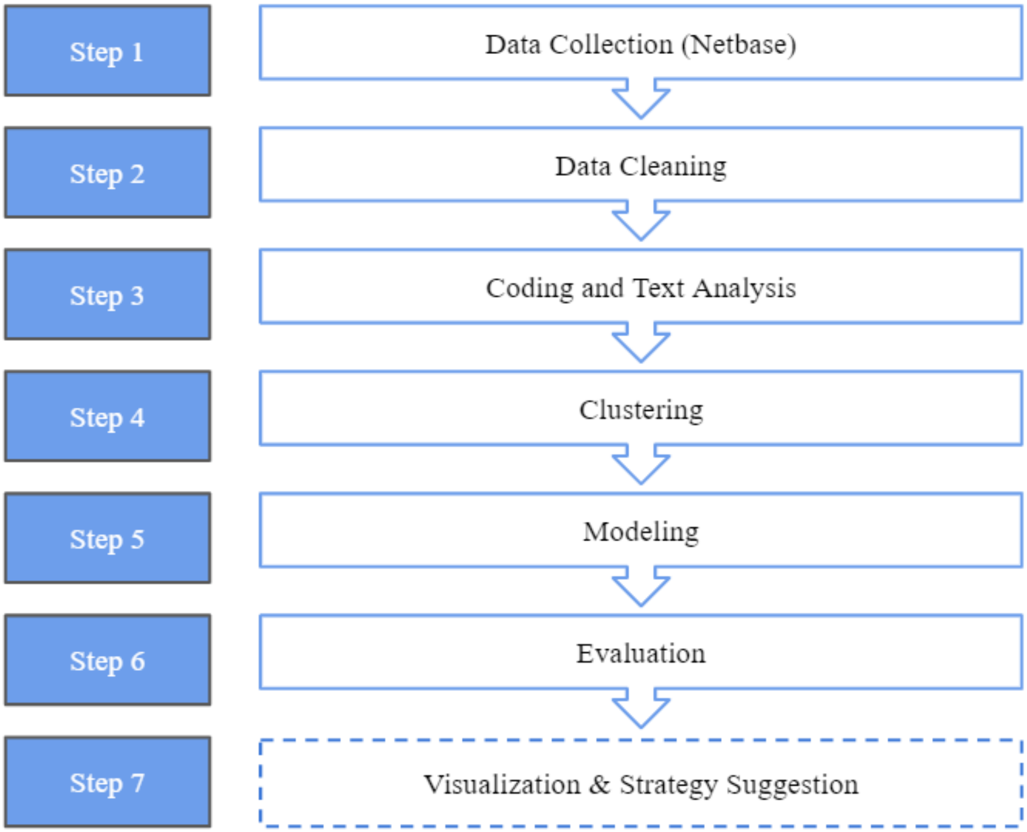

Figure 1 Data Collection, Cleaning and Analysis Process (Mihuandayani, Ramandita, Setyanto, & Sumafta, 2018)

Research Methods and Design

A great number of researchers have tried to model, analyze social media behavior, and predict the social media trends.

Qualitative research methods play an important role in figuring out the influence of social media. Vaterlaus, Patten and their team explored how social media impact people's health behaviors in 2015. They invited 34 young adult participants at a midwestern university in the United States to participate in the research and put them into eight focus groups (Vaterlaus, Patten, Roche, & Young, 2015).

Content analytic approach is commonly employed in the analysis of Instagram images. Content analysis is a replicable method for analyzing both visual and written content (Holmberg, E. Chaplin, Hillman, & Berg, 2016). Most of the social media content is multimedia, which containing both visual and written content. Therefore, it's necessary for me to consider both the images and written descriptions provided by the social media conversation contributors.

Holmberg, Chaplin and their team (2016) explored how adolescents (teenagers) communicate food images on Instagram. Content analysis was their main research method. After collecting the Intagram posts about food from teenagers, the research team created a categorizing frame of food items that were shared by the research targets. In addition, they also explored how food was displayed, and the context of food were shared.

In 2016, Pila, Mond and their team did a content analysis on #cheatmeal images on social media. The team "extracted more than 1.6 million images marked with the #cheatmeal hashtag on Instagram, coded the photographic and textual elements of a sample (n5600)" In their analysis, they explored the correlation of people's gender and body types, and the volume and type of food they posted (Pila E. , Mond, Griffiths, Mitchison, & Murray, 2017). Their research provides a method of categorize social media content.

Modeling and other quantitative analytics methods are also widely used. In 2016, Ying Hu, Changjun Hu and their team developed a model in predicting and evaluation popularity of online information in 2016. “Two directions of popularity evolution: popularity evolution patterns and popularity evolution prediction” were related to their research (Hu, Hu, Fu, Fang, & Xu, 2017). By analyzing a large sample of popular and trending hashtags on Twitter (3000 most popular ones from 3.3 million Twitter hashtags), they designed a new evaluation metric to evaluation the “burst, peak and fade” social conversation circle and found most popular topics “bursts suddenly, peaks very soon, and then fade quickly (Hu, Hu, Fu, Fang, & Xu, 2017).”

Natural Language Processing (NLP) model is one of the most frequently used models in social media analysis. For instance, Fried and his team (2015) gathered over three million food-related tweets as the original dataset, then applied the NLP model to make predictions of people’s political views and community characteristics based on language people used to describe food.

The research and papers above inspired me of the research methods. I will use Pila, Mond and their team’s preliminary method to identify the categories and then apply that to the full raw data, referring to Fig. 2. I also calculated the correction between demographic variables and the mention patterns of veganism food items. Content analysis, Clustering and modeling (NLP and multiple linear regressions) are three major research methods I used in this research.

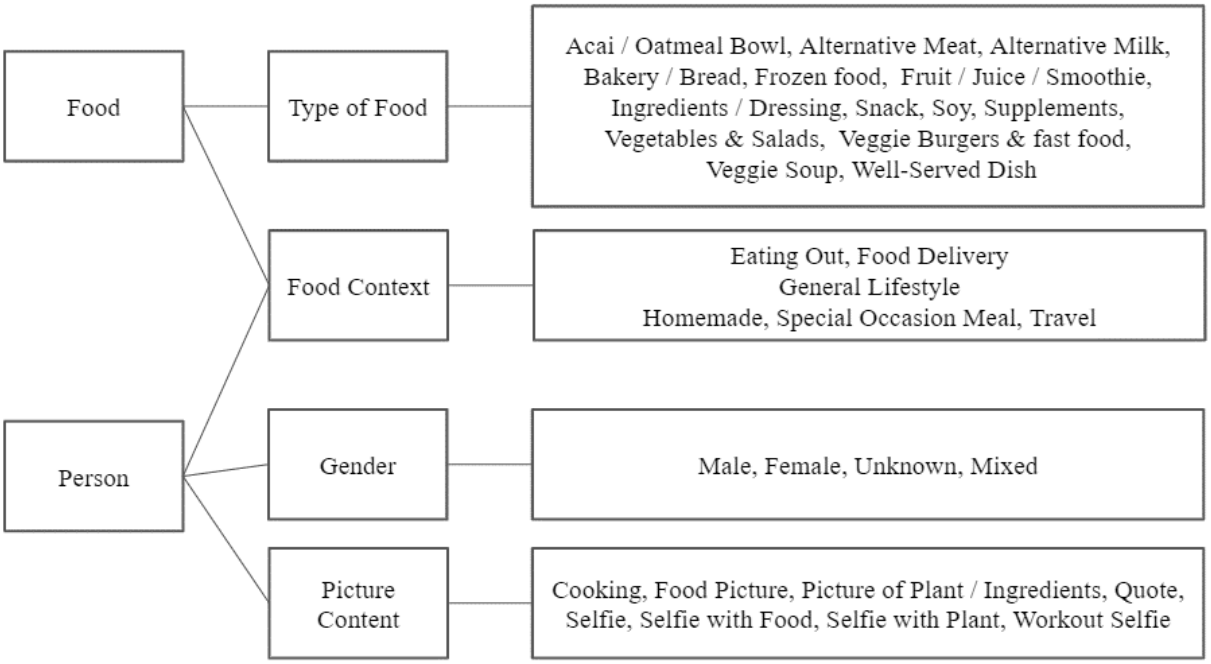
Food
Type of Food
Acai / Oatmeal Bowl, Alternative Meat, Alternative Milk, Bakery / Bread, Frozen food, Fruit / Juice / Smoothie, Ingredients / Dressing, Snack, Soy, Supplements, Vegetables & Salads, Veggie Burgers & fast food, Veggie Soup, Well-Served Dish
Food Context
Eating Out, Food Delivery
General Lifestyle
Homemade, Special Occasion Meal, Travel
Person
Gender
Male, Female, Unknown, Mixed
Picture Content
Cooking, Food Picture, Picture of Plant / Ingredients, Quote, Selfie, Selfie with Food, Selfie with Plant, Workout Selfie

Figure 2 Coding Frame (Pila E. , Mond, Griffiths, Mitchison, & Murray, 2017)

Here are the list of research question and hypotheses.

Question: What variables are affecting people's engagement with the social media posts related to vegan conversation?

Hypothesis 1: The sentiment toward the topic is significantly related to the number of engagements.

Hypothesis 2: The time the social media post gets published is significantly related to the number of engagements.

Hypothesis 3: The geo location of the social media post author is significantly related to the number of engagements.

Hypothesis 4: The media type the social media contains is significantly related to the number of engagements.

Hypothesis 5: The profession of the social media author is significantly related to the number of engagements.

Hypothesis 6: The gender of the social media author is significantly related to the number of engagements.

Hypothesis 7: The number of followers of the social media author is significantly related to the number of engagements.

Hypothesis 8: The content of the social media post is significantly related to the number of engagements.

Population and Sample

Social media provides an accessible source of data with broad demographic penetration across ethnicities and genders making it well-suited for examining the dietary habits of individuals on a large scale.

The population of this research was all the posts, tweets, comments, blogs, pictures and videos about vegan on social media. The population also includes the demographic information of anyone who posts content on social media, which is related to vegan.

I utilized a user generated vegan relate words and hashtag(vegan, #vegan, #veganofig, etc.) in social media to identify the posts for analysis. The sample was pulled in by NetBase, while they scanned 10% of the public posts on Facebook, relevant content in the forums, Instagram, YouTube and tweets. The time frame was set to January 1st, 2018 to December 31st, 2018. After the posts were collected and cleaned, I randomly selected 1,000 of them as a pool for content analysis. Not only the content of posts (texts, pictures or videos) were pulled in and collected, but also the demographic information the content contributors shared on social media.

Operational Definition of Variable

The volume of the most frequently mentioned terms together with vegan: This is the independent variable. The variable is in the format of interval. I calculated the numbers of the terms appeared together with the vegan terms and present the top 100 words or terms. The volume refers to the number of times the terms showing up in the overall post sample. If one term shows up for several times, it also counts as multiple times.

Sentiment of vegan topic: This is the independent variable. The variable is in the format of category. Using naïve Bayes classifier (Bromberg, 2013) I calculated the sentiment of the vegan topic and provided a sentiment score based on the calculation. The sentiment scores then converted into categories (positive, negative and neutral), describing the overall attitude towards vegan on social media.

Mentioned food type: this is a categorical variable. The variable describes the type of food and mentioned in the 1,000 sample posts or in the pictures. The food is divided into these categories: fruit / juice / smoothies, soy products, vegetables & salads, supplements and veggie-burgers and other veggie fast foods, frozen food, veggie soup and vegan snacks. The categories are designed based on Wikipedia's definition of vegan. The food type is used as an independent variable.

Food Context: this variable describes in which context the food is mentioned. This is a categorical variable. The five types of contents is: eating out meal, homemade meal, special occasion meal (event, holiday or birthday party) (Holmberg, E. Chaplin, Hillman, & Berg, 2016), food delivery and general lifestyle.

Picture Content: If the poster of the social media post shares any kind of actions together with the food (in picture), the action will be coded into these categories: ingredients /

dressing picture, food picture, regular selfie, workout selfie, selfie with food (Pila E. , Mond, Griffiths, Mitchison, & Murray, 2017). The variable is one of the independent variables used to make predictions.

Gender of poster: this is a categorical variable, describing the gender of the posters (if available). When it comes to social media analysis, not only the content of the posts matters, but also the posters information too. If the posters share his/her demographic information, it will be coded into Male or Female. This is an independent variable.

Profession of poster: this is a categorical variable. If the posters share their profession information on social media, they will be put into these groups: Banking and Finance, Blogging, Hospitality, Creative Arts, Education, Entrepreneurship, Executive Management, Health and Medicine, Homemaker, Hospitality, Journalism, Law and Order, Personal and Home Services, Sales and Marketing, Science and Research, Social Services; Creative Arts, Sports, Student, Technology. This is an independent variable.

Posting time: this is an ordinal variable. I apply the coding frame based on Pila, Mond and their team's research (2017). The posting time is divided into breakfast (5 a.m. – 9 a.m.) , lunch (10 a.m. – 4 p.m.) , dinner (5 p.m. – 9 p.m.) , late night time (10 p.m. – 4 a.m.). This is an independent variable.

Posting location: this is a categorical variable. Netbase provides the geo-location of each social media posts (if available). The posting location is at three levels: country (the US), state and city. Most of the analysis is done at state level. This is an independent variable.

Engagements of the posts: NetBase captures Comments and Likes as two kinds of engagements. The totally Engagements will be the sum of Likes and Comments. The variable is in the type of interval. It is one of the dependent variables.

Data Collection, Processing, and Analysis

Collecting data was the first step of the research. There are two commonly applied ways to collect social media data. The first one is using the posts, content and other kinds of data previously stored in the existing data pools, so called the "historic data sets" (Batrinca & Treleaven, 2014). The other one is real-time feeds data collected from the platform Application Programming Interfaces (APIs). I chose the second one, since the research need real time data. A social media data collecting tool, NetBase, was applied to gather posts and other data. NetBase provides access to data on most of the major social media channels, including Twitter, Instagram, Facebook, Reddit and forums. Using a Boolean query, NetBase collects the data related to the veganism topic published in the United States from January 1st of 2018 to the December 31st of 2018. These data were various tweets or Instagram posts of the consumers who gave their opinion towards veganism or relevant food. Only public posts were available to crawl and collect from API, while the private posts are not accessible. The collecting data included the post text content, pictures or videos (if available), the account owners' demographic information, geo-location and post time in the Eastern Time zone.

After the data obtained from social media channels, the next stage was cleaning data. It played an extremely import role in social media analysis, due to the dirtiness of social media data. NetBase provides an algorism to select the consumer generated content. In addition to the textual posts, consumers also uploaded images with identifying tags or hashtags, for sharing their opinion towards veganism. Textual comments by others in response to the photo were excluded (Pila E. , Mond, Griffiths, Mitchison, & Murray, 2017). In addition, all the brand posts (the content with the main purpose was product advertisement), posts from spamming accounts (bots) or memes (parody posts), or duplicate posts ("reposts" or "retweets") were excluded using

NetBase filters. Therefore, only the relevant content from real consumers were selected as the data sample. After the data is filtered for cleaning, a random selected sample of 1,000 posts were downloaded.

These 1,000 posts were coded using the coding frame according to Fig. 2. During the coding process, the posts which were deleted, or not relevant to vegan topic were removed (some Instagram users are using #vegan or other vegan related hashtags to gain more impressions, but the actual content has nothing to do with vegan). After removing irrelevant posts, the cleaned sample size was reduced to 714. The coded sample were removed symbols and information that was not needed for the next step, and a "corpus" were created using programming tool R.

The next step was preparing the data for statistical analysis. The posting time was based on EST time zone. I converted the time to author's local time base on their location. The time is also divided into date, day of week and hour. I encoded the categorical data: hour was converted into day parts, then encode to Weekday (1) and Weekend (0); state (West Coast ->1, Others ->0) was encoded into binary code; post type was changed to video (1) and non-video (0). The final step was creating dummy variables for all the categorical variables. The outliers (number of engagements larger than 550) and posts not existing anymore (text as *Post Deleted by Author*) were also removed. The cleaned sample size is 691.

With the cleaned data, I firstly tried NetBase's built-in text analysis algorithm to classifying the trendy veganism terms. Applying a Natural Language Processing (NLP), which can not only pick the top conversation drivers, but also the hashtags, people and brands in the conversation, "lexical analysis to study word frequency distributions, pattern recognition, tagging, annotation, information extractions, data mining techniques including link analysis, visualization and predictive analytics" (Batrinca & Treleaven, 2014). It is also capable to perform sentiment analysis, which means the application to compute text to identify and extract subjective information in the data (Batrinca & Treleaven, 2014). I used it as a reference and converted the data in a coding framework and self-generated bag of words matrix to be applied machine learning algorithm.

Content analysis was applied among this random sample. I combined Holmberg, Chaplin and their team's and Pila, Mond and their team's preliminary method to identify the categories and then apply that to the full raw data, referring to Fig. 2. Both the methods considered the text captains as well as the image content: Food items identified in the photos were categorized and captions and hashtags were also used if they provided additional information about the food item than was clearly visible in the image (Holmberg, et al., 2016). According to Holmberg and his team's research (2016), not only the food is import, the context of food display also matters. They created categories "reflecting the context of the situation, how food was displayed, and how the food was described by the uploader. (Holmberg, et al., 2016)." Table 1 shows that the context referred to what kind of food was mentioned/ posted and how food was prepared. The third part focused on the description of food. In this part, I categorized how the social media users describe their food on the platforms in text, hashtag and captains, as well as their attitude towards it.

I also calculated the correction between demographic variables and the mention patterns of veganism food items. I applied Holmberg, Chaplin and their team's text analytics method to estimate post authors' gender, textual profile information (bio-information).

K-mean clustering was applied to evaluation the social media trend prediction. To cluster the top food terms using social media data, I applied the following procedure: firstly, the more frequent two food items are mentioned together, the closer these foods are perceived by consumers. I created a co-mention matrix. Two food items are considered similar or alternative when they are mentioned together frequently (Kang, Yong, & Hwang, 2016). I grouped the food items using K-Means cluster analysis, to find the nearest mean of each cluster, since the data on social media is very divisive. After finding out the clusters, I identified the characteristics of each cluster for better understanding of the clusters. I ranked the food categories using SAW method (Mihuandayani, Ramandita, Setyanto, & Sumafta, 2018).

Statistically modeling, including **ANOVA** and **multiple linear regressions** were applied for model building. The model was designed to make predictions of Engagements of social media posts related to vegan. In this process, extra data processing was required.

The evaluation processing compared the actual number of social media engagements and the predicted number of engagements. This step is necessary because this project was proposed to provide practical suggestions for social media marketers to communicate with their targeting customers.

The final step was data visualization. Again, this research is designed to benefit marketer, advertisers, strategies and a lot of other people in the business but not doing statistics work. Data visualization helps them understand the research and apply the insights into business decision-making and strategy-planning better. Geographic Heatmaps, Temporal Histograms and Parallel Word clouds are three major types of data visualization used in this research.

Assumptions

It was assumed that what the social media conversation contributors posted on social media are what they were doing or eating. It was also assumed that with the keywords in Boolean Query, NetBase captured the all the social media conversation about vegan. Another assumption is the location captured by NetBase was accurate. I also assumed that the random selected sample of posts was select randomly. In addition, all the posts I pulled was supposed to be organic posts, without any marketing effort (investment) behind it. The final assumption was the Twitter and Instagram algorithm did not change significantly during the year of 2018. Sometimes the social media platform changed their algorithm of how many often the followers can see the posts from the people they are following. For instance, a few years ago, Facebook changed the way their timeline displayed.

Limitations

One of the major limitations of this research is the access of sales data. It is very difficult to get access to the actual sales data of each vegan food and products. This research was proposed to provide suggestions to the marketers of vegan products and brands. Due to the same problem, the original plan of making prediction of brand sales based on social media conversation.

Sample size is also a limitation. This research is based on a randomly selected sample of 1,000 posts. Another limitation is the platform breakdown. Most of the posts were from Instagram. If the project were using a more diverse source sample, it could be more representative. In my future research, I plan to test the methodology in a larger size of sample.

Sentiment of each posts was assigned by the naïve Bayes classifier (Bromberg, 2013). The sentiment of each post was calculated based on the frequency of each sentiment words appearing in the post. However, the human language is much more complicated than that. The research was originally designed to figure out the sentiment of the post and the sentiment toward vegan products as two independent variables. However, with naïve Bayes classifier, it is not possible to do so. In the future studies, I plan to improve the NLP process of sentiment analysis.

Another limitation of this project came from social media platform's policy change. Facebook stopped its API sharing content with any social listening tools (WallarooMedia, 2019).

Therefore, nobody in the market can get access to Instagram posts and users' data since December 20, 2018. I was planning to apply the time series modeling to my project. However, this change made the time series analysis impossible, since there would be no accuracy social volume data over time.

Findings

Content analysis

Word cloud is a good way to show the frequency of the word is used in the conversation.

Figure 4 shows the word cloud, which indicates that people used word like “love”, “delicious”, “healthy”, “yummy” to describe their vegan dining experience.

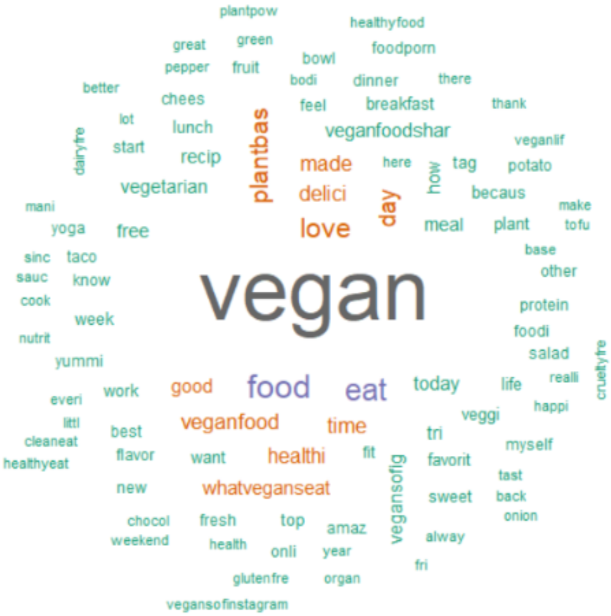

Figure 4 Word Cloud: The Words Mentioned Together with Vegan

“Protein” , “Salad”, “Taco”, “Cheese”, ‘Tofu” and “Chocolate” were also mentioned frequently together with vegan. “Sweet” is one of the 20 most

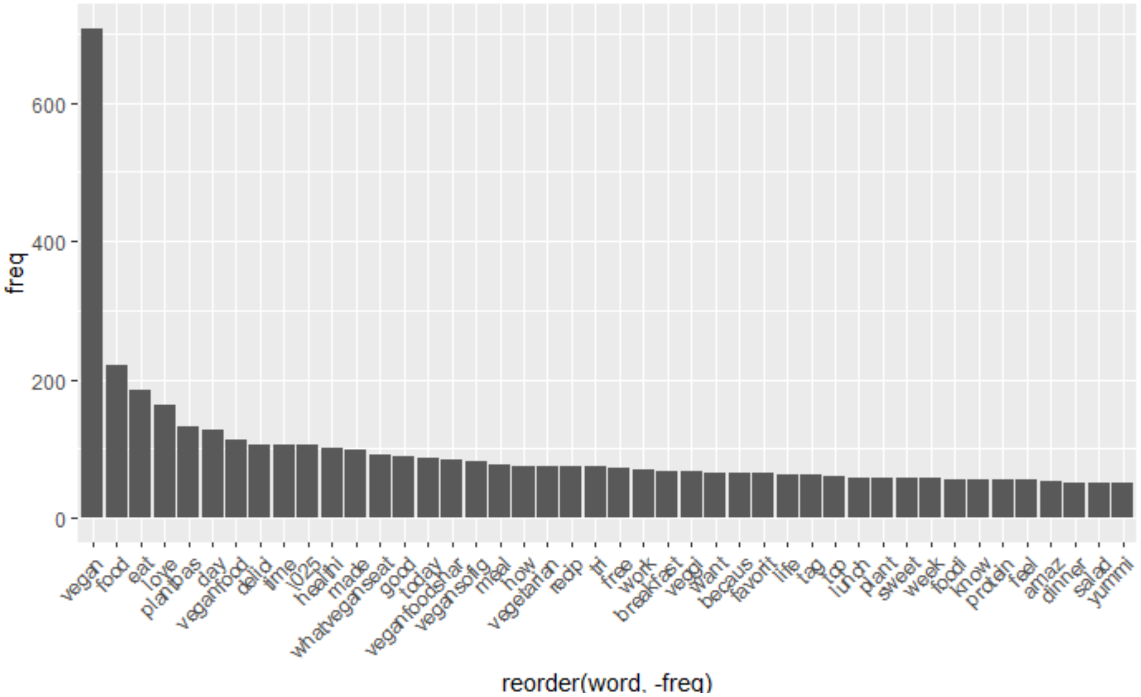

Figure 3 Word Frequency: The Words Mentioned Together with Vegan

frequently mentioned words in the sample. This indicated that veganism people might have been looking for sweet treats or desserts.

While it comes food context, most of the vegan people on social media are sharing about their experience of cooking at home (n=256, 36%), followed by Eating out at restaurant (n=219, 31%). About 12% (n=84) of the posts were not only focus on eating but take vegan as part of their life style. Only 5% (n=35) of the posts mentioned eating or cooking for special occasion.

As most of the posts were collected from Instagram, most of posts were with picture or video. For all the pictures analyzed, about 73% (n=516) of them were food picture. About 117 pictures contained the person or selfie, which took about 16% of the total sample. About 4.5% of the samples were with pictures of cooking process (n=28).

Vegetables & Salads was the most frequently mention vegan food type (n=170, 24%). Veggie Burgers and other fast food (pizza, taco, sandwich) (n=117, 16%) were also very popular among vegan eaters, as well as fruit or juice (n=105, 15%), baked products (n=87, 12%) and Alternative milk / dairy products (n=59, 8%) One posts could mentioned one or more types of food.

Generally, those who talked about vegan on social were holding a positive emotion toward the topic. Approximately 52% (n=371) of the sample were leaning towards positive, while 321 posts (45%) of the sample were with a sentiment of neutral.

In the selected random sample, nearly 97% (n=693) of the social media posts were published on Instagram, which suggests that social media users were more likely to share their experience and opinions about vegan on Instagram.

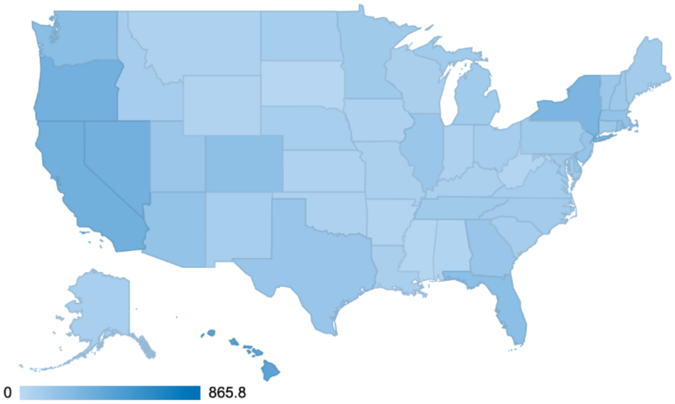

Figure 5 Indexing Score: Share of Social Media Posts against Share of Population

In this random sample, more than 21% of the posts were from California, followed by New York and Texas. However, California is the state with largest population in the US, 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 (Figure 2), which indicates that vegan social media conversation was more likely to happen in the states on west coast.

In this random sample, nearly 80% (n=567) of the conversation were from female audience, while males contributed to about 13.5% (n=96) of the social vegan volume, which indicates that women conversation contributors are more likely to share and post about vegan on social media.

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.

Statistical analysis

To answer the first research question of what variables are affecting people's engagement, this section explored the relationship between the independent variables and the number of Engagements of social media posts related to vegan.

The data was randomly split into two parts: training dataset (80% of the overall sample) and testing dataset (20% of the overall sample).

ANOVA. For reach categorical variable, I first used ANOVA to check their influence on the number of engagements in training dataset.

Hypothesis 1: The sentiment toward the topic is significantly related to the number of engagements.

With a p-value of .799 (>.05), the variable is not significantly related to dependent variable. The hypothesis is rejected.

Hypothesis 2: The time the social media post gets published is significantly related to the number of engagements.

With a p-value of .201 (>.05), whether the post was published on weekend is not significantly related to dependent variable. The hypothesis is partly rejected.

However, the day part saw a significant correlation with the number of engagements, with a p-value of 0.003 (<.05). The posts published in the morning hours (5-9am) saw an average number of engagements of 97.29, which is significantly higher than in other day parts.

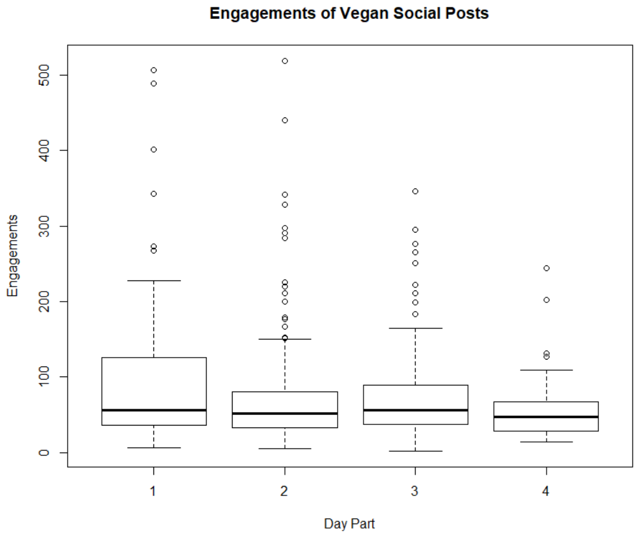

Figure 6 Boxplot: Avg. Engagements by Day Part

Hypothesis 3: The geo location of the social media post author is significantly related to the number of engagements.

With a p-value of .47 (>.05), the variable of whether the post author's location is in West Coast does not see a significantly impact dependent variable. The hypothesis is rejected.

Hypothesis 4: The media type the social media contains is significantly related to the number of engagements.

With a p-value of .144 (>.05), the variable of whether the post contained a video is not significantly corrected to dependent variable. The hypothesis is rejected.

Hypothesis 6: The gender of the social media author is significantly related to the number of engagements.

With a p-value of .17 (>.05), gender of the post author is the variable did not see the significant relation with dependent variable.

Hypothesis 8: The content of the social media post is significantly related to the number of engagements.

None of the food type (p-value=0.457 >.05), food context (p-value=0.503>.05) and picture content (p-value=0.327>.05) were correlated with the number of engagements. The Hypothesis is rejected.

Multiple Linear Regressions: after running the ANOVA, the model building process was started.

I first tested all the dummy variable, numeric independent variables (followers) and their correction with the dependent variable (Engagements of social media posts) in training dataset. To make each variable at the same level, I also did a log transfer of the number of followers.

The model sees an overall R squared of 0.46.

Hypothesis 1: The sentiment toward the topic is significantly related to the number of engagements.

With a p-value around .70 (>.05), the all three dummy variables (Negative, Positive, Neutral, Mixed) are not significantly related to dependent variable. The hypothesis is rejected.

Hypothesis 2: The time the social media post gets published is significantly related to the number of engagements.

With a p-value of .50 (>.05), whether the post was published on weekend is not significantly related to dependent variable. The hypothesis is partly rejected.

However, the day part 1 (5-9am, breakfast hours) is significantly corrected to the engagements (p-value= 0.0178<0.05). With an average engagement of 97.29, the posts published in this time were more likely to get engagements than other time of the day.

Hypothesis 3: The geo location of the social media post author is significantly related to the number of engagements. The hypothesis is accepted.

With a p-value of .55 (>.05), the variable of whether the post author's location is in West Coast does not see a significantly impact dependent variable. The hypothesis is rejected.

Hypothesis 4: The media type the social media contains is significantly related to the number of engagements.

With a p-value of .45 (>.05), the variable of whether the post contained a video is not significantly corrected to dependent variable. The hypothesis is rejected.

Hypothesis 5: The profession of the social media author is significantly related to the number of engagements.

With a p-value of .00716 (<.05), the profession of the post author (whether this person is working in creative arts or not) is the variable sees the significant relation with number of engagements of the posts.

Hypothesis 6: The gender of the social media author is significantly related to the number of engagements.

In the regression model, the variable sees an p-value of 0.18-0.6, which is larger than 0.05. The significance is not strong enough.

Hypothesis 7: The number of followers of the social media author is significantly related to the number of engagements.

With a p-value of nearly zero, gender of the post author is the variable sees the significant relation with dependent variable. The number of followers is positively related to the number of engagements of the post.

Hypothesis 8: The content of the social media post is significantly related to the number of engagements.

Neither of the food type (p-value=0.457 >.05) nor food context (p-value=0.503>.05) were correlated with the number of engagements. The Hypothesis is partly rejected. However, the picture content was correlated with the number of engagements. Cooking picture saw a p-value lower than 0.05. With an average engagement of 108, the post with cooking pictures are more likely to get more engagements.

After the first model building attempt, I cleaned the model up and remove those variables which are not significantly related to the dependent variable, only kept day part, professions, number of followers and picture content.

Evaluation of Findings

After the model updating, I applied the model to the test dataset. The Figure 7 shows the predicted results on the test dataset. Figure 8 shows the actual engagement numbers of the social posts around vegan.

From Figure 9 we can tell that in most of the cases, the predicted numbers of engagements are seeing a difference between -100 to +100.

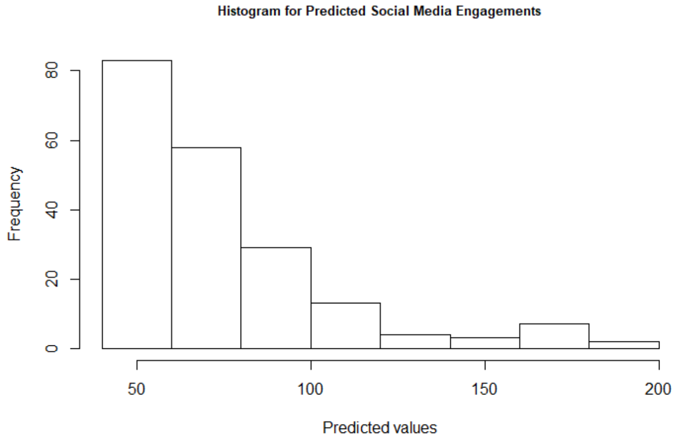

Figure 7 Predicted Numbers of Engagements

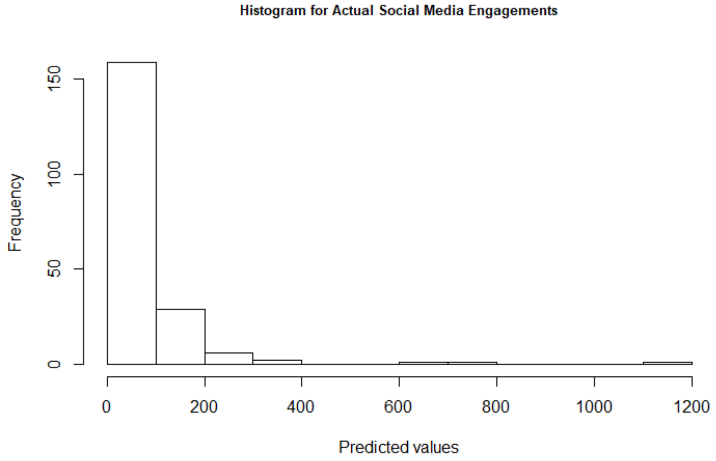

Figure 8 Actual Numbers of Engagements

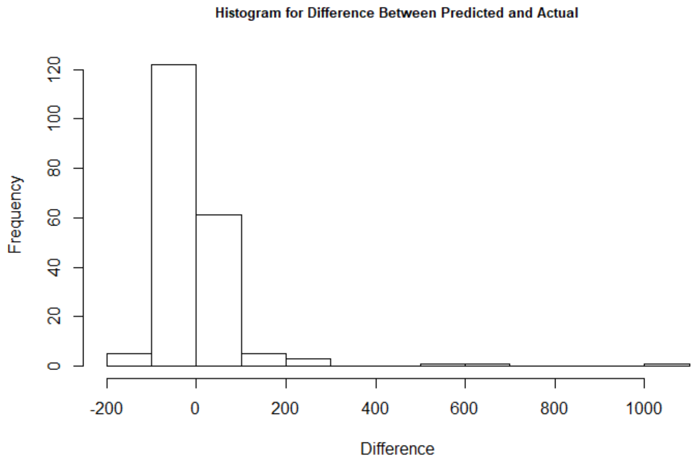

Figure 9 Difference between Predicted and Actual Numbers of Engagements

Recommendations and Conclusions

Conclusions

According to validation, the linear regression model supports part of the hypothesis that the demographic factors (professions), 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. However, the R square of the regression model is lower than 65%, which is a little bit blow the industry benchmark.

Recommendations

For the marketers of vegan products, they could work together closely with social media influencers. With a larger number of followers, the posts from influencers will see a better performance from the perspective of engagements. The influencers who work in creative art may provide a better social media performance and get more consumers engaged. They could also try to 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. They could also share more recipes for vegan fooders to cook at home.

I expect this research will assist marketers in not only food industry but also other industries, including CPG, Cosmetics and entertainment, to develop additional insight into customer preferences.

Social media data is becoming increasingly important in the business decision-making process. However, it is also facing a few challenges: 1) unstructured data; 2) no standardized process; 3) difficult to evaluate. This proposal suggests a practical process of social media data analysis, from data collection, cleaning, to analysis and evaluation, as answers to the above challenges. The methodology can be applied to a various of industries, from food to movie box office. Furthermore, applying both content analysis, bag of words and k-mean clustering analysis, this proposal also provides a solution of combining qualitative and quantitative analysis.

Limitation

However, my study has some limitations and further research needs to be performed. First of all, the optimized method of cleaning social media data is under debating. The accuracy of bag of words can be improved. How to exanimate pictures on social media is another import issue waiting to be solved. Secondly, how to deal with real time data is still a huge challenge. Consumer posts is not the only kind of content available on social media. A lot of brands are also running their accounts, to build a connection with the customers. This research is pretty much focused on analysis based on consumer voice. A brand voice research can also be created, to add another layer of insights and help brands learn how their social media accounts are catching or leading trends. Another problem is language. Currently, most of the analysis is based on English. However, other languages, such as Spanish and Chinese, are also used among large population. The development of bag of words for languages besides English is in urgent need. Last but not least, in this proposal, I advise using alternative data instead of real sales data to evaluate the prediction. A better evaluation methodology needs to be developed.

Reference

Fried, D., Surdeanu, M., Kobourov, S., Hingle, M., & Bell, D. (2015). Analyzing the language of food on social media. In *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*. https://doi.org/10.1109/BigData.2014.7004305

Holmberg, C., E. Chaplin, J., Hillman, T., & Berg, C. (2016). Adolescents' presentation of food in social media: An explorative study. *Appetite*. https://doi.org/10.1016/j.appet.2016.01.009

Vaterlaus, J. M., Patten, E. V., Roche, C., & Young, J. A. (2015). Gettinghealthy: The perceived influence of social media on young adult health behaviors. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2014.12.013

Xu, J., Van Der Schaar, M., Liu, J., & Li, H. (2015). Forecasting Popularity of Videos Using Social Media. *IEEE Journal on Selected Topics in Signal Processing*. https://doi.org/10.1109/JSTSP.2014.2370942

Yoo, S. Y., Song, J. I., & Jeong, O. R. (2018). Social media contents based sentiment analysis and prediction system. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2018.03.055

Altshuler, Y., Pan, W., & Pentland, A. (2011, November 20). *arXiv.org.* Retrieved from arXiv.org: https://arxiv.org/abs/1111.4650

Batrinca, B., & Treleaven, P. C. (2014). Social Media Analytics: A Survey of Techniques, Tools and Platforms. *AI and Society*, 89-116.

Beyza Çizmeci, S. G. (2018). Predicting IMDb Ratings of Pre-release Movies with. *3rd International Conference on Computer Science and Engineering UBMK*, (pp. 173-178).

Brandwatch. (2018, August). *The Social Outlook: The State of Social Report 2018.* Retrieved from Brandwatch: https://www.brandwatch.com/reports/the-social-outlook/view/

Brandwatch, Pi Datametrics. (2018, December). *Integrating Data on the Fashion Industry Online*. Retrieved from Brandwatch: https://www.brandwatch.com/reports/integrating-data-in-fashion/view/

Bromberg, A. (2013). *First shot: Sentiment Analysis in R.* http://andybromberg.com/sentiment-analysis/.

Gemma, J. (2018, February 8). *#PlantBased: An Exploration of Online Vegan Communities and Conversations*. Retrieved from Brandwatch: https://www.brandwatch.com/blog/react-online-vegan-communities/

Gemma, J. (2018, August 21). *Food Influencers: The Biggest Food Trends of 2018*. Retrieved from Brandwatch: https://www.brandwatch.com/blog/react-food-trends-2018/

Hu, Y., Hu, C., Fu, S., Fang, M., & Xu, W. (2017). Predicting Key Events in the Popularity Evolution of Online Information. *PLoS ONE 12(1)*.

Kang, H., Yong, H., & Hwang, H. (2016). Brand Clustering based on Social Big Data: A Case Study. *International Journal of Software Engineering and its Applications*, 27-36.

Mihuandayani, Ramandita, H. D., Setyanto, A., & Sumafta, I. B. (2018). Food Trend Based on Social Media for Big Data Analysis Using K-mean Clustering and SAW: A Case Study on Yogyakarta Culinary Industry. *2018 International Conference on Information and Communications Technology.* Jeju Island, Korea: ICOIACT.

Pila, E., Mond, J. M., Griffiths, S., Mitchison, D., & Murray, S. B. (2017). A Thematic Content Analysis of #cheatmeal Images on Social Media: Characterizing an Emerging Dietary Trend. . *International Journal of Eating Disorders*, 698-706.

Pila, E., Mond, J., Griffiths, S., Mitchison, D., & Murray, S. (2017). A Thematic Content Analysis of #cheatmeal Images on Social Media: Characterizing an Emerging Dietary Trend. *International Journal of Eating Disorders*, 698-706.

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The Power of Prediction with Social Media. *Internet Research*, 528-543.

Scott Coyne, P. M. (2017). Forecasting Stock Prices using Social Media Analysis. *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing* (pp. 1031-1038). Orlando, FL, USA: IEEE.

WallarooMedia. (2019). *Facebook Newsfeed Algorithm History.* https://wallaroomedia.com/facebook-newsfeed-algorithm-history/#seven.

Wang, X., Yu, C., & Wei, Y. (2012). Social Media Peer Communication and Impacts on Purchase Intentions: A Consumer Socialization Framework. *Journal of Interactive Marketing*, 198-208.

Yu, S., & Kak, S. (2012, March 7). *arxiv.org.* Retrieved from arxiv.org: https://arxiv.org/abs/1203.1647

Zhu, Y. Q., & Chen, H. (2015). Social Media and Human Need Satisfaction: Implications for Social Media Marketing. *Business Horizons*, 335-345.

Appendix

Author Location - State		% Percentage
California	153	21.43%
New York	83	11.62%
Florida	51	7.14%
Texas	50	7.00%
Illinois	32	4.48%
Pennsylvania	25	3.50%
New Jersey	23	3.22%
Oregon	23	3.22%
Arizona	20	2.80%
Georgia	18	2.52%
Colorado	17	2.38%
Massachusetts	17	2.38%
Ohio	16	2.24%
Michigan	15	2.10%
Nevada	15	2.10%
Washington	15	2.10%
Maryland	13	1.82%
North Carolina	13	1.82%
Tennessee	12	1.68%
Hawaii	9	1.26%
District of Columbia	8	1.12%
Minnesota	8	1.12%
Utah	7	0.98%
Virginia	7	0.98%
Connecticut	6	0.84%
Indiana	6	0.84%
Oklahoma	5	0.70%
Delaware	4	0.56%
Iowa	4	0.56%
Kentucky	4	0.56%
Missouri	4	0.56%
New Hampshire	4	0.56%
Rhode Island	4	0.56%
Wisconsin	4	0.56%
Louisiana	3	0.42%
South Carolina	3	0.42%
Alabama	2	0.28%
Idaho	2	0.28%
Kansas	2	0.28%
Wyoming	2	0.28%

Author Location - State		**% Percentage**
Alaska	1	0.14%
Arkansas	1	0.14%
Mississippi	1	0.14%
New Mexico	1	0.14%
South Dakota	1	0.14%

Author Gender	*COUNT of Posts*	**% Percentage**
Female	567	79.41%
Male	96	13.45%
Mixed	5	0.70%
Unknown	46	6.44%

Source Type	*COUNT of Posts*	**% Percentage**
Instagram	693	97.06%
Twitter	21	2.94%

Sound Bite Sentiment	*COUNT of Posts*	**% Percentage**
Positives	371	51.96%
Neutrals	321	44.96%
Mixed	12	1.68%
Negatives	10	1.40%

Media Type	*COUNT of Posts*	**% Percentage**
Image	642	89.92%
Image; Link	24	3.36%
Video	20	2.80%
Image; Video	11	1.54%
No Media	11	1.54%
Link	6	0.84%

Food Context	*COUNT of Posts*	**% Percentage**
-	278	38.94%
Homemade	167	23.39%
Eating Out	139	19.47%
General Lifestyle	94	13.17%
Special Occasion Meal	27	3.78%
Food Delivery	7	0.98%
Travel	2	0.28%

Picture Content	*COUNT of Posts*	**% Percentage**
Food Picture	516	72.27%
Selfie	52	7.28%
Selfie with Food	36	5.04%
	33	4.62%
Cooking	28	3.92%
Workout Selfie	28	3.92%
Picture of Plant / Ingredients	13	1.82%
Quote	7	0.98%
Selfie with Plant	1	0.14%

Food Type	*COUNT of Posts*	**% Percentage**
--	635	89%
Vegetables & Salads	170	24%
Veggie Burgers & fast food	117	16%
Fruit / Juice / Smoothie	108	15%
Well-Served Dish	103	14%
Bakery / Bread	87	12%
Alternative Milk	59	8%
Soy	37	5%
Snack	28	4%
Veggie Soup	24	3%
Supplements	21	3%
Alternative Meat	18	3%
Acai / Oatmeal Bowl	9	1%
Frozen food	8	1%
Ingredients / Dressing	4	1%