

Predicting Geo-Located Food Based Sentiment Analytics using Twitter for Healthy Food Consumption across India

Akriti Saxena, Madhuri Tibra, Ronnie D. Caytiles* and Iyengar N.Ch.S.N

School of Computer Science and Engineering. VIT University, Vellore, T.N., India

**Multimedia Engineering department, Hannam University, Daejeon, Korea*

akritisaxena.cs@gmail.com, madhurutibra141193@gmail.com

nchsniyengar48@gmail.com, rdcaytiles@gmail.com

Abstract

Records by means of social platforms not only offer a stage to express opinions and portray visions, but also facts related to conducts in which individuals have differences. An approach to work on these dissimilarities is on basis of statistics acquitted based on the geo-locations. Formerly, research work has been carried for studying and predicting the standard of living of people, searching the occurrences of naturally and artificially occurring lakes and various other geographic positions. This paper focuses on the work done by combining the data from social media (such as Twitter, Facebook) platforms and the real-world geographic locations. This leads to extracting the data from the virtual world, followed by text mining. The paper proposes, machine learning techniques for data mining, trailed by Feature-based sentimental analysis to observe popularity of the healthy and the unhealthy food throughout India. The findings can be utilized for focusing on areas missing a healthy environment, however areas with healthy practices must be supported to persist.

Keywords: *Healthy & unhealthy food consumption, Feature-based sentimental analysis, Geo-located Twitter data, Sentiment Analytics, Data Mining, Twitter based Prediction.*

1. Introduction

Virtual world acts as dependable platform for present and earlier generations. This virtual-reality provides visualizing opportunities for undone along with connecting individuals with their companions. Individuals have started utilizing these sources for expressing their opinions and thoughts. So it can be conferred that social media these days source to a huge database of real world information, collected by access to situations, for which persons act having some positive or pessimistic opinions. The work done in the paper will focus on data that are the tweets, extracted from Twitter.

Twitter maintains a database which has a large collection of quantitative along with categorical data, chronological data and spatiotemporal data regarding myriad topics such as education, artificial intelligence, sports, medicine among others. They help in providing insights to way of living, the manner of approach concerning things, different perceptions in addition to how they effect individuals decision. The good part regarding this virtual world is that by extraction of data on basis of the geo-locations, the information regarding user specific activities can be gained. After performing extraction of data, text mining and sentiment analysis is carried out. Sentiment analysis specifically is not anything apart from some random thought or perception, or outlook which aroused in mind at times, when the people come to know something, either by reading or looking about it.

Learning from habits is a suitable example. Placing psychological lessons aside, behavioural patterns directly influencing the personal as well as public health can be tracked. Up to now, significant dietary studies relating food consumption made use of

surveys and food records to retain track of regular activities of people, who became member of their group. These activities can be undesirable and expensive when it comes to implementation. The idea that made use of social media more popular is not wasting money on mechanisms for investigating food consumption traits that are difficult and mostly expensive while implementing in practice. Thus, the direction of the research intended towards using something which is readily available, such as social networks like Twitter. Variety of associations is often found, like women usually tweet related to less caloric diet in comparison to men, tweets from people with higher level of education often mention consuming food with fewer calories, users from urban areas tweet more often about alcoholic beverages than the population residing in the rural areas.\

Research has been carried out by people by means of twitter for studying changes, which these social networks, & other online platforms bringing about in society. Specified huge amount of obtainable data, along with researchers' ability of handling "Big Data," its essential to employ and analyze the data for addressing the issues related to health. The goals hence, for the work include, i) Proposing of some meaningful agenda for the health based information, acquitted from social media. ii) Executing sentiment analysis, on important features as well as attributes. iii) Based upon polarity and geo-locations of tweets, spatial distributions may be analyzed. iv) Later, after acquisition of information about tweets and locations, prediction regarding the areas which are more subjective to healthy or unhealthy consumption of food, can be carried out.

The paper hence focuses on the extraction of tweets from Twitter, and performing text mining on the same. The keywords collected are analyzed to perform sentiment analysis, and finding the polarity of the keyword. This will reflect the positive or negative sentiment about an item, which will later be analyzed to determine whether a particular region has healthier or unhealthier dietary supplements as per the figures acquitted from Twitter database.

Health of person can be determined not only by his food habits but in addition with that it is also influenced by the lifestyle, such as food habits, daily routine and the time spent in sedentary manner or actively also contributes to person's health. The study regarding whether the person is leading active and healthy lifestyle, the work progresses by extraction of tweets via twitter, which covers the extraction phase. One necessary factor for consideration is list of the food items must be accurate. Likewise, Blackberry is a fruit as well as cell phone company. So, if the extracted tweet comprises of the word 'Blackberry', other words in the tweets must also be considered. Further, if accompanying words, relate to transactions, bills, launch, model no., features, payments etc., then this will result in elimination of that particular tweet. Hence, process of the text mining shows up the importance. The later phase i.e. analysis phase is progressed by the text mining by the use of appropriate algorithms and further classifying results, thereby, predicting whether the people at particular place (by using location), have healthy or non-healthy eating habits.

2. Related Work

Gaining access to data based on geo-locations offers broader aspects for being, which depends not only on the society's opinions, also advising need for the actions that have to be reserved for some particular area, dependent on of what its deprived of. Already researchers initiated greater need as well as interest in exploring sectors that can be retrieved on basis of locations. Twitter is used by some researchers as it provides better approximation regarding public sentiments against the conventional internet article and other web blog when going for the sentiment analysis. All this is result of huge amount of relevant records that can be acquitted from twitter, in comparison to the traditional blogging platforms.

Widenera & Li [1], proposed paper on, use of geolocated Twitter data for monitoring prevalence of the healthy and the unhealthy food orientations across US, and also consumption of food throughout the country. The researchers here, explored the relationships between the content and sentiments of the tweets and location of tweets i.e. spatial distribution of the tweets. A novel approach is introduced for adding the semantics as some additional feature in training dataset considered for sentiment analysis [2]. For every extracted unit from tweets, they added its semantic conception as some additional feature, and the correlation for the representative concept with negative/positive sentiment was measured. The approach was applied to predict sentiment for three different Twitter datasets.

David M. Mixon [3] et al. published their work on the geo-locations of the man-made reservoirs throughout the terrains of fluctuating complexity by using GIS, in which they deployed a process, scripted in the ESRI's AML (ARC Macro Language) for locating reservoirs on DEM (Digital Elevation Models), by use of information provided in RESIS. According to their study the script also delineated the contribution of several watersheds and compiled a number of hydro-logically important constraints for each of the reservoirs. [4] worked on the classification of sentiment regarding the twitter messages by the use of distant supervision. Main idea of paper is using of tweets with the emoticons for the distant supervised learning. However their research is lagging because it does not consider neutral tweets at the same time it does only consider English language sentences.

A technique for analysis of large-scale geo-location related information via social sites for the inference of any individual's activity pattern for the transportation for a whole week along with few hours of any day [5]. A Pappu Rajan & S.P. Victor [6], published work concentrating on Web Sentiment Analysis regarding positive or the negative word scores by using the Twitter data. The objective was show the importance on determination of sentiment of text, further classifying it as negative or positive, which extends to proving information regarding strength of polarity. Daniele Sacco [7] *et al.*, work in his paper for smart cities & urban sensing, along with the big data, on basis of the geo-locations from the social network, proposed the reference model regarding the exploitation of Big Data as far as urban sensing is considered and later performing validation of it by case study. Here, they have worked to find out the urban areas that are not covered by transport service during peaks hours. Previous research in sentiment analysis [8] looked at the attainment of different classifiers focusing regarding movie reviews. The work has served as a baseline to many authors and there is evidence that many researchers have used techniques provided in their paper across different domains. Pang et al. also make use of a similar idea, by the use of star ratings as the polarity signals in the training data. Pak & Paroubek proposed an approach, which retrieves sentiment concerned tweets API of twitter and classify their sentiment alignments [9]. The results show that classifier that used bigram features is producing maximum classification accuracy because it's achieving good balance.

3. Problem Identification

In the paper, we have come across the problem of extracting the sentiments or opinions of various healthy, unhealthy food items in various regions of India. These food items are selected based on their nutritive values from various sources like online websites. Lexicon (dictionary) based approach of sentiment analysis has been used to classify the sentiment of the views of people shared on the social media platform. Previously, a similar type of work has not been done for particularly Indian subcontinent. The geo-locations of people who are tweeting are extracted to analyse the spatial patterns. In the below given figure, the overall flow of proposed architecture is shown.

4. Motivation

Browsing through various social platforms, provided idea for the use of available information for some useful work. The researchers done earlier in the field of sentiment analytics if implemented in proper manner can prove useful for providing effective business solutions as well. The business can expand well if the one know what factors will influence the in increasing the revenue for the business. Moreover, it also helps people to know if society is following the right direction in order to tackle health issues. The work done in this paper may help in realization of unbalanced diet intake, and unhealthy food consumptions and the problems caused related to that.

5. Methodology

5.1. Data Collection

From March 1, 2017 to April 15, 2017, the tweets are collected and streaming API of twitter has been used using search queries of healthy and unhealthy food stuff. After inserting consumer keys and access tokens (by registering on twitter) users can connect easily to Twitter API. Now, real time twitter data can be extracted downloading suitable packages (like TwitterR and RCurl). Thousands of tweet information is collected on both healthy and unhealthy food items. The geo- locations of the tweets are also extracted separately to look at the spatial patterns occuring across Indian subcontinent.

The geo-codes we have used for searching are latitude-22.7679, longitude-78.8718 and radius- 1650km. These combinely are the geo-codes of Kirahkota (487441), Madhya Pradesh, India. This place is a near centre of India and that is the reason of choosing this geocode for data collection. This large radius has been taken because all the parts of India should be covered. Later on the locations that are falling outside India will be discarded.



Figure 1. Reverse Geocoding of Latitude, Longitude (Used in Search Query)

A list of search queries regarding healthy and unhealthy food stuff is made using online websites like www.choosemyplate.gov. Other sources like list of common food chains in India are also considered for search queries. Those food items are taken into consideration

that we consume more in our day to day life. Quadrat analysis (point pattern analysis) divides the study region into quadrats of similar dimensions. The conclusions about the arrangement of tweets in these divided areas have been made. Kernel density estimation is used for getting the density estimate about the regions which are actively participating and contributing in sharing their views about healthy and unhealthy food-stuff. KDE is helping in visualization of geo-spatial data.

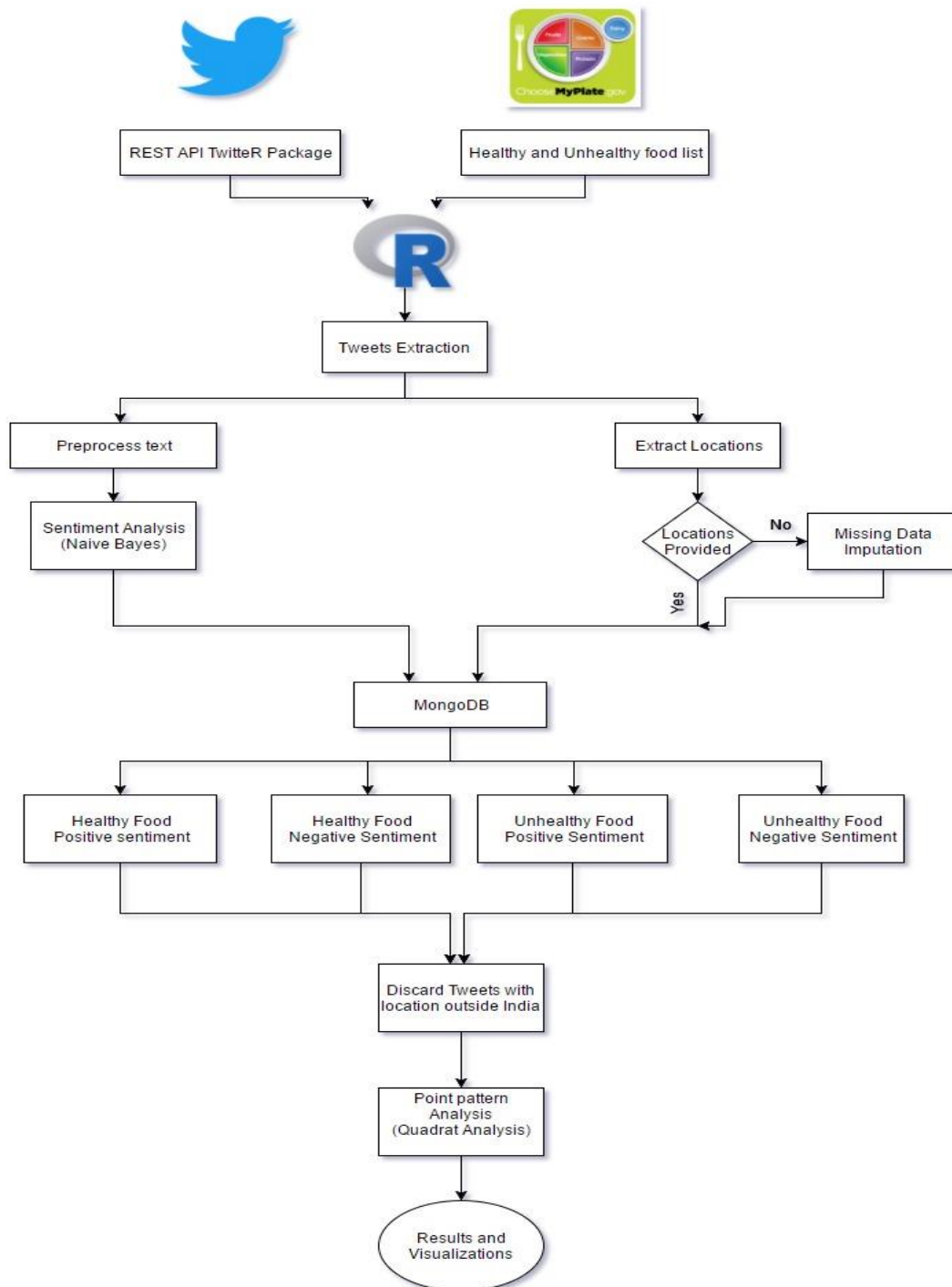


Figure 2. Proposed Flow Diagram

A sample list of the food items considered is:

Table 1. Sample List

HEALTHY FOODS		UNHEALTHY FOODS	
FRUITS	VEGETABLES	GENERAL	FAST FOOD
Apple	Carrots	Bakon	Domino's
Plum	Cucumber	Cheese	Dunkin' Donuts
Lemon	Cabbage	Cookies	Papa John's
Pineapple	Ladyfinger	Donuts	Pizza Hut
Grapes	Broccoli	energy drinks	Starbucks
Mangoes	Corn	Pizza	McDonald's
Papaya	Eggplant	Sausages	KFC
Orange	Spinach	Pasta	CCD
Dates	beet root	Soda	Taco Bell
Bananas	soy beans	Hot dog	Subway
Almonds	black pepper	Fries	Burger King
Coconut	green peas	Chips	
Dates	Cauliflower	Cheese	
water melon	Potatoes	Burger	
Peaches	Tomato		
Cherries	tomato juice		
Raisins	Pumpkin		

5.2. Sentiment Analysis

Sentiment and subjectivity adds quality to the communications done by humans, whether in direct conversations or social networks or online posts. For example, when customer sentiments are electronically captured- that conveys mood, emotion and opinion- carries colossal value in business. Classification of polarity of the text given is most basic task of sentiment analysis. The text can be in a form of a sentence, or a document. At aspect/feature level- whether the opinion that is expressed in the document, entity or a sentence based on that feature/aspect is negative, positive or neutral. As advancement it focuses beyond polarity for instance, at some of the emotional states like “happy”, “sad” or “angry”. There are many techniques of sentiment analysis and some of them are:-

1. Naïve Bayes
2. Maximum Entropy
3. Support Vector Machines
4. Semantic Orientation

In this study, Naïve Bayes is used for analysis of sentiments so as to analyse the tweets as positive, negative or neutral. The classifier assumes that features used in performing classification are all independent. The NB classifier is a model that passes a single tweet document and calculates products of the probabilities for every feature occurring in the tweet that is for each of the three sentiment orientations, positive, negative and neutral. The Naïve Bayes method is greatly used for categorizing text [11].

Naïve Bayes works on the notion of Bag-Of- Words, that is, a disordered set of words without considering their positions, their frequency only matters. It is a probabilistic classifier, which of all the classes $c \in C$, for a document d , return class \hat{c} , having maximum posterior probability.

$$\hat{c} = \operatorname{argmax}_{c \in C} P\left(\frac{c}{d}\right) \quad (1)$$

Bayesian classifier uses Bayes' rule to transform the above equation into other useful probabilities. The presented Bayes' rule, gives a direction to split any conditional probability $P\left(\frac{x}{y}\right)$ into other 3 probabilities:

$$P\left(\frac{x}{y}\right) = \frac{P\left(\frac{y}{x}\right)P(x)}{P(y)} \quad (2)$$

By substituting both the equations, we get-

$$\hat{c} = \operatorname{argmax}_{c \in C} P\left(\frac{c}{d}\right) = \operatorname{argmax}_{c \in C} \frac{P\left(\frac{d}{c}\right)P(c)}{P(d)} \quad (3)$$

The denominator will be same for all the classed and therefore can be ignored. Thus, a class can be chosen based on the maximization of a more simple formula-

$$\hat{c} = \operatorname{argmax}_{c \in C} P\left(\frac{c}{d}\right) = \operatorname{argmax}_{c \in C} P\left(\frac{d}{c}\right) P(c) \quad (4)$$

Thus, the highest product of probabilities: the prior probability and likelihood help in computing the most probable class:

$$\hat{c} = \frac{\overbrace{P\left(\frac{d}{c}\right)}^{\text{likelihood}} \overbrace{P(c)}^{\text{prior}}}{P(c)} \quad (5)$$

5.3. Kernel Density Estimation

Observing a data sample, building an estimate of the underlying pdf (probability density function) is density estimation. It can be either parametric (data from known family), or non parametric. It attempts to easily estimate unknown distribution. Some common methods of density estimation include histogram and kernel methods. Histogram is the earliest and least revealing method of density estimation. Main advantages are that it is remarkably simple and with a high computation speed.

The discreteness that is contained in histogram approaches is overcome by Kernel density estimation. A smooth kernel function is centred at each data point and then summation is done to get an estimate of density [2].

A basic kernel estimator formula can be given as:

$$\widehat{f_{kde}}(x) = \frac{1}{n} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) \quad (6)$$

Where k is the kernel and h is bandwidth.

The techniques of density estimation have been extensively applied in disciplines which involve the data that is spatially referenced. A fundamental utilization of density estimation is conversion of set of spatially discrete events, like crime incidents, disease cases, animals or road accidents locations to a density surface or a continuous field to reveal the spatial distribution and the layout of events geographically. In spatial analysis, in kernel density estimation symmetrical surfaces are placed over the data points and then the evaluation of distances from a reference location to the points is done based on a mathematical formula and finally summing the values. The process is repeated for the following points. This helps in placing kernel on all observations, and summation of these kernels gives the density estimate of the given distribution [3].

$$f(x,y) = \frac{1}{nh^2} \sum_{i=1}^n K\left(\frac{d_i}{h}\right) \quad (7)$$

$f(x,y)$ = Estimate of density at the particular location (x,y)

n = No. of observations

h = size of kernel

K = kernel function

The main motive behind placing these kernels over data points is to create a continuous and smooth surface. At each and every data point a circular area of an already

defined bandwidth is created. This is the reason it is called as kernel density estimation. It works on the principle of Kernel Density estimation. Point pattern analysis has been done here to study and analyze the spatial patterns of the point on the geographical map. Quadrat analysis has been used here to determine the areas where the frequency is more or less. A set of quadrats of cells is super imposed on the study area and number of point in each cell is determined. By analysing the distribution of cell frequencies, the point pattern arrangement can be described.

6. Results

Firstly, the results have been presented quantitatively telling us about the total sentiment of the collected data. According to our study, the positive score about unhealthy food is overcoming the positive score for healthy food. Taking complete data into consideration, there is only 70.81% positive sentiments about healthy food as compared to 80.79% positivity about unhealthy food.

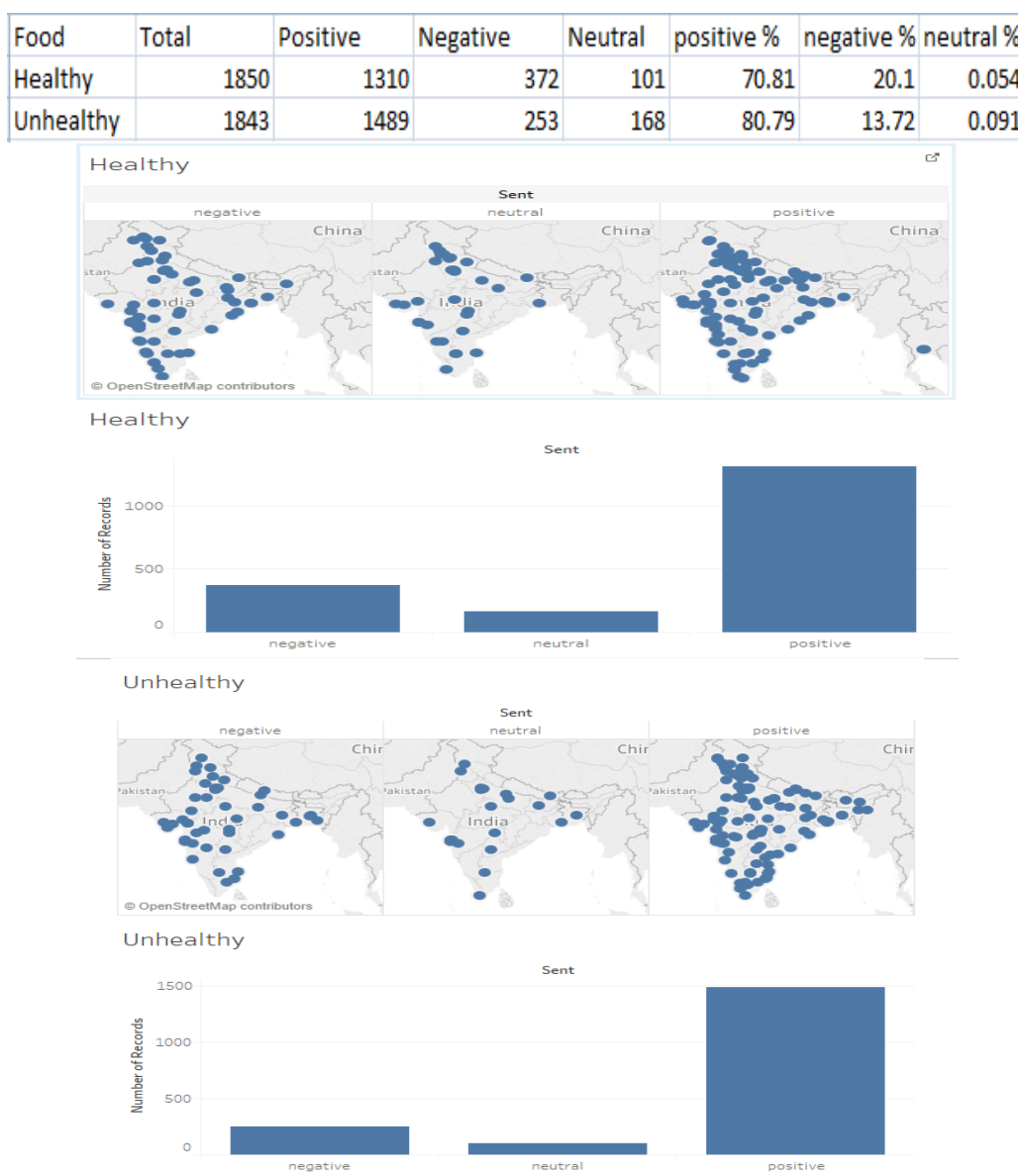


Figure 3. Sentiment Scores with Geographical Locations for Healthy & Unhealthy Food Consumption

Secondly, data is mapped geographically to analyse spatial patterns in the data. The technique of kernel density estimation is used here to show the implementation of methodology or can say proposed framework.

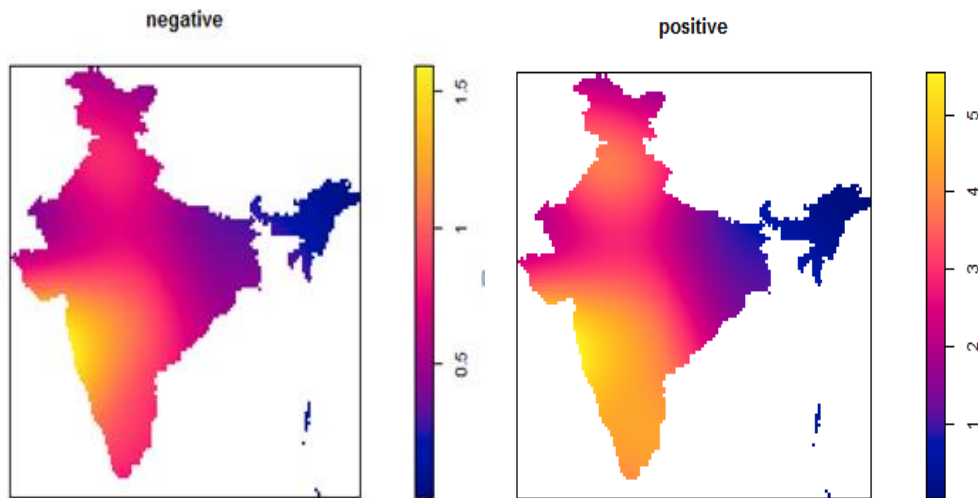


Figure 4. Kernel Density Estimate of a Probability Distribution from a Vector of Numbers (Healthy)

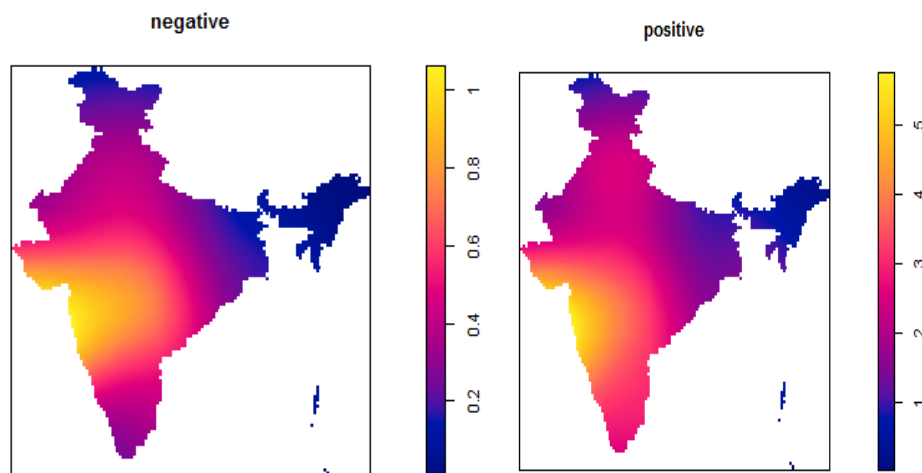


Figure 5. Kernel Density Estimate of a Probability Distribution from a Vector of Numbers (Unhealthy)

From Figure 4. Planar point pattern: 1118 points
Intensity 3.841475 points per square unit
Enclosing rectangle: [68.1202, 97.41516] x [6.75426, 37.13564] units
Window area = 291.034 square units
Fraction of frame area: 0.327

```
Multitype:
frequency proportion intensity
negative      211 0.18872990 0.7250012
neutral       96 0.08586762 0.3298584
positive      811 0.72540250 2.7866160
```

Figure 6. Summary (Healthy)

This density estimation for healthy foods clearly shows the areas where the consumption is more or less.

From Figure 5 Planar point pattern: 901 points

Average intensity 3.095858 points per square unit

Enclosing rectangle: [68.1202, 97.41516] x [6.75426, 37.13564] units

Window area = 291.034 square units

Multitype:			
	frequency	proportion	intensity
negative	127	0.14095450	0.4363751
neutral	66	0.07325194	0.2267776
positive	708	0.78579360	2.4327050

Figure 7. Summary (Unhealthy)

Looking at the point patterns of the data it can be seen that positive tweets about healthy foods are mostly coming from areas that falls under Maharashtra, Andhra Pradesh, Telangana. In contrast, the positive tweets about unhealthy tweets are mostly coming from Gujarat, Punjab, UP, Maharashtra. Several conclusions can be made with this analysis like, the areas where the people are completely inactive or very active. Next, the paper represents the results for the Quadrat Analysis done for finding the frequency of food consumptions in various regions.

The plots show how frequent tweets are popping up from the particular area, which are extracted to find out the inclination of people towards particular food item.

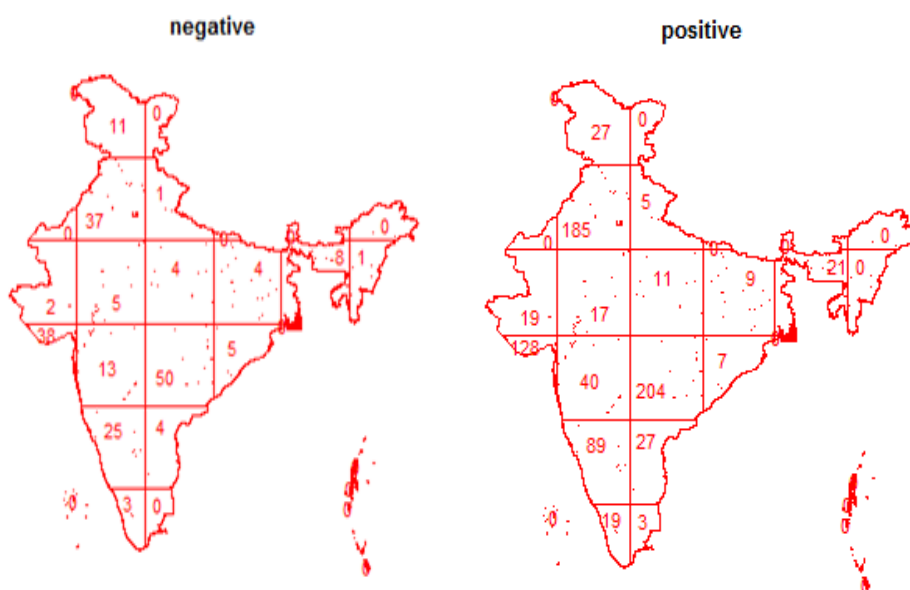


Figure 8. Quadrant Analysis (Point Pattern Analysis) for Healthy Food Consumption

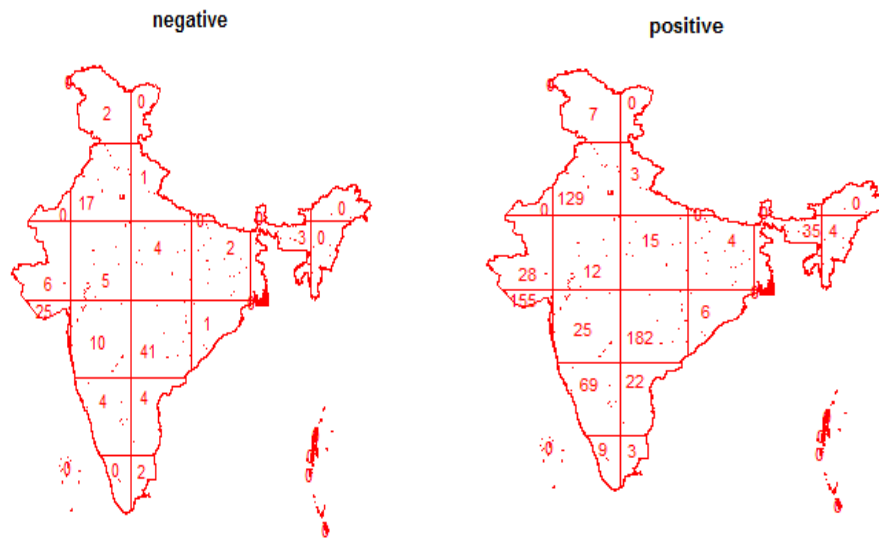


Figure 9. Quadrant Analysis (Point Pattern Analysis) for Unhealthy Food Consumption

7. Conclusion

In this study, we analysed thousands of tweets regarding healthy and unhealthy food stuff for predicting the areas where the consumption of healthy or unhealthy is more. Through the quadrat analysis it can be easily visualized the areas where the people are more active on twitter about food. If they are tweeting more about it that means they are concerned about their health. The mere number of tweets in some areas reflects lack of interest in talking about food stuff, inactiveness of people on social media because of absence of technologies.

Table 2. Healthy Food tweet Frequency in some States

No.	Tile.number	States	negative. Freq	neutral. Freq	positive. Freq
1	Tile row 1, col 2	J&K	11	4	27
2	Tile row 2, col 2	Delhi,Punjab,Haryana	37	25	185
3	Tile row 3, col 1	Gujarat	2	7	19
4	Tile row 3, col 2	Madhya Pradesh	5	3	17
5	Tile row 3, col 3	Uttar Pradesh	4	2	11
6	Tile row 3, col 5	Assam	8	2	21
7	Tile row 4, col 1	Gujarat	38	17	128
8	Tile row 4, col 2	Maharashtra	13	2	40
9	Tile row 4, col 3	Telangana, Chattisgarh	50	17	204
10	Tile row 5, col 2	Karnataka	25	13	89
11	Tile row 5, col 3	Andhra Pradesh	4	1	27

Table 3. Unhealthy Food tweet Frequency in some States

No.	Tile.number	States	negative. Freq	neutral. Freq	positive. Freq
1	Tile row 1, col 2	J&K	2	1	7
2	Tile row 2, col 2	Delhi,Punjab,Haryana	17	15	129
3	Tile row 3, col 1	Gujarat	6	1	28
4	Tile row 3, col 2	Madhya Pradesh	5	0	12
5	Tile row 3, col 3	Uttar Pradesh	4	2	15
6	Tile row 3, col 5	Assam	3	1	35
7	Tile row 4, col 1	Gujarat	25	17	155
8	Tile row 4, col 2	Maharashtra	10	2	25
9	Tile row 4, col 3	Telangana, Chatisgarh	41	21	182
10	Tile row 5, col 2	Karnataka	4	3	69
11	Tile row 5, col 3	Andhra Pradesh	4	0	22

From this result of Quadrat analysis, it can be easily concluded that in Telangana and Chattisgarh the tweets about Healthy foods are maximum. Delhi, Punjab or we can say collectively as central India is talking more negative about healthy as compared to unhealthy. People are moving more towards fast food. Various things can be concluded looking these tables of healthy and unhealthy food items. With respect to research work carried out so far, we have made one major conclusion as a whole by looking the complete percentage of healthy and unhealthy sentiments of people that they are talking more positive about unhealthy food. They are moving towards it.

8. Future Work

The study done in the paper may extend to the mapping of diseases, caused by intake of unhealthy food. The comparisons may be listed concerning occurrences of the diseases in the healthy Vs non-healthy areas. These things will surely reflect differences why several people are becoming prey to the diseases such as Atherosclerosis (relating to clogged arteries) that may cause strokes or heart attacks, obesity, hypertension, damage of liver and even Cancer. However, the unhealthy food though might not result in cancer, then too it may give birth to the conditions such as weight gain which can precipitate cancer. Hence, the model when extended may prove beneficial, when the information about these diseases, the prone areas related to the diseases as well as the lifestyle and eating habits of the people in a particular area, are mapped.

References

- [1] M. J. Widenera, "Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US", Department of Geography, University of Cincinnati, Cincinnati, OH, USA, Wenwen Li, GeoDa Center for Geospatial Analysis and Computation, School of Geographical Sciences and Urban Planning.
- [2] H. Saif, Y. He and H. Alani, "Semantic Sentiment Analysis of Twitter", Knowledge Media Institute, The Open University, United Kingdom.
- [3] D. M. Mixon, D. A. Kinner, R. F. Stallard, J. P.M. Syvitski, "Geolocation of man-made reservoirs across terrains of varying complexity using GIS"
- [4] A. Go, L. Huang and R. Bhayani, "Twitter Sentiment Classification using Distant Supervision", (2010).
- [5] S. Hasan, S. V. Ukkusuri, "Urban activity pattern classification using topic models from online geo-location data"
- [6] A PappuRajan and S.P.Victor, Web Sentiment Analysis for Scoring Positive or Negative Words using Tweeter Data.
- [7] D. Sacco, G. Motta, L. You, N. Bertolazzo, C. Chen, "Smart cities, urban sensing, and big data: mining geo-location in social networks", University of Pavia, Pavia, Italy,

- [8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?Sentiment classification using machine learning techniques", Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), (2002), pp. 7986.
- [9] P. Alexander, and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", LREC., vol. 10, (2010).
- [10] J. H Martin and D. Jurafsky. "Speech and language processing." International Edition 710, (2000).
- [11] H. Deng and H.Wickham, "Density estimation in R." Electronic publication, (2011).
- [12] T. K. Anderson, "Kernel density estimation and K-means clustering to profile road accident hotspots", Accident Analysis & Prevention, vol. 41, no. 3, (2009), pp. 359-364.
- [13] A. F. Dugas, Y.-H. Hsieh, S. R. Levin, J. M. Pines, D. P. Mareiniss, A. Mohareb, et al., "Google Flu Trends: correlation with emergency department influenza rates and crowding metrics", Clinical Infectious Diseases, vol. 54, no. 4, (2012), pp. 463e 469.
- [14] R. Ehrenberg, "Twitter kept up with Haiti cholera outbreak social media can track disease spread even in poorest countries", ScienceNews, vol. 181, no. 4, (2012), p. 16.
- [15] S. Farber, M. Z. Morang and M. J. Widener, "Temporal variability in transit- based accessibility to supermarkets", Applied Geography, vol. 53, (2014), pp.149-159.
- [16] R. A. Forshee and M. L. Storey, "Demographics, not beverage consumption, is associated with diet quality", International Journal of Food Sciences and Nutrition, vol. 57, 7e 8, (2006), pp. 494-511.
- [16] D. Ghosh and R. Guha, "What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System", Cartography and Geographic Information Science, vol. 40, no. 2, (2013), pp. 90-102.
- [17] M. F. Goodchild and J. A. Glennon, "Crowdsourcing geographic information for disaster response: a research frontier", International Journal of Digital Earth, vol. 3, no. 3, (2010), pp. 231-241.
- [18] K. He, F. Hu, G. Colditz, J. Manson, W. Willett and S. Liu, "Changes in intake of fruits and vegetables in relation to risk of obesity and weight gain among middle-aged women", International Journal of Obesity, vol. 28, no. 12, (2004), pp. 1569-1574.
- [19] A. Herdagdelen, W. Zuo, A. Gard-Murray and Y. Bar-Yam, "An Exploration of Social Identity: The geography and politics of news-sharing communities in Twitter(arXiv:1202.4393)", (2012).
- [20] J., Higdon, B. Delage, D. Williams and R. Dashwood, "Cruciferous vegetables and human cancer risk: epidemiologic evidence and mechanistic basis", Phar- macological Research, vol. 55, no. 3, (2007), pp. 224e 2.
- [21] X. Wan, "Using bilingual knowledge and ensemble techniques for unsu- pervised Chinese sentiment analysis", Proceedings of EMNLP08, Honolulu, HI, (2008), pp. 553e 561.
- [22] M. J. Widener, S. Farber, T. Neutens and M. W. Horner, "Using urban commuting data to calculate a spatiotemporal accessibility measure for food environment studies", Health & Place, vol. 21, (2013), pp. 1-9.
- [23] M. J. Widener, S. S. Metcalf and Y. Bar-Yam, "Dynamic urban food environ- ments: a temporal analysis of access to healthy foods", American Journal of Preventive Medicine, vol. 41, no. 4, (2011), 439-441.
- [24] N. Wrigley, "Food deserts' in British cities: policy context and research priorities", Urban Studies, vol. 39, no. 11, (2002), pp. 2029-2040.
- [25] C. Xu, D. W. Wong and C. Yang, "Evaluating the "geographical awareness" of individuals: an exploratory analysis of Twitter data", Cartography and Geographic Information Science, vol. 40, no. 2, (2013), pp.103-115.

Authors



Madhuri Tibra, she did her B.Tech in Computer Science and Engineering at BK BIET (Pilani), and currently pursuing MTech in Computer Science and Engineering (with Specialization in Big Data) at VIT University. The area of interests are Data Analytics and Programming.



Akriti Saxena, did her B.Tech in Computer Science and Engineering from Institute of Engineering & Technology (Faizabad), and currently pursuing M.Tech in Computer Science & Engineering (with Specialization in Big Data Analytics) at VIT University, Vellore. The area of interests include Machine Learning & Data Analytics.



Ronnie D. Caytiles, he had his Bachelor of Science in Computer Engineering- Western Institute of Technology, Iloilo City, Philippines, and Master of Science in Computer Science– Central Philippine University, Iloilo City, Philippines. He finished his Ph.D. in Multimedia Engineering, Hannam University, Daejeon, Korea. Currently, he serves as an Assistant Professor at Multimedia Engineering department, Hannam University, Daejeon, Korea. His research interests include Mobile Computing, Multimedia Communication, Information Technology Security, Ubiquitous Computing, Control and Automation



N. Ch. S. N. Iyengar (b 1961), he currently Senior Professor at the School of Computer Science and Engineering at VIT University, Vellore-632014, Tamil Nadu, India. His research interests include Agent-Based Distributed Computing, Intelligent Computing, Network Security, Secured Cloud Computing and Fluid Mechanics. He had 30+ years of experience in teaching and research , guided many scholars, has authored several textbooks and had nearly 200+ research publications in reputed peer reviewed international journals. He served as PCM/reviewer/keynotespeaker/ Invited speaker for many conferences. He serves as editorial board member for many international journals, reviews papers for many conferences with an interest of serving to the education community