# Cuisine mining and Analysis of Twitter Data

**Harish Annavajjala**
Indiana University
Bloomington, India
hannavaj@iu.edu

**Abhinandan S**
Indiana University
Bloomington, India
asampath@iu.edu

**Chirag Chheda**
Indiana University
Bloomington, India
cbchheda@iu.edu

**Siddharth Jayasankar**
Indiana University
Bloomington, India
sidjayas@iu.edu

## Abstract

In this paper we use the diversity of opinions, preferences and information of people present on twitter to analyze how varied are the preferences of people on the choice of cuisine. We also used twitter data to extract interesting information about different dishes in different cuisines around the world, get an insight on what kind of food is eaten over the course of the day in different parts of the world. In this project we intend to address and provide results for all these analysis in the form of rich visualizations which could be easily interpreted.

This analysis on food also increased our curiosity to find the similarities between cuisines by finding the common ingredients between different cuisines. For this we collected data from BigOven. We address these similarities by providing word clouds of ingredients for different cuisines.

## 1 INTRODUCTION

The type of food is not the same throughout the world. People in different countries prefer different kinds of food. Each country and place has its own signature dishes. The choice of food also varies with season .These are some of interesting phenomena which could give rise to a lot of information and interesting insights. Through this project we would like to study how peoples choice of food varies with geographic location and through the course of the day and try to get to know peoples food habits and diet in different parts of the world.

For these analysis we planned get the required data by mining social media. We chose to mine Twitter, a famous micro blogging social networking service, by using its API. We also collected information about dishes from BigOven, which is an online repository containing recipes of various dishes from a wide verity of cuisines.

Though a lot of analysis are being performed on Twitter data like presidential election analysis, very little amount of analysis are being carried out on food using twitter data. Through this project we can also test if twitter data can be reliably used to perform analysis on food.

## 2 LITERATURE REVIEW

A lot of research has been done using twitter data in different fields. For this project we had referred few papers where Twitter was used as a major platform to perform their analysis. Some authors studied the variation of use of language on twitter (H. Schwartz, J. Eichstaedt, M. Kern, L. Dziurzynski, M. Agrawal, G. Park, S. Lakshmikanth, S. Jha, M. Seligman, L. Ungar ,2013) .Research on smoking habits and tobacco products were also done using Twitter (M. Mysln, S.-H. Zhu, W. Chapman, and M. Conway ,2013). Twitter platform was also used to visualize and study the sentiments of people around the world (V. D. Nguyen, B. Varghese, and A. Barker,2013). There has been some research work done on food and health related fields using twitter data(Daniel Fried, Mihai Surdeanu, Stephen Kobourov, Melanie Hingle, Dane Bell,2014) and (M. J. Paul and M. Dredze,2011).

## 3 DATA

Data for this project was collected from different sources. In this project we have considered the top 10 cuisines of the world and from each cuisine top dishes where selected.

The cuisines considered in this project are 1) Indian Cuisine 2)Italian Cuisine 3)Spanish Cuisine 4)Mexican Cuisine 5)Chinese Cuisine 6)Thai Cuisine 7)Australian Cuisine 8) French Cuisine 9)Japanese Cuisine and 10 )Lebanese Cuisine
In these top 10 cuisines of the world an extensive

research was conducted and top dishes were considered. The dishes that were considered in these cuisines are as follows
Indian Cuisine : Biryani, Pani puri, Butter Chicken, Tandoori chicken, Idli, Dosa, Vindaloo, Chole Bhature, Malai kofta, Pav bhaji and Filter coffee.
Italian Cuisine: Pizza, Gelato, Pasta, Risotto, Ravioli, Prosciutto, Gnocchi, Bruschetta and Tiramisu
Spanish Cuisine: Pulpo a la Gallega, Gazpacho, Jamon, Chorizo, Empanada Gallega, Rabo de Toro, Arros Negre, Fabada Asturiana, Pimientos de Padron, Gambas al Ajillo and Paella
Mexican Cuisine: Chilaquiles, Tacos, Tostadas, Elote, Guacamole, Tamales, Tortillas, Pozole, Barbacoa, Burrito and Nachos
Chinese Cuisine: Sweet and Sour Pork, Gong Bao Chicken , Ma Po Tofu, Dumplings, Chow Mein , Peking Duck, Spring Rolls and Sweet and Sour Chicken
Thai Cuisine: Tom Yam Goong, Pad Thai , Kang Keaw Wan Kai , Tom Kha Kai, Tom Yam Kai, Moo Sa-Te, Som Tam, Panaeng, Por Pia Tord, Kuay Tiew, Gai Med Ma Moung and Kao Phad
Australian Cuisine: Barbecued shrimp, Lamington, Dim sim, Potato cakes, Pavlova, Tim Tam, Meat pies, Grilled kangaroo, Crab sticks, and Fantales
French Cuisine: Bouillabaisse, Quiche Lorraine, Steak-Frites, Coq au vin, Beef Bourguignon, Cassoulet, Escargots de Bourgogne, Moules marinres, Choucroute Garnie and Sole Meunire
Japanese Cuisine: Sushi, Ramen, Unagi, Tempura, Kaiseki, Soba, Shabu-Shabu, Okonomiyaki, Tonkatsu and Yakitori
Lebanese cuisine: Falafel, Tabouleh, Fattoush, Shawarma, Kebab, Manakeesh, Hummus, Fattoush, Shanklish, Dolma, Mansaf, Baklava and Knafeh
Tweets for these dishes were collected from twitter by using Twitters streaming API and Rest API. Twitter provides an accessible source of data with broad demographic penetration across ethnicities, genders, and income levels, making it well-suited for examining the dietary habits of individuals on a large scale.

Also a rich data set containing close to 8 million tweets from the authors of the paper Analyzing the Language of Food on Social Media Daniel Fried, Mihai Surdeanu, Stephen Kobourov, Melanie Hingle, and Dane Bell is collected. Data for this paper was collected by mining twitter for posts containing hashtags related to meals. The data was collected from the period between October 2, 2013 and May 29, 2014

For data related to recipes we collected data using the API of BigOven. By giving the cuisine and dish names this API provided us with different recipes for the same dish. We considered only the ingredients from the top 10 recipes for a particular dish.

Using the above data sources we filtered data to around 400 thousand tweets which has atleast one dish name present in the tweets.

## 4 DATA STORAGE

As mentioned earlier data was collected from different sources and these data were brought in to a standard form which was consistent. The data related to dishes were stored in a MySQL database and the data related to ingredients were stored in MongoDB.

For the twitter data we used three different types of information. 1) Information about tweeting person 2) Information about the tweets and 3) The tweet text. The general table schema for dish related tweets is as below:

| Colum Name | Type |
|---|---|
| twitter_handle | varchar(100) |
| username | varchar(100) |
| user_id | bigint(20) |
| location | mediumtext |
| timezone | varchar(45) |
| user_timestamp | varchar(100) |
| language | varchar(10) |
| tweet_timestamp | varchar(100) |
| geocode | varchar(75) |
| geo_data | mediumtext |
| tweet | longtext |
| food_in_tweet | mediumtext |

Figure 1: Schema of table storing tweets

## 5 APIs Used

In the project, our team was working on numerous programming for extracting large mount of data of various cuisines by utilizing different APIs like Tweepy , REST, BigOven API etc. This section mainly discusses how to use Tweeter API and BigOven API to get tweets and recipe of any particular dish.

### 5.1 Tweepy API

Using tweepy streaming API we have mined live tweets and images related to the dishes we have chosen. The advantage that we get when we mine data using tweepy streaming API is that there is no rate limiting, and we can get real time data. Below is a sample code snippet.

```
class StdOutListener(StreamListener):
    imgcount=0
    functioncount=0
    def on_data(self, data):
        self.functioncount=self.functioncount+1;
        tweet = json.loads(data)
        tags=["#biryani", "#burger", "#taco", "#pizza", "#pasta", "#dosa", "#sushi"]
        khaana =''
        if u'entities' in tweet:
            tweettext=tweet[u'text']
            tweettext=tweettext.lower()
            for tag in tags:
                if tag in tweettext:
                    khaana=tag[1:]
                    break
            if u'media' in tweet[u'entities']:
                if  u'media_url' in tweet[u'entities'][u'media'][0]:
                    print tweet[u'text'].encode("utf-8")
                    url = tweet[u'entities'][u'media'][0][u'media_url']
                    self.imgcount=self.imgcount+1
                    name=khaana+str(self.imgcount)+".jpg"
                    urllib.urlretrieve(url, name)
        return True
    def on_error(self, status):
        print status

if __name__ == '__main__':
    #This handles Twitter authetification and the connection to Twitter Streaming API
    listener1 = StdOutListener()
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    stream = Stream(auth, listener1)
    stream.filter(track=["#biryani", "#burger", "#taco", "#pizza", "#pasta", "#dosa", "#sushi"])
```

Figure 2: Script which uses Tweepy API to mine live stream twitter data

### 5.2 BigOven API

As explained on the BigOven official web- site , BigOven offers an API for all its developers which can be used to get ingredients of any particular dish of any particular cuisine.

This is accomplished in two parts. First we have to hit a URL , which is a HTTP API service of BigOven, with the dish name - for which the ingredients have to be mined. This gives the recipeId of the dish. Now using the recipeId of the dish we have to hit another URL and this returns the recipe results in an XML format which we later convert into JSON format. And we save this JSON object into mongodb database.

```
import urllib
import urllib2
import json,xmltodict
from pymongo import MongoClient
dishName = 'biryani';
cuisineName = 'Indian';
recipeId = '00000000'
response = urllib2.urlopen('http://api.bigoven.com/recipes?title_kw='+urllib.quote(dishName)+
                           '&api_key=0gQlzMap35EV1JQhdOZXr&pg=1&rpp=50').read()
response = xmltodict.parse(response)
response = json.dumps(response)
response = json.loads(response)
if response[u'RecipeSearchResult'][u'Results'] is not None:
    for recipeInfo in response[u'RecipeSearchResult'][u'Results'][u'RecipeInfo']:
        if recipeInfo.has_key(u'Cuisine'):
                if recipeInfo[u'Cuisine'] is not None:
                    if recipeInfo[u'Cuisine'].lower() == cuisineName.lower():
                        if recipeInfo.has_key(u'Title'):
                            if recipeInfo[u'Title'] is not None:
                                if recipeInfo[u'Title'].lower() == dishName.lower():
                                    recipeId = recipeInfo[u'RecipeID']
                                    break
    print "RECIPEID="+recipeId
    recipeResponse = urllib2.urlopen('http://api.bigoven.com/recipe/'+recipeId+
                                     '?api_key=Map35EV1JQhdOZXr').read()
    recipeResponse = xmltodict.parse(recipeResponse)
    recipeResponse = json.dumps(recipeResponse)
    recipeResponse = json.loads(recipeResponse)
    if recipeResponse.has_key(u'Recipe') and recipeResponse[u'Recipe'] is not None and \
                recipeResponse[u'Recipe'][u'Ingredients'] is not None :
        if recipeResponse[u'Recipe'][u'Ingredients'][u'Ingredient'] is not None:
            JSONObj = { "cuisineName ": cuisineName, "dishName":dishName ,
                        "ingredients":recipeResponse[u'Recipe'][u'Ingredients'][u'Ingredient'] }
            client = MongoClient('localhost', 27017)
            db = client['twitter_db']
            collection = db['twitter_collection']
            collection.insert(JSONObj)
            print 'Ingredients for '+dishName+' are below:'
            for ingredient in recipeResponse[u'Recipe'][u'Ingredients'][u'Ingredient']:
                print ingredient[u'Name']
```

Figure 3: Script which uses BigOven API to get the list of ingredients

## 6 VISUALIZATIONS AND RESULTS

- We have found out the dishes that are famous in a country depending on the number of tweets that have been made on a particular dish and we plotted a world map with this data. This visualization is achieved using google charts.

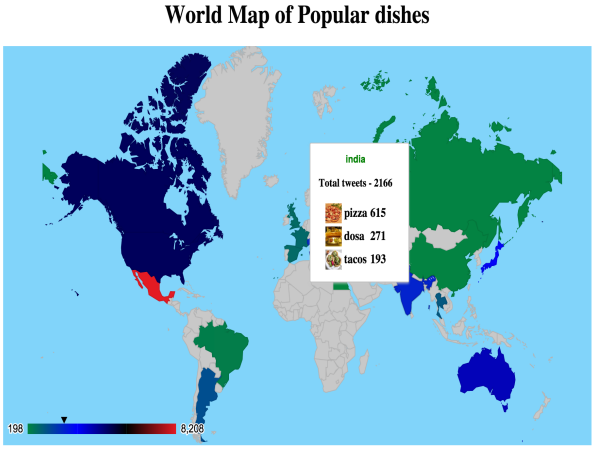


Figure 4: Geo map showing the top 3 dishes in each country

- We have found out the share of each dish in a particular cuisine and visualized the results in a pie chart. Below is an example.

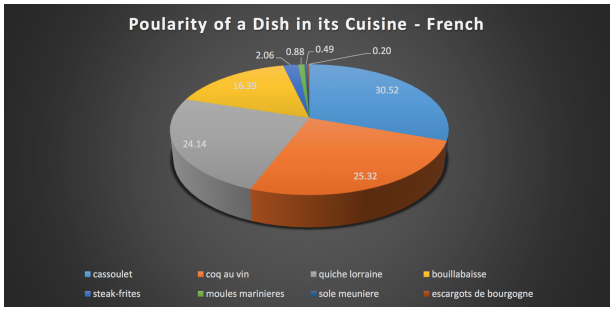


Figure 5: A pie chart showing the share of all the dishes in french cuisine

- We have found out the meal wise preference of each dish in all the cuisines and visualized the results in a clustered column graph. Below is an example.

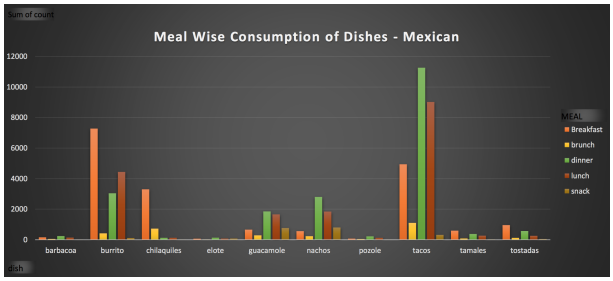


Figure 6: A clustered column graph showing the meal wise distribution of all the dishes in Australian cuisine

- We have found out how each dish of a particular cuisine is preferred in different countries and visualized the results in a stack bar graph. Below is an example.

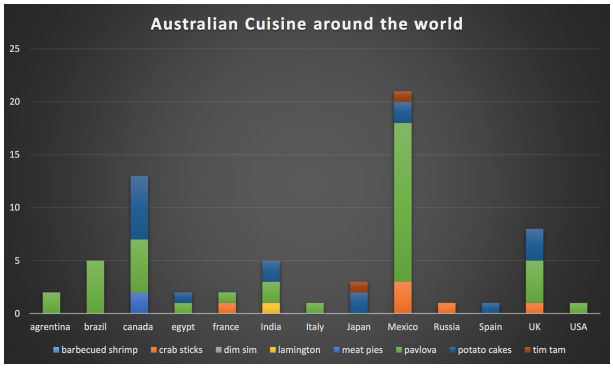


Figure 7: A stacked bar graph showing how each dishes in Australian cuisine are preferred in different countries

- We have mined the ingredients that are used in many dishes across all the cuisines and we plotted a word-cloud for ingredients that are used in all the dishes in a particular cuisine. Below is an example.


Figure 8: Word cloud containing ingredients used in Chinese cuisine.

## 7 INTERESTING FINDINGS

- Few Popular dishes like Pizza and Sushi are popular all around the world. Pizza is eaten all over the world and our results showed that Italian cuisine dominates world food, and 80 percent of the contribution is by pizza
- Mexicans explore a lot of cuisines and are also very active on Twitter. US and Canada too are extremely active and try out dishes.
- There is a clear preference to some dishes in each cuisine . Example: Chorizo in Spain.
- In each cuisine ,not all dishes are eaten during different meals. Example: Dosa for breakfast , Biryani for lunch and Butter Chicken for dinner.
- More common ingredients are present in cuisines of neighboring countries. Influence of dishes across neighboring countries is high.

## 8 ACKNOWLEDGEMENTS

This work was submitted as term paper in Social Media Mining Class (Fall-2015) at Indiana University. We would like to extend our thanks and express our sincere gratitude to our adviser/Prof. Muhammad Abdul-Mageed for the continuous support and guidance through out ths course of the project.

## 9 REFERENCES


1. L. Barth, S. G. Kobourov, and S. Pupyrev. Experimental comparison of semantic word clouds. In SEA, pages 247 – 258, 2014.

2. M. Hingle, D. Yoon, J. F. S. G. Kobourov, M. Schneider, D. Falk, and R. Burd. Collection and visualization of dietary behavior and reasons for eating using a popular and free social media software application. Journal of Medical Internet Research (JMIR), 15(6): 125 – 145, 2013.

3. M. Mysln, S.-H. Zhu, W. Chapman, and M. Conway. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res, 15(8):e174, Aug 2013.

4. V. D. Nguyen, B. Varghese, and A. Barker. The royal birth of 2013: Analysing and visualising public sentiment in the uk using twitter. In Int. Conf. on Big Data, 2013, pages 46 – 54. IEEE, 2013.

5. M. J. Paul and M. Dredze. You are what you tweet: Analyzing Twitter for public health. In ICWSM, 2011.

6. H. Schwartz, J. Eichstaedt, M. Kern, L. Dziurzynski, M. Agrawal, G. Park, S. Lakshmikanth, S. Jha, M. Seligman, L. Ungar, et al. Characterizing geographic variation in well-being using tweets. In 7th Intl. AAAI ICWSM, 2013.

7. Daniel Fried, Mihai Surdeanu, Stephen Kobourov, Melanie Hingle, Dane Bell ,Analyzing the Language of Food on Social Media arXiv:1409.2195v2 [cs.CL] 11 Sep 2014


## 10 BIO

- Harish Annavajjala

  M.S. in Computer Science - School of Informatics and Computing, Indiana University, Bloomington (2015-2017)

  Contribution: He learned how to use BigOven API and he coded a python script in which he used the BigOven API to retrieve the list of ingredients for all the dishes in all the cuisines and also puts that data into MongoDB. He also helped generate wordclouds

  which are used to visualize the most common ingredients used in a cuisine. He also contributed to mining twitter data using Tweepy API.
  Number of SMM classes not attended : 1

- Abhinandan SampathKumar

  M.S. in Computer Science - School of Informatics and Computing, Indiana University, Bloomington (2015-2017)

  Contribution: He learned how to use the Tweepy API and retrieved live tweets with hash tags containing the dishes we need to mine. He also contributed in coding a python script which mines images relating to dishes from twitter. He has also generated the visualizations the pie charts that depict the popularity of a dish in a particular cuisine .
  Number of SMM classes not attended : 0

- Chirag Chheda

  M.S. in Data Science - School of Informatics and Computing, Indiana University, Bloomington (2015-2017)
  Contribution: He experimented with using the REST API and contributed in mining tweets of dishes from twitter using REST API. He generated the visualizations the clustered bar graphs which depict the meal wise preference of different dishes in a cuisine , the stack bar graphs which depict the

affinity of a particular dish among different countries. In addition, he orchestrated the preparation process of the in-class presentation and is also partly responsible for making PowerPoint slides. He also performed as a good communicator between the project group members.
Number of SMM classes not attended : 0

- Siddharth Jayasankar

  M.S. in Data Science - School of Informatics and Computing, Indiana University, Bloomington (2015-2017)

Contribution: He was the database administrator/owner for this project. He is responsible for collecting and processing the data that has been acquired. He also extracted the data from database in the format required for visualizations. He has generated the visualization of Geo maps using Google Charts, which illustrates the most famous dishes in a country. He compiled the presentation and managed all the documentation of the project. He helped in developing the contents for the final paper. He also contributed to mining the twitter using REST API.
Number of SMM classes not attended : 0