Big Data Project

# Title - Improving Restaurant performance by performing Sentiment classification on online food reviews

# Abstract

Online food reviews are an important asset for users as well as for the restaurants serving the food. Most of the online reviews are in a free text format, which is difficult for the computer systems to analyze. In this project we will take the raw data from different social media websites, clean it to get the data with which we can work on and perform sentiment classification on the data using some famous classification algorithms to improve the quality of food from certain restaurants which has bad reviews and increase the overall performance of the restaurant. This project can further be expanded to predict what kind of food from a restaurant is likely to be ordered and if it is a healthy consumption.

# Related domain of study

This project aims to improve the business performance of restaurants and fast foods by detecting what kind of food item on their menu brings their ratings down. This can be achieved by using Naïve Bayes algorithm.

# Algorithms

# Naïve Bayes

Naïve Bayes is a popular algorithm for classification of text. We will be using this algorithm along with Natural Language processing tools for classifying and predicting the sentiment of the food reviews.

We need to transform the raw data collected into numeric representation, for which we will use tokenization. After that, we filter the reviews and remove all the stop words, punctuations, numbers from the review and feed it to the algorithm.

We will be using the stored data from yelp in this project. (Batch processing).

Depending upon the data received after performing sentiment analysis, we will classify the food item with negative review from the ones with positive review and link them to the restaurant which made this food item. By this data the restaurants will be able to improve the food items which have negative review, thus providing customer satisfaction and improving the business performance of the restaurant.

# Data sources

I will be using data from yelp.com

* <https://www.yelp.com/dataset/download>
* [https://www.kaggle.com/yelp-dataset/yelp-dataset#yelp\_academic\_dataset\_review.json](https://www.kaggle.com/yelp-dataset/yelp-dataset%23yelp_academic_dataset_review.json)

The data provided by yelp is a huge file, but we will be using only the “business id” – to identify the restaurant, “stars” and “text” columns of the data from the review.json file.

Sample data:

**root:**{} 9 items

**review\_id:**x7mDIiDB3jEiPGPHOmDzyw

**user\_id:**msQe1u7Z\_XuqjGoqhB0J5g

**business\_id:**iCQpiavjjPzJ5\_3gPD5Ebg

**stars:**2

**date:**2011-02-25

**text:**The pizza was okay. Not the best I've had. I prefer Biaggio's on Flamingo / Fort Apache. The chef there can make a MUCH better NY style pizza. The pizzeria @ Cosmo was over priced for the quality and lack of personality in the food. Biaggio's is a much better pick if youre going for italian - family owned, home made recipes, people that actually CARE if you like their food. You dont get that at a pizzeria in a casino. I dont care what you say...

**useful:**0

**funny:**0

**cool:**0

# ---- ?show or describe the steps you will take to transform the sample and make it ready for the analysis

# Graphics

Word cloud- for showing the food with most negative/positive reviews.

Bar graph – to plot the restaurant vs number of negative reviews.

(or)

Joined Histogram – to show the frequency of food item vs negative review and frequency of food item vs positive review.

# Current challenges

* Handling such huge data.
* Trying to understand if Naïve Bayes is the correct algorithm for this type of classification.

# Reference URLs planning to cite

1. <https://onlinelibrary.wiley.com/doi/full/10.1111/cgf.13217>
2. <https://medium.com/@martinpella/naive-bayes-for-sentiment-analysis-49b37db18bf8>
3. <https://www.kaggle.com/ngyptr/python-nltk-sentiment-analysis>
4. <https://itnext.io/sentiment-analysis-concept-analysis-and-applications-8b2c1c6fd77a>
5. <https://www.upgrad.com/blog/sentiment-analysis-what-is-it-and-why-does-it-matter/>

# Link to my current project -