SENTIMENT ANALYSIS OF RESTAURANT REVIEWS

A TECHNICAL REPORT

Prepared by
Vejey Subash Gandyer

OBJECTIVE

The objective of this project is

To design and develop a foolproof Sentiment Analyzer system capable of

(i) Classifying positive and negative restaurant reviews

(ii) Identifying Top-k dishes in every restaurant, outputting those with their ratings

that can be deployed in mobile and web applications.

SIGNIFICANCE / NOVELTY

The significance of this project is the fact of providing best dishes (reviews given by food critics in websites like Zomato, Foursquare, Yelp, etc) of any restaurant depending on the app users' favorite cuisine, dishes and calorie intake for the day.

INTRODUCTION

This Sentiment Analyzer project requires 5 stages namely

1. Data Pre-processing,
2. Data Exploration,
3. Data Modeling,
4. Model Validation, and
5. Model Optimization

1. Data Pre-processing stage

In Data Pre-processing stage, there are 3 main steps to be done.

1. Data Acquisition
2. Data Cleaning
3. Feature Preparation

Here, the data is acquired from Zomato already (90,956 reviews of almost all Chennai eat outs). It should be cleaned (inside some reviews, the contents are repeating twice and thrice). Feature preparation should be done for the cleaned data. Feature preparation is a process in which certain features that are not going to contribute for our classification of positive and negative sentiments will be removed and certain new features will be created with the existing features for better classification accuracy.

In our dataset, the following is the data format for all the reviews

city | review_text | reviewer_name | review_num | r_name | no_of_reviewer_rev | following | reviewer_url | no_of_follower | popular | rest_review | rate_of_review | date

Take for example review #3

chennai|Have been to Double Roti a couple of times and I must say that the menu is very catchy. I had visited twice in the initial stages of their opening have had good experiences. Being a vegetarian, I can only comment on their veg dishes though I've heard that their illegal burger and tandoori chicken fries are good. I've personally tried a veg burger (don't remember which one) and a panini pesto sandwich, again don't remember the exact one but it wasn't really out of the world. It was just good enough. Having said this, their sides are definitely good. Oreo shake being a speciality, good! Meloncholy (?) - a melon based mocktail is my personal favourite! Cheesy fries - good Fries with salsa and sour cream - great!! There are a few things I would like to comment on 1. On a crowded day, noise is pretty bad. It is exactly like being in a fish market. 2. Their packaging for delivery/ take away is good 3. Had recently ordered for a quick take away and ordered for their fries with sour cream and salsa. My suggestion would be to give the cream and salsa separate as the fries become soggy by the time the food gets home. All in all a good place for decently priced food! Rated Have been to Double Roti a couple of times and I must say that the menu is very catchy. I had visited twice in the initial stages of their opening have had good experiences. Being a vegetarian, I can only comment on their veg dishes though I've heard that their illegal burger and tandoori chicken fries are good. I've personally tried a veg burger (don't remember which one) and a panini pesto sandwich, again don't remember the exact one but it wasn't really out of the world. It was just good enough. Having said this, their sides are definitely good. Oreo shake being a speciality, good! Meloncholy (?) - a melon based mocktail is my personal favourite! Cheesy fries - good Fries with salsa and sour cream - great!! There are a few things I would like to comment on 1. On a crowded day, noise is pretty bad. It is exactly like being in a fish market. 2. Their packaging for delivery/ take away is good 3. Had recently ordered for a quick take away and ordered for their... read more — with Vishant Vibhaker|Kalyani|1|Double Roti| 20||https://www.zomato.com/users/kalyani-3018295|6|351|667|Rated 3.5|2016-02-09 11:53:15

Here, the underlined data is repeated word by word in the same review. This is the similar case with lot of the other reviews as well. I would say, this is an error that has happened while scraping the data from the website. This is not useful (redundant) data for our classification. It has to be removed, if not the feature space will go double and we will unnecessarily increase the computation time in classifying this into positive and negative sentiments later.

Next, the unwanted features should be removed from the working dataset.

city | review_text | reviewer_name | review_num | r_name | no_of_reviewer_rev | following | reviewer_url | no_of_follower | popular | rest_review | rate_of_review | date

Here, city, reviewer_name, review_num, r_name, following, no_of_reviewer_rev, reviewer_url are not important for our classification.

The most important feature for us is review_text. It goes without saying.
So is the rate_of_review feature that is numeric datatype with the range of 1.0 – 5.0 with 0.5 increments in between.
Certain features like date will give us information about the recent reviews that will be very useful than the review that was written 3 to 5 years ago.
Popular and no_of_follower features will give us information about the reviewer's popularity in the zomato website. In other words, it talks about the authencity of the ratings.
rest_review is also important in some cases.

The goal of this stage is to acquire, clean and prepare features for further exploration of data.

2. Data Exploration Stage

In this stage, the pre-processed data is taken and explored further to understand the data better. It goes without saying that this is the most important stage in data analysis.
Three main steps in Data Exploration involves the following,

1. Asking the right questions
2. Hypotheses generation
3. Visualizing data

Asking the right question involves a bit of brainstorming with all the possible questions that one can come up by looking at the dataset for the first time with the final goal in mind. In our case, the final goal(s) is/are (i) To classify the positive and negative sentiments from zomato reviews dataset and (ii) To identify the Top-K dishes in every restaurant in chennai and output their ratings in %. Keeping this mind, asking the right questions will be of immense help to create a right direction for data exploration.

The right questions would be the following:

I. How to convert this into a supervised learning as we have a dataset without any labels attached?
II. How to create features that will discriminate well between positive and negative sentiments?
III. How to identify the features for classification?
IV. What are the words responsible for a positive sentiment?
V. What are the words responsible for a negative sentiment?
VI. What is the most famous restaurant?
VII. What is the least famous restaurant?
VIII. Which restaurant has the most number of reviews?
IX. Which restaurant has the least number of reviews?

These are a few questions to name a few.

Once these questions are asked, hypotheses should be generated with final goal in mind again. In our case, one hypothesis would be

Dictionary of positive and negative words will classify the review into positive and negative sentiments => ***Dictionary of words leads to sentiment classification***

Once this hypothesis is generated, it is required to test this hypothesis with some basic assumptions in mind. The next logical step is to create a dictionary of words with two columns named positive and negative and filling the columns with words that we encounter in our review dataset. After creating this dictionary of words from the review vocabulary, build a machine learning model that classifies teviews into positive and negative sentiments using this dictionary. Thus, a hypothesis is tested. If it is a failed hypothesis, tweak the parameters of the model or change the model as well to gain some insights into the dataset. Doing the same procedure for all the hypotheses generated in the previous step will solidify our model in the later stages.

Visualizing the data with different points of view yields a overall picture of the data. It is very important to visualize the whole dataset with the questions and the hypothesis in mind. For this project, I am going to use Pandas for data exploration and Matplotlib and Seaborn for data visualization.

3. Data Modeling Stage

This stage involves building a Machine Learning model that discriminates the features (reviews here) well and classifies them into positive and negative reviews. Models can be of different forms like Logistic Regression, Decision Trees, k-Nearest Neighbors or Support Vector Machines. Advanced models like Random Forests, Extreme Gradient Boosting or Bagging can also be chosen.

For an ideal sentiment analysis system to be developed, each review should be split into tokens. The general NLP idea would be to take only the tokens we would be interested from already existing vocabulary and discard the rest of the words. In any Text classification, every word is a feature by itself. Our primary aim should be to reduce the number of features as much as possible. This should be attained without compromising the classification accuracy. The entire procedure of Data Modeling is given below.

Illustration:
Let us take for example Review#90864

chennai|The nearest hotel to my place. A nice place to have a good meal. Most of the dishes are decent...however. I love their masala dosa and the rava masala dosa....especially the fresh coconut chutney they serve with it...good value for money|Ritabrata Bhattacharya|2|New Sri Balaji Bhavan| 8||https://www.zomato.com/users/ritabrata-bhattacharya-3069266|120|2|2|Rated 4.0|2015-09-05 01:46:58

At first, the review_text is extracted from the full review.

The nearest hotel to my place. A nice place to have a good meal. Most of the dishes are decent...however. I love their masala dosa and the rava masala dosa....especially the fresh coconut chutney they serve with it...good value for money

It is fed into the ML model that we will build and discussed later.

One table with Words with weights are needed here for classification.

Words	Weights
love	3.1
good	2.1
Yummy	3.2
Decent	1
Nice	1.5
Terrible	-3
Bad	-2
Crap	-2.5
Lousy	-1.5
Poor	-1.25
Steep	-1.3
Low	-1

These above shown weights are learned by training the classifier model on training set of reviews split already according to train_test_split methodology of Machine Learning theory.

The test review (one sample from the testing set) is fed into the trained ML model for prediction. The classifier predicts it as either positive or negative sentiment. Depending on the actual class that this sample review belongs to, its accuracy is computed. This same procedure is done for all the reviews in the testing set. Overall accuracy, precision and recall is computed and tabulated.

This is the usual basic procedure followed in any Sentiment Analysis.

But this project is not the usual basic Sentiment Analysis project. The objective is not only to predict whether the review is positive or not at the review level but at a finer level (dishes). It is our objective to identify top dishes in every restaurant with its percentage of positive reviews for each dish.

Here, it is not advisable to build a ML model with the whole reviews as features. A different approach is needed to meet the objectives. To accomplish this, the following four steps are needed. One, finding the dishes in the restaurant and keeping a tab on them on the fly. Second, each found dish to be analysed for positive or negative sentiment. Three, computing the dish popularity index (DPI) for every dish found in the restaurant. And finally, Sort the dishes in ascending order of DPI.

As it is mandatory to first find the dishes in the restaurant for identifying the top dishes, we need to build a Dishes Table / List that contains all the possible dishes that Chennai restaurant provides the customers. Secondly, send these sentences to a trained classifier for predicting it as a positive or a negative sentiment. Thirdly, compute how many positive reviews and negative reviews exist. All the dishes of every restaurant is then subjected to a simple formula,

Dish Popularity Index (Percentage) = # of positive reviews / Total # of reviews

Finally, sort the dishes with this Dish Popularity Index to give the Top dishes in ascending order of DPI.

Illustration:
Let us take the same example Review#90864 again for comparison of this different methodology

chennai|The nearest hotel to my place. A nice place to have a good meal. Most of the dishes are decent...however. I love their masala dosa and the rava masala dosa....especially the fresh coconut chutney they serve with it...good value for money|Ritabrata Bhattacharya|2|New Sri Balaji Bhavan| 8||https://www.zomato.com/users/ritabrata-bhattacharya-3069266|120|2|2|Rated 4.0|2015-09-05 01:46:58

At first, the review_text is extracted from the full review.

The nearest hotel to my place. A nice place to have a good meal. Most of the dishes are decent...however. I love their masala dosa and the rava masala dosa....especially the fresh coconut chutney they serve with it...good value for money

Splitting into tokens (with '.' or '|' or other delimiters) gives,

nearest hotel to my place
A nice place to have a good meal
Most of the dishes are decent
however
I love their masala dosa and the rava masala dosa
especially the fresh coconut chutney they serve with it
good value for money

It is assumed that there are tables with the following items existing.

1. Dish Table – Contains all possible dishes in all the restaurants
2. Aspect Table – Contains all possible aspects in restaurant industry
3. Word Table – Contains all the positive and negative words in its corresponding columns

Dish Table

Dishes
Masala Dosa
Rava Dosa
Idly
.
.
.

Aspect Table

Aspects
place
ambience
service
cost
.
.

Words Table

Words	**Weights**
love	3.1
good	2.1
Yummy	3.2
Decent	1
Nice	1.5
Terrible	-3
Bad	-2
Crap	-2.5
Lousy	-1.5
Poor	-1.25
Steep	-1.3
Low	-1
Fresh	1.3

These above shown weights are learned by training the classifier model on training set of reviews split already according to train_test_split methodology of Machine Learning theory.
Comparing the above split sentences with these words table and identifying these words in them and computing the score will give us whether that will be a positive or a negative sentiment.
Identifying the dishes in the review sentences is a simple task if we have the dishes table with an exhaustive list of dishes listed in it already. Similarly, identifying the aspects is also simple provided we have the exhaustive list.

For example,
Let us consider the first sentence

nearest hotel to my place

Here, there are no words that match our words table. Hence the Score is 0.
Similarly for the other sentences in the review,

Sentences	Dish / Aspect	Score
A nice place to have a good meal	place, meal	Score is +1.5 +2.1 = 3.7
Most of the dishes are decent	dishes	Score is +1
however	nil	Score is 0
I love their masala dosa and the rava dosa	masala dosa, rava dosa	Score is +3.1
especially the fresh coconut chutney they serve with it	coconut chutney	Score is +1.3
good value for money	value, money	Score is +2.1

In the above step, not only the score is computed if the algorithm finds a word from the words table, it also finds what dish and what aspect of the dish or the restaurant is positive or negative. This differentiates it from being an ordinary sentiment analysis problem.

In the first sentence, the words 'nice' and 'good' are identified from the words table, their score can be computed. But the problem is to assign this score to some dish or aspect, only then it is possible to identify the top dishes in a restaurant. To do that, we need to identify if that scored sentence contains any dish or any aspect of the restaurant. In other words, find any dish appears in the sentence comparing it against the dish table and any aspect appears against the aspect table. If it a match of either of these two, then assign this score to that dish or an aspect and store them in a table with the following format.

Review #	Restaurant Name	Dish	Aspect	Features	Score	Sentiment
90864	New Sri Balaji Bhavan	-	hotel	Place, nearest	0	Neutral
90864	New Sri Balaji Bhavan	-	place	Nice, good	3.7	Positive
90864	New Sri Balaji Bhavan	-	dishes	decent	1	Positive
90864	New Sri Balaji Bhavan	Masala dosa, rava masala dosa	-	love	3.1	Positive
90864	New Sri Balaji Bhavan	Coconut chutney	serve	fresh	1.3	Positive
90864	New Sri Balaji Bhavan	-	value, money	good	2.1	Positive

The last column 'Sentiment' is obtained by looking at the 'Score' column and checking if that score is above a threshold value. If it is above a pre-determined threshold value, predict the sentiment as positive. If it is below the threshold value, predict the sentiment as negative.

Depending on the actual class that this sample review belongs to, its accuracy is computed. This same procedure is done for all the reviews in the testing set. Overall accuracy, precision and recall is computed and tabulated.
After finding the sentiments of the dishes and the aspects found in the above table for all the reviews of all the restaurants available, the total number of reviews and the positive and negative reviews are also tabulated.
Then for each dish, computing the dish popularity index (DPI) in that restaurant. And finally, Sort the dishes in ascending order of DPI.

All the dishes of every restaurant is then subjected to a simple formula,

Dish Popularity Index (Percentage) = # of positive reviews / Total # of reviews

Finally, sort the dishes with this Dish Popularity Index to give the Top dishes in ascending order of DPI.

4. Data Validation Stage

Once the model is created and tested for the initial accuracy, it is mandatory to check how it will be classifiying the unseen data in terms of new reviews coming to the ML model. This contains steps of train_test_split, cross validation and more. I skipped this portion for now.

5. Data Optimization Stage

Once the model is validated, we need to optimize it for its improved accuracy. There are vast numbers of techniques available to improve the accuracy of a chosen model. One technique that I would be using is GridSearch. This will give the best set of parameters for a model that I can use finally. There are various hyperparameters to tune in the models I explored. This GridSearch will be an effective tool in this case.