SAN JOSÉ STATE UNIVERSITY

PROJECT REPORT
CMPE 255

Improving Restaurants by Analysis of Yelp Reviews

Instructor: David C. Anastasiu

Project Team

Gowtham Katari (011815557)
Mitesh Kumar(011524604)
Sai Supraja Malla(012055888)

1. Introduction:

1.1 Motivation:

In today's world of technology, businesses pay a lot of attention to what people say about them. People are just seconds away from seeking information and expressing their opinion using social networking apps. This is especially true for restaurants. This is because public opinion and reviews about businesses drive more customers or turn them into loyal customers and directly impact their fame and revenue. In our project, we plan to give suggestions to restaurants for improving their rating and making them successful by analysing reviews on yelp.

1.2 Objective:

Through our project, we would like to answer the following questions by using various scikit-learn methods on the Yelp restaurant reviews and attributes:

- What are the features that the clients are not happy about?
- How can we improve the rating and satisfaction level of the clients?
- What features does this restaurant lack to be competitive with other restaurants in that area?
- What are the mandatory driving features any new restaurant should have to be successful?

2. System Design and Implementation

We focus on one common text categorization task, sentiment analysis, the extraction of sentiment, the positive or negative orientation that a writer expresses toward some object. A review of a restaurant expresses the author's sentiment toward the service and quality of the restaurant. Automatically extracting consumer sentiment is important for marketing of any sort of product. The simplest version of sentiment analysis is a binary classification task, and the words of the review provide excellent cues. Consider, for example, the following phrases extracted from positive and negative reviews of restaurant. Words like *great*, *richly*, *awesome*, and *pathetic*, and *awful* and *ridiculously* are very informative cues.

2.1 Algorithms used:

Naive bayes

We assume word's position doesn't matter, and that the word has the same effect on classification whether it occurs as the 1st, 20th, or last word in the document. Thus we assume that the features $f1, f2, ..., fn$ only encode word identity and not position. naive Bayes assumption The second is commonly called the naive Bayes assumption: this is the conditional independence assumption that the probabilities $P(fi|c)$ are independent given the class c and hence can be 'naively' multiplied as follows:

$$P(f1, f2, ..., fn|c) = P(f1|c) \cdot P(f2|c) \cdot ... \cdot P(fn|c)$$

The final equation for the class chosen by a naive Bayes classifier is thus:

$$cNB = \text{argmax } P(c)\ P(f|c)$$

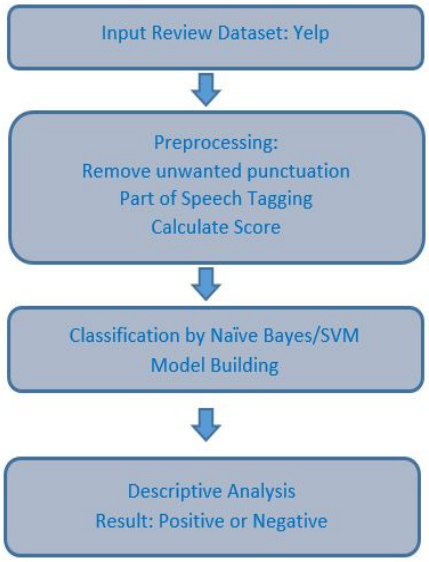
Support Vector Machine

Support vector machine is non probabilistic algorithm which is used to separate data linearly and nonlinearly. Here dataset D = {Xi, yi} where Xi is set of tuples and yi is associated class label of tuples. Class labels are -1 and +1 for no and yes category respectively. The goal of SVM is to separate negative and positive training example by finding n-1 hyperplane. Quadratic Programming (QP) problem is needed to be solved in linear data. This problem is transformed using the Lagrange Multipliers theory and Optimal Lagrange coefficients sets are obtained. A separating hyperplane is written as: W*X + b = 0 (1) where W = { w1, w2, w3, … , wn}, wn is weight vector of n attributes and b is bias. Distance from separating hyperplane to any point on H1 is 1/|W| and Distance from separating hyperplane to any point on H2 is 1/|W| so maximum margin is 2/|W|. The MMH is rewritten as the decision boundary according to Lagrangian formulation

2.2 Technologies used:

Programming Language: Python 2.7
Libraries: Pandas, Scikit-Learn, NLTK, matplotlib, seaborn
Tools: Jupyter Notebook

2.3 Workflow



3. Experiments/ Proof of concept evaluation

3.1 Dataset:

The data set is made available by Yelp Inc. to access local business information to develop an academic project as part of an ongoing course of study. This dataset is originally put together for

the Yelp Dataset Challenge which provides an opportunity for students to analyse and research on local businesses data. It has data from customers in 11 metropolitan cities. It is available in CSV format in kaggle website.
The data includes:
5M reviews by 1M users on 170K business
1.2M attributes like star rating, availability, parking and ambience
200K pictures
Yelp reviews file has 9 features and 5261668 samples
Yelp business file has 13 features and 174567 samples
Source: https://www.kaggle.com/yelp-dataset/yelp-dataset

3.2 Methodology

I. Data preprocessing:

Data preprocessing refers to any type of data processing performed on raw data to prepare the data for analysis. This will transform the data into a format that will more easy and useful for the analysis.

Data Filtering:

Yelp dataset is very huge dataset and includes many categories of business. For our specific problem data is filtered so that it only includes restaurant business and excludes others. There are two csv files: yelp_review.csv - this contains data about the restaurants and their reviews, yelp_business.csv - this contains data about different businesses. We have merged the two csv files on business_id feature.
Dataset contains data from 11 metropolitan cities, we have decided to analyze restaurants in Las Vegas city identified by business_id. After filtering the dataset we are left with 59125 reviews.

Feature Selection:

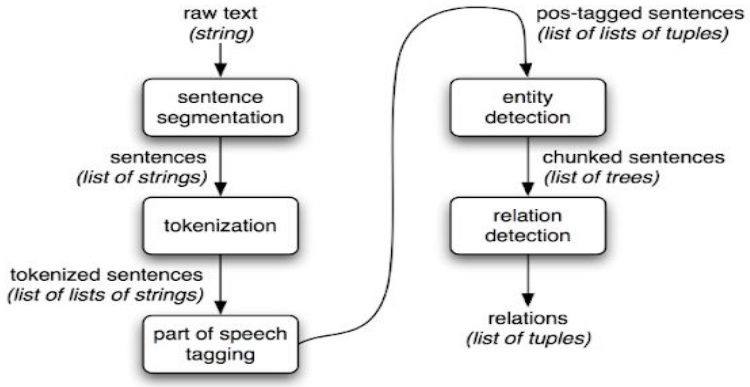
The data after merge has various features like business_id, name, neighborhood, address, stars_x, stars_y, text, etc.,. We have selected business_id, name, postal_code, categories, stars_y, text. Features that are not required are dropped from the dataframe.Only the features bussiness_id, name, postal_code, categories, star rating, text are left in dataframe.

```
In [17]: Las_vegas_full_data_df.head()
```

Out[17]:

	business_id	name	postal_code	categories	stars_y	text
182	kCoE3jvEtg6UVz5SOD3GVw	"BDJ Realty"	89128	Real Estate Services;Real Estate;Home Services...	4	I have been renting with BDJ for the past 2 ye...
183	kCoE3jvEtg6UVz5SOD3GVw	"BDJ Realty"	89128	Real Estate Services;Real Estate;Home Services...	1	Beware!!! I had the UNFORTUNATE experience of ...
184	kCoE3jvEtg6UVz5SOD3GVw	"BDJ Realty"	89128	Real Estate Services;Real Estate;Home Services...	5	I have been living in a BDJ managed home for 5...
185	kCoE3jvEtg6UVz5SOD3GVw	"BDJ Realty"	89128	Real Estate Services;Real Estate;Home Services...	5	I've rented from BDJ in the past and based on ...
186	kCoE3jvEtg6UVz5SOD3GVw	"BDJ Realty"	89128	Real Estate Services;Real Estate;Home Services...	5	I have been renting my awsume condo for 2.5 yr...

II . Phrase Extraction:



The most important step of this project is to extract informative and qualitative evaluation of the restaurants. Parts-of-Speech Tagger reads reviews as input and assigns part of speech to each word or tokens like below: Noun, Verb, Adjective, Preposition, Conjunction etc. Only the phrases in Parts-Of-Speech format are extracted. After the words are tokenized based on their tags, all the phrases that do not satisfy the pattern are eliminated.

{<JJ> <NN>|<JJ> <NNS>|<RB> <JJ>|<RBR> <JJ>}.

This pattern is used for filtering the phrases that are most meaningful i.e. described by adjectives and nouns. The phrases were tokenized and then retrieved based on POS tags using Chunking process with regular expressions.

To extract phrases Chunker package in NLTK is used. For example <JJ> <NN> tag where J-adjective and N-Noun, extracts a phrase that is in the form adjective followed by a noun

III. Polarity Determination:

Polarity of the word is a value which determines how positive or negative the word is. Pattern package is used in determining the polarity of the words that are extracted in the phrases. Pattern Package has method sentiment which returns the polarity value of the word between +1 and -1. Threshold for positive adjectives is set as 0.2 and positive nouns as 0.5. Words with

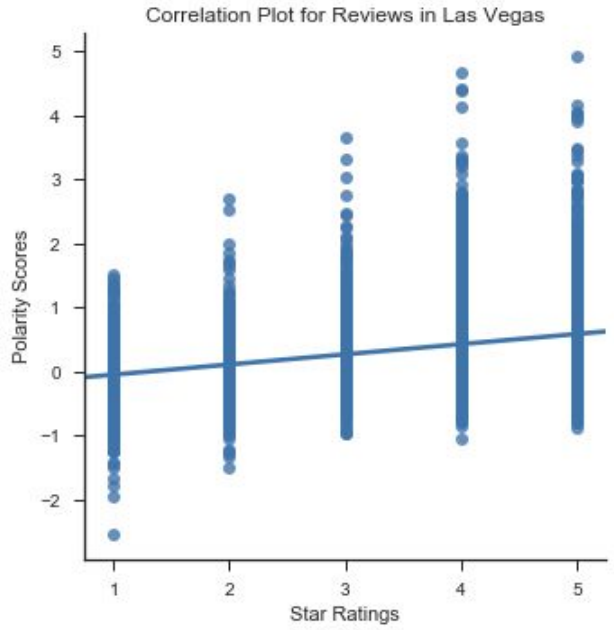
value greater than this threshold will be considered as positive phrases. Similarly, threshold for negative phrases is -0.1 for adjectives and 0.4 for nouns. Adjectives less than -0.1 and nouns having score greater than 0.4 are negative. Positive and negative polarity of phrases from the given review are calculated. The threshold values are selected by trial and error method. Table 1 shows a few examples of phrases that were selected from the reviews with their polarity scores.

Phrase	**Polarity Score**
bad experience	-0.700000
delicious pizza	1.000000
always great	0.800000
Good stuff	0.700000

IV. Correlation between average polarity scores and star rating

After the polarity values were calculated for each review as explained. Average polarity score for each of each review is computed and added as a new features in data frame. This average score is checked against the corresponding star rating of the review. This step is important as it is important to check if the phrases extracted are good enough and are representative of the reviews. Figure 1. shows a correlation plot that states if the polarity score is negative then they mainly fall in the rating (1,2) and positive scores fall in the rating (3,4,5). Graph shows linear correlation between star rating and average polarity score.

```
Text(0.5,1,u'Correlation Plot for Reviews in Las Vegas')
```



V. Classification :

City	Baseline	Accuracy
Las vegas	91.83%	99.6%

Star ratings are classified into two groups:
"Good" - (3 Stars, 4 Stars, 5 Stars)
"Bad" - (1 Star, 2 Stars).
Reviews with positive polarity scores are classified as Good and reviews with negative polarity scores are classified as Bad.

Firstly Naive-Bayes algorithm is used for classification task. Samples are divided into training and testing sets using test-train split with test size of 0.2. 80% of data is used for training and 20% of data is used for testing. Training and testing data has features average polarity score and star rating. Labels for the data are "good" and "bad"

Using naive bayes gave any accuracy of 91.83%

Using Support Vector Machine (SVM) algorithm, the accuracy score was 99.6% for predicting the reviews into Good and Bad using the star rating and the average polarity score as features. This is higher than the model trained using Naive-bayes.

VI. Restaurant Evaluation:

Goal is to compare with other local restaurants that are performing well and also that are not performing well. For this using postal code field restaurants in an area are extracted. The data is further narrowed down to the a specific kind of food like pizza restaurants are offering.

From this restaurant's the ones with rating 3,4,5 are separated from the ones with 1,2. By fetching the positive phrases for restaurants with rating above 3 it is will clear why these restaurants are performing well. Similarly by fetching negative phrases for non-restaurants with rating below 3 it will be clear why they are not performing well and what they are lacking in. This will help provide qualitative feedback on restaurants serving the same food which may not be very clear by the star rating alone.

3.3 Analysis of results:
By extracting Positive phrases from successful restaurants it is clear why there are performing well.
For example for the Business Id: fwZzpkSH4ZOJGzDzI2rSaQ the phrases great combination, excellent quality, pretty friendly, different kind clearly answers why it is successful.

```
**************************************************************************
('Business ID', 'fwZzpkSH4ZOJGzDzI2rSaQ', 'performing well due to')
--------------------------------------------------------------------------
548    [ great pizza,  so many,  free crust,  many me...
549                                                   []
550                                  [ great combination]
551                  [ Friendly stuff,  delicious pizza]
552                                  [ excellent quality]
553                           [ -Pretty good,  own pizza]
554                                                   []
555    [ top-notch quality,  very friendly,  great pi...
556                      [ good pizza,  Cheap ingredients]
557                                             [ so happy]
558    [ own pizza,  that many,  pretty friendly,  ve...
559             [ Good service,  own pizza,  good value]
560                                       [ great choices]
561          [ own pizza,  different kind,  full variety]
562                                     [ friendly staff]
563    [ pretty much,  pretty good,  fresh pineapples...
```

By extracting Negative phrases from unsuccessful restaurants it is clear why there are performing well.
For example for the Business Id: fwZzpkSH4ZOJGzDzI2rSaQ the phrases terrible service, so small, thin crust shows why the restaurant is not performing well.

```
Business ID:  fwZzpkSH4ZOJGzDzI2rSaQ not performing well due to
----------------------------------------------------------------
549           [ small corner,  not many]
553          [ thin crust,  not amazing]
554                  [ Terrible service]
555                        [ thin crust]
560                        [ thin crust]
563                       [ really thin]
567                       [ very little]
569              [ really disappointed]
570                       [ thick crust]
572                      [ casual place]
573                     [ small kitchen]
574                        [ thin crust]
575    [ extremely difficult,  so small]
579                       [ thick crust]
```

4. Discussion and Conclusions

4.1 Decisions made:

- Yelp dataset has many different businesses in it. Including restaurants, fast food centers and others. The analysis was made only on regular restaurants.
- 5 new features are extracted Positive Phrases, Negative Phrases, Negative Polarity, Positive Polarity, average polarity score.

- Model is trained using Naive bayes and Support Vector Machines
- To analysis local businesses, business only in a specific postal code are extracted. Data is furthermore reduced to only restaurants that serve pizza in that location.

4.2 Difficulties faced

- Even after filtering only restaurant businesses the data set was still huge and took too much time for processing
- Reviews are in text format which are typed by users. These texts has spelling errors, Repeated letters, some sentences which are not meaningful and grammatical errors. Such text cannot be used for qualitative analysis
- Initially we used python 3.6 but later found out that pattern library was only available for python 2.7. Because we had to uninstall the entire setup environment and reinstall for python 2.7.

4.3 Things that worked and didn't work well

- Naive Bayes algorithm gave baseline for the classification problem
- Using Naive Bayes the accuracy score was 91.83%
- Using Support Vector Machines the accuracy was 99.8%

4.4 conclusion

This analysis using yelp data set will help in getting quality feedback of the local restaurants which are successful and also restaurants that are unsuccessful. This suggestions will help in starting a new business or improving current business. We have trained a model to categorise a review as “good review” or “bad review” based on polarity score and star rating. This model can also be applied to other businesses where online reviews help drive their business.

5. Project Plan and Task Distribution

S.No	**Task**	**Assigned person**
1	Data Selection	All
2	Data Preprocessing	Gowtham, Supraja
3	Research on classification algorithms	All
4	Polarity determination	Gowtham
5	Parts of speech tagging	Supraja

6	Correlation determination	mithesh
7	Naive bayes Algorithm	Supraja
8	SVM	mithesh
9	Descriptive Analysis	Gowtham
10	Documentation	All

References:

- https://www.scribd.com/document/328359729/jadav-2016-ijca-910921
- https://www.kaggle.com/yelp-dataset/yelp-dataset
- https://web.stanford.edu/~jurafsky/slp3/6.pdf