CMPE 256 - LARGE-SCALE ANALYTICS

Group Project - Team PREDICT THIS!

By:

Sri Harsha Vanga

Jainul Patel

Manasa Yedire

Guided By:

Professor Magdalini Eirinaki

Project Link:
https://github.com/manasacsy/CMPE--256-SENTIMENT-ANALYSIS-AND-RECOMMENDATION-SYSTEM-BASED-ON-HOTEL-REVIEWS/settings/collaboration

Ch.1 Introduction

Motivation:

The motivation to do this project is to build a system that automatically suggests the Hotel owner on areas of improvement and business growth. Customer reviews are very important for any industry. We all are aware of Yelp which provides the reviews for local businesses and these reviews are sometimes so influential that it can affect the brand value. So we are using the customer reviews and performing Sentimental Analysis, suggesting the hotel owners which qualities are most liked by their customers and what are they key areas for improvement. The reviews in dataset are classified into positive reviews and negative reviews , however we had extended our project to provide the feature of classifying the review into positive or negative automatically. We also used recommendation system by finding the similar hotels nearby the particular hotel and recommending the hotel owner to add more amenities or provide the services which are not offered by the nearby hotels. As each hotels will be having thousands of reviews, it is not feasible to manually read and analyze the reviews so this automatic analysis would be very effective.

Objectives:

1. Build an analytical tool for hotel owners to assess their actions and current position in industry using Sentimental Analysis and Natural Language Processing
2. Perform analysis on top hotel reviews to get better understanding on the factors that led the top hotels to perform significantly better over rest.
3. Finding the similar hotels and merging their reviews to analyse and comprehend the positive and negative reviews of similar hotels
4. Auto Classify the review into positive and negative using a classifier model
5. Perform an evaluation of a classification model
6. Visualize the data through various graphs to understand the market in more intuitive and lucid manner.

Ch.2 System Design & Implementation details

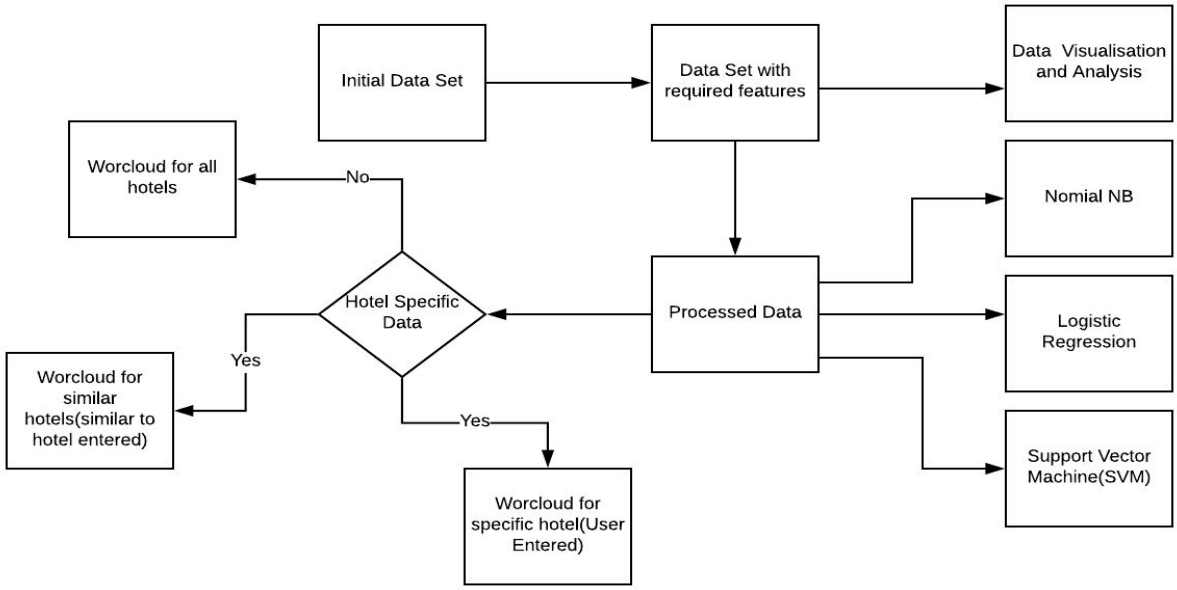
Algorithms considered:

To get the optimized results, we considered multiple algorithms for each system developed.

Technologies & Tools used:

1. Python 3 with Jupyter Notebook
   Libraries: pandas, numpy, nltk, wordcloud,TextBlob, matplotlib , sklearn,folium,spatial
2. Wordcloud for representing the most important words from a document
3. Github for repository and collaboration

System design:



Ch.3 Experiments / Proof of concept evaluation

Dataset:

The dataset contains 515,000 customer reviews and scoring of 1492 luxury hotels across Europe. Data is originally owned by Booking.com. The dataset has 17 fields which are Hotel_Address, Review_Date, Average_Score, Hotel_Name, Reviewer_Nationality, Negative_Review, Review_Total_Negative_Word_Counts:, Positive_Review, Review_Total_Positive_Word_Counts, Reviewer_Score, Total_Number_of_Reviews_Reviewer_Has_Given, Total_Number_of_Reviews, Tags, Days_since_review, Additional_Number_of_Scoring, lat, lng

Link: https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe

Data Preprocessing:

The Data preprocessing is done as series of steps

- The hotel names and reviews are converted to lowercase and the white spaces are removed to match the users text input entry
- All the reviews which doesn't add value- (for example "No positive" or "No negative") are removed
- The review text is converted to a uniform format (lowercase to achieve uniformity in comparison and analysis)
- Stopwords are taken from the default stopword list of wordcloud. generic words like hotel and room which doesn't add value are added as stopwords. Stopwords that are marked as default (eg: "not") but may add value for sentiment analysis are removed.
- The input reviews are stemmed for achieving consistency between similar text/words
- POS tagging-Adjectives and nouns are considered and is appended with _(underscore) to get more sense from wordcloud (Example: friendly_staff)

Methodology followed:

The three main goals for the project:

Analysing reviews and recommending the hotel owner for areas of improvement

- The input csv file is read as pandas dataframe for making the operations over data easy and convenient
- The hotel name is read as input from the user and this field is used to filter data from all the hotels to hotel specific data
- The positive review column is updated from considering both the positive and negative reviews using textblob .The actual positive comments are filtered from specified negative reviews columns. Similar operation is performed for getting actual negative reviews
- **Wordcloud for user specific hotels-**After the data preprocessing all the positive reviews are combined as one large sentence and this is fed as input for generating the wordcloud for performing analysis for hotel specific positive reviews. Similar operation is performed for hotel specific negative reviews.
- **Wordcloud for best performing hotels-**Using the 99 percentile for average score the top performing hotels are identified and similar wordclouds are generated for the top performing hotels to update the hotel owner about the best performing hotels in europe to recommend corrective measures
- **Wordcloud for similar hotels in proximity-**Using the latitude and longitude of each hotels the distance from target hotel is measured and the cosine similarity is measured to find the similar performing hotels and analysis for best performing similar hotels are generated and recommended.

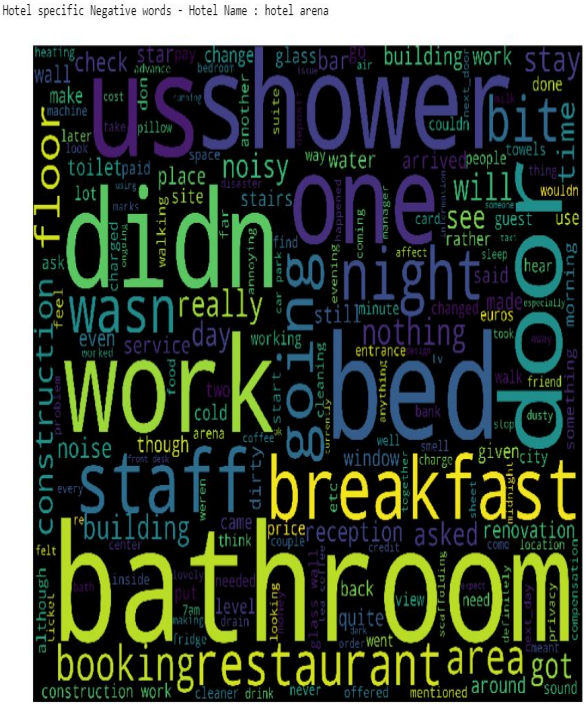
Auto Classifying a review as positive and negative by training the model with the positive and negative text review inputs:

- Extracting only Positive and Negative Reviews from the main data set.
- Merging all the positive reviews and all the negative reviews and using them to train the model.
- **Multinomial Naive Bayes** : Bag of Words transformer and CountVectorizer on preprocessed data the sparse matrix is created and the MultinomialNB model is trained with positive reviews as positive sentiment and negative reviews as negative. The model is now used to classify the text automatically into positive and negative.

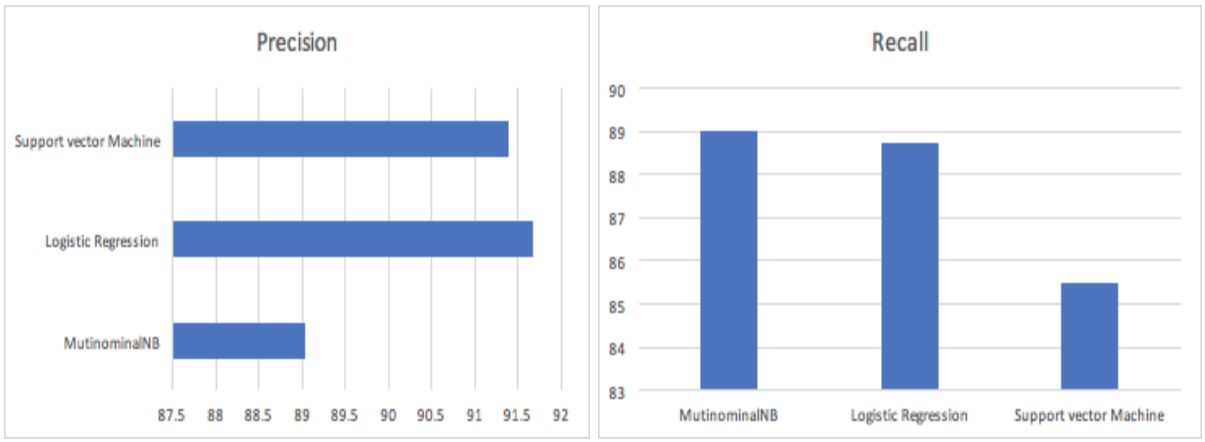
- **Logistic Regression** : SGD Classifier and Grid Search cross validation is used to train and evaluate the model. Cross Validation parameter is set to 5 and the sentiment is being predicted for the training data set and compared to test sets.
- **Support Vector Machine :** The Support Vector Machines and Grid Search cross validation to train and evaluate the model. Here also the CV value is set to 5 and model is trained and sentiment values are predicted.
- For all the three different methods , the precision, recall and accuracy are calculated and the results are compared. Logistic regression is giving the best precision and accuracy among these 3 models.

Analysis of results:

Word Cloud - Recommendation



Auto Classifying the sentiment of a review

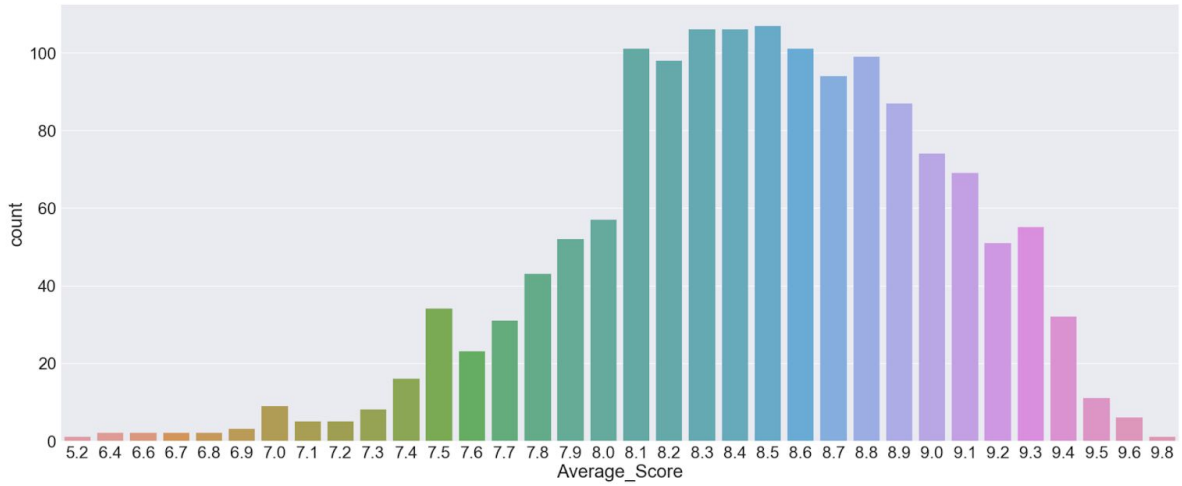


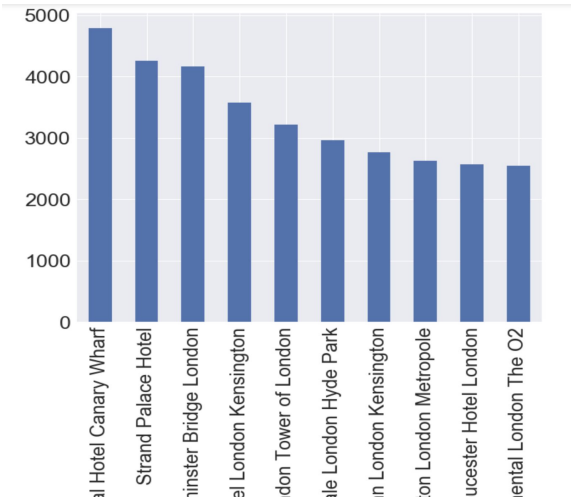
Visualisation of data for better insights and confident predictions

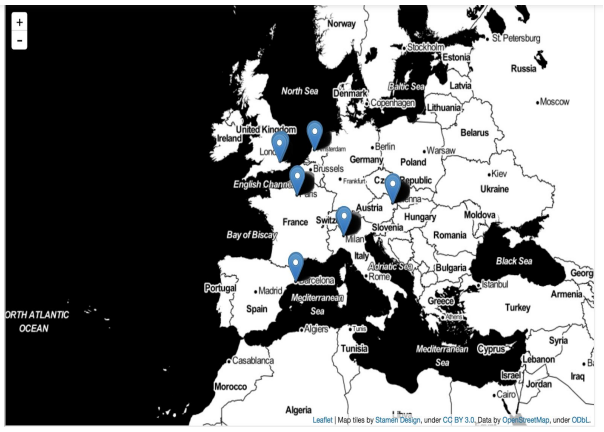
- Missing values of latitude and longitude(17 hotels) are replaced by checking the values manually.The duplicate hotels are removed
- Using Matplotlib bar graph are generated by grouping the hotels according to the average score(X-axis) and the count of hotels(Y-axis)
- The top 10 hotels with most reviews are plotted and to analyse which hotel is more popular among reviewers
- The locations are plotted as map using folium library to provide the user a better and intuitive way of data analysis

Graphs:

Average Scores vs hotel counts







(L-R)Top 10 hotels based on most positive reviews

(L-R)Top hotels based on average scores

Ch.4 Discussion & Conclusions

Decisions made:

We have built a recommendation system and analytical tool for hotel owners to assess their actions and position in market. The idea is to help hotel owners make a vigilant decisions by making them aware of the needs of the market. The analytical tool could be used to understand the customers in a better way to serve them better. We decided to implement this idea in 3 modules. The 3 modules is discussed in detail below.

MODULE 1:

- Notify hotel owners about the positive and negative aspects of the hotel
- Analyse the good and bad features of the top performing hotels
- Notify hotel owners about the better and poorly performing aspects of similar hotels to improved business and increase customer base.

MODULE 2:

This module deals with Auto Classifying a review as positive or negative using sentiment analysis.

MODULE 3:

This module deals with visualizing the data through various graphs to understand the market data in more intuitive and lucid manner.

Difficulties faced:

The main feature of the dataset is the review columns and it has discrepancies. The negative review has few positive comments and vice versa. The data preprocessing is also challenging, like finding the adjectives and nouns and processing them as required and also in case where there are null values in longitude and latitude column. We also faced difficulty with the processing power of our laptops.

Things that worked: The missing values in longitude and latitude column are filled with the values derived from the address.After analysing various ways of data visualisation word cloud seemed to be better option rather over plain text.

In generating the word clouds after careful analysis we found out generating the word cloud for better performing hotels and most similar hotels would be very useful for the individual hotel owner.

Things that didn't work well: We tried Opinion Mining for getting the important words and their positive and negative weights but it was not showing accurate results. We also tried getting the similar hotels considering the "Tags" but since each reviewer has tagged the trips differently, we had to drop the idea of taking this column into consideration for finding similarity.The idea of accurately handling the positive and negative reviews is compromised to certain tradeoff owing to limitations with dataset.Splitting a review into multiple simple sentence is not possible with the data set as there is no separate sentences for the data

Conclusion: We were successful in accomplishing the objectives mentioned.

Ch.5 Project Plan / Task Distribution

Task	**Assignee**	**Justification (if any)**
Project Idea / Dataset Selection	All	Completed as planned
Project Analysis/ Research	All	Completed as planned
Module 1	Harsha & Jainul	Completed as planned
Module 2	Manasa & Harsha	Completed as planned
Module 3	Jainul & Manasa	Completed as planned
Project Report/ Presentation	All	Completed as planned