SENTIMENT ANALYSIS AND RECOMMENDATION SYSTEM BASED ON HOTEL REVIEWS

By:

Sri Harsha Vanga

Jainul Patel

Manasa Yedire

Guided By:

Professor Magdalini Eirinaki

DATASET

- The dataset contains 515,000 customer reviews and scoring of 1493 luxury hotels across Europe.
- Data is originally owned by Booking.com.
- The size of the dataset is ~227 MB

- 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

PROBLEM STATEMENT

MODULE 1:

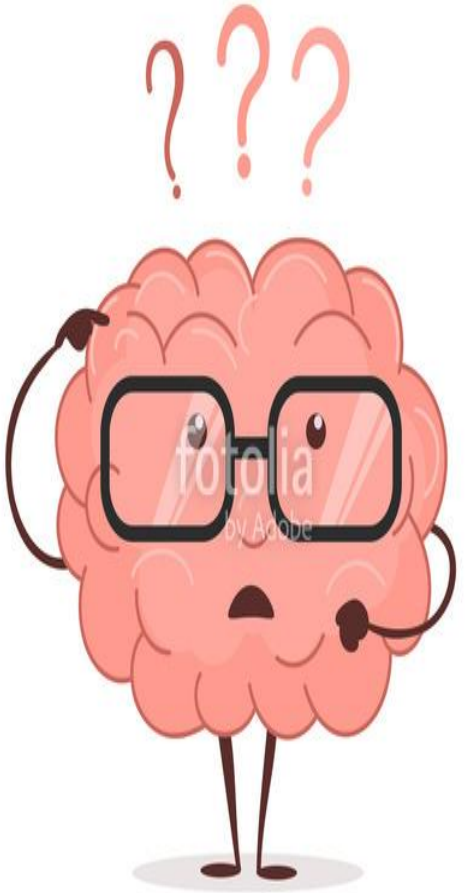
Analysing reviews and recommending the hotel owner for areas of improvement

MODULE 2:

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

MODULE 3:

Visualisation of data for better insights and confident predictions



DATA PREPROCESSING

- Uniform Data-(Lower case and whitespace removal)
- Data filtering - Removing text that does not add value(Eg : “No negative”)
- Stopwords- Customizing as per the data(addition of some stopwords and removing some)
- Stemming
- POS-tagging

SOLUTION - ALGORITHMS?

- Input from User - Hotel Name
- Output to user - Visual depiction of hotel's performance
- The positive review column is updated from considering both the positive and negative reviews using textblob.The actual positive comments are filtered Similar operation is performed for getting actual negative reviews
- Wordcloud for user specific hotels-Filtering reviews based on hotel name
- Wordcloud for best performing hotels-Considered 99 percentile by average score
- Wordcloud for similar hotels in proximity-Measured distance and similarity is determined

SOLUTION - ALGORITHMS? CONT..

Auto Classifying a review:

- Training the model with tagging the positive and negative reviews
- Multinomial Naive Bayes
- Logistic Regression
- Support Vector Machine
- 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.

SOLUTION - ALGORITHMS? CONT..

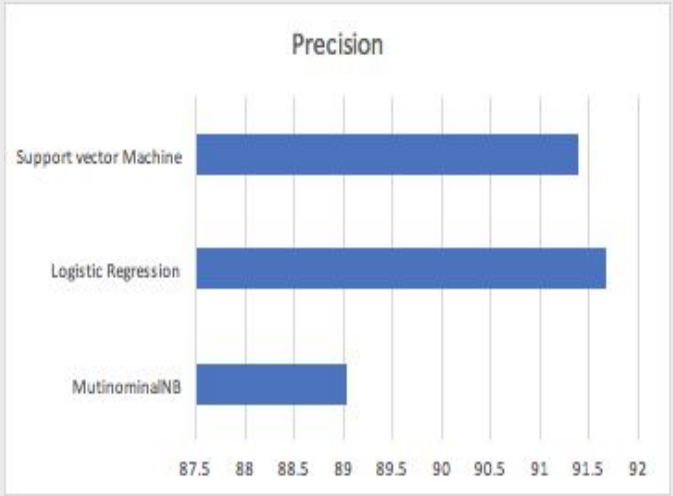
Visualisation of data for better insights and confident predictions:

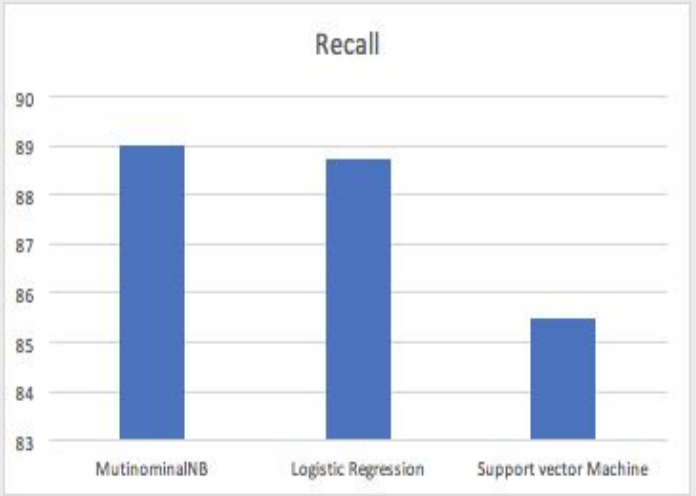
- Missing values of latitude and longitude(17 hotels) are replaced .The duplicate hotels are removed
- Bar graph are generated by grouping the hotels according to the average score(X-axis) and the count of hotels(Y-axis)-Used MatplotLib
- The top 10 hotels with most reviews are plotted and to analyse 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

WHY?

- For eliminating anomalies in positive and negative reviews textblob when considered with subjectivity and polarity is faring better
- Implementation of multiple text classifiers gives us better insights
- The visualisations which gives better insights over many are considered and implemented.So the owners would have clear idea about the market

EVALUATE

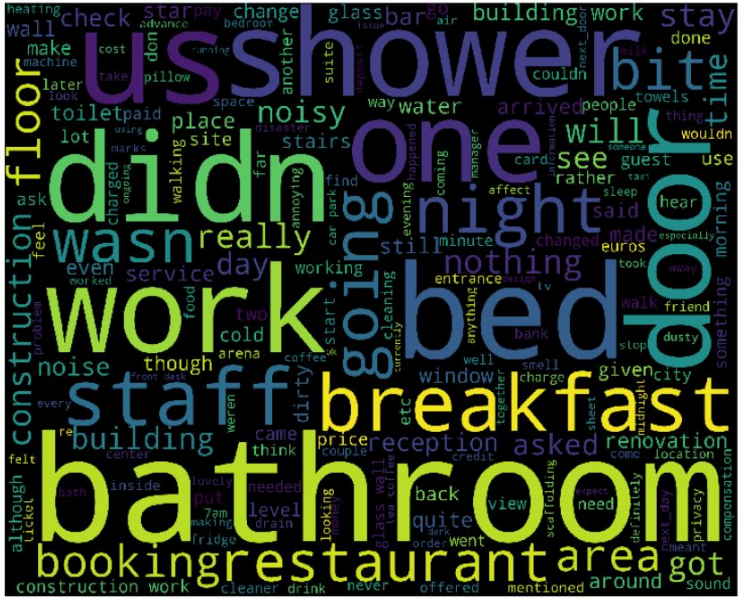

Precision
Support vector Machine
Logistic Regression
MutinominalNB
87.5
88
88.5
89
89.5
90
90.5
91
91.5
92


Recall
90
89
88
87
86
85
84
83
MutinominalNB
Logistic Regression
Support vector Machine

WORDCLOUD

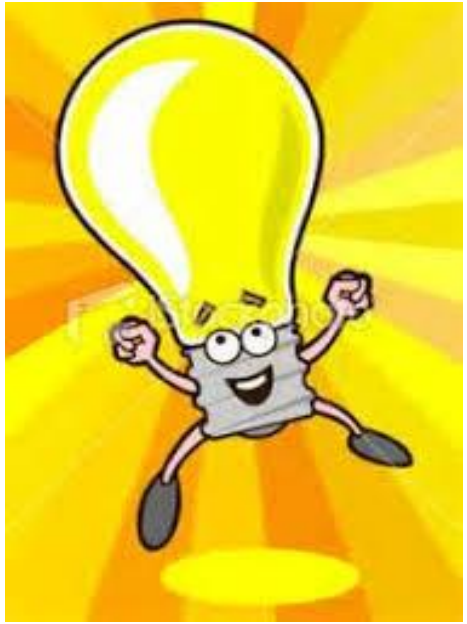
Hotel Specific Positive words - Hotel Name : hotel arena

Hotel specific Negative words - Hotel Name : hotel arena



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.
- 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

- Opinion Mining for getting the important words and their positive and negative weights but it was not showing accurate results.
- 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 sentences is not possible as there is no separate sentences in data set

THANK YOU