Sentiment Analysis of Hotel Reviews

By

Anish Prabhu(UIN: 660261580)

Judhajit Roy(UIN: 669429614)

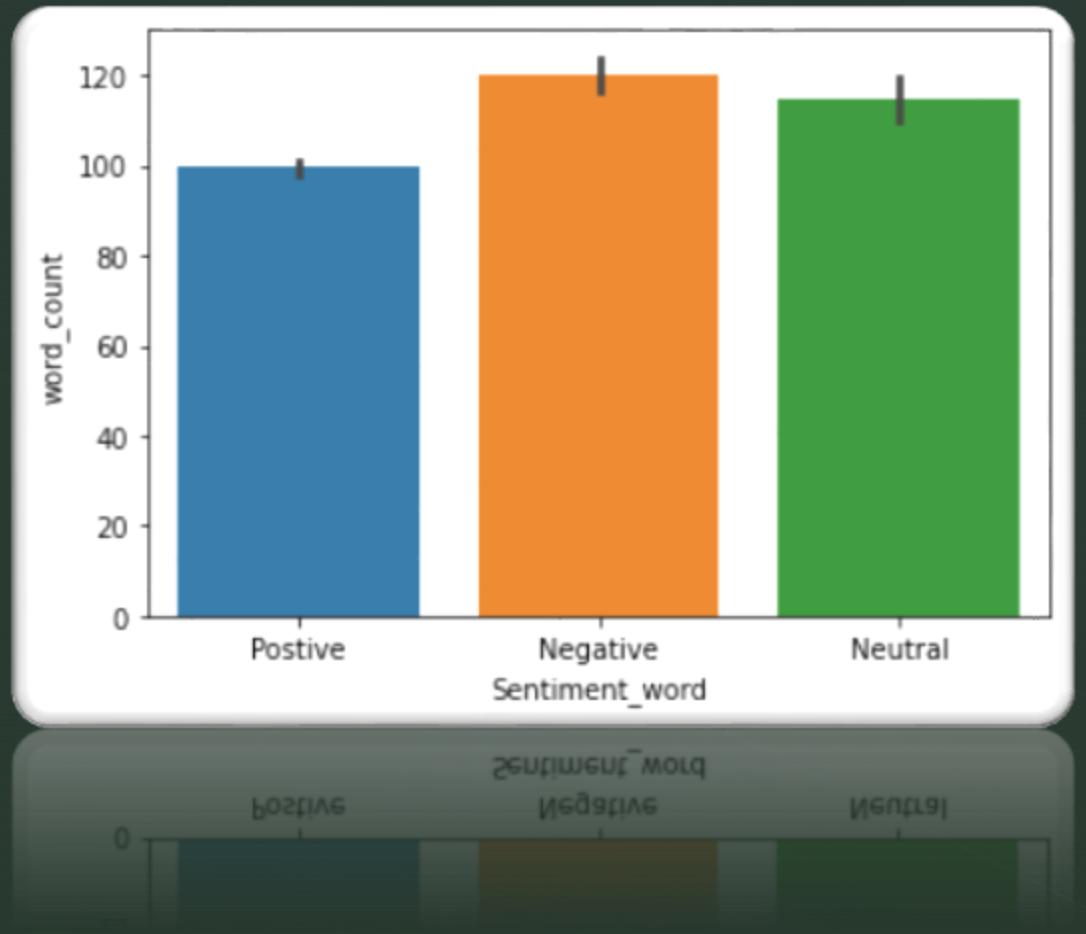
Problem Statement

- Our objective is to analyze the given hotel reviews into sentiments.
- Sentiment Analysis is the process of analyzing text in order to classify the sentiment (positive, negative and neutral) of the expression.
- So our problem is a Multinomial Classification problem.
- It is helpful in getting insights from large volumes of text.

Dataset Description

- Dataset consists of 20k+ hotel reviews from Tripadvisor.
- Tripadvisor is an American travel company that has a website and mobile app for hotel bookings, transportation and other tourism requirements.
- With this dataset it can be explored what makes a good hotel and how to provide a better experience to customers.
- Features in dataset are Reviews and their ratings on the scale of 1 to 5.

Word Count Distribution



Classification

- A classifier maps the input data to the correct target value/class.
- There are different types of classification problems namely Binary (two classes), Multinomial (more than two classes) and Multi-label classification (multiple classes assigned to each input).
- For example: Binary (classes are cat and dog), Multinomial (classes are cat, dog, snake and wolf) and Multi-label (both cat and dog)

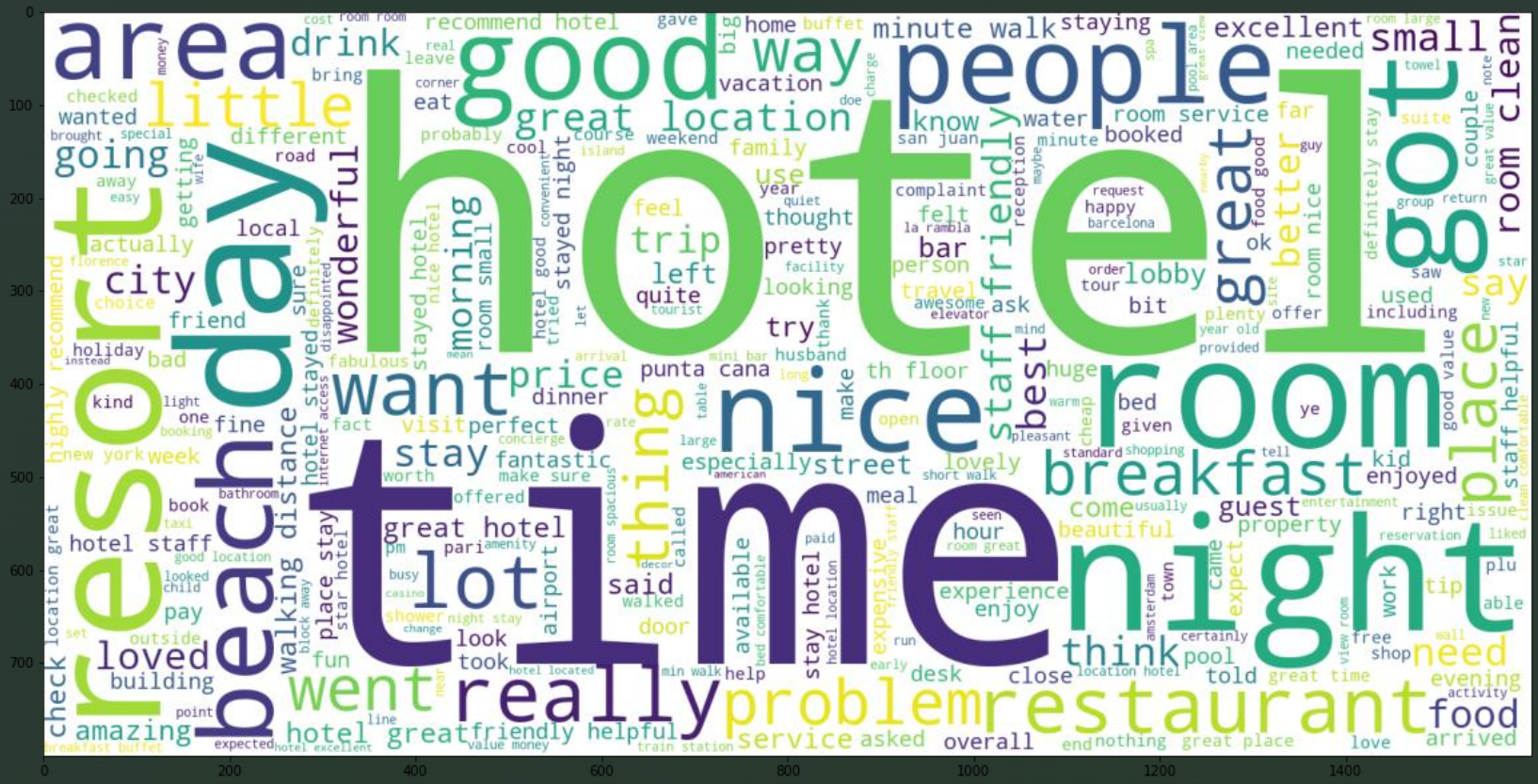
Preprocessing

- Raw textual data contains a lot of irrelevant information which is removed before analysis and training can be done.
- Techniques used for preprocessing include stopwords removal, lemmatization, removing punctuations, POS-tagging etc.
- Also, the ratings on the scale 1 to 5 are categorized into sentiments (1-2 is Negative, 3 is Neutral and 4-5 is Positive).
- For training we converted the words to numeric vectors using the TF-IDF vectorizer.

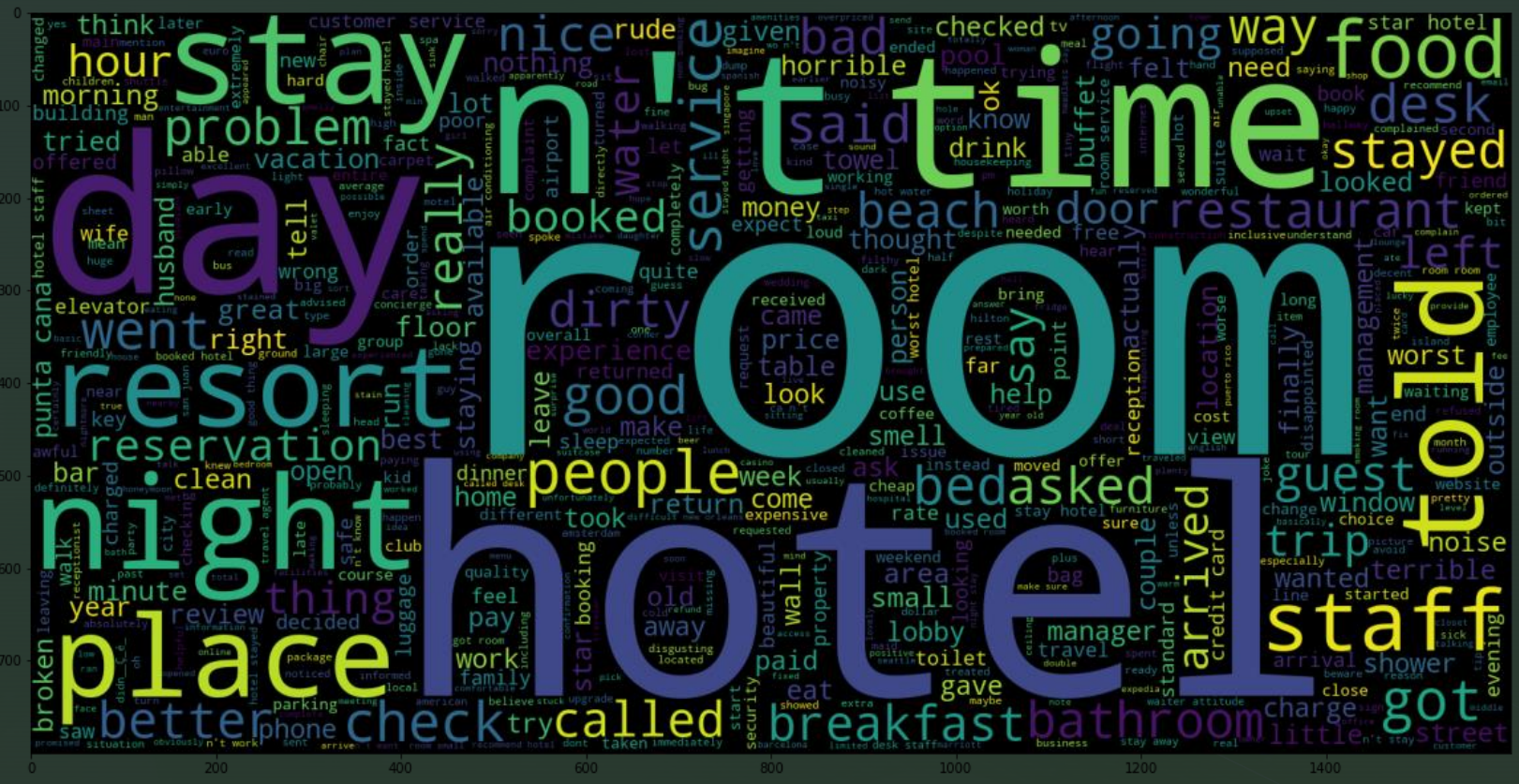
Drawbacks of Stopwords

Before Stopwords	After Stopwords
The product is really very good **(Positive)**	product really good**(Positive)**
The products seems to be good. **(Positive)**	products seems good **(Positive)**
Good product I really liked it**(Positive)**	Good product really liked **(Positive)**
I didn't like the product **(Negative)**	like product **(Positive)**
The product is not good **(Negative)**	product good **(Positive)**

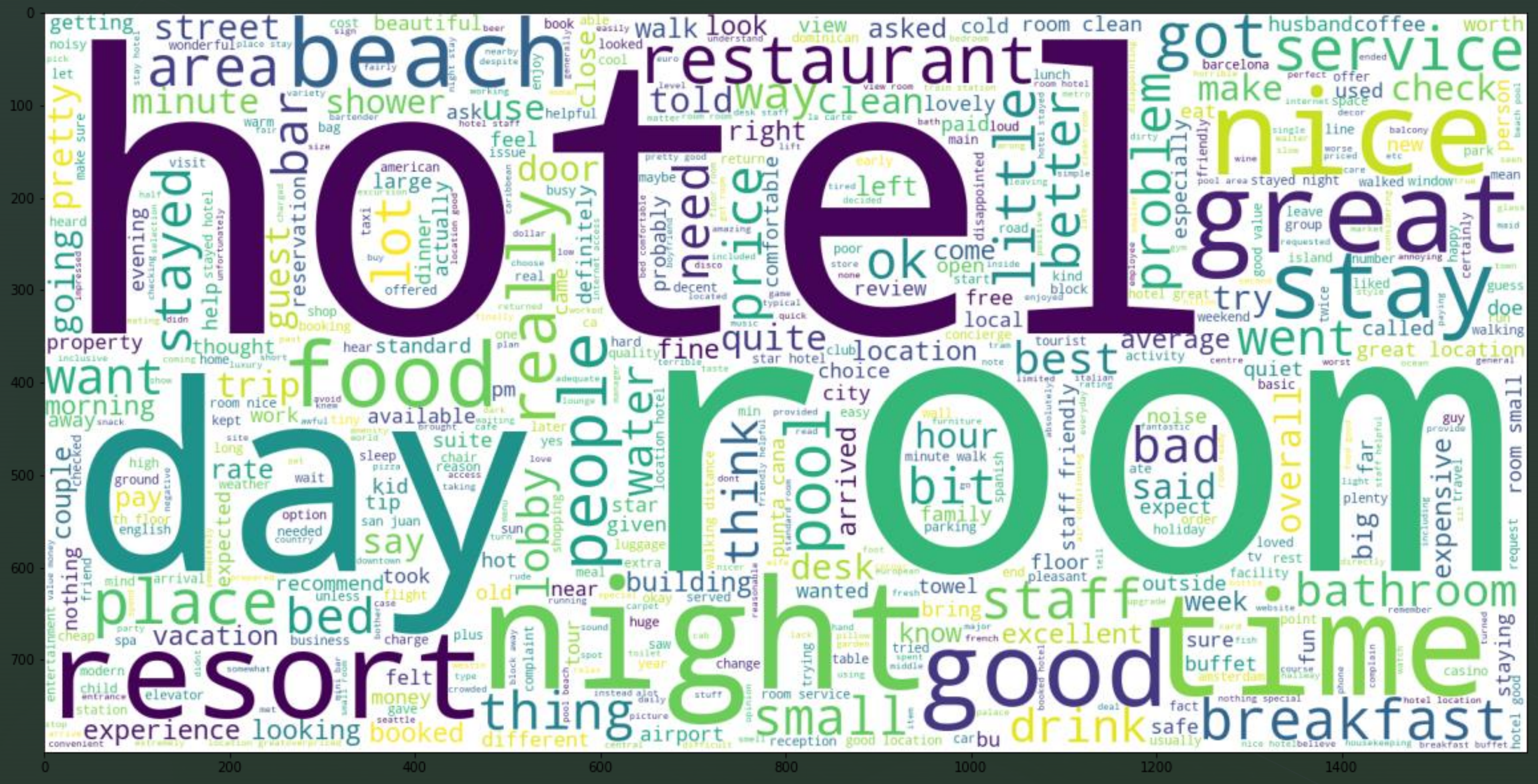
Good Reviews Wordcloud



Bad Reviews Wordcloud



Neutral Reviews Wordcloud



Models Used

Model	Accuracy Score	Train time
Logistic Regression	85.9%	6.42s
Multi-layer Perceptron	86.3%	34.95s
Gaussian Naïve Bayes	66.4%	2.32s
Multinomial Naïve Bayes	82.9%	0.01s

Cross-Validation on Multi-layer Perceptron

With different no. of Hidden layers and neurons

Hyperparameter (No. Of Neurons)	Mean Validation Accuracy
(50,50)	0.86198663
(100,)	0.86641075
(30,30,30)	0.85801697
(100,100)	0.86127031

Impact of cross-validation on Logistic regression

- Logistic regression score training on entire training data was 0.85
- With 5 fold cross-validation, mean score is 0.861

Hyperparameter tuning for Logistic Regression

- We chose to change the penalty and see the testing scores
- Score after using L1 penalty: 0.862
- Score after using L2 penalty: 0.859
- Score after using Elasticnet penalty: 0.858
- Score after using no penalty: 0.841

Future Improvements

- Word2vec and Glove embeddings can be used to improve the performance of model.
- Performance on more hyperparameters like values of L2 penalty and different optimizers can be tried to tune the model.
- Fine tuning the model by testing on tools like GridSearchCV and RandomSearchCV.