Sentiment Analysis on Trip Advisor Hotel Reviews

Done by: Walid El Kassem

1. Collecting data

- The data is taken from Kaggle .
- It consists of two columns (Review, Ratings) and 20491 rows

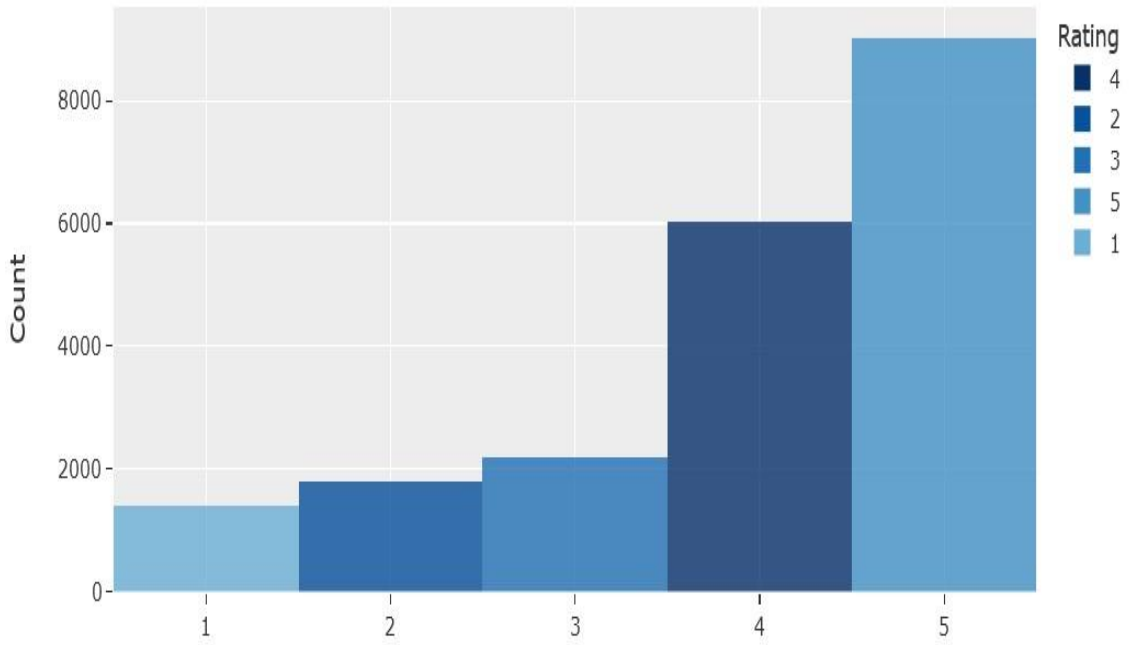
	Review	Rating
0	nice hotel expensive parking got good deal stay hotel anniversary, arrived late evening took adv...	4
1	ok nothing special charge diamond member hilton decided chain shot 20th anniversary seattle, sta...	2
2	nice rooms not 4* experience hotel monaco seattle good hotel n't 4* level.positives large bathro...	3
3	unique, great stay, wonderful time hotel monaco, location excellent short stroll main downtown s...	5
4	great stay great stay, went seahawk game awesome, downfall view building did n't complain, room ...	5
...	...	...
20486	best kept secret 3rd time staying charm, not 5-star ca n't beat, time stayed increased esteem, b...	5
20487	great location price view hotel great quick place sights.directly street space needle downtown t...	4
20488	ok just looks nice modern outside, desk staff n't particularly friendly, corridors dark smelt st...	2
20489	hotel theft ruined vacation hotel opened sept 17 2007 guests week, happy stumble scouting hotels...	1
20490	people talking, ca n't believe excellent ratings hotel, just n't, yes patricia extremely helpful...	2

20491 rows × 2 columns

2. Data Analysis

- no missing entries
- no duplicates
- unbalanced distrubtion of ratings:

Rating	Counts
5	9054
4	6039
3	2184
2	1793
1	1421



3. Create new labels and balance the dataset

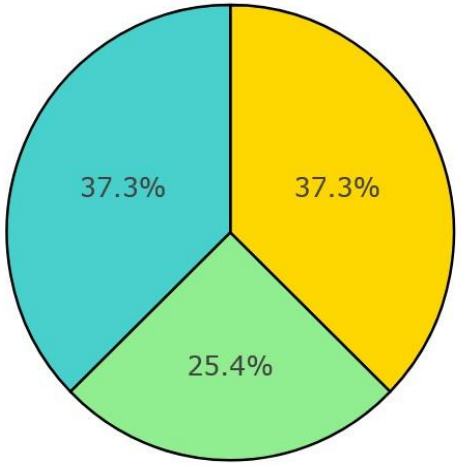
- Reduce the number of labels into 3: Positive, Negative & Neutral

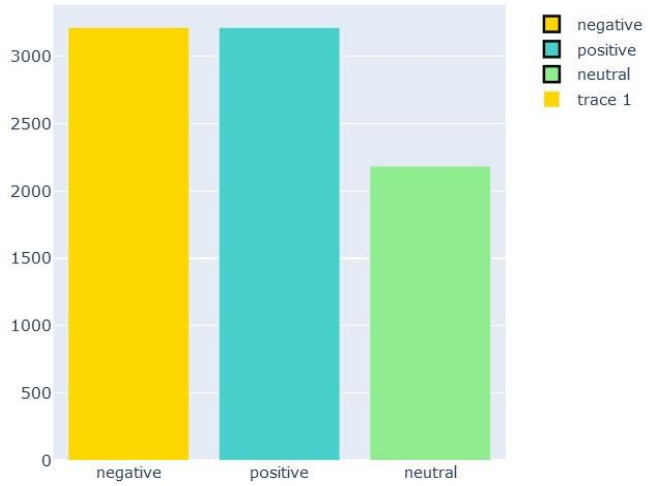
  Labels > 3 are Positive

  Labels = 3 are Neutral

  Labels < 3 are Negative

- Balanced dataset:





Labels	Counts
Positive	3214
Negative	3214
Neutral	2184

4. Cleaning Data

Step 1. lower case

Step 2. remove html and urls

Step 3. remove emojis

Step 4. remove non–ascii characters

Step 5. remove punctuation

Step 7. remove numbers

Step 8. remove stopwords

Step 9. lemmatization with spacy

Step 10. Stemming with Spacy

Before	After
Perfect !!! 👍 , Love Old Hotels 😊, my @ husband & children full satisfied,	perfect love old hotels my husband children full satisfied

5. Train & Test the Models

- use TF-IDF to convert the reviews into vectors

Create 7 Models & use the **Cross-Validation** technique

Model	Accuracy
1. Decesion Tree Classifier	55.04 %
2. Random Forest Classifier	68.93 %
3. Support Vector Classifier	74.78 %
4. Logistic Regression	74.78%
5. KNeighbors Classifier	57.45 %
6. BernoulliNB	67.77 %
7. Gaussian Naive Bayes	60.27 %

6. Use **Spacy** for word embeddings

- use **Spacy** to convert the reviews into vectors

- Create 7 Models:

Model	Accuracy
1. Decesion Tree Classifier	49.61 %
2. Random Forest Classifier	64.86 %
3. Support Vector Classifier	70.47 %
4. Logistic Regression	57.85 %
5. KNeighbors Classifier	72.17 %
6. BernoulliNB	59.48 %
7. Gaussian Naive Bayes	60.10 %

7. Keras (Deep-Learning-Library)

- Universal-Sentence-Encoder-Transformer from Tensorflow (Google)
- 3 layers:

  1st. 256 neurons

  2nd. 128 neurons

  3rd. 3 neurons activation function : sigmoid
- **Accuracy:** 66.80 %

Conclusion

Yes !!!

we can analyse the sentiment of the customers using the Logistic-Regression Model