# Methodology and Data

1. **Data Collection**

   The data was collected directly from the yelp.com website. A total of 50,000 reviews were used. The reviews contained the user information: User_id, busines_id, and date of the review. The other parameters were useful, funny, and stars along with the text review.

   The data file was exported from JSON to CSV for analysis.

2. **Data Cleaning and Filtration**

   It was done to remove incomplete, irrelevant, corrupted data.

   The parameters, user_id, business_id, and date of reviews were dropped.

   The initial count of respective stars was as follows:

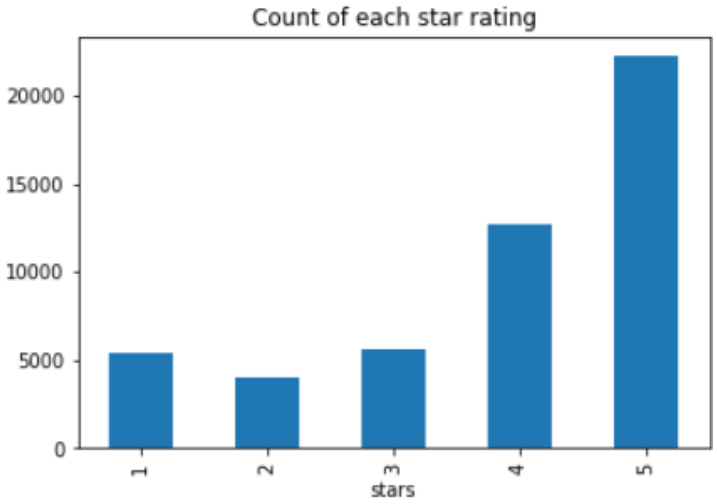



**Fig1: Count of each star**

Stop words were removed from the text reviews through lemmatization.

The following graph shows the review word count vs the stemmed word count in all the text reviews:

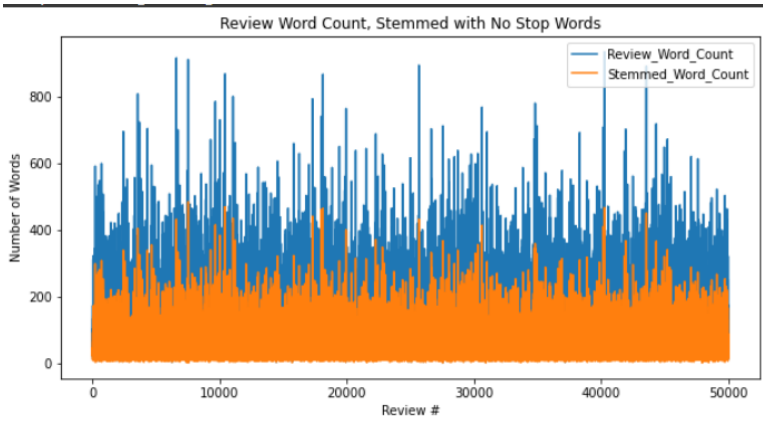


**Fig2: Review word count vs the stemmed word count**

Positive word count and positive word percentage were found from positive lexicons. Similarly, negative words and negative words count were found in negative lexicons.

## 3 EDA

Since the data was unlabeled, the polarity( and hence the sentiments) were assigned using TextBlob's polarity value. The used values were as follows: Polarity >0.4 as positive, Polarity>0.2 & polarity <0.4 as slightly positive; polarity>0 & polarity<=0.2 as slightly negative, and polarity <0 as negative reviews.

The assigned sentiments of the training dataset:

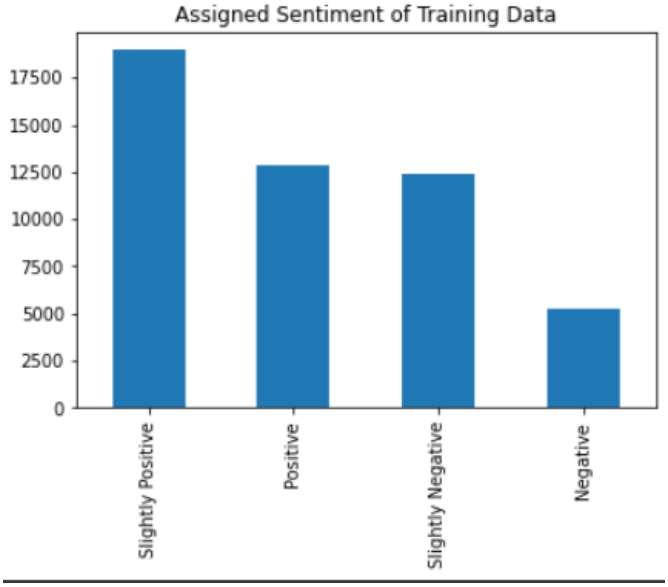



**Fig3: Assigned sentiment of Training Data**

In feature processing: The numerical features, stop word %, word count and stemmed words didn't have a distinct difference between classifiers.

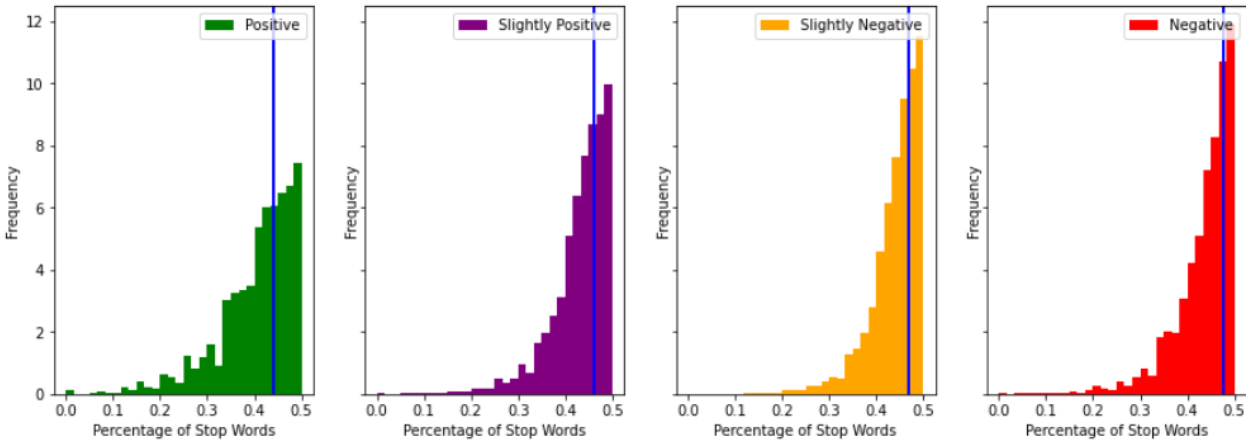



**Fig4: Percentage of Stop Words for each review class**

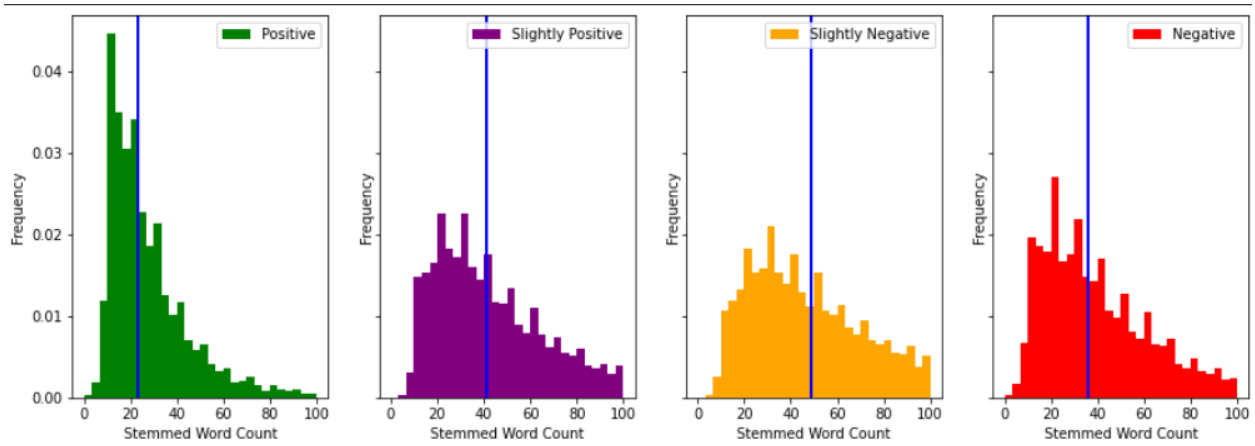



**Fig5: Stemmed word count for each review class**

The percentage of positive words stood out as the average consistently decreases as one goes from positive to negative reviews.

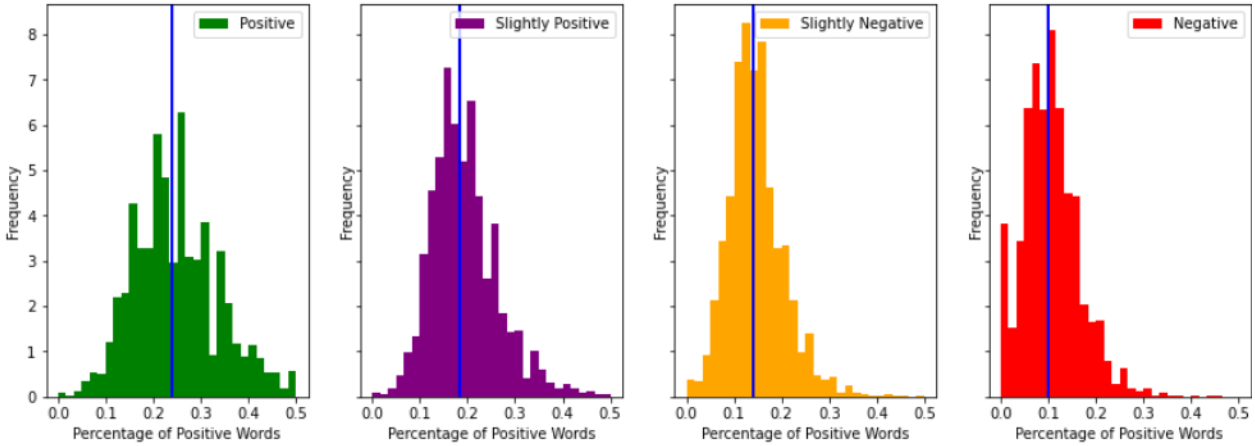



**Fig6: Percentage of positive words for each review class**

Numerical analysis is done

For the text feature-based analysis, the text reviews were converted into vectors using a count vectorizer, and TFIDF(Term Frequency Inverse Document Frequency) to compare the accuracy of both. This is done so that the model could interpret them.

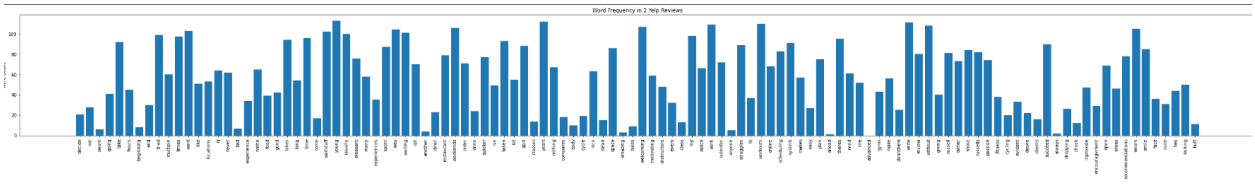


**Fig7: Word Frequency in 2 yelp reviews(using countvectorizer)**

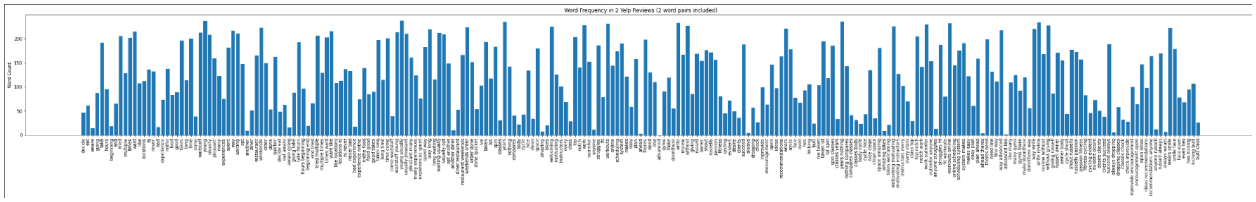



**Fig8: Word Frequency in 2 yelp reviews**

In the model selection, 5 models were tested in total.

Using seaborn, the following correlations were obtained:

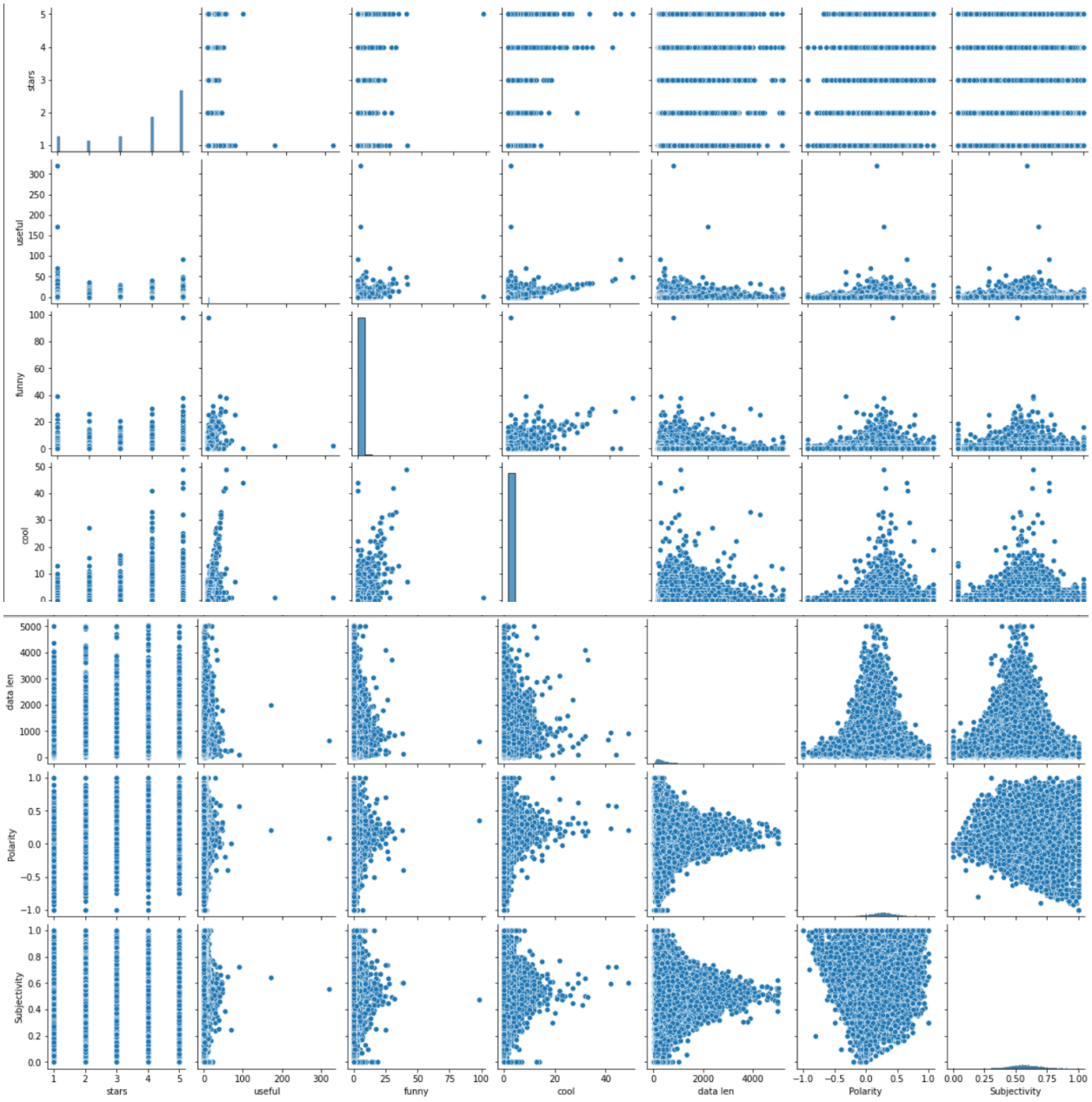
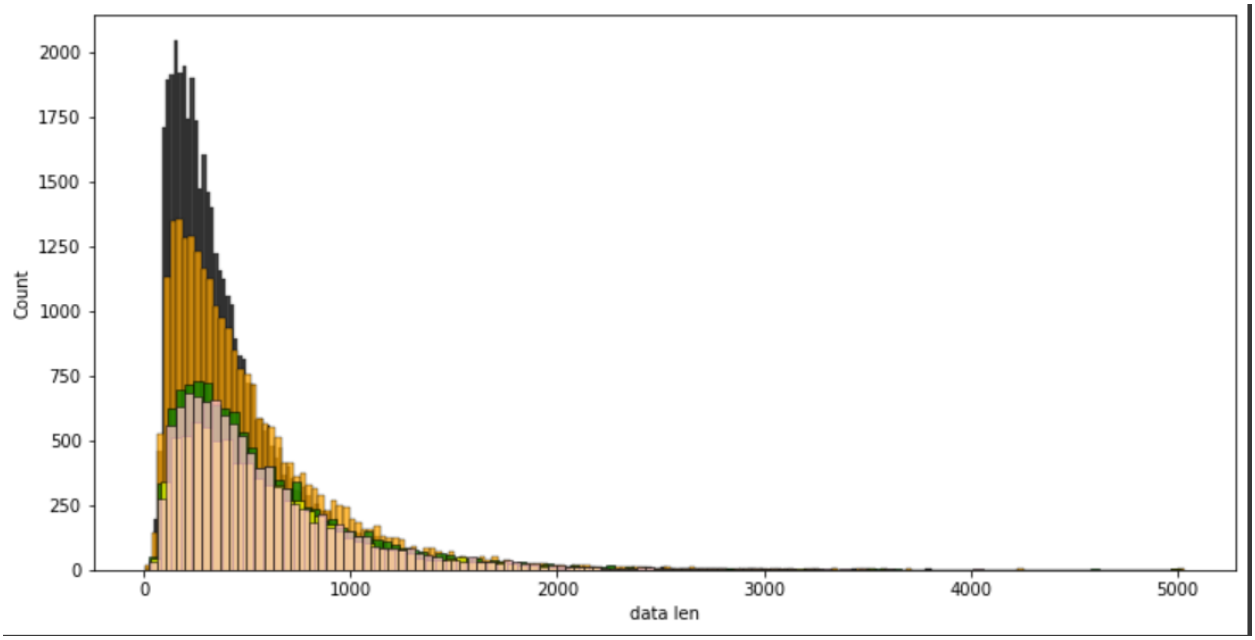



**Fig9:**

Subjectivity vs data len- most points are concentrated from 0 to 2000 data len, implying most variations occur between that range. And avg value of subjectivity across all data len is around 0.6.

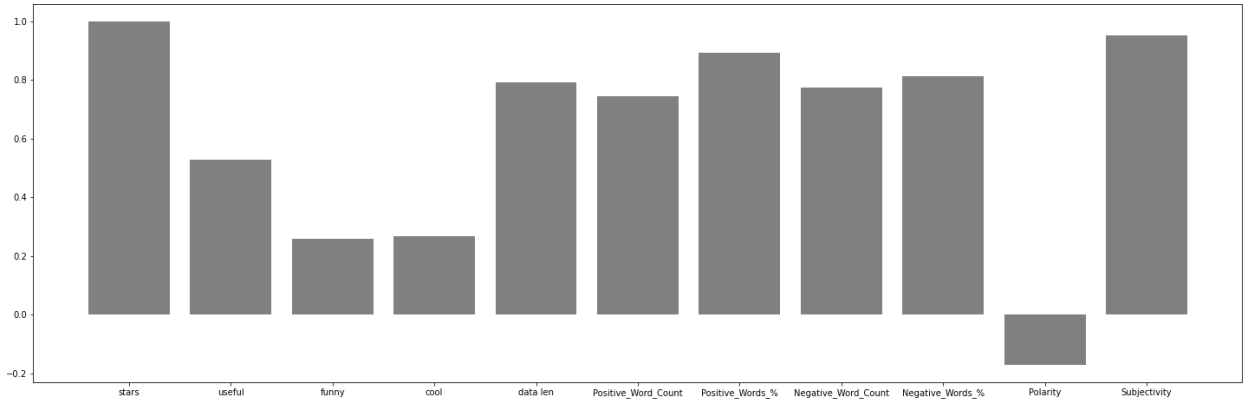


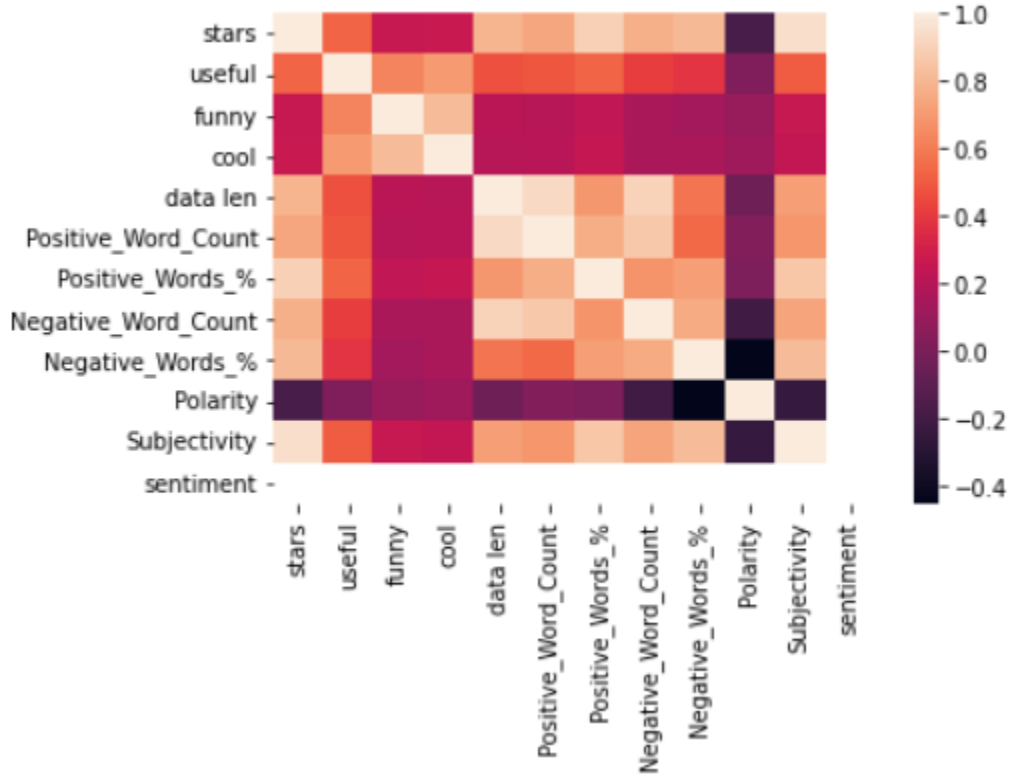

Black- 5 star
Orange-4 star
Yellow-2 star
Pink-1 star
Green-3 star

Most stars have a similar distribution of data length, lying mostly below 1000. However, 5 star is more concentrated towards 500 while 1 star extends till 2000, implying 5-star reviews are shorter than other stars and 1-star is reviews are longer than any other stars.

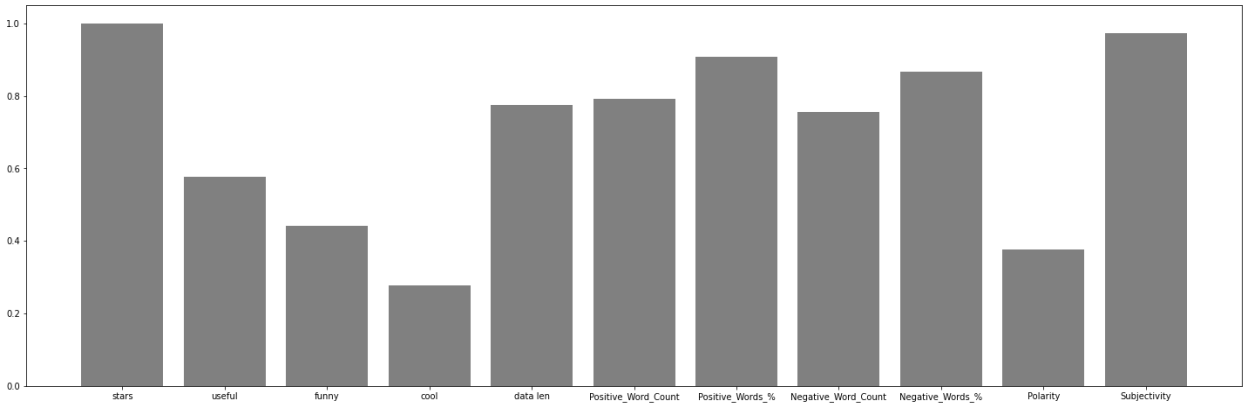
Hypothesis1:
1. There can be a discrepancy between the text sentiment and the rating, which indicates a non-valuable data source for research studies utilizing review texts or rating scores in their solution models.
Hypothesis:  Polarity should be one of the most important parameters and should increase with stars, maximum for 5 stars and minimum for 1 star.
Positive word percentage should be maximum for 5 stars and minimum for 1 star, similarly negative.

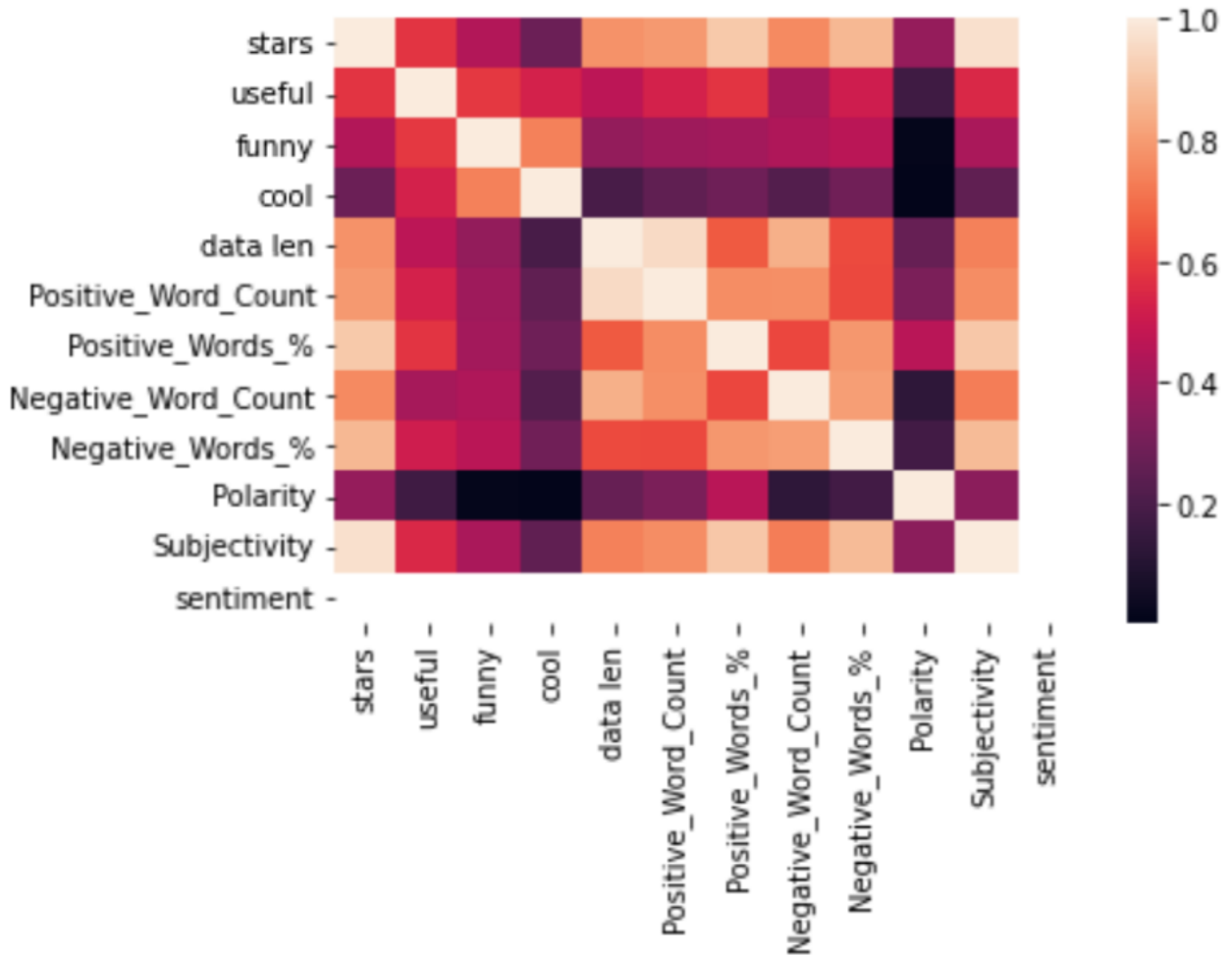
**Graphs showing the effect of each parameter on the stars.**
1 Star


1.0
0.8
0.6
0.4
0.2
0.0
−0.2
stars
useful
funny
cool
data len
Positive_Word_Count
Positive_Words_%
Negative_Word_Count
Negative_Words_%
Polarity
Subjectivity

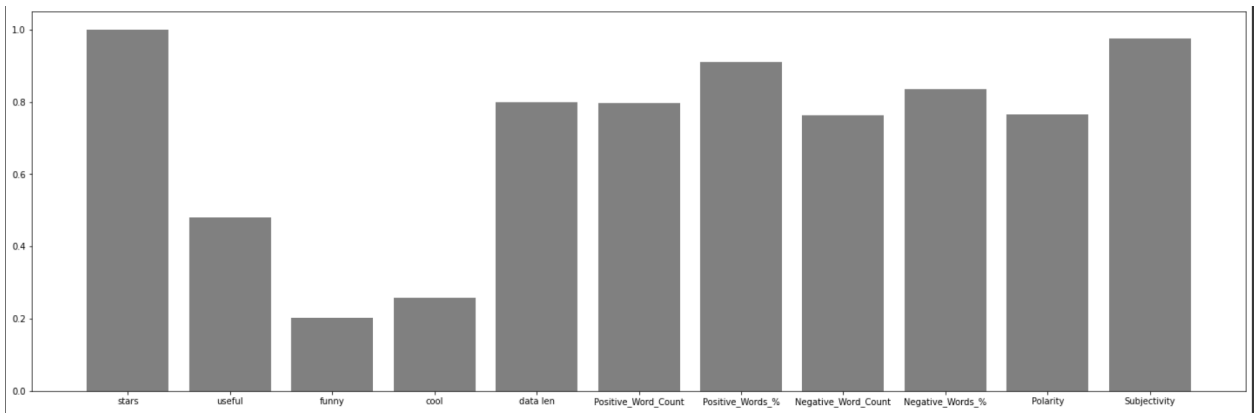

stars
useful
funny
cool
data len
Positive_Word_Count
Positive_Words_%
Negative_Word_Count
Negative_Words_%
Polarity
Subjectivity
sentiment
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

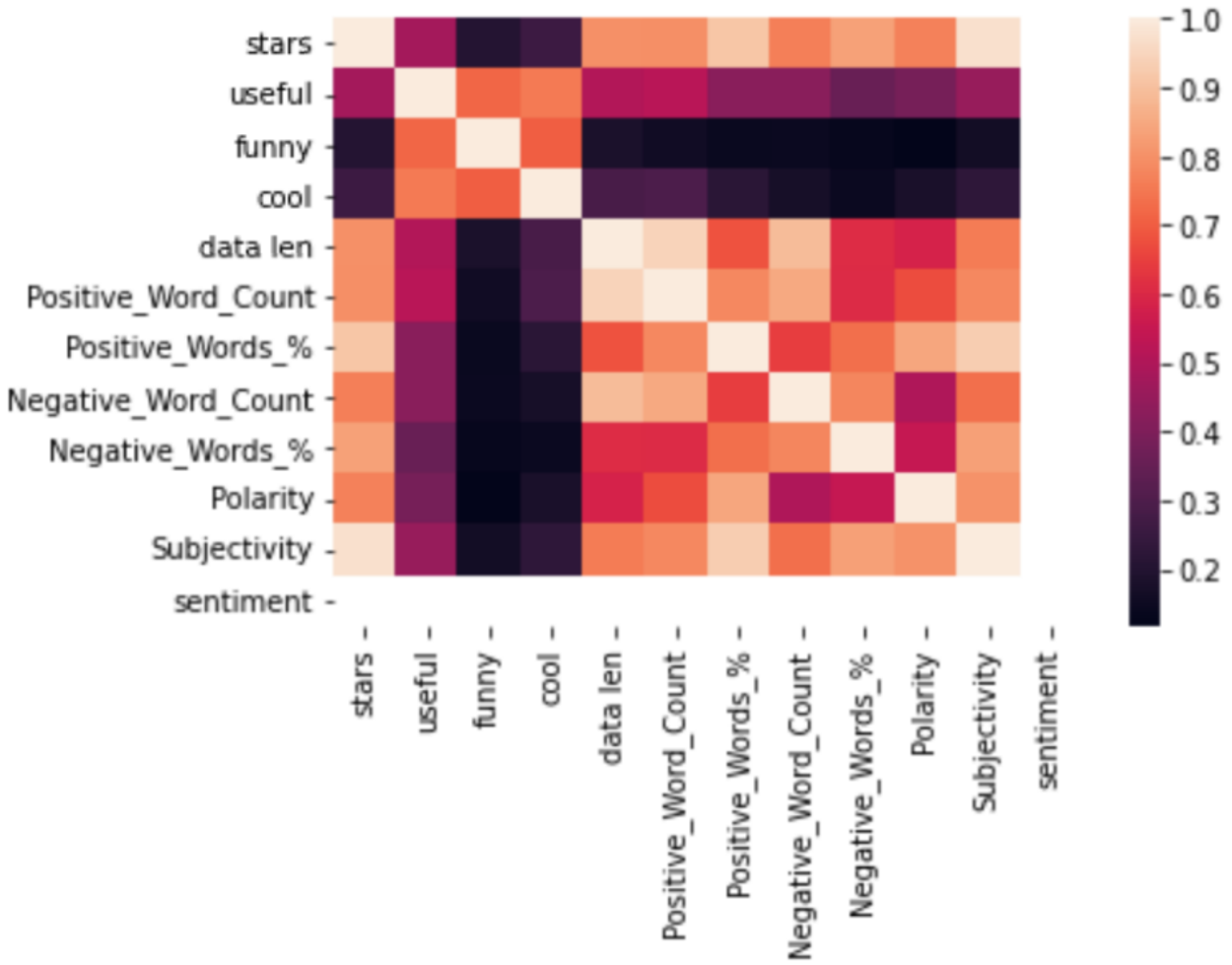
## 2 Star


1.0
0.8
0.6
0.4
0.2
0.0
stars
useful
funny
cool
data len
Positive_Word_Count
Positive_Words_%
Negative_Word_Count
Negative_Words_%
Polarity
Subjectivity

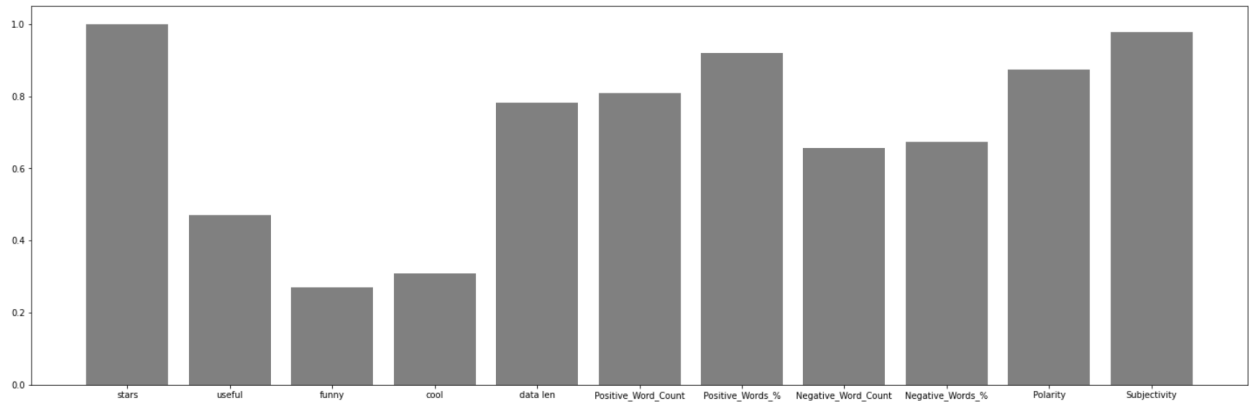

stars
useful
funny
cool
data len
Positive_Word_Count
Positive_Words_%
Negative_Word_Count
Negative_Words_%
Polarity
Subjectivity
sentiment
1.0
0.8
0.6
0.4
0.2

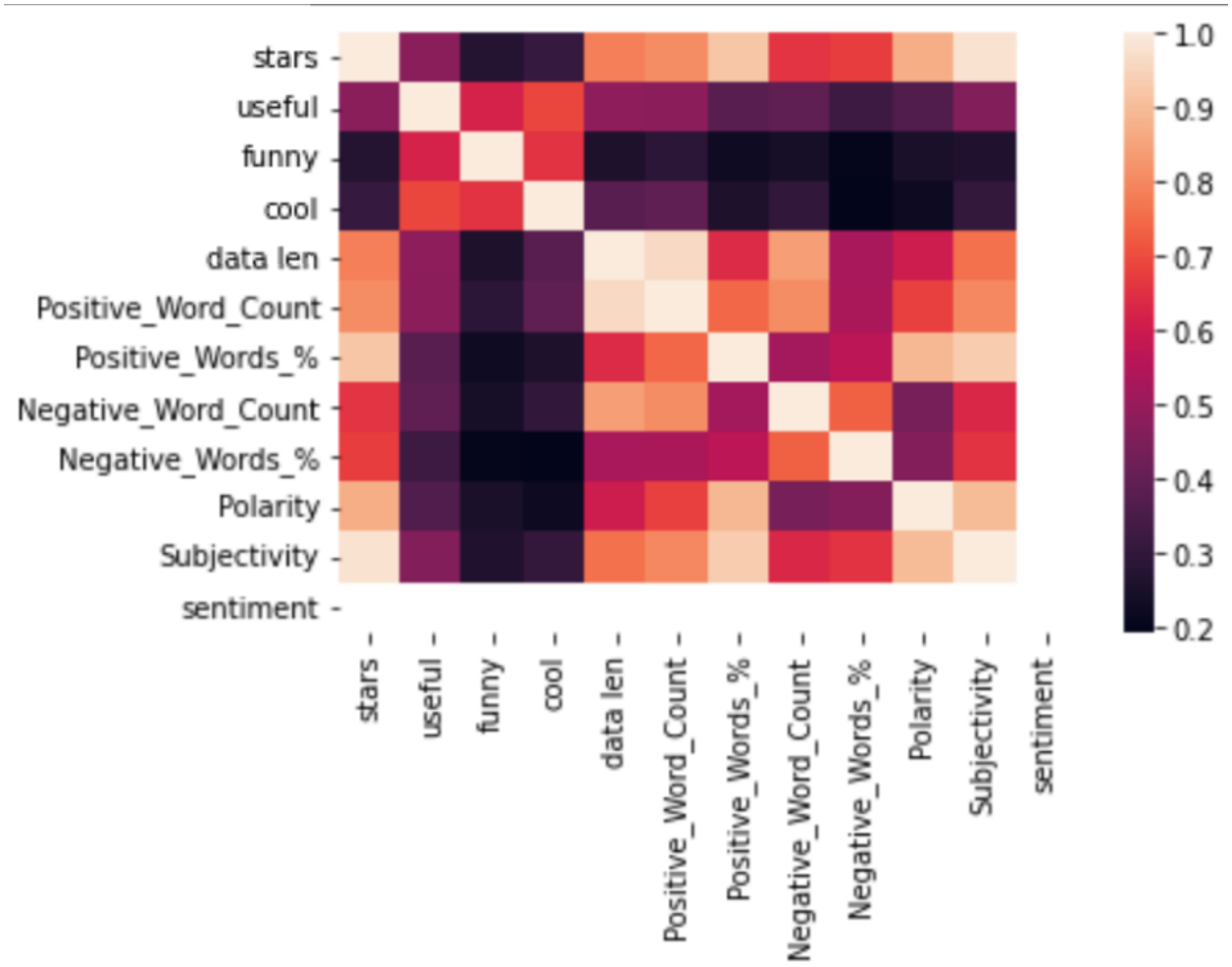
3 Star


1.0
0.8
0.6
0.4
0.2
0.0
stars
useful
funny
cool
data len
Positive_Word_Count
Positive_Words_%
Negative_Word_Count
Negative_Words_%
Polarity
Subjectivity

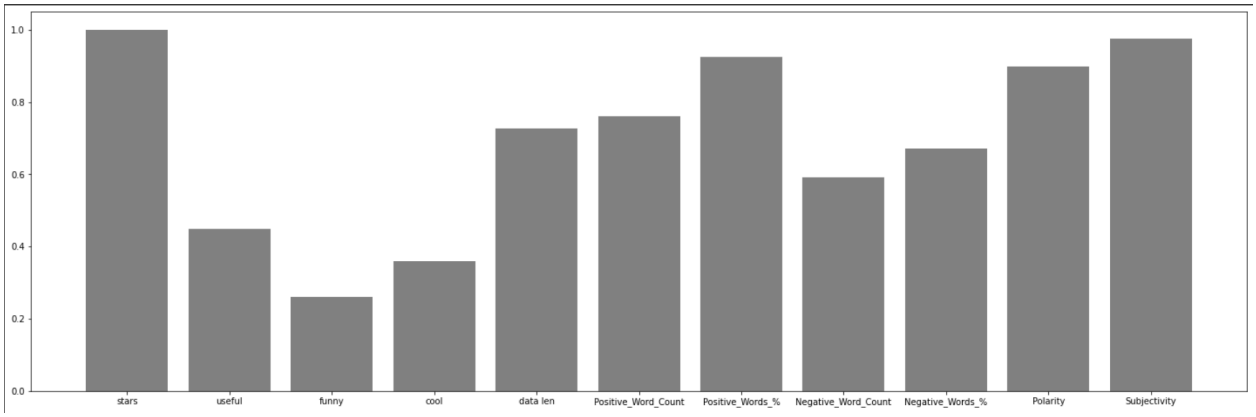

stars
useful
funny
cool
data len
Positive_Word_Count
Positive_Words_%
Negative_Word_Count
Negative_Words_%
Polarity
Subjectivity
sentiment
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

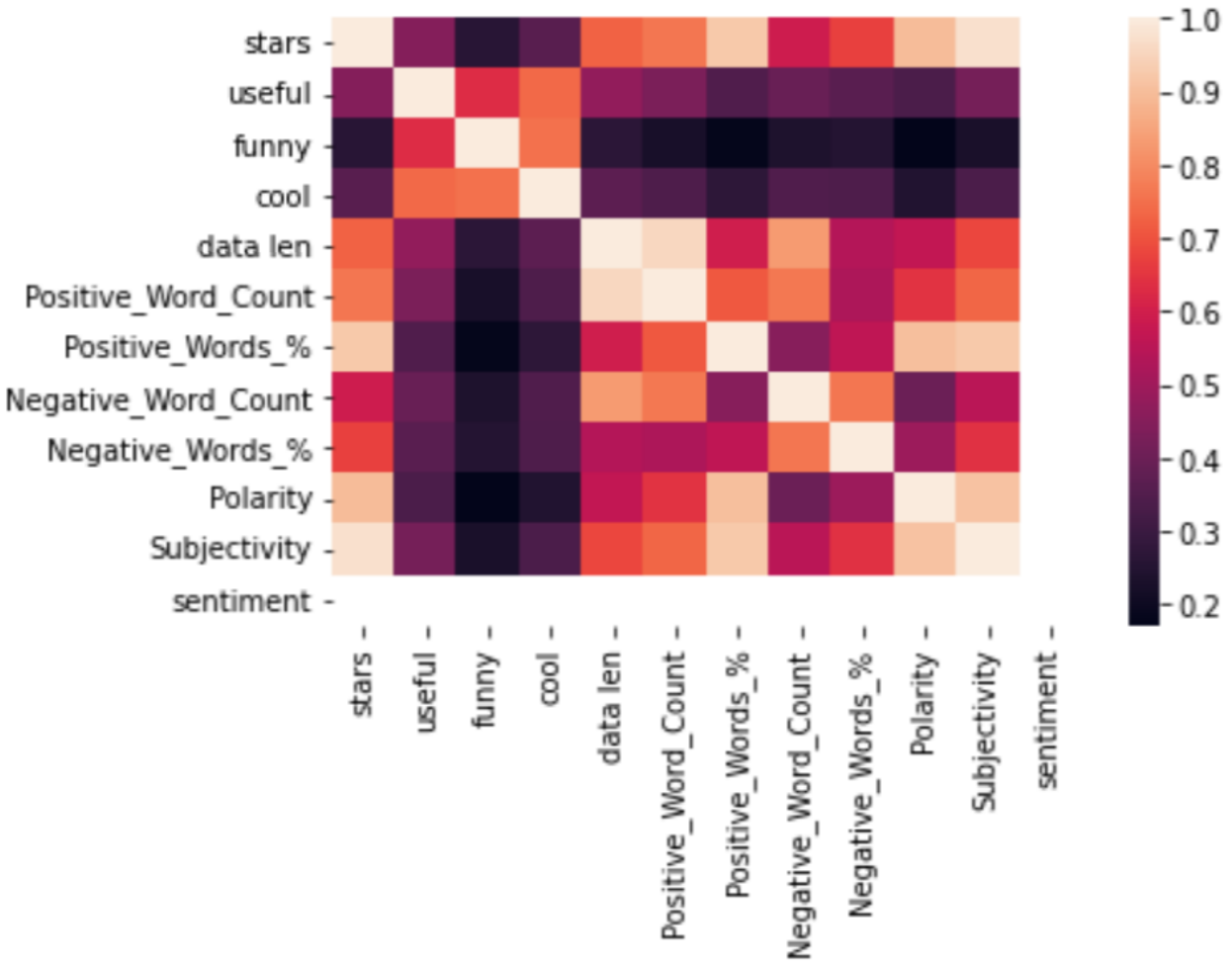
4 Star


1.0
0.8
0.6
0.4
0.2
0.0
stars
useful
funny
cool
data len
Positive_Word_Count
Positive_Words_%
Negative_Word_Count
Negative_Words_%
Polarity
Subjectivity


stars
useful
funny
cool
data len
Positive_Word_Count
Positive_Words_%
Negative_Word_Count
Negative_Words_%
Polarity
Subjectivity
sentiment
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

5 Star


1.0
0.8
0.6
0.4
0.2
0.0
stars
useful
funny
cool
data len
Positive_Word_Count
Positive_Words_%
Negative_Word_Count
Negative_Words_%
Polarity
Subjectivity




**Correlation data of different parameters.**

As expected polarity is the most important parameter and positive word percentage increase with star and negative word percentage decrease.

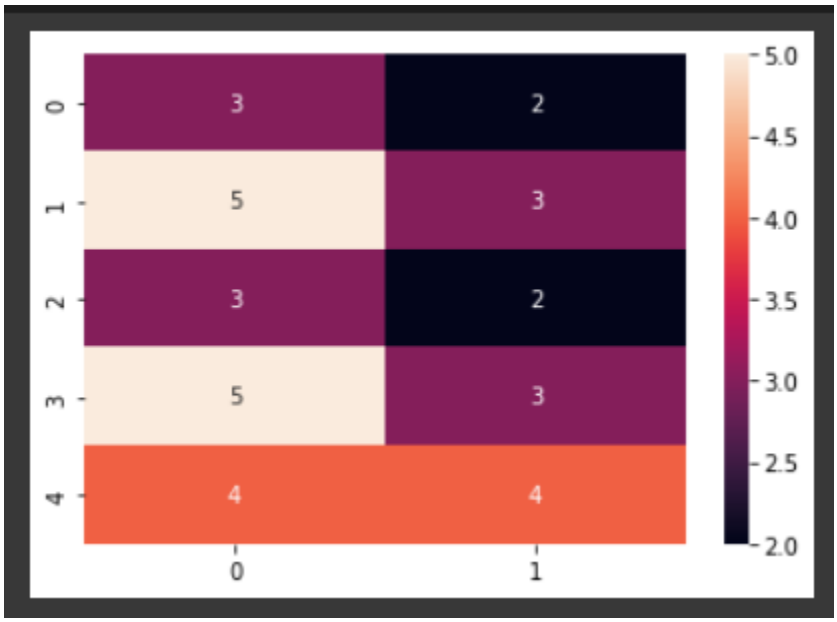
| | Star | Useful | Funny | Cool | Data Len | Positive Word Count | Positive Word % | Negative Word Count | Negative Word % | Polarity | Subjectivity |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0.528512 | 0.258078 | 0.266260 | 0.791706 | 0.743263 | 0.892943 | 0.774518 | 0.811898 | -0.172063 | 0.952494 |
| 1 | 2 | 0.577413 | 0.440597 | 0.277031 | 0.775597 | 0.791820 | 0.908275 | 0.756555 | 0.866381 | 0.377546 | 0.972260 |
| 2 | 3 | 0.479355 | 0.203296 | 0.258719 | 0.798731 | 0.795530 | 0.910717 | 0.763686 | 0.834040 | 0.765904 | 0.975112 |
| 3 | 4 | 0.470583 | 0.269488 | 0.308412 | 0.782916 | 0.809538 | 0.918484 | 0.656568 | 0.672329 | 0.872962 | 0.978033 |
| 4 | 5 | 0.448894 | 0.260891 | 0.360126 | 0.726693 | 0.760488 | 0.924099 | 0.591330 | 0.671112 | 0.897068 | 0.974537 |

**Mean Data of different parts of sentences such as adverbs, noun,s etc collected by performing pos tagging on the dataset.**

| | Stars | PROPN | VERB | NOUN | NUM | ADP | ADV | ADJ | Data length |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 11.171645 | 9.801814 | 19.362975 | 2.853901 | 1.589710 | 3.115026 | 6.423401 | 574.427961 |
| 1 | 2 | 12.092915 | 9.002036 | 19.696917 | 2.741413 | 1.597990 | 2.911285 | 7.558175 | 586.989508 |
| 2 | 3 | 12.555048 | 8.225380 | 19.075159 | 2.532081 | 1.633180 | 2.830480 | 7.943057 | 570.838119 |
| 3 | 4 | 12.428653 | 7.245049 | 17.461502 | 2.254105 | 1.563914 | 2.696652 | 7.460883 | 518.461049 |
| 4 | 5 | 10.515186 | 6.328096 | 14.346980 | 1.936184 | 1.458499 | 2.497055 | 6.044570 | 429.763411 |

Observation- data length decreases with star, a possible explanation could be that negative reviews are generally longer than positive reviews.
5-star reviews contain the least amount of adverbs, nouns, numbers etc. which is to be expected since 5-star reviews are the shortest among the five.

Stars vs Sentiments




Limitations:

1. Unlabeled Data(sentiment basis):
2. Not much information(parameters) was available for text reviews
3. Uneven distribution of stars(5 stars reviews were highest and 2 stars were least)
4.

References:

1. Unsupervised Sentiment Analysis of Yelp Reviews using Natural Language Processing, Ben Chamblee
2. Positive Lexiocns, Negative Lexicons | https://ptrckprry.com/course/ssd/data/positive-words.txt https://ptrckprry.com/course/ssd/data/negative-words.txt
3.