IDS 572

Assignment -4

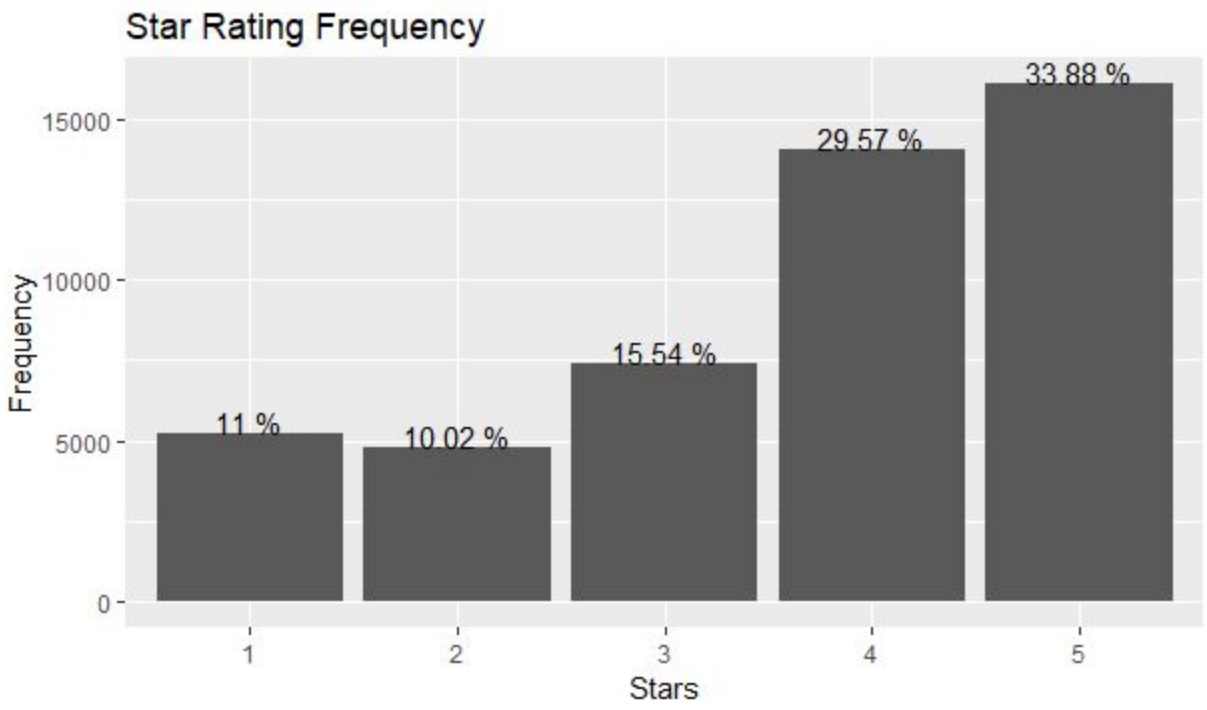
Team Members:

Aditi Dhawan - 657569917
Gnana Teja Peddi - 674047493
Shubham Sharma - 675565126

A.1 Explore the data. How are star ratings distributed?

On exploring the dataset, we noticed that the Star variable was numeric, ideally, it should be a factor variable with 5 levels. Thus, we transformed the variable to a factor.

A better understanding of the distribution of star ratings can be inferred from the histogram below :



The count for each star levels

STARS	1	2	3	4	5
COUNT	5524	4757	7380	14042	16091

- From the above barplot, we can conclude that the majority of Star ratings are for Star Rating category 4 and 5, consisting of **63.45%**.
- Total percentage for Star rating categories 1, 2 and 3 is **37.56%**
- The highest reviews are for Star rating 5 and fewest reviews are for Star rating 2.
- Average Star Rating - 3.653
- Standard Deviation - 1.328

A.2 How will you use the star ratings to obtain a label indicating 'positive' or 'negative' – explain using the data, graphs, etc.?

- Since the majority of Star Ratings are for Ratings 4 and 5, we will consider ratings below 4 as 'negative' and ratings 4 and 5 as 'positive'.
- We have made a new variable "label" to indicate whether the Star Rating is positive (1) or negative (0).

A.3 Does star ratings have any relation to 'funny', 'cool', 'useful'? (Is this what you expected?)

To check if there is a relation between ratings and 'funny', 'cool', 'useful', we conducted and chi-square test with a confidence interval of 95%. Below are our findings :

Explanatory Variable	P value	Conclusion
Funny	2.2e-16	Statistically significant
Cool	2.2e-16	Statistically significant
Useful	8.637e-11	Statistically significant

Based on the results of the chi-squared test, we concluded that there is a relation between 'funny', 'cool', 'useful' and Star rating. Yes, we expected this conclusion because restaurants with high ratings have a higher probability of getting comments with words like "Funny", "Cool" and "useful".

Useful most frequently occurs with 3 stars rated reviews. A useful comment could be labelled as useful with respect to identifying positive and negative features of a certain restaurant. Thus we can say that it is useful is a neutral label.

Funny occurs across all star rated reviews and has a very similar distribution to Cool. Thus the two can be categorised as 'positive' tags. It's the highest occurrence is with 4 star rated reviews (144 times).

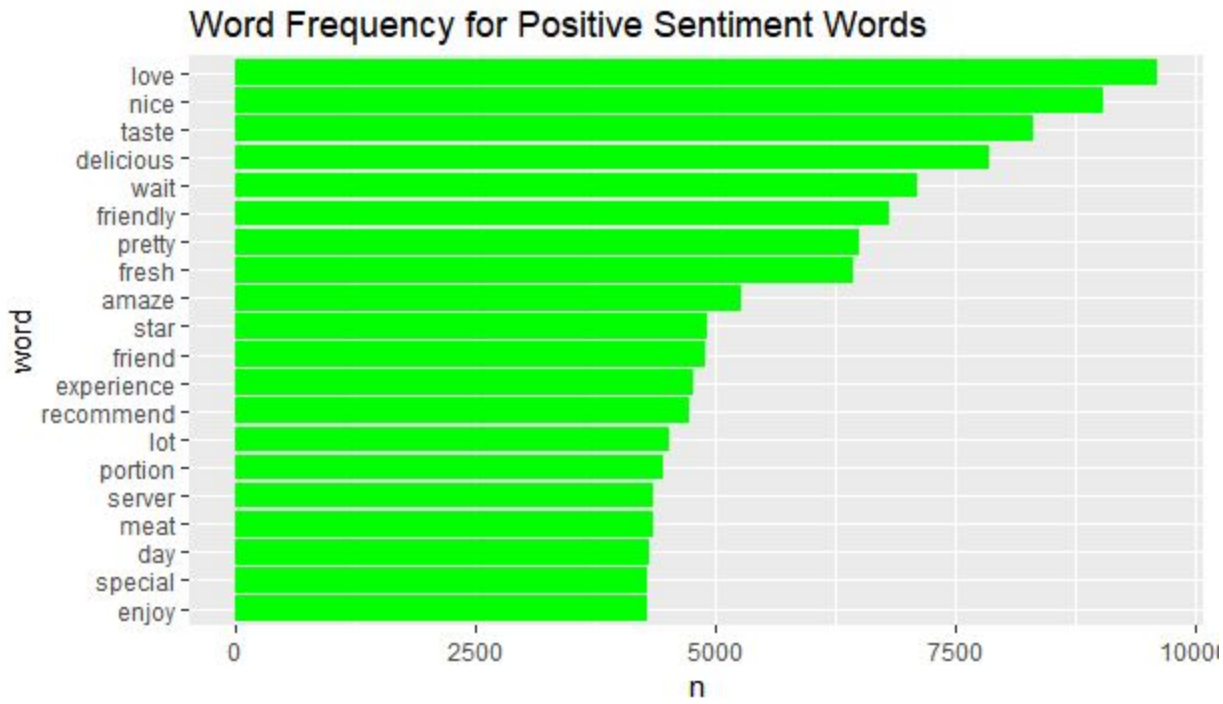
Cool occurs across all star ratings. It occurs more frequently across star ratings 3, 4 & 5. It most frequently mentioned in 4.00-star ratings (201 times).

B.1 What are some words indicative of positive and negative sentiment? (One approach is to determine the average star rating for a word based on star ratings of documents where the word occurs).

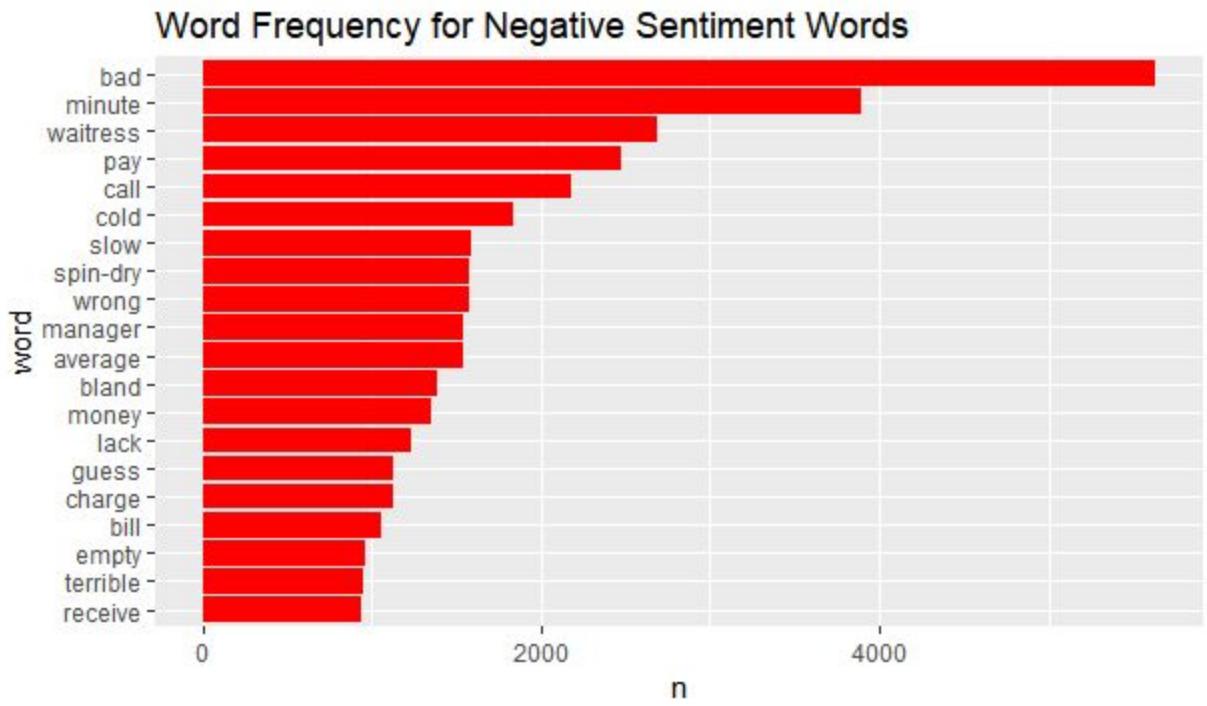
To answer this question, we performed the below-mentioned techniques on our data:

1. Tokenization
2. After tokenization, we removed stopwords and neutral words like Food items(Sandwich, Burger, Cheese, Chicken...) name and general items(Table, Meal, Service, Fry)
3. Lemmatization
4. After Lemmatization, we removed all the words with frequency less than 3.
5. Next, we calculated the average of stars for each word based on the star reviews for a rating.

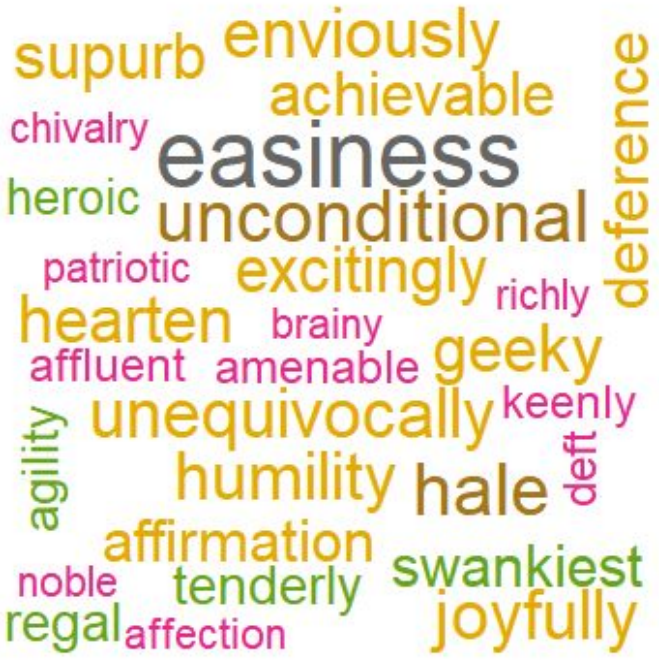
Top 20 words Positive Words



Top 20 Negative Words



POSITIVE WORD CLOUD



NEGATIVE WORD CLOUD

B.2 Do these 'positive' and 'negative' words make sense in the context of user reviews?

Yes, positive words do indicate positive sentiment and the same goes with the negative words.

Positive words: From our graph indicating the most frequent positive words- love, nice tasty, delicious are the most used words. These words indicate the positive sentiments of the customer and occur most commonly in higher star rated reviews.

One might notice that these words are specifically names of food items/dishes that are served at a restaurant. These could indicate that dishes that contain Chicken/Cheese/Fries/ Pizza are enjoyed by a higher population of people and thus most commonly occur with high star rated reviews. This helps us categorise them as indicative of 'positive' sentiment. Also words like 'good', 'great' & 'love' are some of the top 'positive' words present in our word-set. These are more commonly relatable to positive sentiment as they are colloquially used in context with positive speech.

Negative words: The most commonly occurring negative words are - bad, minute, waitress and pay. These words indicated a sense of dissatisfaction and negative sentiments of the customer. Also, these words are most commonly present in the least rated star reviews.

These words indicate a sense of dissatisfaction and are thus present in reviews that have been indicative of 'negative' sentiment. These tell us that a user found a certain restaurant not up to par with his/her expectations.

(C) How many matching terms are there for each of the dictionaries?

We will consider three dictionaries - the Harvard IV dictionary of positive and negative words, the extended sentiment lexicon developed by Prof Bing Liu of UIC-CS, and the AFINN dictionary which includes words commonly used in user-generated content in the web. These dictionaries contain a different number of positive and negative words, a summary of which is given below:

DICTIONARY	POSITIVE WORDS	NEGATIVE WORDS
Sentiment Lexicon	2006	4783
AFINN	878	1599
Harvard	1636	2006

C.1 Consider using the dictionary based positive and negative terms to predict sentiment (positive or negative based on star rating) of a movie. One approach for this is: using each dictionary, obtain an aggregated positive score and a negative score for each review; for the AFINN dictionary, an aggregate positivity score can be obtained for each review.

The number of matching terms along with the prediction accuracy is given below:

DICTIONARY	PREDICTION ACCURACY	MATCHING TERMS
Sentiment Lexicon	70.50%	1550
AFINN	71.85%	785
Harvard	59.31%	1500

Are you able to predict review sentiment based on these aggregated scores, and how do they perform? Does any dictionary perform better?

Yes, We were able to predict the review sentiment based on the aggregated scores of the 3 different dictionaries. The performance is listed in the above table.

This process gave us a comparative performance evaluation for the three different dictionaries. It was found that AFINN performed the best with 785 matching terms and an accuracy of 71.85%.

C.2 Compare this approach with the use of SentiWordNet. Describe how you use SentiWordNet.

DICTIONARY	PREDICTION ACCURACY	MATCHING TERMS
Sentiword Net	62.35%	461

We found that prediction accuracy for Sentiwordnet dictionary is 62.35%, which is less that Sentiment Lexicon and AFINN. We have used "**Lexicon**" package in R to use Sentiwordnet.

(D) Develop models to predict review sentiment. For this, split the data randomly into training and test sets. To make run times manageable, you may take a smaller sample of reviews (minimum should be 10,000).

(i) Develop models using only the sentiment dictionary terms (you can try individual dictionaries or combine all dictionary terms).

DTM Modeling:

- Dictionary: Sentiment Alixon - Because it did well in predicting review sentiments and has good matching with our sample.
- Size of the DTM: 20000*2526
- Measure: Tfidf

SENTIMENT LEXICON

MODEL	TRAINING ACCURACY	TEST ACCURACY
SVM	87.15%	78.18%
NAIVE BAYES	65.91%	66.48%
RANDOM FOREST	82.78%	79.65%

AFINN

MODEL	TRAINING ACCURACY	TEST ACCURACY
SVM	87.21%	79.10%
NAIVE BAYES	66.41%	67.36%
RANDOM FOREST	71.23%	68.41%

HARVARD

MODEL	TRAINING ACCURACY	TEST ACCURACY
SVM	85.01%	75.71%
NAIVE BAYES	68.3%	68.52%
RANDOM FOREST	72.56%	74.24%

Do you use term frequency, tfidf, or other measures? What is the size of the document-term matrix?

We have used Tf-idf as the measure for weighing words because TFIDF is often used in information retrieval and text mining. We ran separately with "Term Frequency" and "tf- idf". The size of the document term matrix using the combined dictionaries is (19782X2436).

(ii) Develop models using a broader list of terms – how do you obtain these terms? Will you use stemming here?

To get a broader list of terms, we have changed the sentiment threshold to 3 from 4. This means that more words are now grouped into positive sentiment related group. For the first and second model, we have considered the AFINN dictionary because of its performance in predicting the review sentiments.

We have not used the stemming method here as it may affect our accuracy because we are using dictionaries here.

Report on the performance of the models. Compare performance with that in part (c) above. For models in (i) and(ii): Do you use term frequency, tfidf, or other measures, and why? Do you prune terms, and how (also, why?). What is the size of the document-term matrix?

Maximum term length = 17

Sparsity= 100%

Size of DTM: 19782*2436

SENTIMENT LEXICON

MODEL	TRAINING ACCURACY	TEST ACCURACY
SVM	88.18%	76.28%
NAIVE BAYES	67.54.%	67..48%
RANDOM FOREST	83.52%	77.65%

AFINN

MODEL	TRAINING ACCURACY	TEST ACCURACY

SVM	85.21%	78.10%
NAIVE BAYES	68.41%	66.42%
RANDOM FOREST	73.56%	69.54

HARVARD

MODEL	TRAINING ACCURACY	TEST ACCURACY
SVM	86.01%	77.71%
NAIVE BAYES	69.3%	67.54%
RANDOM FOREST	71.25%	68.21%

We have used Tf-idf as the measure for weighing words because TFIDF is often used in information retrieval and text mining.

We have used pruning to reduce the complexity of the model and to improve the performance of the models.

Size of the Document-Term Matrix : (19782X2436)