TEXT MINING – sENTIMENT ANALYSIS

IDS 572 – DATA MINING

ASHOK BHATRAJU – 670248723

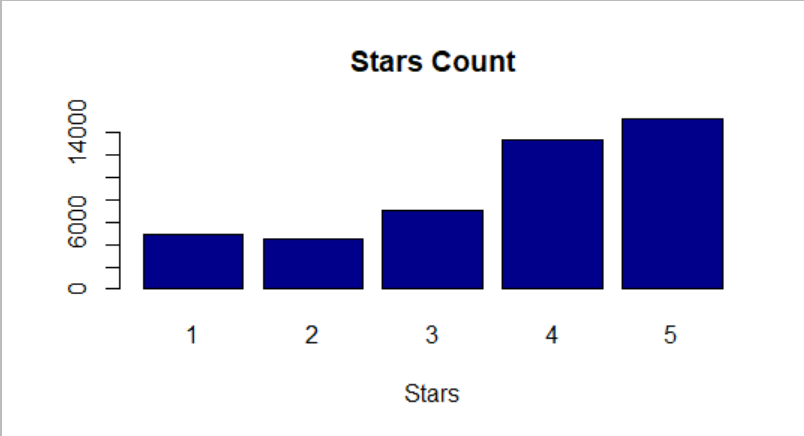
SHOURYA NARAYAN –

VIVEK KUMAR -

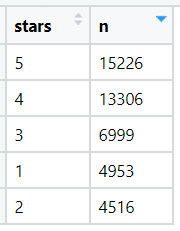
1. **Explore the data. How are star ratings distributed? How will you use the star ratings to obtain a label indicating ‘positive’ or ‘negative’ – explain using the data, graphs, etc.? Do star ratings have any relation to ‘funny’, ‘cool’, ‘useful’? (Is this what you expected?)**

The dataset is customer’s review data for various restaurants from the Yelp Database. Along with review and ratings from the customer, we also have features like cool, funny and useful for these reviews.

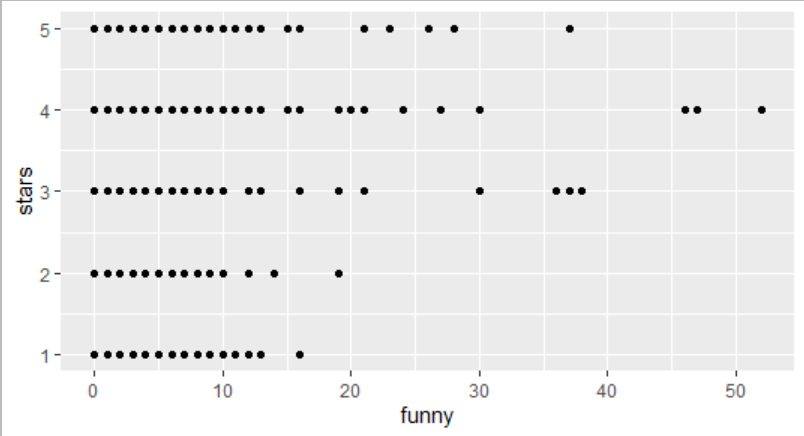
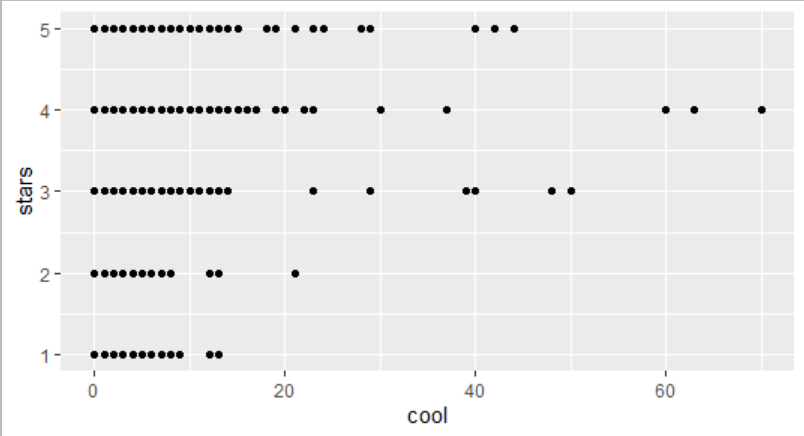
The distribution of star rating is shown in the table and the plot below, we can see that we have maximum count for ratings 5,4 (~ 65% combined) and it decrease gradually thereafter.

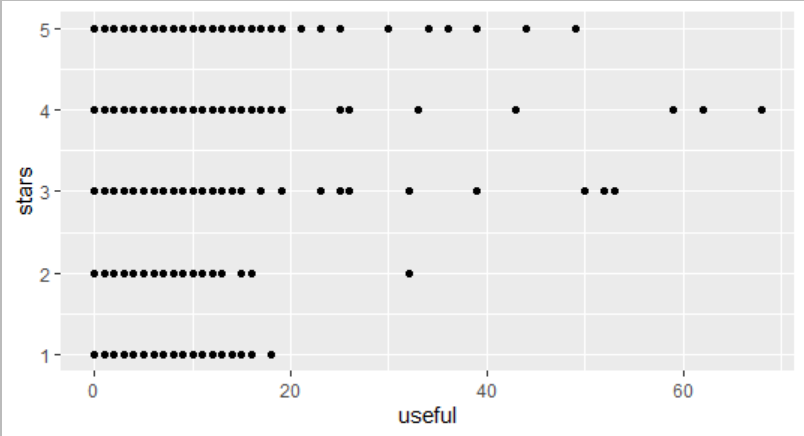


The star rating can be used to indicate the review as positive or negative, while it is obvious that star rating 4 & 5 indicate can be labeled as positive and star rating 1 & 2 can be labeled as negative but it is not clear whether Star Rating 3 shall be labeled as positive or negative. For this if we look at the plot of Star Rating against ‘funny’, ‘cool’ and ‘useful’, then reviews with star-rating 3 also tend to have higher values of these feature, hence we can conclude that reviews with **star rating 3,4, & 5** can be labeled as **positive** and reviews with **star rating 1 & 2** can be labeled as **negative**.

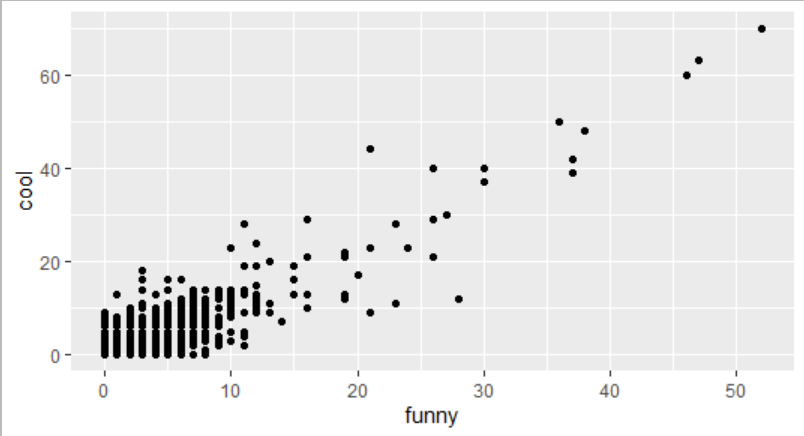
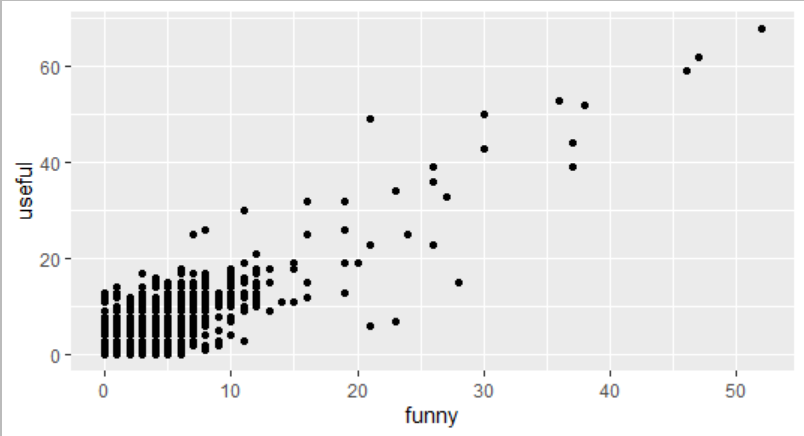


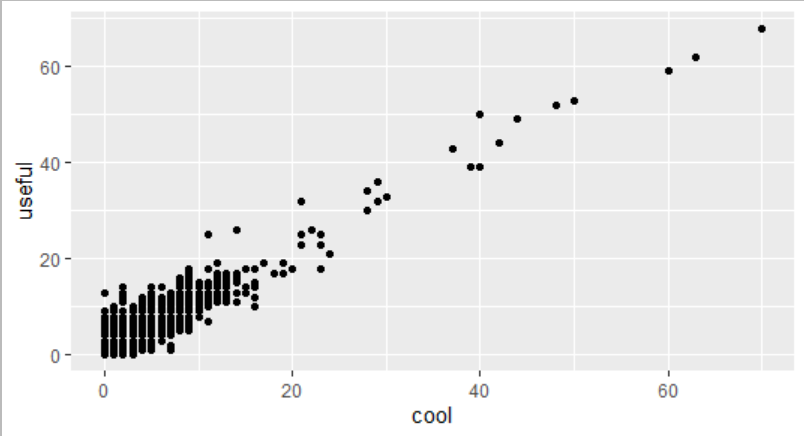
The scatterplots below show the distribution of ‘funny’, ‘cool’ and ‘useful’ reviews across different star ratings. All these three features have the similar pattern w.r.t star-ratings i.e. reviews with star-rating 4 and 5 have higher value of these features whereas it remains 20 or less for star rating 1 & 2. This pattern was expected from the data set as these features have more inclination towards positive sentiment which is likely to have higher value for higher star rating.



The following graph shows correlation between ‘cool’, ‘funny’ and ‘useful’. This correlation plot shows that reviews which were found to be more ‘funny’ were also more ‘cool’ and likewise for ‘funny’ & ‘useful’ and ‘cool’ & ‘useful’.

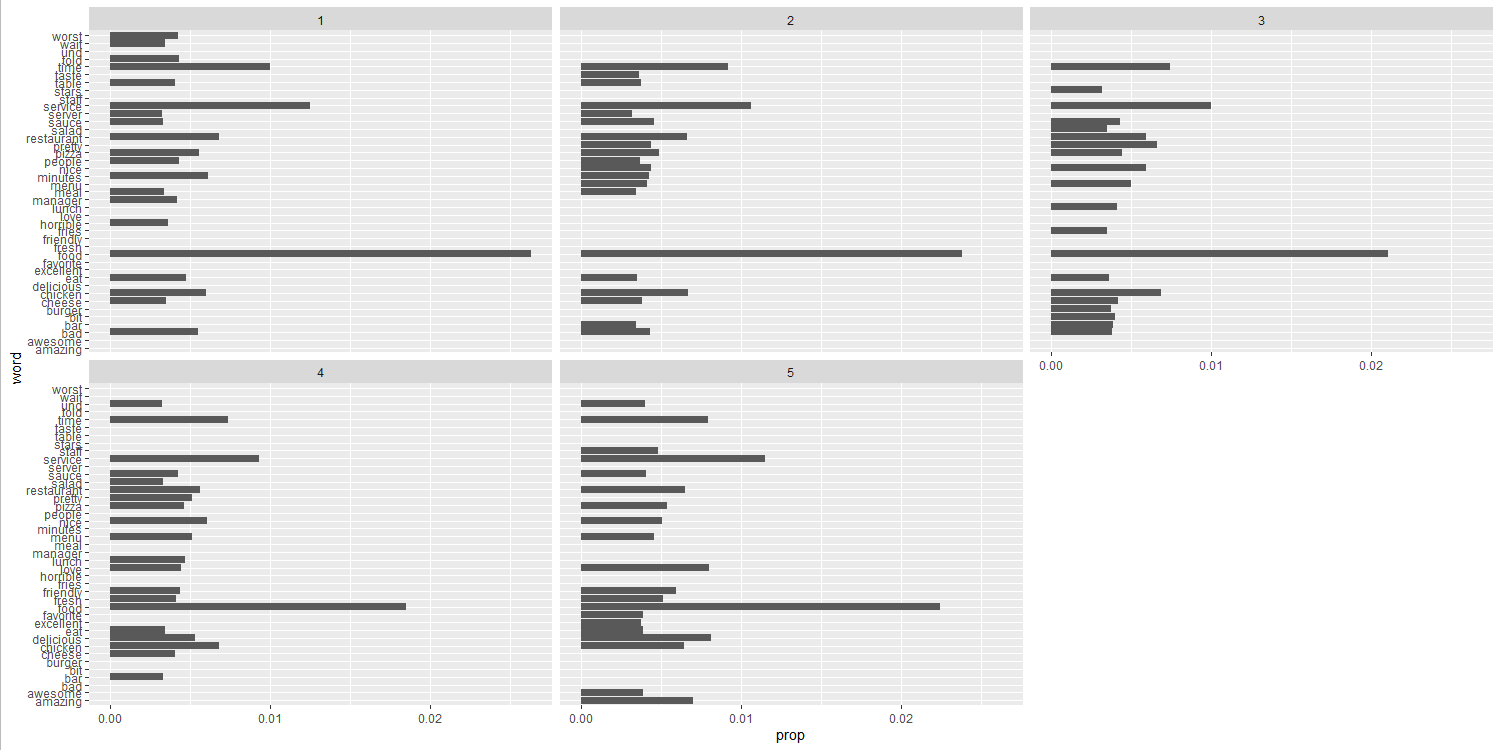
 



1. **What are some words indicative of positive and negative sentiment? (One approach is to determine the average star rating for a word based on star ratings of documents where the word occurs). Do these ‘positive’ and ‘negative’ words make sense in the context of user reviews? (For this, since we wish to get a general sense of positive/negative terms, you may like to consider a pruned set of terms -- say, those which occur in a certain minimum and maximum number of documents).**

While giving review for a restaurant a lot of words are used to convey the feeling of a user. Not all the words used gives an insight about user’s perception about the restaurant, but some gives a great hint about what user wants to convey. Therefore, we try to find out the words which gives a sense of feeling about review.

For finding the sentiments that a word can convey we perform a series of steps. As we have done the data cleaning part in the last question, we group the words according to their stars and sort them. Then, we calculate the proportion of the frequency of every word per star rating, using which we create a dataset to collect the top 20 words of each star rating as per their frequency. The top 20 words for each star rating are shown below:



Further we analyzed words as per their average occurrence in every star rating and collected the top 20 and lowest 20.

|  |  |
| --- | --- |
| S.no | Bottom 20 |
| 1 | arguing |
| 2 | blech |
| 3 | bullshit |
| 4 | coffee |
| 5 | disgust |
| 6 | disrespectful |
| 7 | dolmas |
| 8 | mask |
| 9 | neven |
| 10 | nnwas |
| 11 | patronizing |
| 12 | received |
| 13 | rubio's |
| 14 | santi |
| 15 | tipping |
| 16 | understands |
| 17 | unedible |
| 18 | unwilling |
| 19 | useless |
| 20 | vehicle |

|  |  |
| --- | --- |
| S.no | Top 20 |
| 1 | amazing |
| 2 | cheese |
| 3 | chicken |
| 4 | delicious |
| 5 | eat |
| 6 | food |
| 7 | fresh |
| 8 | friendly |
| 9 | love |
| 10 | lunch |
| 11 | menu |
| 12 | nice |
| 13 | pizza |
| 14 | pretty |
| 15 | restaurant |
| 16 | salad |
| 17 | sauce |
| 18 | service |
| 19 | staff |
| 20 | time |

In here, we can find that words such as amazing, delicious, fresh, friendly, love, pretty and nice are the ones which give some positive impact or positive sentiments about the review. These words give us an idea about the perception of a particular review, like we would give 5-star rating to a restaurant which is having a delicious food or where food is fresh.

For these sets of words, we find that words such as arguing, blech, bullshit, disgust, disrespectful, patronizing, unedible, unwilling and useless are the ones conveying a negative sentiment or bad response about a review. For example, words such as disgust can be used to convey the harshest of experience or words such as blech can express the distaste a reviewer felt.

This can be better understood using the plots below:

**Interpretation:**

* Words that resonate positive sentiment occur more frequently in star ratings 4 and 5.
* Words that resonate negative sentiment occur more frequently in star ratings 1 and 2.
* This is something we predicted.
* As positive sentiment is associated with better rating and negative sentiment is associated with lower rating.

1. **C) We will consider three dictionaries, available through the tidytext package – the NRC dictionary of terms denoting different sentiments, the extended sentiment lexicon developed by Prof Bing Liu, and the AFINN dictionary which includes words commonly used in user-generated content in the web. The first provides lists of words denoting different sentiment (for eg., positive, negative, joy, fear, anticipation, …), the second specifies lists of positive and negative words, while the third gives a list of words with each word being associated with a positivity score from -5 to +5.**

**How many matching terms are there for each of the dictionaries?**

**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 positiveScore and a negativeScore for each review; for the AFINN dictionary, an aggregate positivity score can be obtained for each review. Are you able to predict review sentiment based on these aggregated scores, and how do they perform? Does any dictionary perform better?**

The three different dictionaries provided to us tell us about sentiments of words in different approach. The NRC dictionary provides us with a more descriptive meaning about the sentiments of the words by denoting different sentiments of the words like whether it is positive or negative or it expresses joy or fear, etc. The Bing Liu dictionary directly specifies whether a word is positive or negative and at last AFINN dictionary scores each word with a positivity score which ranges from -5 to +5.

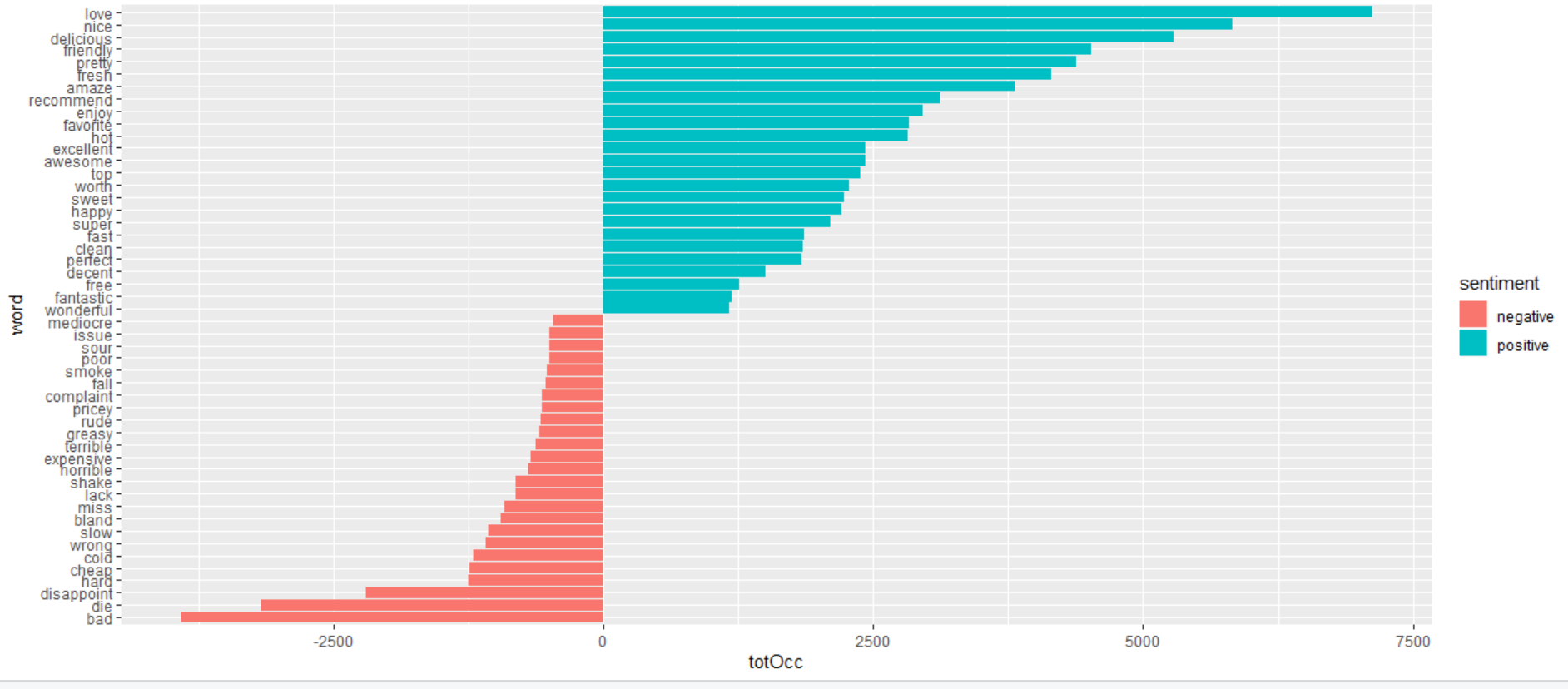
To get the matching terms for each dictionary we performed the join between the dictionary dataset and Yelp restaurant review dataset. Using left join, we get all the records, even those which doesn’t have matching records but by using inner join we get only those records which have matching records. Thus, to find matching term for each dictionary we will look for inner join process.

The table below depicts the count of words in the three dictionary and the number of matching terms with our review data:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No. | Dictionary | Count of Words in Dictionary | Total Words in Review Data | Total Count of Matching Terms | Unique Count of Matching Terms |
| 1. | Bing Liu | 6786 | 927684 | 161695 | 935 |
| 2 | NRC | 13901 | 927684 | 619488 | 1308 |
| 3. | AFINN | 2477 | 927684 | 129310 | 518 |

With these dictionaries it becomes an easy task to label a review as negative or positive but before using this dictionary we have also labelled the words as positive and negative basis them appearing more in positive reviews than negative reviews. So, we can compare the derived sentiment and the sentiment from the dictionary to check which dictionary performs better on this data. This will not be an exhaustive comparison of accuracies of these dictionaries because the data set is not able to match all the keywords in all these dictionaries.

**BING LIU:**



This graph shows top 25 most occurring positive and 25 most occurring negative sentiments word in Bing Liu Dictionary.

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| Actual | -1 | 1 |
| -1 | 5100 | 1570 |
| 1 | 3534 | 17269 |

**The accuracy for predicting the sentiments for a review using Bing Liu dictionary was 81%**

**NRC:**

A screenshot of a cell phone

Description automatically generated

This graph shows the top 25 most occurring word according to their positive and negative sentiment analysis using NRC dictionary.

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| Actual | -1 | 1 |
| -1 | 2996 | 3946 |
| 1 | 2766 | 18595 |

**The accuracy for predicting the sentiments for a review using NRC dictionary was 76%.**

**AFINN:**

A picture containing screenshot

Description automatically generated

Top 25 words expressing positive and top 25 words expressing negative sentiments using AFINN dictionary.

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| Actual | -1 | 1 |
| -1 | 4203 | 2327 |
| 1 | 2535 | 17854 |

**The accuracy for predicting sentiments of a review using AFINN dictionary was 82%**

Since we have got the words available in the dictionary, we now need to find out whether a review given by user is positive or negative. For that, using the dictionary we will get an aggregated positive or negative score for each review.

**Comparison of Accuracies of the three dictionaries:**

|  |  |  |
| --- | --- | --- |
| S.No. | Dictionary | Accuracy |
| 1. | **Bing-Liu** | **81%** |
| 2. | **NRC** | **76%** |
| 3. | **AFINN** | **82%** |

The above table shows that AFINN and Bing-Liu Dictionary works best on this review data set. This analysis gives an accuracy score for the dictionaries by comparing the sentiment of review / words from dictionary to that derived using an aggerate and frequency words appearing in different star ratings.

**Further we compared Average sentiment score for each dictionary**

Bing

|  |  |
| --- | --- |
| Stars | Average Senti Score |
| 1 | -0.405 |
| 2 | -0.107 |
| 3 | 0.193 |
| 4 | 0.475 |
| 5 | 0.622 |

NRC

|  |  |
| --- | --- |
| Stars | Average Senti Score |
| 1 | 0.0238 |
| 2 | 0.0897 |
| 3 | 0.144 |
| 4 | 0.204 |
| 5 | 0.232 |

AFINN

|  |  |
| --- | --- |
| Stars | Average Senti Score |
| 1 | -2.35 |
| 2 | 0.714 |
| 3 | 3.14 |
| 4 | 5.55 |
| 5 | 6.46 |

As we can see from the tables above, average sentiment scores for all the dictionary are as expected. The value is lowest for star rating one and highest for star rating 5. The value increased from one through five.

1. **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).**

**One may seek a model built using only the terms matching any or all of the sentiment dictionaries, or by using a broader list of terms (the idea here being, maybe words other than only the dictionary terms can be useful). You should develop at least three different types of models (Naïve Bayes, and at least two others of your choice ….Lasso logistic regression (why Lasso?), xgb, svm, random forest,…) [Note: those working alone can compare with at least two types of models]**

1. **Develop models using only the sentiment dictionary terms – try the three different dictionaries; how do the dictionaries compare in terms of predictive performance for rating ? Then with a combination of the three dictionaries, ie. combine all dictionary terms. Do you use term frequency, tfidf, or other measures, and why? What is the size of the documentterm matrix? Should you use stemming when using the dictionaries?**
2. **Develop models using a broader list of terms (i.e. not restricted to the dictionary terms only) – how do you obtain these terms? Will you use stemming here?**

**Report on performance of the models. Compare performance with that in part (c) above.**

**How do you evaluate performance? Which performance measures do you use, why.**

In this question we took a sample of 10000 rows to reduce the running time of models. Further we make sure that the proportion of star ratings is similar.

Term tf-idf was used to know how important a word is to a document collection.

We will not use stemming when using the dictionary since we are going to compare words in different dictionaries.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No. | Model | Data | Model Performance | | AUC |
| Training Accuracy | Test Accuracy |
| 1. | Naïve Bayes | Train 10990\*937  Test 13736\*937 | 44% | 45% | Training 0.70  Test 0.71 |
| 2. | SVM-1 | Train 10990\*937  Test  13376\*937 | 92% | 89% | Train  0.87 Test 0.81 |
| 3. | SVM-2 | Train 10990\*937  Test  13376\*937 | 96% | 88% | Train 0.94 Test 0.80 |
| 4. | Random Forest  Number of Trees = 500 | Train 10303\*935  Test  10303\*935 | 91% | 86% | Training: 0.98  Test:  0.91 |

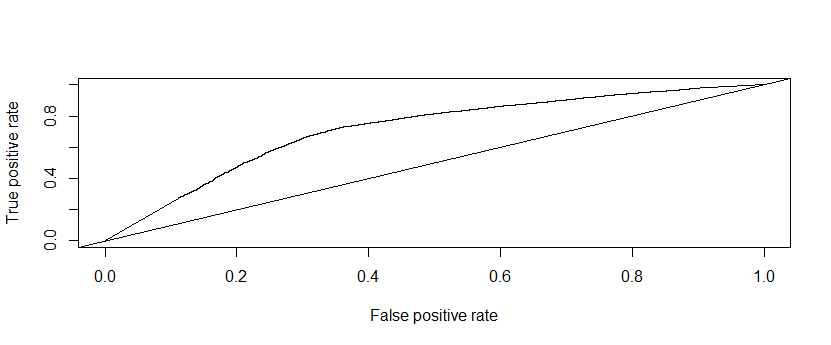
We performed SVM, Naïve Bayes and Random forest on the above three dictionaries. Along with that we also combined all the three dictionaries and performed all the above-mentioned models on that data aswell.

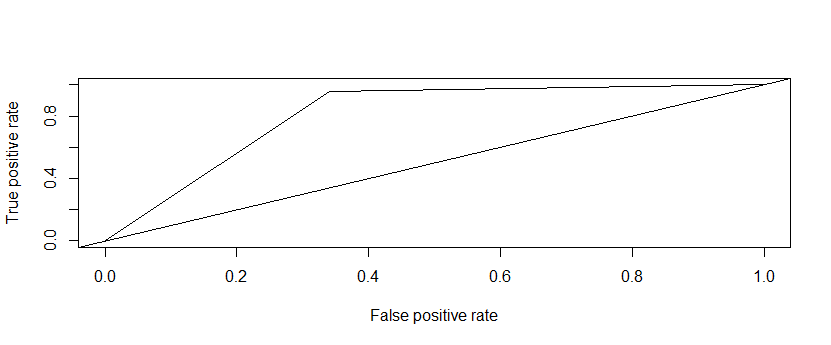
Based on the above comparison, SVM-2 model performed best on training data and SVM-1 model performed best on test data.

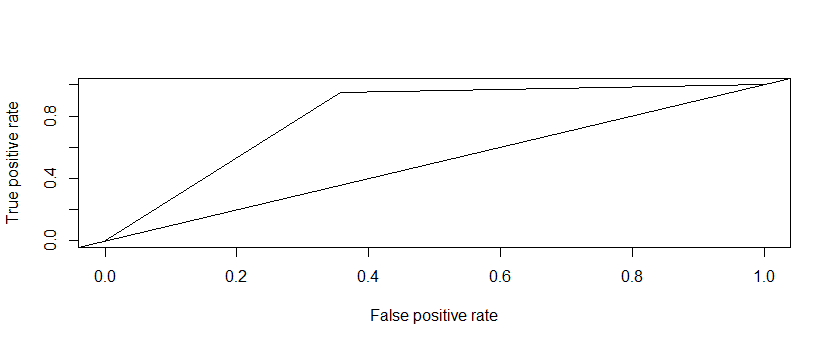
From the table above, \_\_\_\_\_\_\_\_\_\_\_ model performed better in Bing dictionary, \_\_\_\_\_\_\_\_\_\_\_ model performed better in NRC dictionary and \_\_\_\_\_\_\_ model performed better in AFINN dictionary.

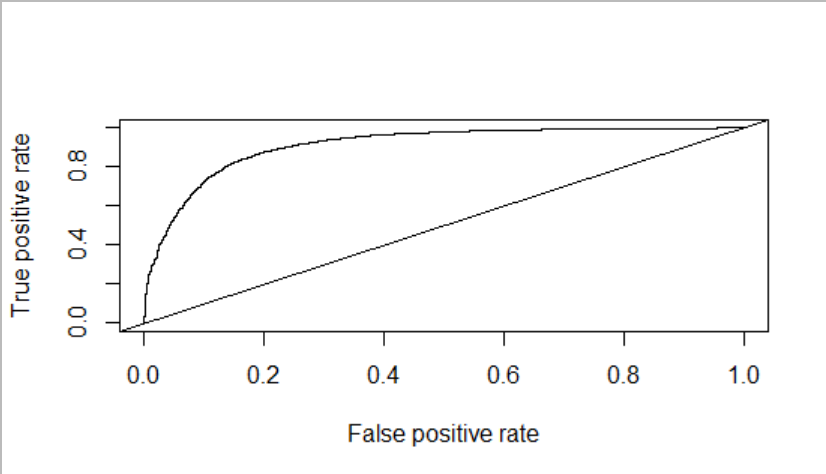
Overall, SVM-2 gave us the best results among all the dictionaries.

When combined, \_\_\_\_\_\_\_\_\_ model performed best and the performance of the other models increased a bit.

**ROC Curve for Naïve Bayes:** 

**ROC for SVM Model-1:**

**ROC of SVM-2 for Bing Dictionary:** 

**ROC Curve for Random Forest:** 

Random forest is the best model in terms of Area under curve, it has a training AUC of 0.98 and test AUC of 0.91.

ii)

In this comparison, we are going to use sentiments other than the ones that are part of the dictionaries we used so far.

\_\_\_\_\_ model performed the best on the test data and \_\_\_\_\_ model performed the best on training data.

\_\_\_\_\_\_\_ model performed the worst among all the models compared above.

Models performance is better on this as we used broader list of sentiments to compare. The reason being, this model which is built is not specific to any business. The ones we used earlier are used for restaurant business. The one we built is something that can be used for any business.