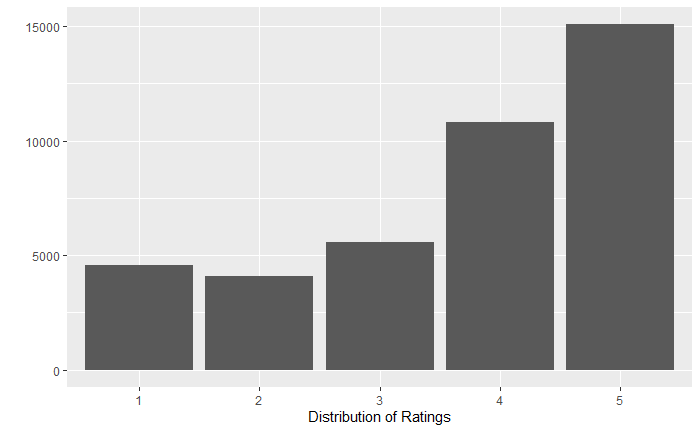
**Assignment 3 – Text Mining**

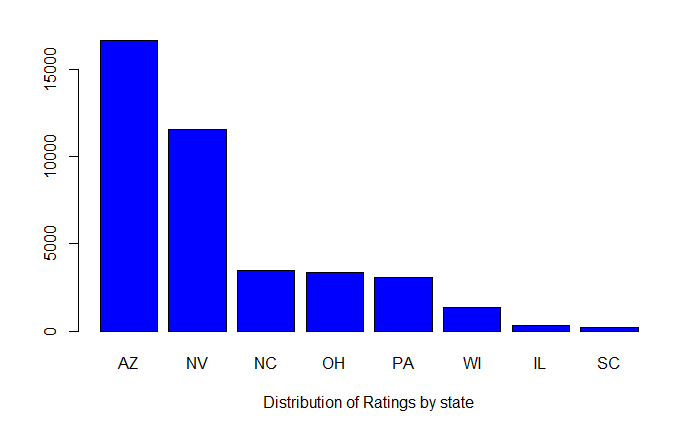
1. **Explore the data.  
   (i) How are star ratings distributed? How will you use the star ratings to obtain a label indicating  
   ‘positive’ or ‘negative’ – explain using the data, graphs, etc.?  
   Do star ratings have any relation to ‘funny’, ‘cool’, ‘useful’? Is this what you expected?  
   (ii) How does star ratings for reviews relate to the star-rating given in the dataset for businesses  
   (attribute ‘businessStars’)? (Can one be calculated from the other?)**

The Star rating in the data has five different levels from 1 to 5 with 1 being the lowest and 5 being the highest rating. Overall, the number of reviews across different ratings has uneven distribution with most numbers of reviews receiving a 5-star rating.

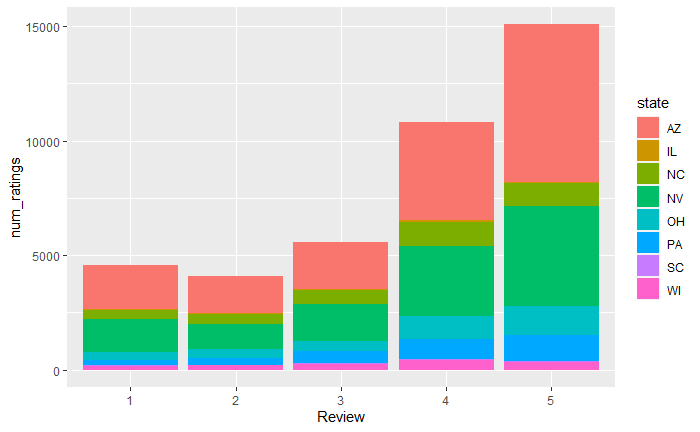


We looked at the start ratings for different dimensions of the data. Let’s look at the distribution of star ratings by different states.

Overall, Arizona seems to have the maximum number of reviews received. This is evident due to the fact that Arizona has the highest number of businesses in our data.

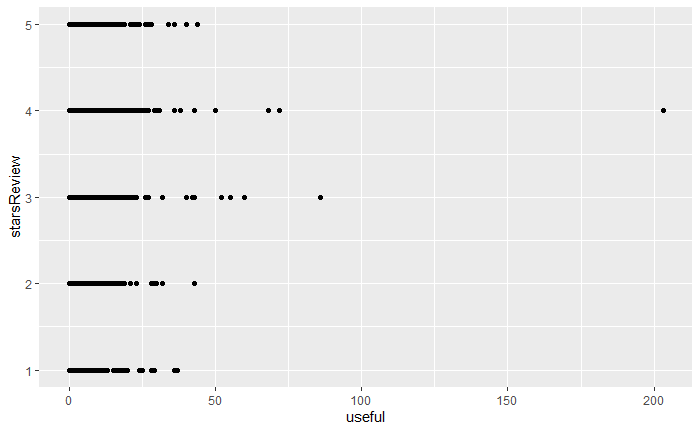


Let’s slice the reviews for each state by their star ratings.



From the above graph, we can also observe that ‘Arizona’ has the maximum number high quality restaurants.

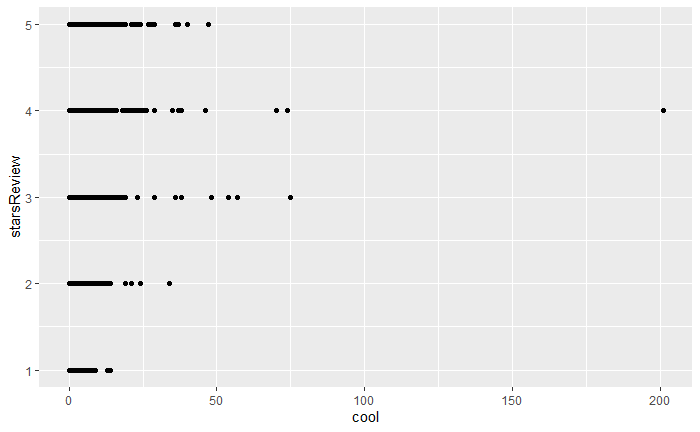
Next, we will explore the relationship between the star ratings and ‘useful’, ‘funny’, ‘cool’ dimension in our data. These columns indicate that for a particular review how many users of the ‘Yelp’ have identifies it as either ‘useful’, ‘funny’ or ‘cool’.



From above graph we can see that the maximum number of ‘useful’ tags has been received for the revies having 3 and 4 rating.



From the relationship between star rating and ‘funny’ tag, we can observe that the maximum number of ‘funny’ tags has been received for 3- & 4-star reviews.



We can see a similar distribution of number of ‘cool’ tags for different star ratings. From the relationship between star rating and ‘cool’ tag, we can observe that the maximum number of ‘cool’ tags has been received for 3- & 4-star reviews.

The above relationship between ‘useful’, ‘funny’, ‘cool’ with star ratings is highly expected as the reviews which have 3- and 4- star ratings are written by customers with careful consideration and their analysis of the business. This is also evident from the fact that 3- and 4-star rating have the highest average number of word count. Hence, the reviews with 3- and 4- star ratings have the highest number of different tags received.

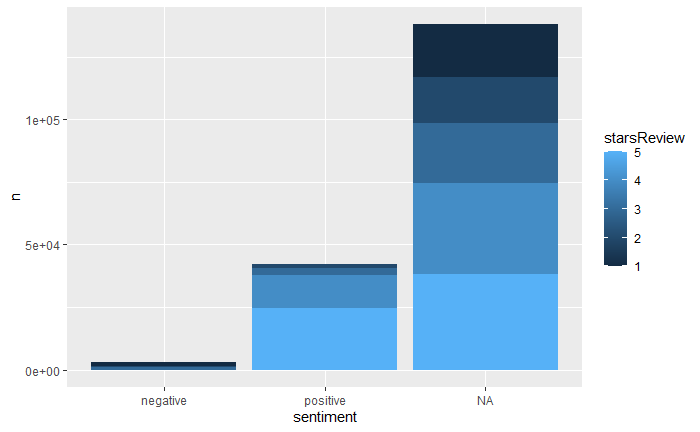
According to the data we can see the star ratings is highly correlated with business star ratings. However, the interval scale of star ratings is 1,2,3,4,5. And the interval scale of business star ratings is 1.5,2,2.5,3,3.5,4,4.5,5.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Stars Ratings | | | | | |
| businessStars | **num\_of\_ratings** | **num\_5\_ratings** | **num\_4\_ratings** | **num\_3\_ratings** | **num\_2\_ratings** | **num\_1\_ratings** |
| 5.0 | 267 | 93 | 6 | 1 | 0 | 1 |
| 4.5 | 6362 | 67 | 21 | 6 | 3 | 3 |
| 4 | 13799 | 44 | 30 | 12 | 7 | 6 |
| 3.5 | 11138 | 28 | 31 | 17 | 12 | 12 |
| 3 | 5820 | 18 | 25 | 20 | 17 | 20 |
| 2.5 | 2131 | 11 | 20 | 19 | 20 | 30 |
| 2 | 405 | 6 | 11 | 15 | 20 | 49 |
| 1.5 | 165 | 3 | 2 | 10 | 12 | 73 |

Though it seems highly difficult as to predict either of the ratings from other, it can be calculated using expected value across different weights for each ratings. For ex. – the number of star reviews for business stars is distributed across all the different star ratings. Hence using different weights for each star ratings based on the distribution on number of reviews across each category, we can come up with a probabilistic estimate for business star and star ratings vice-versa.

We can use star ratings to obtain a label indicating ‘positive’ and ‘negative’ for each review. There can be two ways to do this.

1. Using sentiments from dictionary – We can use the sentiments from different available dictionaries like bing and nrc to get the label for each word. Then we can use a label of 1 for ‘positive’ sentiments and 0 for ‘negative’ sentiments for each word. Then we can take the average of these scores for each review. And based on the average score, we can label each review as either ‘positive’ or ‘negative’. Positive means average score >= 0.5 and negative means average score < 0.5.
2. Another way to achieve the ‘positive’ and ‘negative’ label would be to use hardcoded scores. For ratings >= 4 we can label the review as ‘positive’ and for ratings <=2, we can label the review as ‘negative’.

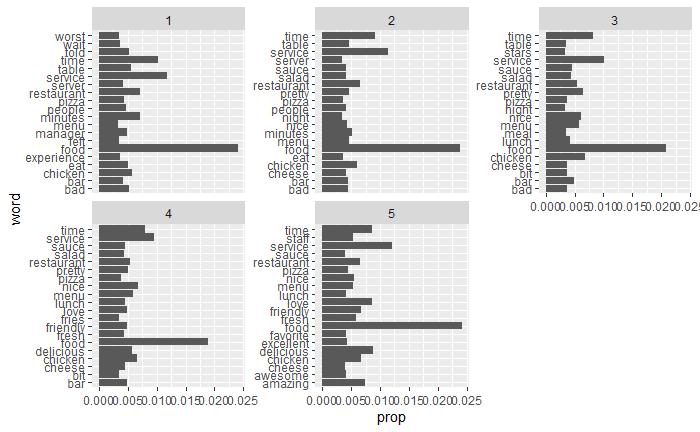


The above graph shows the distribution of ‘negative’ and ‘positive’ label based on ‘BING’ dictionary across different star ratings. Hence, the star ratings can be used to obtain a label.

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 being considered?  
   (For this, since we’d like 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).**

For further analysis, we will do some data cleaning before. We have executed different techniques as below for cleaning our data:

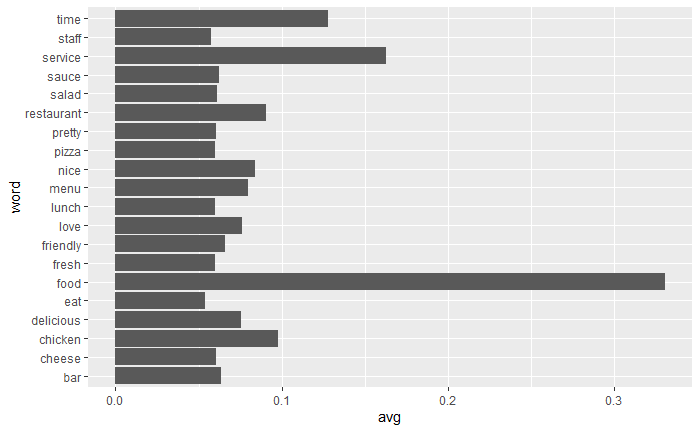
- Performed Tokenization  
- Transformed case (to all lower/upper)  
- Filtered stopwords  
- Filtered tokens by length - min 3 and max 15.  
- Filtered tokens by content – if they matched a ‘dictionary’ of terms  
- Used Lemmatization



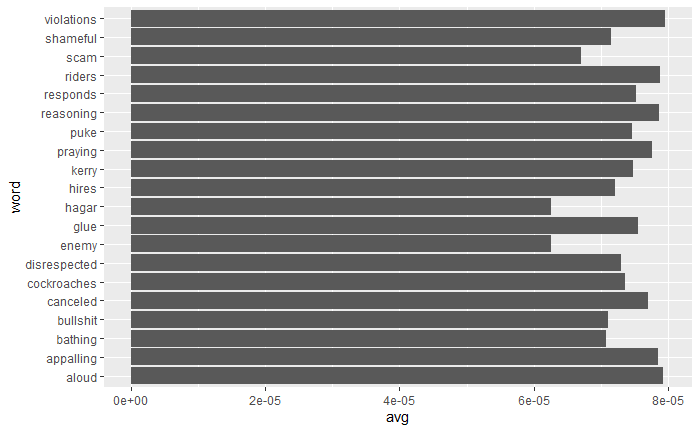
The above graph shows words related with different star ratings. We can see which words are more indicative for each rating based on the proportion of their occurrence within each rating. Next, we will look at the average star ratings for each word. For calculating the average star rating associated with each word we take two steps:

* Sum the star ratings associated with reviews where each word occurs in
* And consider the proportion of each word among reviews with a star rating

**Top 20 words based on their average STAR ratings**



**Bottom 20 words based on their average STAR ratings**

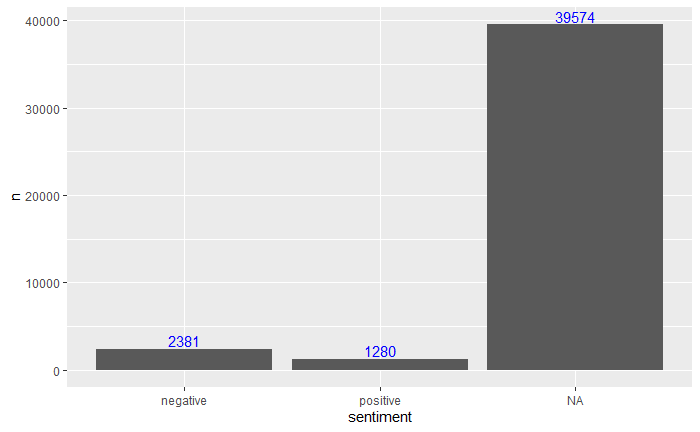


So, the presence of words from top 20 bucket are highly indicative of a positive segment. And the bottom 20 words bucket can be taken as the word’s indicative of negative sentiment. These ‘positive’ and ‘negative’ words do make sense in the context of user reviews being considered. The positive words are ‘nice’, ‘fresh’, ‘delicious’ etc. which are used positively in English language. Similarly, the negative words, which are ‘violations’, ‘shameful’, ‘enemy’ are used in a negative sense in English language.

1. **How many matching terms are there for each of the dictionaries? Describe how you obtain predictions  
   based on aggregated scores. Are you able to predict review sentiment based on these aggregated  
   scores, and how do they perform? Does any dictionary perform better?**

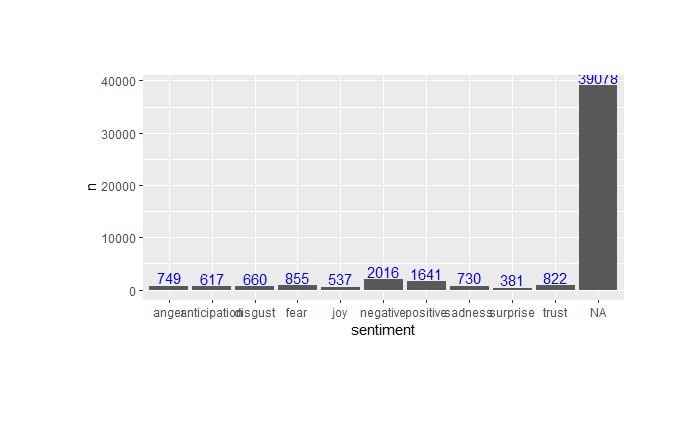
We will consider three dictionaries, available through the tidytext package – the NRC dictionary of terms denoting different sentiments, the extended sentiment lexicon developed by Prof Bing Liu, and the AFINN dictionary which includes words commonly used in user-generated content in the web. The first 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.

**Matching Rate with Bing:**



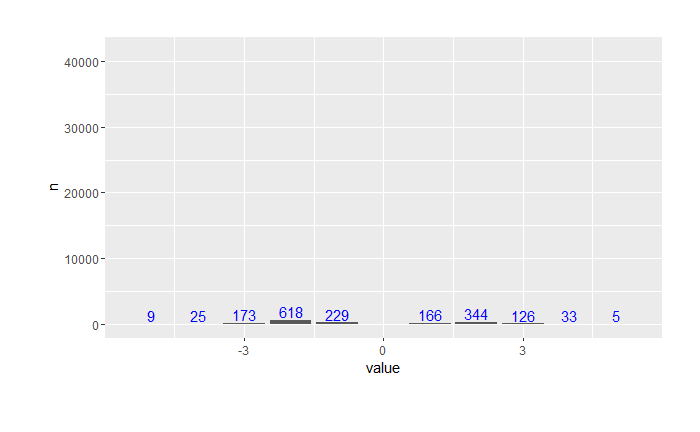
Out of all the distinct words that is in our data after performing data cleaning, there are 3661 words that match with the bing dictionary.

**Matching Rate with NRC:**



Out of all the distinct words that is in our data after performing data cleaning, there are 4157 words that match with the NRC dictionary.

**Matching Rate with Afinn:**



The matching rate is lowest for the Afinn dictionary with only 1728 words being matched to different score in Afinn.

Using proportion of occurrence for each words and across documents we calculate the TF-IDF values for each words. Next, we used the scores from different dictionaries to perform prediction on reviews and label them as either ‘positive’ or ‘negative’. Using each dictionary, we obtained an aggregated positiveScore and a negativeScore for each review and for the AFINN dictionary we derived an aggregate positivity score for each review.

For calculating the positive and negative scores for each review we performed few calculations for each words using the three different dictionaries.

**For BING**: For each words we had calculated its ‘occurrence rate’. Next we get our sentiments from bing and then if the sentiment for a particular word is positive we used positive value of the ‘occurance rate’ for that word. And if the sentiment is negative for a word we took negative value of the ‘occurance rate’ for that word. Next, for each review, we calculate the number of words within it, the sum of ‘occurance rate’ for positive sentiments words and the sum of ‘occurance rate’ for negative sentiments words. We then calculated the positive and negative proportion for each review dividing the sum of ‘occurance rate’ by number of words within each review. We then calculated the ‘Senti Score’ which is the difference between the positive and negative scores. And at the last, for each star rating we take the average positive, negative and senti score.

starsReview avgPos avgNeg avgSentiSc

1 0.3093150 0.6906850 -0.3813701

2 0.4475654 0.5524346 -0.1048691

3 0.6100337 0.3899663 0.2200675

4 0.7552362 0.2447638 0.5104725

5 0.8325914 0.1674086 0.6651827

**For NRC**: For NRC, we considered {anger, disgust, fear sadness, negative} to denote 'bad' reviews, and {positive, joy, anticipation, trust} to denote 'good' reviews. For each words, we had calculated its ‘occurrence rate’. Next we get our sentiments from based on above positive and negative bag created, and then if the sentiment for a particular word is positive we used positive value of the ‘occurance rate’ for that word. And if the sentiment is negative for a word we took negative value of the ‘occurance rate’ for that word. Next, for each review, we calculate the number of words within it, the sum of ‘occurance rate’ for positive sentiments words and the sum of ‘occurance rate’ for negative sentiments words. We then calculated the positive and negative proportion for each review dividing the sum of ‘occurance rate’ by number of words within each review. We then calculated the ‘Senti Score’ which is the difference between the positive and negative scores. And at the last, for each star rating we take the average positive, negative and senti score.

**For Afinn**: AFINN itself assigns negative to positive sentiment value for words matching the dictionary took the sum of sentiment value for words in a review and then calculated the average senti score for each star ratings.

stars avgLen avgSenti

1 5.131027 -2.543253

2 5.148071 0.729226

3 5.060825 3.812583

4 4.940350 6.608025

5 4.399350 7.372450

Based on the above scores that we have created using each dictionaries, we can use it to perform predictions.

Using each of the three dictionaries, for each reviews we have two things: 1. Star Ratings and 2. Senti Score. Using the above two we can make our predictions.

**Actual Class**: For each review, if its rating is 4 and above, we will classify them as 1 and if the rating is 2 and below we will classify them as -1. We will remove the reviews with 3 star ratings to handle ambiguity. This will be our actual class for each reviews.

**Predicted Class**: Next, using the senti scores that we calculated, if the senti score > 0, we can predict that review as 1 else 0.

Now, we have our actual class and predicted class for each different reviews using the three different dictionaries. We can look at the confusion matrix for each of the dictionaries.

**Using BING: predicted Using Afinn: predicted**

**actual -1 1 actual -1 1**

**-1 6371 1992 -1 5122 3076**

**1 3440 21794 1 2130 22574**

**Using NRC: predicted**

**actual -1 1**

**-1 3185 5412**

**1 1711 23955**

From above table we can see that the NRC dictionary is performing well in predicting the negative class and the Bing dictionary is doing better than others for positive class predictions.