Investigation Into Predicting Yelp Ratings Based on Review Text

Introduction

This report describes an investigation into the following question.

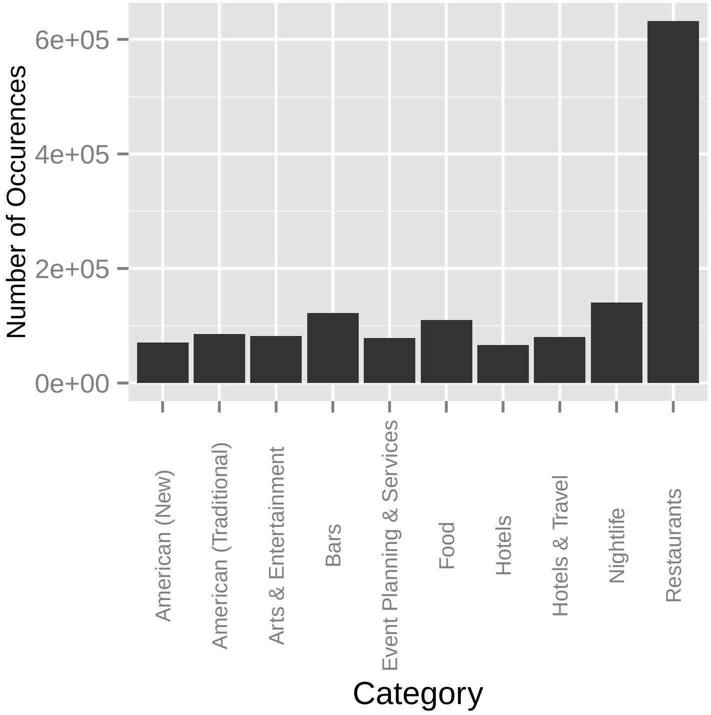
Is it possible to predict the number of stars given to a business based on analysis of the text in a Yelp review? Is the accuracy of the model affected by demographics, for example does it work less effectively in Germany because of the language difference. Given the fact that Spanish is increasingly becoming the language of the US, is that important in answering this question?

This question would definitely be of interest to yelp.com and their customers as it might provide a basis for judging reviewer sentiment in situations where they have comments but no ratings. It might also form the basis of a means of generating estimated ratings from Social Media posts regarding businesses.

Methods and Data

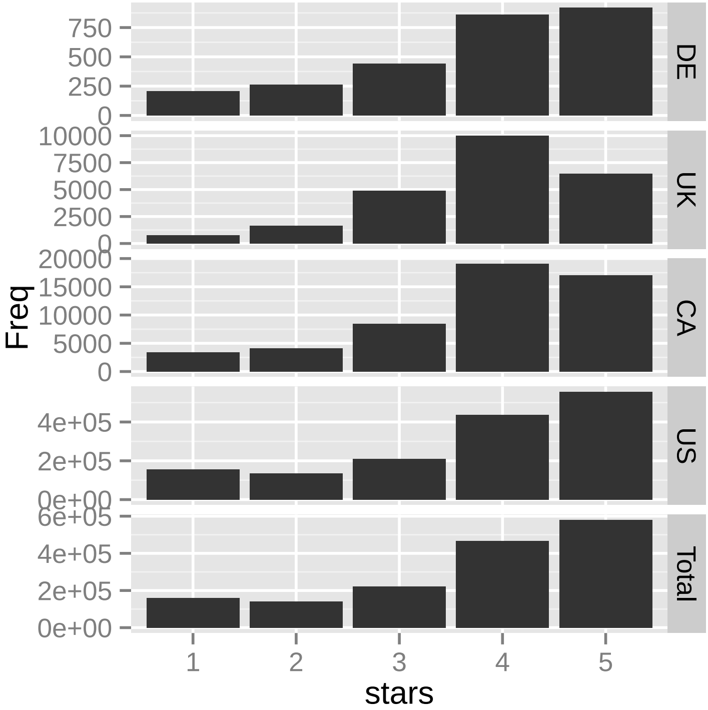
Exploratory Analysis

Initially we investigate the distribution of business categories to see if one dominates in order to see if there was potential to limit the scope of the search to make this huge dataset more manageable. As can be seen from the results below, restaurants dominate all other categories so we limited the scope of this investigation to restaurants.

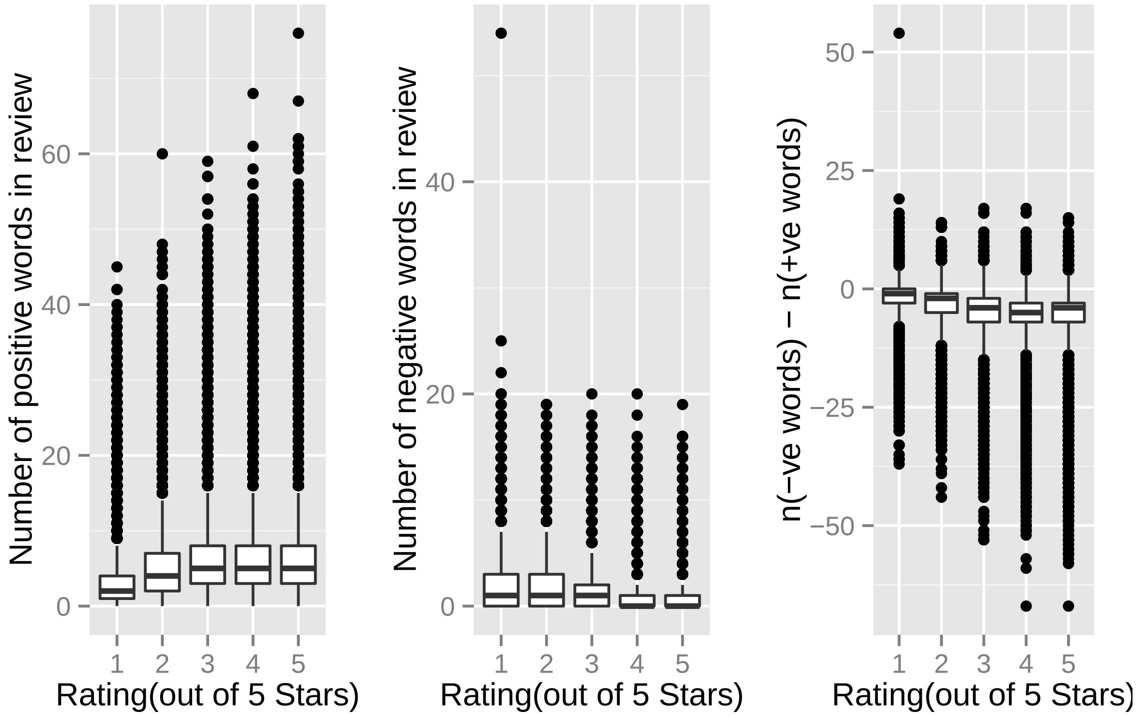


We also wanted to understand the geographical spread of the establishments in question, so we extracted this from the data. This information was gleaned from the state variable in the data which we discovered wasn't entirely regular or always mapping to official state designations so we manually built a lookup table to clean this and then evaluated it.

country	length(country)
DE	2692
UK	23751
CA	52168
US	1490653



As discussed in more detail later in this report, the conclusion was reached that modelling using the whole bag of words was not realistic computationally. The alternative approach then employed was based on sentiment analysis. More specifically it was based on the thesis that higher ratings would tend to have more positive words in them overall and that lower review scores would tend to have more negative words. The boxplot below based on the training data allows the investigation of this idea.



On the whole, this plot seems to at least partially support this thesis. Higher ratings tend to get more positive words and lower scores tend to get more negative reviews. However there do appear to be significant numbers of outliers that do not follow this trend. Its also not clear if the trend is pronounced enough to make it possible to distinguish between for example four and five star reviews.

Preparation

All code, along with a ReadMe to describe processing steps is on github in order to facilitate reproducability.

The streaming JSON files as published by Yelp were downloaded and an ER Diagram was built to relate all the quantities and the relevant keys. R code was written to pull the relevant information from the files. Some preprocessing was done including limiting the analysis to Restaurants and then the data was split into training, validation and test sets in a 60:20:20 ratio. As this was a lengthy process, the data was saved to three RDA files.

The summarised data then contains the following categories

```
##  [1] "user_id"      "stars"        "date"         "text"
##  [5] "business_id"  "full_address" "categories"   "city"
##  [9] "name"         "longitude"    "state"        "latitude"
## [13] "country"
```

The tm Text mining package was then used to extract the Corpus from the text. We then proceeded to apply standard transforms to the corpora in order to facilitate analysis. Punctation, numbers and whitespace were removed, all text was converted to lower case and standard English “stop-words” were removed using the facility provided by the tm framework.

The Corpora were then transformed into a Term Document Matrix in order to give a “bag of words” representation and only words that occurred in more than 1% of reviews were retained. Initially the approach was to attempt to model directly using the bag of words but it became readily apparent that attempting to model a large number of observations with such a large number of features was computationally prohibitive with the available resources.

Faced by this problem we switched tack and attempted a basic sentiment analysis of the text following Breen et al using lists of known positive and negative words originally provided by Hiu and Lu. Following this approach, we simplified the problem to one of modelling the dependency of the star rating on three features. These were number of positive words, number of negative words, and review length in words.

Model Development

During the training phase, multiple models were evaluated including rpart, ctree (this was used because the rpart tree that was seen, would never predict 2 or 3 star reviews), Random Forest, Naive Bayes and a Linear Model. The efficacy of these models was evaluated by investigation of the Root Mean Square(RMS) error and percentage of exact matches "in Sample".

	PctMatches	RMSValues
ctree	42.98	1.42
RF	42.87	1.45
NB	40.25	1.49
rpart	40.09	1.48
lm	23.90	1.30

Based on these results, the ctree, Random Forest and Naive Bayes models were selected to be used to predict the star ratings for the validation data set in order to get an estimate of the out of sample errors in an attempt to preclude potential overfitting.

	VPctMatches	VRMSValues
ctree	42.86	1.41
RF	42.64	1.44
NB	21.22	2.07

Based on these results the ctree based model was chosen to be the one that we would use to predict from the test data.

Results

Model Accuracy

The figure below, shows the Confusion Matrix which was used to judge the accuracy of the model in predicting the outcomes of the test data set.

	1	2	3	4	5
1	18436	9054	6536	5725	4991
2	1101	1129	971	852	486
3	665	928	918	870	499
4	5016	8204	14681	25057	21024
5	6860	8691	21437	60815	88905

As can be observed, the model works fairly well for predicting the extremities of the review scores (1 and 5) but is less effective for entries where the observed scores were 2,3 and 4 stars. One could conceive that this could be the case, if the reviewer is giving mixed reviews, maybe because some things were good and some things were not. They then might use a mixture of sentiments in their review whereas a 1 star review is likely to be predominantly negative and a 5 star review is more likely to be predominantly positive. The sensitivity results back this up with 1 star = 0.57, 2 stars = 0.04, 3 stars = 0.02, 4 stars = 0.27 and 5 stars = 0.78. This also demonstrates the effect of the dataset being skewed with many more five star reviews than any other kind.

Geographic Dependency

	%Predicted	RMS Error
NA	40.404540	1.391928
CA	39.833248	1.375968
UK	37.438381	1.338695
DE	1.165931	2.275953

As can be seen, the only strongly evident dependency on the regional nature of the data is for the German data. The rate of successful prediction and the RMS errors are much worse in Germany than other areas - this clearly shows that the fact that our model used English words to identify sentiment prevents it being useful in Germany.

Discussion

Based on the results for the accuracy of the predictions from the model detailed in the previous section the answer to the first part of the question that I set myself is yes we can to some extent predict the star ratings from the Review text as I see a success rating of approximately 40% which is far in excess of the 20% success rate one would expect from randomly selecting a star rating.

We also see that the selected model is much more successful for very negative and very positive reviews. This is something that was seen across all the algorithms that we applied to the sentiment analysis. The thesis that this model was based upon is clearly at least partially flawed. The box plots earlier in the document show that we see considerable outliers where 1 star ratings have many positive words or 5 star ratings have many negative words. A number of factors could be at play here. Firstly, we consider each word in isolation - the context of these words could change totally when considered with preceeding and following words, for example good -> not good. The model could potentially be improved by the inclusion of NGrams in some way. There are also potentially less important effects where different parts of the community might use the word to mean something other than its dictionary meaning, for example, the use of the word bad to mean good amongst youths. There is clearly more work to be done in extending the algorithm but 40% success is not to be sniffed at.

The second question around if language is important is clearly supported. Now with the model that I used based on positive word lists this is not at all surprising, indeed it is expected. We would only expect to identify an indication of sentiment where the reviewer happened to use English (maybe as a foreign visitor or where the words happen to be the same between languages or have been adopted into one language from the other). What would have been much more interesting to look at was to see if this effect had also been seen modelling the bag of words. Obviously as I was not able to evaluate that model due to computational constraints, I am not able to answer that question.

To summarise, this work is a good start but there a number of other avenues that it would be helpful to pursue in order to further develop this area.