**Introduction**

Yelp is a popular website which allows users to post reviews of businesses and organizations with which they have had some form of interaction. In an effort to help academics and educational institutions, Yelp published a dataset which included 10 metropolitan areas in the United States and Canada. Within the published dataset, there were 6 subsets: Business Information, Checkins, Photos, Reviews, Tips and Users.

*Question: is it possible to predict if a user’s rating will be positive or negative based on language used in the reviews they post on Yelp?*

**Initial review of the data**

Initial review of the Yelp dataset revealed information for the following metropolitan areas: Arizona, Illinois, Nevada, North Carolina, Ohio, Pennsylvania, Wisconsin, Alberta (CA), Ontario (CA) and Quebec (CA). Given the voluminous number of available records in each subset, it was necessary to use only a portion of the subset in order to analyze the data since only one workstation was available for processing and computations.

Original dataset

|  |  |
| --- | --- |
| False (Not a negative review) | True (Negative review) |
| 230,218 | 69,782 |

230218/(230218+69782) = 0.7673933 76.7%

**Taking a subset of the data**

Libraries used: *dplyr, ggplot2, jsonlite, tidytext, wordcloud*

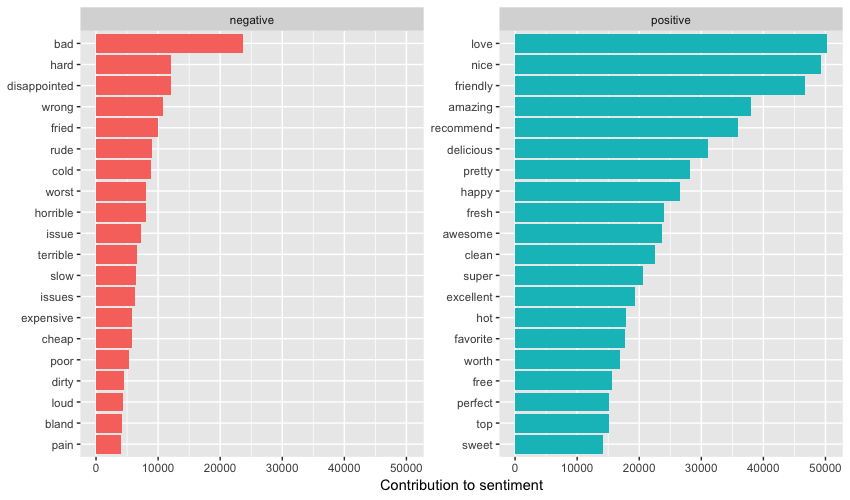
For the new subset to be created, two of the original subsets were used: Business Information and Reviews. The subsets from Yelp came in JSON format but upon further inspection, the JSON files were NDJSON (New Delimited JSON) files meaning they contained multiple JSON values which were considered independent variables. The *jsonlite* library contains the stream\_in() function which allows R to read the NDJSON correctly.

Since the Reviews subset did not contain geographical location, it was necessary to combine the two subsets into a new dataset. Using the business\_id column as the key, the inner\_join feature in the *dplyr* library was used to find reviews associated with the Arizona metropolitan area.



Using the Bing lexicon from the *tidytext* library, the most common negative and positive words from all the reviews in the new dataset were identified. Using the *ggplot2* library, the top 20 recurring positive and negative words as well as their overall frequency was created. These results will not be used in the creation of the model but serve

Most Common Negative and Positive Words in Reviews



After the records were combined into the new dataset, a conditional column was created to identify if the review was positive (3 or more stars) or negative (less than 3 stars) to be used during the creation of the training model.

Wordcloud generated using negative reviews



Wordcloud generated using positive reviews



Initially, all of the records in the new dataset were used to create the training model but heavy computational requirements of textual analysis precluded use of the entire dataset. Instead, a random subset of 300,000 reviews from the Arizona metropolitan area was exported as a CSV file to make working with the data more manageable.

**Creating a training model**

Libraries used: *caTools, dplyr, rpart, rpart.plot, SnowballC, tm*

Assumptions:

*Reviews with less than 3 stars considered “negative”*

*Reviews with 3 or more stars considered “positive”*

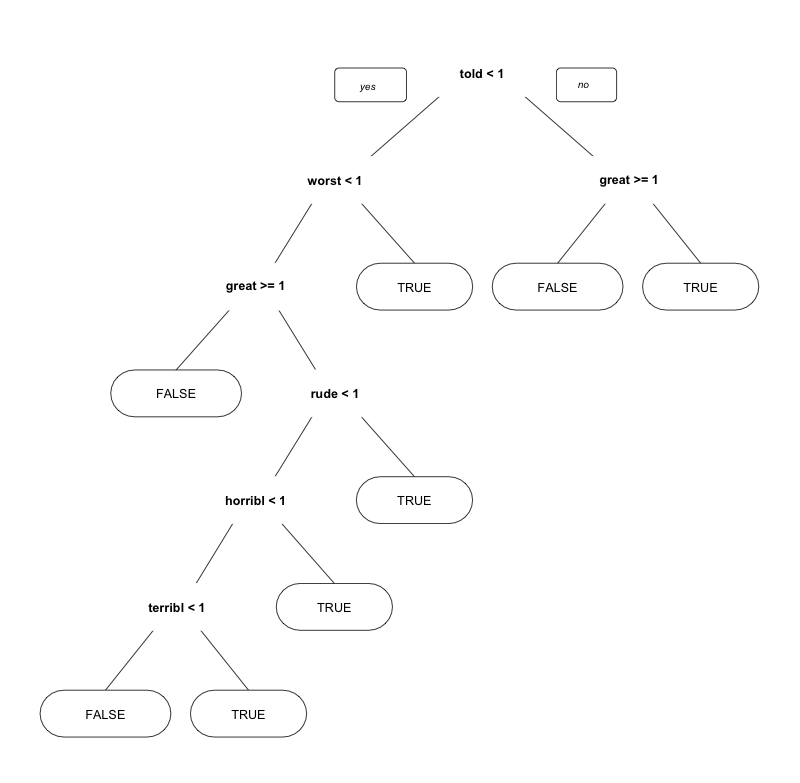
Given the breadth of data contained in the Yelp dataset, assumptions were required to limit the scope of the project. The assumption for this project was that differentiating between a positive and negative review was more imperative than identifying the differences between different star ratings. Taking a more focused approach allowed the use of logistic regression given the binary outcome.

With the new subset finalized, the logistic training model could be created given that there were only two outcomes. Each of the 300,000 reviews was loaded into the corpus or set of data being analyzed. Upon loading the data, all letters were converted to lowercase, punctuation was removed, stopwords defined in the *tm* library were removed and all text was stemmed for consistency using the *SnowballC* library.

Terms occurring in 2% or more of the reviews were used to establish the data frame which resulted in 453 terms being applied as columns using the Bag of Words model. To avoid problems with naming conventions in R, all columns were renamed using the make.names functionality found in the R base package.

The 300,000 reviews were split 70/30 – 210,000 reviews for training and 90,000 reviews used for testing. A logistical regression model was created using the *rpart* library using the calculated Negative column which was compared to all other columns in the matrix.

From the training data, the following decision tree was created using *rpart.plot*:



**Testing the data and interpreting the results**

Libraries used: *dplyr*

Using the model created from the training data, the test data was run through the model to determine its accuracy. With the results from running the testing data, a confusion matrix could be created.

|  |  |  |
| --- | --- | --- |
|  | False | True |
| False | 66,812 | 2,253 |
| True | 13,212 | 7,723 |

(66862+7723)/(66862+2203+13215+7723) = 0.8281667 82.8%

Given the information of the confusion matrix, the following information was used to calculate the model accuracy. Using logistical regression, it is possible to predict a positive or negative review in a Yelp review based on language used with an 82.8% accuracy.

Leveraging the Model

Solution: Help businesses identify common terms used in negative reviews to know where to focus their efforts to improve customer satisfaction.

Implementation: With the model created and the common terms identified, Yelp can build the logistic regression model into its website. Whether on a periodic or as-needed basis, Yelp can provide the data to businesses in an effort to help said businesses improve customer satisfaction. Yelp can provide this information free of charge to businesses or at a fee, depending on the Yelp’s business model and at the discretion of executive management.

Solution: Identify reviews in which the person erroneously selected a negative rating (less than 3 stars) even though their experience was generally positive.

Implementation: At its core, Yelp provides information to people about businesses in a given area. If reviewers are providing star ratings that are inconsistent with the language used in the review, it can cause skewed averages leading to a loss in credibility. Verification of language consistent with a positive or negative review either at the time of submission of the review or after the fact, can provide an additional level of quality control.