Capstone Report for Data Science Specialization

Coursera & Johns Hopkins University

John Slough II

22 November 2015

Abstract

This report is concerned with a sentiment analysis of text-based reviews from the Yelp Challenge Dataset. Using a bag of words model, the sentiment of text-based reviews was successfully predicted with logistic regression. In addition, a multi-class model predicting the star rating (1-5) of the text reviews was explored, with less successful results.

Introduction

This capstone report for the Coursera Data Science Specialization from Johns Hopkins University will answer the question: Can we predict the sentiment of a textual review (positive or negative) from a corpus of restaurant and food service businesses reviews. In addition, a more granular prediction model will be explored using a subset of the data, i.e. predicting the number of stars from 1 to 5 given to a food service business from the review's text.

Methods and Data

The dataset was provided by Yelp, a website where users can rate business with a textual review and a 1-5 star review (whole stars). The dataset was provided on an academic license agreement. All analyses performed are purely for academic purposes.

Analysis Software

Because the main software used in this specialization was R, I chose to perform the data processing and exploration using R. However, Python proved to be much faster in building the prediction model with textual data. Dato's GraphLab platform was used to build the models. We are not limited to using R in this project so Python was used to build the machine learning classification models.

Data Processing

The data was originally in JSON form, in a total of 5 datasets connected by identifiers. The data contained reviews from businesses in 10 cities around the world. The 6 cities in the USA were selected to be analyzed to remove reviews in other languages, and other dialects of English. We will also limit the analysis to food service businesses, as the features of the reviews would not be similar for different kinds of businesses. Using the function 'grep' the variable 'categories of business' was searched for restaurants and business serving food.

Review Text Processing

After the subset of the data and specific variables of interest were obtained some processing of the text was necessary.

Cleaning of the Review Text

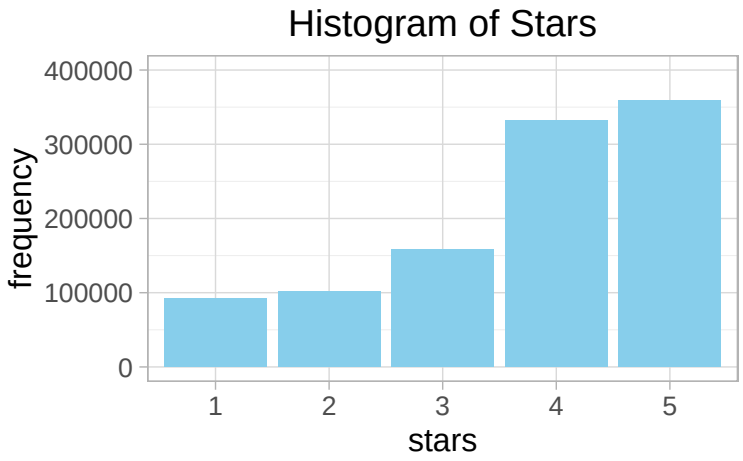
Taking advantage of the 'tm' package in R, the text was cleaned by removing numbers, punctuation, line break markings, and to convert everything to lower-case letters. Stopwords were removed as well. This is a common practice in natural language processing which removes common words such as the and I which contain little information for prediction.

Emoticons

One feature of the text reviews is that some of them contain emoticons. These were incorporated in the building of the prediction model. A list of positive and negative emoticons from the R package 'qdap' was used. For each positive emoticon the word "emotismiley" was substituted and for each negative emoticon the word "emotifrowney" was substituted. In this way emoticons can be incorporated into the word count feature just like any other textual word. There were a total of 42,559 positive emoticons and 9,712 negative emoticons in the reviews.

Data Exploration

The Dataset includes reviews of the restaurants, the user's id, the business' id, and various other attributes which we are not concerned with in this analysis. After processing of the data there were a total of 1043540 reviews, of which 691175 or 66.2% were 4 or 5 stars. 193704 (18.6%) were 1 or 2 stars. A histogram shown below displays the distribution of the stars. It is clear the most reviews are positive.



There were a total of 22675 establishments providing food service which means that there were, on average, 46 reviews per business. The overall average star rating was 3.73.

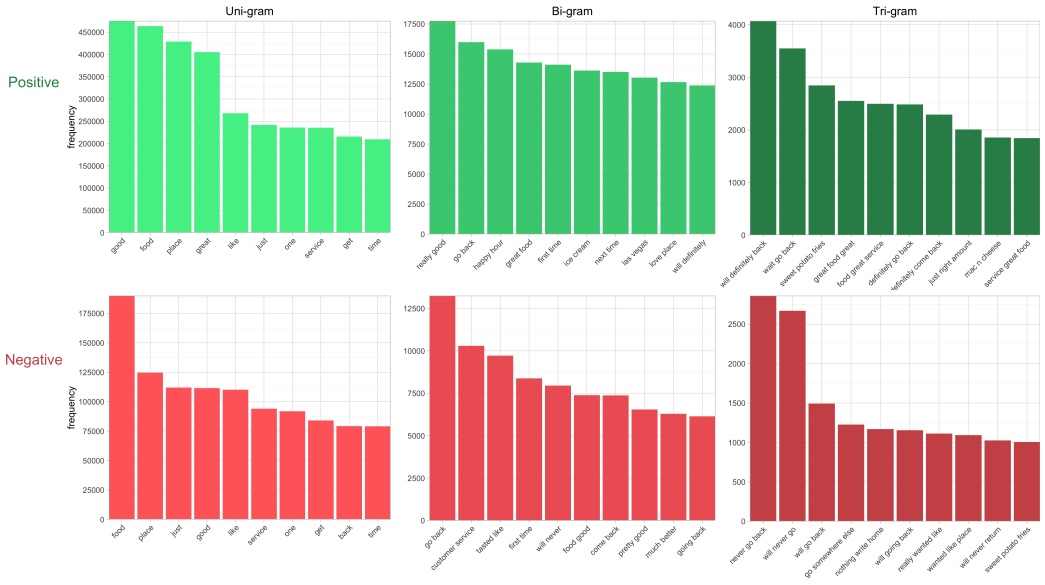
An example of a review is taken from the business with the id: wx2EJUCNOCPrMC0DtKb98A. The name of that establishment is "Brooklyn Bagel Deli" There are 63 reviews of the deli. An excerpt (after text processing) from a review of this restaurant is:

"cream cheese go hand packed large selection flavors think salmon strawberry cheddar bacon veggie large selection beverages snacks daughter loves picking belly washer go gurt go chocolate chip bagel emotismiley. . . "

This reviewer gave the restaurant 5 stars. The mean number of stars for all reviews of this restaurant was 3.54.

N-grams

N-grams were analyzed using the *quanteda* package in R. Below is a chart of the most common n-grams, divided into positive and negative reviews based on their star rating (1 or 2 stars are negative, 4 or 5 are positive). See the full size here.



We can see that it is unclear if the uni-grams are positive or negative, however it is pretty clear that the bi-grams are positive or negative. Tri-grams are easily differentiated ("will definitely back" versus "never go back"). Some of the negative n-grams clearly would have a negation such as "not." Unfortunately some of these are included in the stopwords which were removed. This is a limitation of this kind of modeling.

The Model

The first model we will look at is just a simple binary classifier. This classifier will aim to answer the question: Is the review positive or negative based on the text? Each review was labeled as 1 for positive and 0 for negative, determined by the number of stars. All 4 and 5 star reviews were classified as positive and all 1 and 2 star reviews were classified as negative. All 3 star reviews were removed from the analysis because they were considered to express a neutral sentiment and were therefore not applicable in this model.

The model was trained on data selected by randomly splitting the entire dataset into a 70% training and 30% testing dataset.

Adding features such as n-grams (where $n > 1$) did not increase the accuracy on the training dataset and significantly increased computation time so they were not included in the final model.

The final model uses the 'bag of words' approach or 1-gram counts. This was created using graphlab's function 'text_analytics.count_words,' which counts words in each review.

Multiple algorithms such as random forests, naive Bayes, and support vector machines were explored, however logistic regression gave the most accurate results in the training dataset. This is beneficial because the results are easily interpreted.

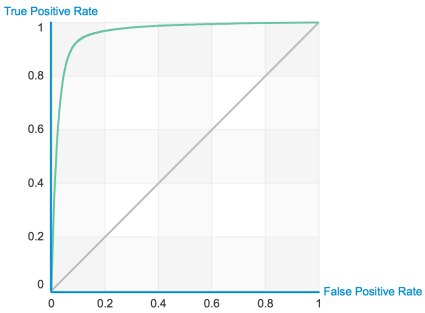
Results

Evaluation of the Model

The trained model achieved an accuracy of 97.3% on the training dataset and an accuracy of 94.2% on the testing dataset. The confusion matirx for the testing dataset is shown below. 1 is positive and 0 is negative sentiment.

		Predicted Label	
		0	1
Target Label	0	48741	9339
	1	6009	201198

The figure below shows the ROC curve for the logistic regression model.



Model Applied to a Specific Business

As an example of the output, the model was applied to the the Bagel Deli discussed above. An excerpt from the table including the reviews and sentiment analysis probability is shown below. Overall, it appears that the classifier does well in predicting the sentiment of the review.

Most Positive Reviews by Predicted Sentiment

text	stars	sentiment	predicted_sentiment
upgrading brooklyn star rating can see check 's ...	5	1	0.999999728882
love place customer almost eight years ba ...	5	1	0.999940649904
love bagel cream cheese passed place years ...	4	1	0.999870012228
going deli last years let tell probably favorite ...	5	1	0.999781595779

Most Negative Reviews by Predicted Sentiment

text	stars	sentiment	predicted_sentiment
terrible service gloves worn handeling food ...	1	0	8.70816216414e-06
seriously worst experience ever ...	1	0	5.89318154212e-05
im breakfast sandwich kinda guy far worst p ...	1	0	0.00461133306956
star customer service way sorry takes couple ...	2	0	0.0109311856362

Multi-class Model

A model to predict the star rating of the review based on the text was created as well, using the testing and training data. All the reviews (all star ratings) were included in this analysis. Random Forest and boosted trees classifiers were explored however the multinomial logistic regression model proved to be most accurate. The same feature were used in this model as in the sentiment model.

Evaluation of the Model

Applied to the test dataset, the model achieved an accuracy of 75.0% on the training dataset and 55.0% accuracy on the testing dataset in predicting the star rating of the review based on the text. The confusion matirx for the testing dataset is shown below.

		Predicted Label				
		1	2	3	4	5
Target Label	1	16077	6581	2153	1626	1161
	2	5685	11719	7931	3801	1334
	3	1820	6708	17541	16876	4499
	4	752	2205	10782	52057	34074
	5	386	804	2966	28844	74591

Discussion

Sentiment analysis is a very important part of natural language processing and has been used in many areas such as predicting the stock market fluctuations, predicting election winners, to analyzing brand sentiment from Tweets. In this analysis we have successfully predicted the sentiment of reviews from their text. Because the prediction accuracies for the training and predicting dataset were very similar, we can assume that the model is not overfitting. Predicting the star rating from the text was less successful, with an accuracy of 75% on the training dataset and 55% on the testing dataset, indicating overfitting. Overfitting could be reduced by choosing a random forest model or performing cross validation. In addition, adding more features to improve the accuracy would be helpful and this will be looked into.

I would also note that this exercise in creating a sentiment analysis has helped me understand how those models are built and has given me a good introduction to the field of Natural Language Processing.

Emoticons

The inclusion of the emoticons in the analysis resulted in coefficients being produced for the sentiment analysis model for positive and negative emoticons. These were 0.969 and -0.585, respecively, which indicates that they were in the logical directions. The emoticon anaylsis was also included in the multi-class model with similar results. While they were not very influental on the predictions due to the relatively small number of reviews that contained emoticons, this kind of analysis could be very beneficial for text with higher instances of emoticons.

Limitations and Further Work

This model was trained on reviews for food service businesses in the USA. As such, it may not be generalizable to other kinds of reviews or in other locations. An exploration of reducing the number of coefficients (only choosing the most important words) would be optimal for the multi-class model, as the computation time was relatively long. In addition, reducing the number of coefficients would speed up the processing time if larger datasets were used. The next step would be to identify the most important words in the reviews for prediction, and perhaps adding other features to increase prediction accuracy.

Conclusion

The prediction model was succesful in classifying reviews as positve or negative, with an accuracy of about 94% on the test dataset. The multi-class model was not as successful, as it achieves an accuracy of 55% on the test dataset and appears to be overfitting. Most importantly, I have gained much more knowledge and appreciation of this field. The code for the entire project can be found here.