Predicting ratings from review text

vulpinae

November 16, 2015

1. Management Summary

Yelp online reviews are a great source of information for users to choose where to visit or what to eat. However, there are so many reviews, that it is difficult for users to quickly obtain the information they are searching for. One quickly turns to the 1-5 star(s) rating from a review. This rating can be subjective and biased toward users personality. In this paper, a review rating is predicted only based on the review texts. This not only provides an overview of plentiful long review texts but also cancels out subjectivity.

From the Yelp dataset, the Las Vegas restaurants are chosen as the base the analysis. The predictions are made from three different feature generation methods:
- bag of words from the top frequent words in the review texts
- bag of words from the top frequent adjecent words (bigrams) in the review texts
- top frequent adjectives from Part-of-Speech analysis of review texts .

The classifier to predict a review star rating is self-built. For each star rating (1-5) a logistic regression model is build to predict the chance of having that star rating. So, the classifier takes the probabilities of the review being a 1 to 5 star rating. The model which has the highest probability is taken as the prediction.

Twelve classifiers are built on a combination of three different feature generating methods with different number of features (100, 150, 250 and 500). The best classifier is the single word classifier with 500 features. The percent of correct predicitons (accuracy) on reviews that are not used to train the model is 48.9% For one and five star reviews it is even above 67%. This indicates that discrepant reviews for one and five stars can be identified. It are those reviews for which the predicted stars are three or four stars different from the actual stars.

The results also indicate that there is room for improvement. Generating more features will increase accuracy, but also increase overfitting, so that sample size has to increase as well.

2. Introduction

Nowadays, people's decisions of where to visit or what to eat are subject to other people's opinions. These opinions are widely spread throughout the internet. Websites like Yelp are a very useful help because reviews and opinions written by everyday people are concentrated in one place. However, user-generated reviews are usually inconsistent in terms of length, content, writing style and usefulness because they are written by unprofessional writers. Important information can be easily obscured unless users are willing to spend a great deal of time and effort on reading the reviews thoroughly. A common solution to provide a brief overview is to show overall rate of a business in form of 1-5 star(s). While Yelp ratings are often considered as a reputation metric for businesses, they may suffer from subjectivity and being biased towards the personality of users Any two users can describe their experience with multiple positive words such as "fabulous","must go", "excellent service" "tasted great" etc. However, their ratings might differ. The goal to the analysis at hand is to identify reviews that have a clear discrepancy between the content (review text) and the actual rating of the review. These discrepant reviews than can be hidden from the Yelp website so that users can obtain their information more quickly.

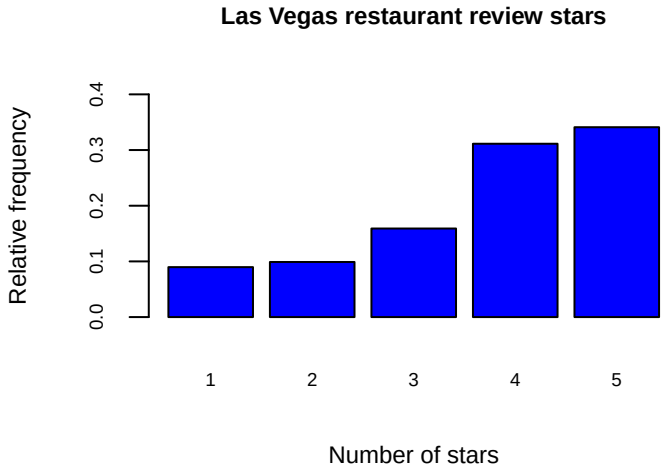
3. Methods and Data

3a. Data The data originates from the from the Yelp Dataset Challenge round 6 [http://www.yelp.com/dataset_challenge] and contains:

- 1.6M reviews and 500K tips by 366K users for 61K businesses
- 481K business attributes, e.g., hours, parking availability, ambience
- Social network of 366K users for a total of 2.9M social edges
- Aggregated check-ins over time for each of the 61K businesses

To overcome large throughput times of analysis the following filters are used:
- since restaurants is by far the largest business category, only restaurants are being analyzed
- only reviews from the city of Las Vegas are being analyzed.

This leads to a dataset of 370,194 reviews. The distribution of the review stars is shown in the following figure.



As can be seen, five star ratings are most common, closely followed by four star ratings.
From this dataset a weighted sample of 10,000 reviews is taken, by sampling 2,000 reviews per star rating. To examine if different words are used in the different star ratings, two wordclouds are presented below. The first is the wordcloud for the most frequent words in a 1 star review, the second is the wordcloud for the most frequent words in a 5 star review.

Word cloud 1 star review <———— ————> **Word cloud 5 star review**

One star reviews frequently contain words like “never”, “don’t”, “bad” and “didn’t”, while five star reviews frequently contain words like “love”, “best”, “ever” and “delicious”. This indicates that frequency of words can be used as features of a review to predict the star rating.

3b. Data preparation Before extracting features a bit of data preparation has to be done. The following steps are taken to prepare the review text for feature generation:
1. removing punctuation
2. removing numbers
3. capitalization (convert to lowercase)
4. removing stopwords

In the R tm-packages there are standard 174 english stopwords. The ones that might have sentimental meaning have been removed from the stopword list, so that they would remain in the text of the review. Examples of those words are "not", "wasn't", "most", "too", etc. 5. removing common word endings (e.g., "ing", "es")
6. stripping whitespace

3c. Feature generation This stage generates features for the different cleansed review text. The endproduct is a matrix of i reviews (rows) and j features (columns). Three methods of generating features from a review text are used. These are bases on the principles 'bag of words' [https://en.wikipedia.org/wiki/Bag-of-words_model] and 'part of speech' [https://en.wikipedia.org/wiki/Part_of_speech]. The bag-of-words model is commonly used in methods of document classification, where the (frequency of) occurrence of each word is used as a feature for training a classifier. Part of speech is used to determine the adjectives in the review text.

Method 1 : top words (bag of words)
This method uses the words that occur most frequently in the review texts as features. The value for review i for feature j is the frequency of the j-most frequent word in the analyzed reviews in the text of review i.

Method 2 : top bigrams (bag of words)
This method is similar to method 1, but uses the top combinations of 2 words (bigrams). The value for review i for feature j is the frequency of the j-most frequent bigram in the analyzed reviews in the text of review i. Rationale behind method 2 is that a combination of two words is more meaningfull than two single words. For instance if the review text contains 'not bad', this might be seen as positive sentiment. The previous method would only see two words that might be seen as negative sentiment.

Method 3 : top adjectives (bag of words, part of speech)
This method uses the adjectives that occur most frequently in the review texts as features. The value for review i for feature j is the frequency of the j-most frequent adjective in the analyzed reviews in the text of review i. The rationale behind method 3 is that adjectives are the most commonly used type of words to describe positivity or negativity.

For each of the three generated feature matrices, the principle of term frequency–inverse document frequency [https://en.wikipedia.org/wiki/Tfidf] is applied. So the initial features (=term frequencies) are weighted by a idf component of log(number of reviews/number of reviews containing term j).Finally, features that highly correlate with other features (> 0.9) are removed.

The features generating proces is done four times for each of the three methods (single words, bigrams, adjectives). Respectively 100, 150, 250 and 500 features were generated for each method.

3d. Methods The estimation of the review stars from the review text is a clear classification problem. So, the random forest method is an obvious choice. However, the throughput time of the analysis was very long, so another method was considered. Naive Bayes classification performs well in terms of throughput time, but the results were not as expected. Almost every prediction was 4 or 5 stars. Therefore a third and final alternative is developed. The classifier works in two stages. In the first stage five logistic regression models are estimated to estimate if a review has a particular star value or not. The outcome of the regression models are chances that a review has a particul star rating. The second stage of the classifier is picking the model that predicts the highest probability for each review and assigning it the stars that that model was predicting. The example below would be classified as a 1 star review, because the prediction for the one star model is the highest of the five models.

Model	Prediction logit model
1 star	0.34
2 stars	0.28
3 stars	0.22
4 stars	0.19
5 stars	0.15

3e. Validation The models are estimated on 70% of the data, randomly selected. The other 30% is used for validation. The research problem is predicting the rating stars from the text. The focus lies on generating the wright prediction, so the results of the classifier are being measured by the accuracy on the validation set. This is actually how well a classifier can predict the actual star rating from a new review text that is not used in training the model.

4. Results

In the following table, the results are shown for the twelve combinations of feature generating methods and number of features. Both the accuracy on the analysis set and the validation set are shown.

Classifier	Accuracy analysis set	Accuracy validation set
Single words 100 features	0.471	0.430
Single words 150 features	0.500	0.453
Single words 250 features	0.551	0.470
Single words 500 features	0.626	0.489
Bigrams 100 features	0.389	0.364
Bigrams 150 features	0.423	0.377
Bigrams 250 features	0.475	0.392
Bigrams 500 features	0.544	0.408
Adjectives 100 features	0.451	0.414
Adjectives 150 features	0.466	0.425
Adjectives 250 features	0.502	0.421
Adjectives 500 features	0.553	0.425

The results show that the best classifier on the used sample is single word with 500 features. It has an accuracy on the validation set of 0.489. Remember that the stars were evenly distributed in the sample, so a random guess would be succesfull in 20%. The chance of being wright is increased by a factor 2.445.

There are some other general findings:
- As the number of features increase there is more overfitting, which is shown by the increasing distance between the accuracy in the analysis set and the validation set.
- Feature generation with simple words works better than bigrams or adjectives on the used sample.
- The adjectives beyond 100 do not contain much information, since the accuracy on the validation remains fairly stable as the number of features increase.

Below is the confusion matrix (predicted stars (rows) versus actual stars (columns)) on the validation set

for the single word classifier with 500 features. In the final row is the accuracy of the reviews that have a particular star rating.

Predicted vs actual stars	1	2	3	4	5
1	409	169	42	27	25
2	97	216	146	40	23
3	47	133	218	109	38
4	17	50	132	220	110
5	27	35	62	204	404
Accuracy	0.685	0.358	0.363	0.367	0.673

There is a remarkable difference in the accuracy of the one and five star accuracy compared to the two, three and four star accuracy. The classifier is far better classifying a one or five star review than another review. This implicates that discrepant reviews for a one star review are those with a four/five star prediction and discrepant reviews for a five star review are those with a one/two star predicion.

5. Discussion

The best classifier found, performs decently on new reviews, certainly the one star and five star reviews can be predicted correctly for 2 out of 3 of those reviews. For the problem at hand, finding discrepant reviews, the classifier might work quite well. Discrepant reviews can be identified if the predicted star ratings differ 3 or more from the actual star rating. On the other hand, the two, three and four star reviews can only be predicted correctly in roughly for 1 out of 3 of those reviews. It clearly demonstrates that is not easy to predict rating stars for the review text alone and their is enough room for improvement in classifying.

Overfitting is an issue in predicting the star rating from the text, because the more features are used the better the accuracy, but also the higher the overfitting.

Further research areas include:
- using bigger samples with more features, keeping in mind the overfitting issue
- combining features from different feature generating methods, preferable with bigger sample sizes to avoid overfitting
- using ensembling from a set of classifiers