Rating Predictions from Cosmetic Reviews

CSE190:Data Mining

Ranjit Ghodke
A09352495
rghodke@ucsd.edu
Spring 2015

Daver Muzaffar
A10290879
dmuzaffa@ucsd.edu
Spring 2015

ABSTRACT
Analyzing text and data can give us great insight on what to expect in the future. Patterns can be noted, with which futures patterns can be predicted. Certain words or phrases can also give indications of possible outcomes or results.

1.0 INTRODUCTION
For this project we will be predicting the star rating given to a product based on reviews. We will also be using features, such as helpfulness to better our prediction.

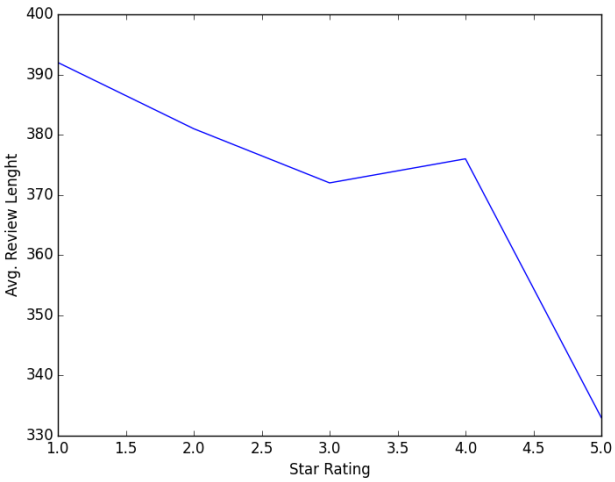
1.1 DATASET
The dataset is of Amazon reviews, ranging from 1996 to 2015. In this report we will be using reviews of cosmetics and beauty products only [1][2]. We will be using a subset of the whole dataset, 80,000.

Each review has the following features:
reviewerID - id of user
reviewerName - username
helpful - was the review helpful
reviewText - the review
overall - star rating
summary - summary of review
reviewTime - when reviewed
asin - product id

First we decided to look at the frequency with which a given rating occurs. This is based on a star rating system. The only options are 1,2,3,4 or 5 stars that a user can give a product. Showing how often a rating occurs:

1	2	3	4	5
7221	4231	6106	11264	51178

1.2 AVERAGE REVIEW SIZE - WORD COUNT
The dataset suggests that users leave longer reviews when giving the product a bad review. On average the longest reviews are given when a product is given one star (rating of 1). The shortest reviews are seen when a product is given a five star rating. It seems that when customers are very unsatisfied with the product, they are more motivated to leave a lengthy review. On the opposite side of the spectrum, when a customer is very happy with a product, they leave a short review.



Average Review Length - Avg

Avg(word count)	392	381	372	376	333
Rating (stars)	1	2	3	4	5

1.3 UNIGRAMS - IN REVIEW

When thinking about how to predict the rating a product will get, one of the ways that seemed obvious is text mining for unigrams that are popular. Ignoring unigrams that would not help predict rating,s we decided that out of all the top popular unigrams, the ones that would be helpful would be 'love', 'hate', 'terrible', 'bad' , 'amazing'. It seems that customers most often use these words to describe the product they are reviewing. In order for this information to be useful, we need to check if there is any correlation between the rating a customer will give and whether the customer will use the above mentioned unigrams. We wanted to test if the rating is lower when the reviewer uses the words 'terrible', 'bad' or 'hate', and if the rating is higher when the review has the words 'amazing' and 'love'.

The graph on the right hand side shows percentages on the y-axis. The percentage is to show what percent of the reviews have the unigrams(with respect the rating). What is indicated by the graph and table is that it is a product was given a high rating then it is more likely that the words 'amazing' and 'love' were used in the review. If the rating for the product was low then it is more likely that words 'terrible', 'bad' or 'hate' were used in the review. The data point that stands out the most is that 23 percent of customers that gave a product a 5 star rating also used either the words 'amazing' or 'love' in their reviews.

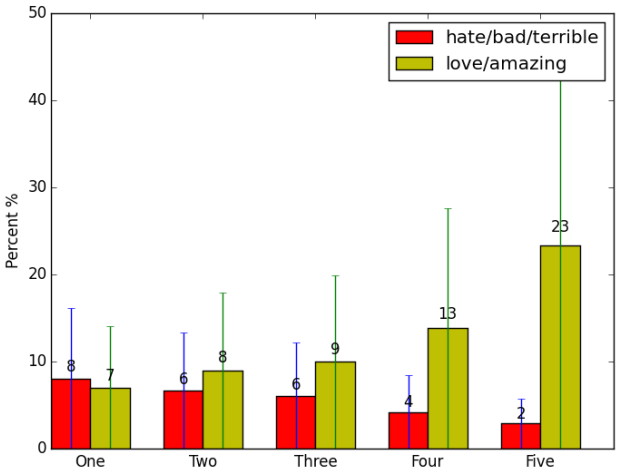
'love' or 'amazing'

%	7.02	8.98	9.95	13.82	23.36
Rating	1	2	3	4	5

'terrible' or 'bad' or 'hate'

%	8.05	6.68	6.07	4.21	2.88
Rating	1	2	3	4	5

Percentage on the graph below is to show how many of the total reviews with a given rating had the unigrams from the two groups. The groups are shown in the legend in the graph.

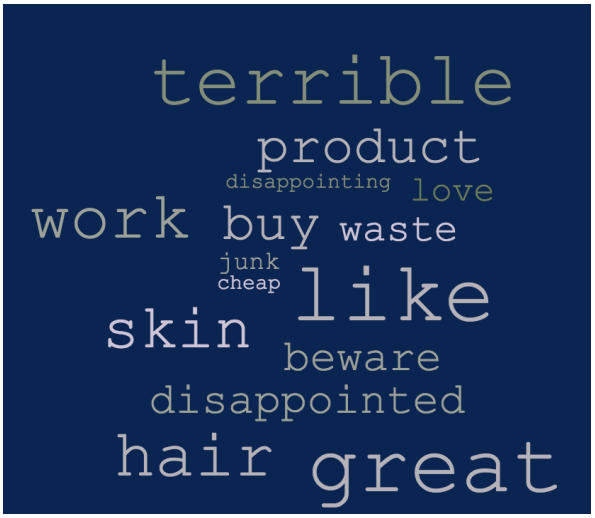


Equations for percentage:

G = Group (Red or Yellow form graph)
X = star rating (One - Five)
N = total X star review
Nu = total X star reviews with bigram from group G
% = (Nu / N) * 100

Most popular words in summary feature

Summary feature - 1 star ratings



Summary feature - 5 star ratings

1.4 SUMMARY FEATURE

A feature that might be worth analyzing for patterns is the summary feature.

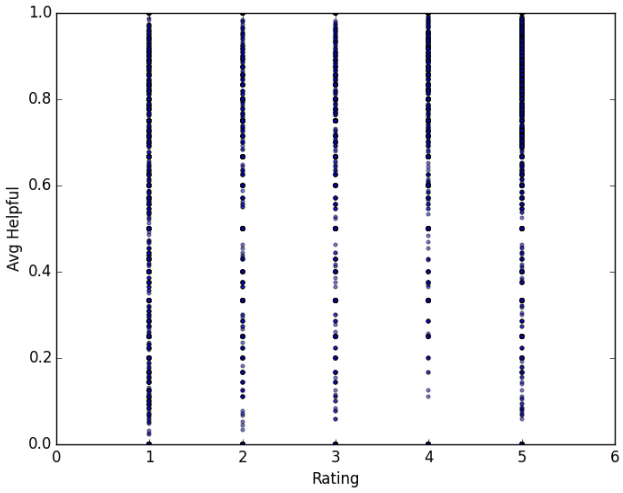
Rating	1	2	3	4	5
Avg Length	22.5	23.7	23.3	21.8	21.6

The table above indicates that we cannot really make any predictions simply based on the length of the summary feature, as the average seems to be the around the same regardless of the rating.

1.5 HELPFUL AVERAGE

When a customer writes a review, other customers who view that review can mark weather or not they found that review to be helpful. What this could possibly translate to is whether or not people agree that the product deserves the rating that the reviewer gave it. This means that there could be some correlation between the average helpfulness[1.5] and the rating. Which could be used in predicting the rating. The scatter graph below shows rating versus average helpfulness of a review.

Avg Helpfulness vs Rating



As the graph above indicates, when the rating is 1 star, the average helpfulness is very scattered and has a lot of variance, thus it does not seem very helpful in predicting the rating. However, when the rating is 4 or 5, there is a heavy concentration between

.7 to 1.0 for the average helpfulness rating, this could prove helpful in our prediction.

[1.5]Equation for Avg Helpfulness:
nH = number of people who found review to be helpful
N = total number of people who rated the review
Avg = nH / N

2.0 PREDICTIVE TASK - BASELINE

For this project we are attempting to make a model to predict the rating a customer will give a product, given a (userID, ProductID) set. To start of, we will use a simple model based on supervised learning. There are two possibilities. Either the userID was seen in the training set or it was not. Because of this we started out with two models, one for each case. This model serves as the baseline. We are ignoring all features other than user ID and Product ID

If user was NOT seen in training data:
$rating(userId, ProductId) = \alpha$
If user was seen in training data:
$rating(userId, ProductId) = \alpha_{USER}$
α_{USER} = Avg rating given by user

2.1 PREDICTIVE TASK - SENTIMENT

One possible way to improve on the model above(baseline) was to try and predict the sentiment of the review based on unigrams. In our evaluation of the dataset, we found the reviews differed a great deal when a customer was happy with the review from when a customer was disappointed by the product. We decided to use the frequency of keywords to see if we can improve our predictor. We once again went with a simple model that might be an improvement on the baseline. To do this we used linear regression mixed with n-grams. Once again we will have 2 sub models within our model. In this model we will look at an additional feature, the review text.

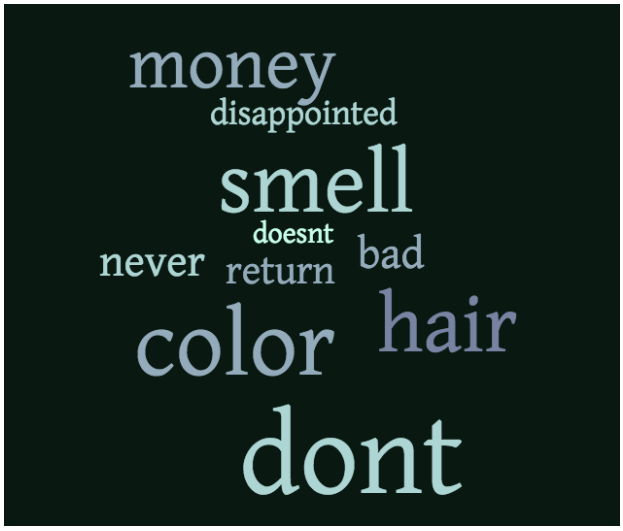
If user was NOT seen in training data:
$rating = \alpha + \Sigma_{w \varepsilon Text} Count(w).\theta_w$
If user was seen in training data:
$rating = \alpha_{User} + \Sigma_{w \varepsilon Text} Count(w).\theta_w$
α_{USER} = Avg rating given by user

2.2 PREDICTIVE TASK - 5 STAR

The rating frequency table on the first page indicates that there is a very high occurrence of 5 star ratings. So we are interested in predicting from the review when there will be a 5 star rating given to a product. To do this we first found the top positive weighted words for 5 star rating reviews and the top negatively weighted words in 1 star reviews. These do not include stop words, however, for 1 star reviews we thought that the unigram "dont" might be a good indication of a customer being disappointed with the product.

Top ten words in 5 star reviews

Top 10 words in 1 star reviews



At this point our model is linear regression plus n-grams. But because there seems to be a pattern of 5 star reviews in the dataset it might be worth it to test a model that specifically tries to predict 5 star ratings from text mining. We found the top unigrams in 5 star reviews, we could possibly just look for high frequencies of these words in the reviews and use that as an indication of a possible 5 star review. We see a potential problem with this, it is possible that we could have many 4 or 3 star reviews with the same positively weighted words that are found from 5 star reviews. To overcome this potential issue, our model will predict a 5 star rating if the positively weighted words are found in the review and if the review has no negatively weighted reviews.

In order to further improve predictions of 5 star rating we will look at another feature from the dataset, helpfulness.

Percent helpful based on rating

Star Rating	Percent- %
1	65.51
2	67.59
3	74.75
4	82.12
5	84.74

The table above indicates that 5 star reviews are more likely to be found helpful by other. We will use this in our model, we reason that a review is more likely to be rated at 5 when the review has a high helpful average. The new model is now linear regression plus sentiment analysis, and now also using using an additional feature, helpfulness.

2.3 PREDICTIVE TASK - SUMMARY

There is one more feature that can prove to be useful. The summary feature. We use the summary feature in the same way are the review feature, however using both the review and summary feature in the same model proved to be to completed of a model, this we will split the models, one use the summary feature and one using the review feature.

3 LITERATURE

3.1 SOURCE

The dataset was available online on a Stanford website that Professor Julian McAuley had created. The website [5] contains datasets of product reviews and metadata from Amazon spanning May 1996 to July 2014. This dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).[2] The data was clearly meant to be used for educational and informational purposes. The raw review data was too big for us to analyze so we decided to analyze the Beauty category of the raw review data.

3.2 TEXT MINING TODAY

One of the most interesting uses of text mining and sentiment analysis today is datasets used by the NSA. While in this class we got a taste of the potential predictions one can make using text mining, with allegations being made against the NSA having access to google search queries and social network graphs, there is a lot of possibilities. The key steps in text mining remain the same regardless of how advanced the models become. Tag and label data, filter out the noise and start making predictors. Filtering out noise is something that we failed to do well, which greatly affected our models ability to predict the rating correctly. This is something that we will explain in our results section at the end of the report.

3.3 HISTORY

In the past we have studied the beer reviews dataset as well as calculated predictions off a Video Game dataset. A modern dataset studied by entertainment corporations is movie and television show reviews to provide recommendations to users. They utilize socially regularized recommender systems as well as performing algorithms on metadata to find similarities. Characteristics such as time of creation, average ratings, and user dependent behaviors will determine the output of the algorithm. Netflix offered a prize of $1,000,000 to the contestant who could provide the best algorithm and it was awarded to the BellKor's Pragmatic Chaos algorithm. [3][4]

3.4 CONCLUSIONS

The conclusions drawn from existing work is similar to the results we have drawn in that metadata is fairly accurate at predicting information. However, as evident in both of the cases, any prediction algorithm can be optimized even further.

4.0 RESULTS - MSE

MSE - Using summary text

Model	MSE
Supervised learning	2.22
Supervised learning + Sentiment analysis	2.88
Supervised learning + Sentiment analysis + Predict 5 star rating	2.08

MSE - Using review text

Model	MSE
Supervised learning	2.94
Supervised learning + Sentiment analysis	3.24
Supervised learning + Sentiment analysis + Predict 5 star rating	2.85

4.1 RESULTS ANALYSIS

The data set used was made of 80,000 reviews, of that 60,000 reviews are used for training, while the other 20,000 were used for training. As mentioned before, and indicated by the tables above on this page. There were two different models,one using the review feature and the other used the summary feature.Both build on the baseline mentioned in section 2.0. First we look at the results for review feature model. The results were a little surprising. The use of sentiment analysis using popular unigrams based off the rating proved to overfit the model. Our reasoning is that because of the products being review (beauty products) using words like “love” to indicate a happy customers and using words like “hate” to indicate a disappointed customer was wishful thinking. Now looking back it seems obvious that it is entirely possible that a person might use the word “love” in a product that they were disappointed in, because they could be describe a similar product they bought that they were happy with. In other words the customer could have been comparing products in their review. For sentiment analysis to work, this problem would have to be solved first. As mentioned in the dataset analysis, 5 star ratings seems to be popular, and because of this the model which used sentiment analysis to ONLY predict 5 star rating worked well, better than the baseline.

Similar results were found using the summary feature based model, however they all worked better. The reason for this is the same reason that review based sentiment analysis failed to work well. In the summary customers did not try and compare products, this there wasn't a mixed use to both positive and negative sentiment unigrams. We spent the most time trying to better the baseline using sentiment analysis on

the review feature based model, which did not work as well as we had hoped.It might have been a better choice to try TF-IDF, this model might have done better job of taking into account the fact that customers compare good and bad products in their reviews.

END REPORT

[1]Inferring networks of substitutable and complementary products J. McAuley, R.Pandey, J. Leskovec *Knowledge Discovery and Data Mining*, 2015

[2]Image-based recommendations on styles and substitutes J. McAuley, C. Targett, J. Shi,A. van den Hengel *SIGIR*, 2015

[3]The Science Behind the Netflix Algorithms That Decide What You'll Watch Next T. Vanderbilt *Wired.com, 2013*

[4]Recommender Systems with Social Regularization H. Ma, D. Zhou, C. Liu, M.Lyu, I. King *Recommender Systems with Social Regularization*

[5]http://snap.stanford.edu/data/amazon/productGraph/