Timmy Chi
Mark Lizada
Ninad More
Abhishikth Nandam
Assignment 2
6/2/2015

Predictive Rating Analysis of Amazon User Review Data Using Sentiment/Opinion Mining

I. Dataset

A. We choose to conduct our experiment on the Amazon user review dataset. After completing homework 4, we were intrigued by what further intuitions a sentiment analysis of the reviews could provide per category. Also, since we worked with this same data for the first couple homework assignments and the kaggle task, our group has become readily familiar with it. The dataset contains user & item pairs, review text, category, time, and helpfulness ratings. After perusing the dataset, we noticed that the majority of reviews were under the clothing category and categories such as pets had less than 100 reviews. Since the categories were skewed in this manner, we could not just randomly pick the data to train on, since we might be missing out on some categories. Therefore when we picked our data, we made sure it was in one of the top 10 categories.

II. Predictive Task

A. Our intention with this assignment is to model a simple predictor, a predictor using sentiment analysis, and a predictor using sentiment analysis per category to predict review ratings. Then we compared the three predictors. Ideally the predictor using the sentiment analysis per category would be the best at predicting ratings, but after completing assignment 1 it seems as if the less complicated models work best, and adding minutiae on top of that only slightly increase the prediction accuracy.

B. Since the train.json file already provides us with the correct ratings, we trained our predictrors on a random 30% of the data, and test on another random 30%. This type of method was suggested by the professor in order to build a model with the lowest error rate.

C. The bag of words model utilizes words with positive or negative connotation to influence whether a person would rate an item high or low. Our implementation involved grabbing the most popular words from the given datas review section and then predicting the rating based on how frequently they were used per review. We calculated the top 1000 words from all 1,000,000 pieces of data overall, as well as per category for the top 10 categories.

III. Literature

A. Practical Text Mining with Perl (Wiley Series on Methods and Applications in Data Mining). We read a summary of this book and skimmed through the chapters to get an idea of how exactly to build our bag of words. Specifically, Bilisoly provides tips on retrieval techniques and which common/useless words to remove from

Timmy Chi
Mark Lizada
Ninad More
Abhishikth Nandam
Assignment 2
6/2/2015

the dataset. We also followed the methods mentioned in class to implement the bag-of-words model, which includes merging different inflections of words, stemming, removing capitalization/punctuation and stopwords, and unigrams. We decided against using bigrams because the top thousand most frequently used words would just display single words anyway; also it did not seem like it added much additional information.

B. Opinion Mining and Sentiment Analysis by Bo Pang and Lillian Lee was useful in providing examples of experiments conducted on Twitter and other news forums. It also provided contextual evidence on why sentiment analysis would even be useful in predictive tasks (ratings of reviews based on text, in our case).

IV. Results

A. We scanned 1,000,000 random amazon reviews, removed punctuation, stop words, and found the top 1000 word stems. The top 25 words with positive and negative influence are shown below.

B. Our basic predictor used just the user and item bias as well as the average rating (3.7) given across all reviews to predict a specific rating a user would give an item. We modeled a linear regressor and this had a relatively high MSE around (2.2).

$$rating(user, item) = \alpha + \beta user + \beta item$$

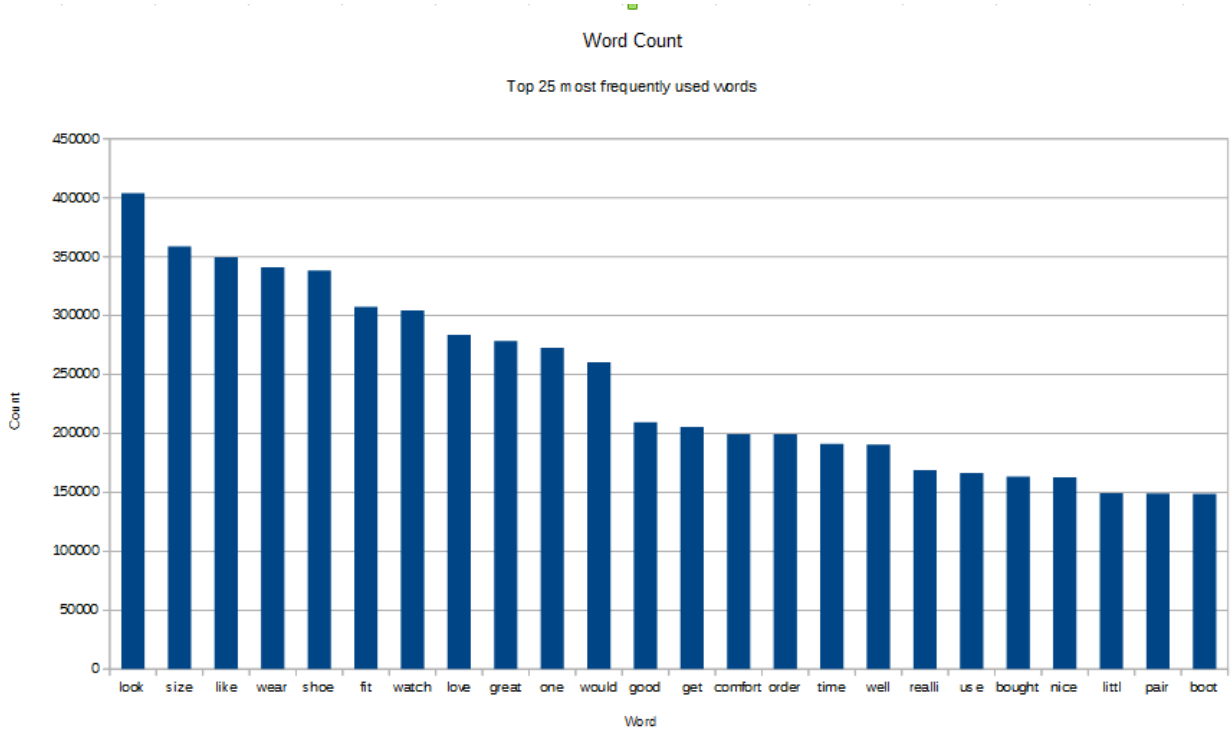


Figure 1: Shows the count of the most popular words across all categories

Timmy Chi
Mark Lizada
Ninad More
Abhishikth Nandam
Assignment 2
6/2/2015

Figure 2: Shows the rating influence of the most popular words in the baby-products category.

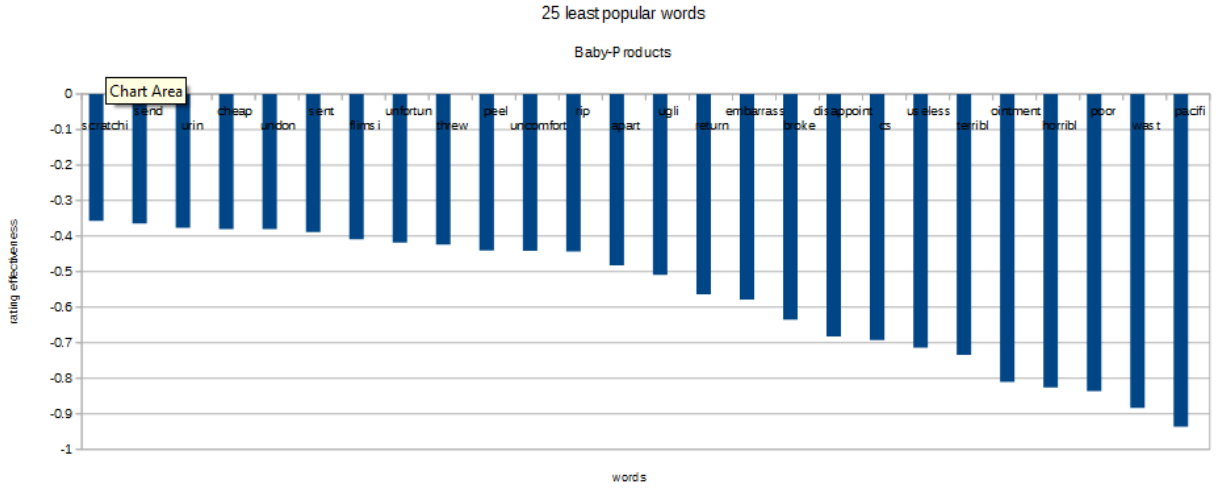


Figure 3: Shows the rating influence of the least popular words in the baby-products category.

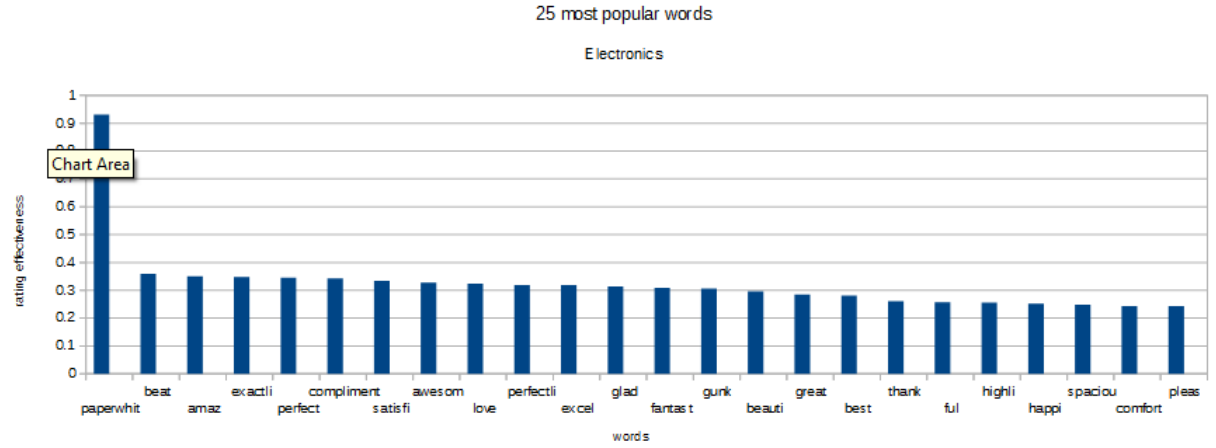


Figure 4: Shows the rating influence of the most popular words in the electronics category.

Timmy Chi
Mark Lizada
Ninad More
Abhishikth Nandam
Assignment 2
6/2/2015

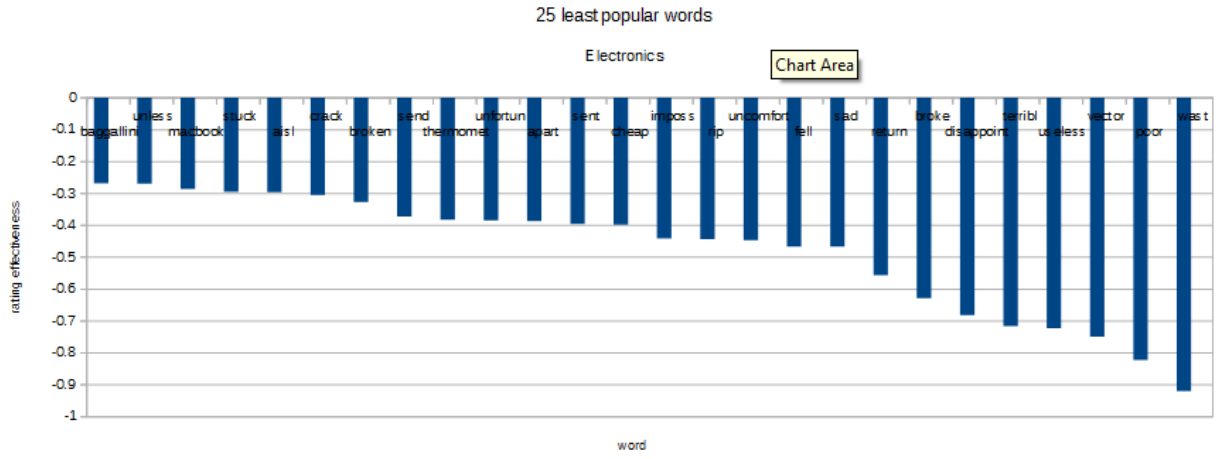


Figure 5: Shows the rating influence of the least popular words in the electronics category

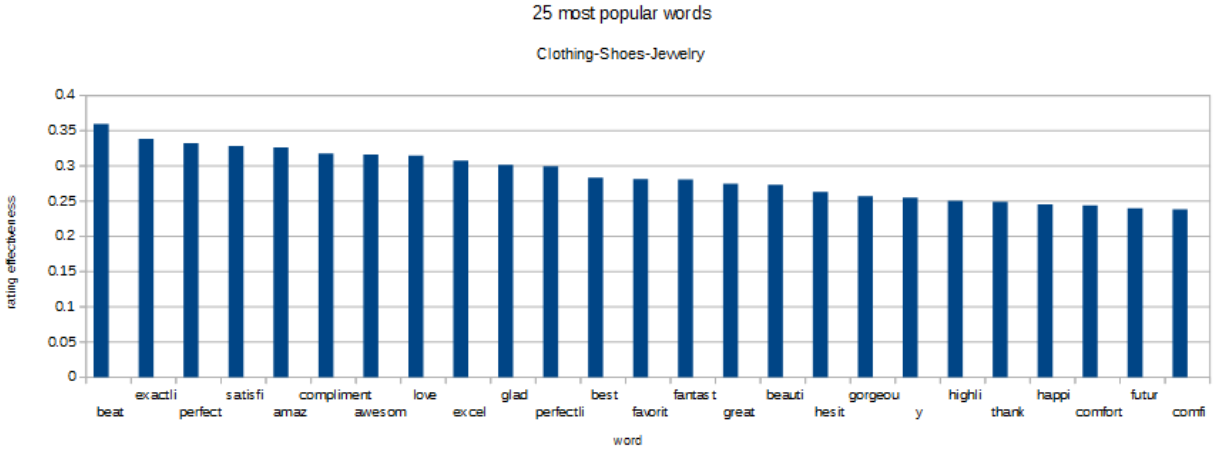


Figure 6: Shows the rating influence of the most popular words in the clothing category

Timmy Chi
Mark Lizada
Ninad More
Abhishikth Nandam
Assignment 2
6/2/2015

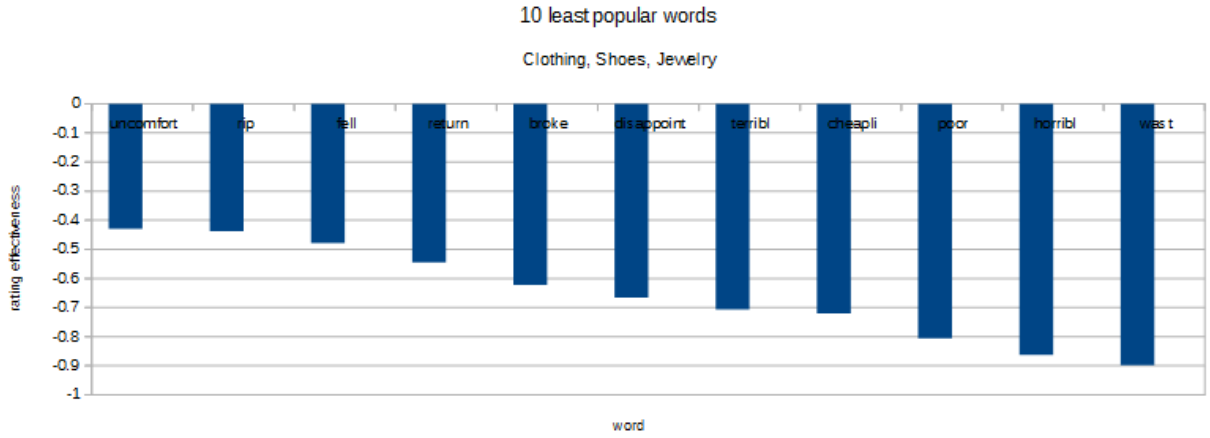
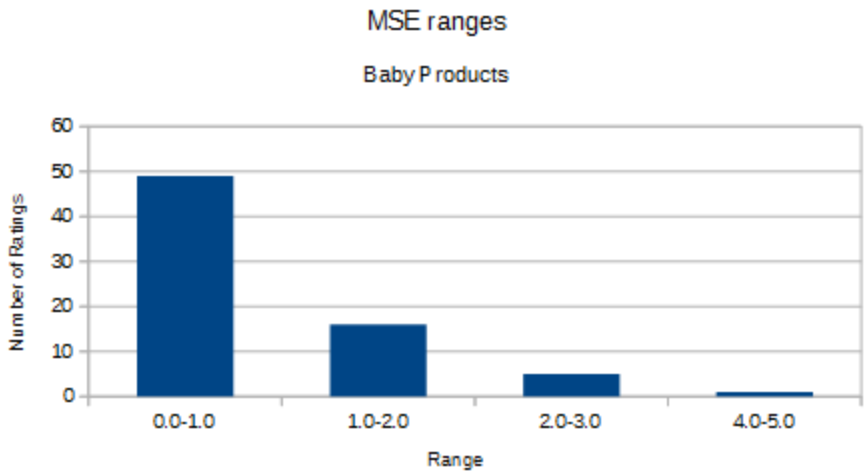
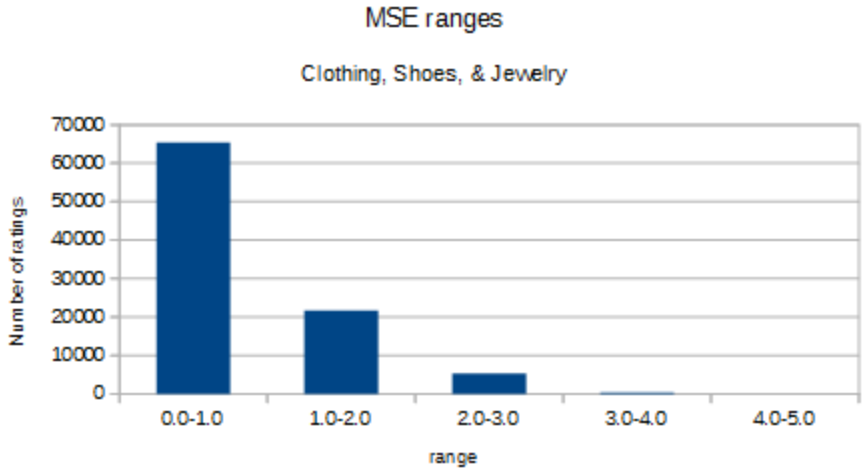


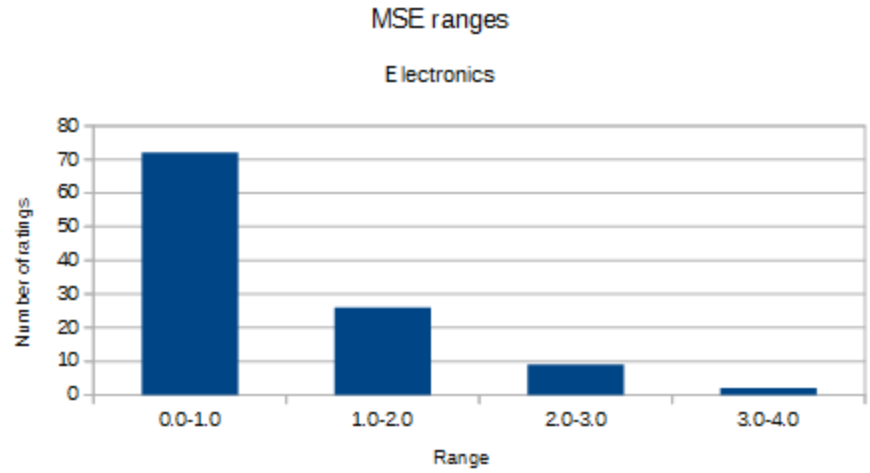
Figure 7: Shows the rating influence of the least popular words in the clothing category.

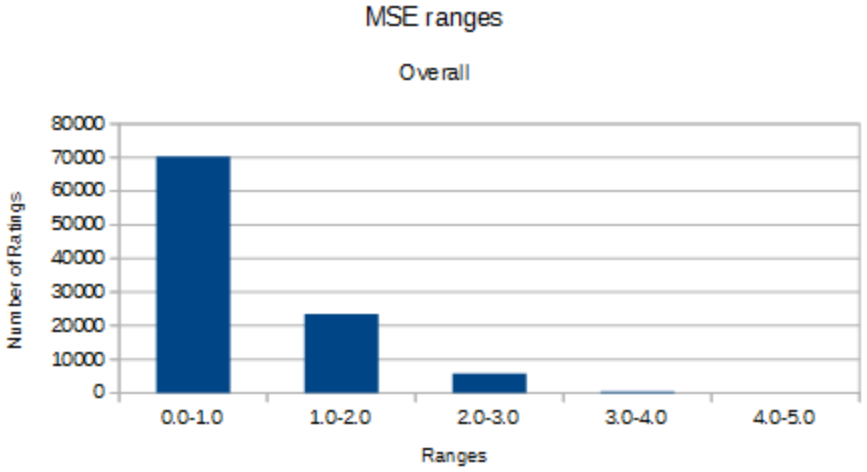
As displayed by the graphs above the most popular words, overall (“look”, “size”, “like”, “wear”, “shoe”, “fit”) all have to do with the clothing category, since this has the largest number of reviews. The 15 most popular categories(with at least 1,000 reviews) are [Arts, Crafts & Sewing; Automotive; Baby Products; Beauty; CDs &Vinyl; Cell Phones & Accessories; Clothing, Shoes & Jewelry; Electronics; Health & Personal Care; Home & Kitchen; Office Products; Patio, Lawn & Garden; Sports & Outdoors; Tools & Home Improvement; Toys & Games]. One key observation we made was that there are outliers when the top word is only shown in one review; the rating effectiveness will either be a really large positive number or really large negative number (i.e. 194612395740 - grovia, 89820352706 - evenflo, -389224791479 - soaker, -930110421406- chicco for the baby-product category).



Timmy Chi
Mark Lizada
Ninad More
Abhishikth Nandam
Assignment 2
6/2/2015



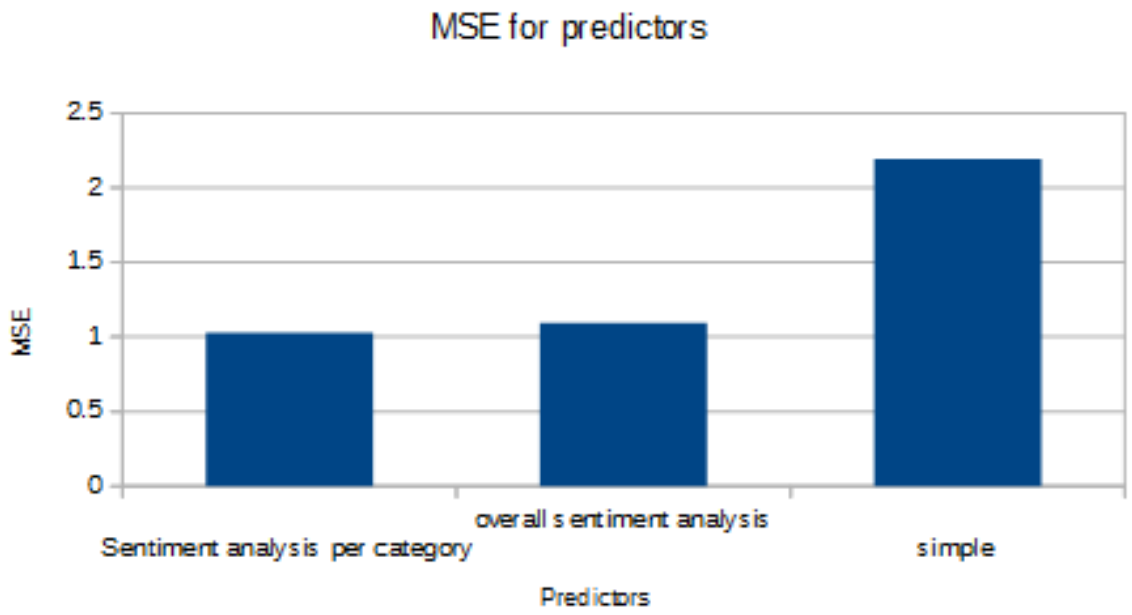




Timmy Chi
Mark Lizada
Ninad More
Abhishikth Nandam
Assignment 2
6/2/2015

MSE Figures:

We only provided the overall, baby-products, electronics, and clothings categories in order to display how varied the data is (the different MSEs produced per category and the different bag of words). Each category contributes a non-negligible amount to predicting the rating a user gives an item.



Significance:

The reason we choose to conduct this experiment is because rating prediction is incredibly useful in predicting what types of products a user will purchase (particularly useful for Netflix or Amazon/Ebay). Performing a sentiment analysis per category allows us to show data that shows a certain level of positive sentiment yields sales or translates to purchase intent. We want to continue to build upon sentiment analysis because the ability of Artificial Intelligence to make these kinds of analysis is not there yet due to the inability of AIs to understand indicators of intent.

Successes & Failures:

We noticed that top words that were only used in one review had either an extremely high positive or negative value, so we had to ensure to exclude those words from our data. Also, instead of using bigram and unigrams, we decided to just use unigrams. We didn't think that using both would decrease the MSE significantly since bigrams are not as frequent as unigrams

and would not show up much In the top 1000 words. We decided instead to dedicate our processing power to unigrams only.

Conclusion:

As shown by the above figures, our predictor is relatively accurate in predicting ratings based on reviews (as the majority of the predicted ratings only deviate from the actual ratings by less than 1). Training predictors by category, the MSE is 1.03, which is much better than the simple predictor (2.19) that we had initially tried in homework 3 (using average rating and user/item bias). In obtaining the MSE, we corrected ratings that were lower than 1.0 or higher than 5.0, since such ratings are not possible. Without doing this, the MSE is substantially higher due to error from some categories having their top thousand words represented by too few reviews. However, doing sentiment analysis overall had an MSE of 1.09. Our sentiment analysis using the most popular words cut the MSE in half and is a much better model than we had initially hypothesized.