CS505 Final Project - Amazon Reviews Analysis - Mid-Progress Report

Eugene Kolodenker Jose Lemus

ABSTRACT

We are analyzing the quality of product reviews on Amazon.com using the Amazon Reviews dataset. Our primary interest is in observing year-to-year degradation. To achieve this we are analyzing each review, and creating a set of features that describe subjectivity, polarity, and helpfulness of each review. By using clustering techniques on this dataset we wish to observe the data to cluster into categories that are some mix of fake reviews, biased reviews, and good reviews. We believe that we can observe a yearly growth in the size of the 'fake', and 'biased' reviews compared to the 'good' reviews.

INTRODUCTION

The ratings and reviews given to a product have major consequences on how well the product will sell on online websites. Making purchase decisions based on reviews has become a part of the online shopping experience. We believe that reviews are a great way to give customers more information which will help them make better purchasing decisions. However, because of the obvious link between good reviews and product sales, reviews are subject to being manipulated. Online retailers such as Amazon have much reason to care about the quality of reviews because if customers lose faith in the reviews, these customers may turn elsewhere to do their online shopping.

In order to assess the quality of reviews found on Amazon we are using the Amazon Reviews Dataset. We will describe the dataset more in a later section. Our approach to this project was to first analyze the data on a year to year trend based on explicit sentiment by users for product reviews. We presented some preliminary results in our previous progress report that showed that reviews rated as useful have net grown year to year exponentially, but the ratio of useful reviews to total reviews has decreased year to year linearly. In order to take a deeper look at the trend, we are going to apply more sophisticated data analysis methods like sentiment analysis which should help us categorize reviews.

TECHNIQUE

We are analyzing the data with exploratory data mining. Through the technique of clustering we hope to find interesting groups of reviews. Our hypothesis of the data is that we can observe three natural high level clusters of the reviews: 'fake', 'biased', and 'good'.

Analysis model

We define a 'fake' review to be one that is written by the product's vendor (to increase sales), or a competitor (to decrease sales). We define a 'biased' review to be one in which the reviewer has ulterior motives besides giving a fair review, this is typically the case when the review is sponsored. A 'good' review is one that is unbiased, not fake, and helpful. Because, we have discrete clusters that we wish to observe, we are leveraging *K-Means* clustering which will group the reviews into discrete groups. Ideally, we wish to observe three natural clusters, however, we are using the *silhouette coefficient* to validate our clustering internally. Additionally, we are using the silhouette coefficient to determine the optimal number of clusters.

Features

In order to perform clustering, we must define some features to cluster by. Because we are interested in separating out the reviews based on essentially their 'sentiment', a powerful tool to perform this is, *sentiment analysis.* By performing sentiment analysis of the review text we obtain two features: *polarity*, and *subjectivity.* Polarity determines if a review is 'positive', or 'negative'. Our hypothesis is that overly positive and overly negative reviews are either biased or fake. Subjectivity determines if a review is 'subjective', i.e., opinion-based, or 'objective', i.e. fact-based. Our hypothesis is that overly subjective, and overly objective reviews are either biased or fake (i.e., biased praise/insult).

By leveraging the full dataset, we can also do deep analysis of the reviewers themselves. We believe that the history of a reviewer also adds features to the reviews themselves, i.e., a review written by a a biased reviewer who rates everything 5 stars is itself biased. We add three more features through deep parsing of the reviewers: *positivity of reviewer*, and *similarity of reviewer's reviews.* An overly positive or negative reviewer is potentially biased, or fake. A reviewer whose reviews are too similar, is potentially fake as well. To measure the similarity of reviews written by a reviewer we will use cosine similarity.

Finally, another feature that we will incorporate is the reviews *helpfulness* (found in the dataset), as determined by user votes.

DATASETS AND EXPERIMENTS

The Amazon Reviews Dataset can be found here (http://jmcauley.ucsd.edu/data/amazon/). This dataset contains product reviews from May 1996 - July 2014. It comes as a gzipped json file where each review is a json object. The entire file is thus a list of json object.

```
{
 "reviewerID": "A2SUAM1J3GNN3B",
 "asin": "0000013714",
 "reviewerName": "J. McDonald",
 "helpful": [2, 3],
 "reviewText": "I bought this for my husband who
  plays the piano.  He is having a wonderful time
  playing these old hymns. The music  is at times
  hard to read because we think the book was
  published for singing from more than playing
  from. Great purchase though!",
 "overall": 5.0,
 "summary": "Heavenly Highway Hymns",
 "unixReviewTime": 1252800000,
 "reviewTime": "09 13, 2009"
}
```

Figure 1. Example JSON object from the Amazon Reviews Data Set

Dataset

Each review contains information about the product and the reviewer. Variables that will be of interest to us are the “overall”, “helpful” and “reviewText” variables. The dataset can be downloaded in different pre-processed ways. We have chosen to work with a subset of the dataset. Our dataset will contain products from the following categories: Electronics (7,834,166 reviews), Beauty (2,026,943 reviews), Home and Kitchen (4,260,181 reviews). Using these three datasets we can investigate whether there are more fake, or biased reviews in the Electronics category than in the Beauty or Home and Kitchen categories. We believe that we will find evidence to support this theory because, the Electronics category is much more popular than the other two categories and therefore there are more opportunities to write fake reviews.

In order to get the most accurate results we have to first filter out some outliers. The first kind of filtering that we will do is removing products that have less than 5 reviews (this number may change as we analyze the data more). We will also filter out reviews that have a small amount of words in their review text. For example, if a review only consists of the sentence “Its good” or “I liked the product” etc, we will remove this review from our dataset. We may also decide to apply further filters as we do more data analysis.

Experiments

We are in the process of running multiple experiments on the data. First, for each year we are going to cluster the reviews into 3 categories based on a sentiment analysis of the reviewText. To do this, we are going to vectorize the text and find similar reviews based on a similiarity metric. During this process, we will apply LSA in order to reduce the dimensionality of the data and obtain better results. After we do this for each year, we are going to plot the year to year trend which will show whether the quality of reviews have gone up or down. One metric that we will be interested in is the year to year trend of the proportion of sponsored reviews to total reviews. We will be able to get this metrics from the clusters that we compute.

RESULTS AND DISCUSSION

Working with this large dataset is challenging. We initially began research with the full 30GB dump of all reviews. However, processing the data onto our computers proved to be difficult. We are now working with a smaller subset of categories. We wish to go back to the full set, after some thought into better memory management, and perhaps dimensionality reduction. For each review, we’ve begun our feature extraction. So far we have, polarity, subjectivity, helpfulness, and positivity of reviewer. Our next step is to perform clustering. So far, we observe that users tend to rate reviews more ‘helpful’, if their subjectivity is lower, and if their polarity is lower. Users seem to find more negative reviews as helpful. Additionally, we first helpfulness of reviews as going up in our initial research year-to-year; however, with these new categories we observe it to actually be about steady.

CONCLUSION

Although we have not yet done any deep clustering analysis of the dataset we are confident in the approach that we have taken with this project so far. We plan on obtaining strong results once we start applying the techniques explained above that will show whether or not the quality of reviews has decreased.