Status Versus Utility Classification and Prediction Algorithm for Amazon Products

Linda Wogulis, Yijun Zhang

July 12, 2015

1 Background

Text Mining The first concept and attempts at text mining originated in the 1980s, but it was a large task with the limited computational power available at the time. The past decade has seen a large rise in interest in and research about data mining and sentiment analysis. Bo Pang [3] explains the reasoning behind getting involved in sentiment analysis, which is the widespread demand to know other people's opinions. This can manifest as a desire to sell products and appeal to customers, or it can be as impactful as influencing an election. Pang's research is representative of much of what people are doing related to our project. The central idea is to make accurate predictions of people's opinions based on information from text, whether that be reviews, search engine entries, or one of many other sources of data.

Natural Language Processing In order for these predictive tasks to be accomplished, researchers must process and manipulate the given data to extract the desired information. There are many techniques for doing so, such as removal of stop words, using n-grams, and stemming. The latter was introduced as early as 1968, with Julie Beth Lovins' *Development of a Stemming Algorithm* [2], where she describes the process by which all conjugations and forms of the same word root are reduced to a single representation. Research in the field of text mining requires the use of at least some of these techniques, and many others. These processes are all options to be explored in our work.

2 Project Description

Objective The main goal for our project is to be able to discern the difference between a product that is being rated or reviewed based on its functionality versus one that is judged for its status. We wish to determine which features of a written review indicate its utility or status.

Preparation In the process of solving the posed problem, we must learn about techniques for modeling textual data. Some of the methods, such as stemming and using n-grams, are necessary in order to organize the data in meaningful ways. The most important method to learn, however, is that of linear regression. We will be using linear regression to fit a model to the given information. In order to accomplish this, we must learn about and practice solving linear regression equations by extracting information about textual connections with user ratings. We additionally have to decide on a universal definition for what a status or utility good is, and we will need to label some of the data manually in order to train and test the model.

Modeling As mentioned above, we will be using linear regression to model the connection between user reviews and product categorization as status or utility goods. Part of the question is, what elements of a review are associated with this categorization? It could be largely dependent on keywords ('effective', 'pretty', etc.) or pairs of keywords ('not useful', 'less appealing', etc.). It could also depend on other aspects of the text, such as length of the review or percentage of punctuation used. In order to determine which factors influence a categorization of functionality or status, we will test multiple models and observe the given results.

Analysis We will collect information on the results of our models, and we will need to understand what the feedback means. In order to do this, we must analyze the results with statistical methods such as obtaining the Mean Squared Error. We will compare the different models, and in identifying the most accurate models, we can determine which factors are the most significant in predicting a product's appeal.

3 Challenges

At this stage, there are four expected challenges that we identify:

Manually labeling data We will be using Amazon product data provided by Prof McAuley. The Amazon product dataset contains product reviews and metadata from Amazon, including 143.7 million reviews spanning May 1996 - July 2014. [1] Since we are interested in discerning the difference between a product being sold for aesthetic or utility, we need to manually label some data. Only one or two categories will be chosen for us to train and test our model. Still, this can be a very exhausting task.

Defining the problem in the right way Whether a product is sold for aesthetic or utility is more of a subjective question. Therefore, we need to define aesthetic and utility to classify products. In addition, we also need to determine a proper representation of the products, which can be binary, or on a scale of one to ten, where 10 means the product is sold purely for utility.

Scaling models to large datasets In order to develop a scalable model, we may need to learn techniques to handle large datasets.

Statistical significance To determine whether our finding has statistical significance, we need to learn to perform statistical hypothesis testing on our model.

4 Tools/Knowledge to learn

Linear predictor We decide to use linear predictor functions as an approach to analyze a training set of data to develop a hypothetical relationship between features of products and parameters to predict real-valued outputs, ie. people's tendency of buying the product for status or utility. The functions appear in classification, the problem of identifying to which category an item belongs, and regression. We may need to learn both models to learn exactly how are we going to apply the linear predictors.

Bag-of-words Bag-of-words is a model representation in processing natural language by treating text as merely a collection of words, neglecting grammar and punctuations. By looking at customer reviews as a bag of words, we seek to identify keywords that are indicative of product appearance or utility based on frequency, and use them as features for regression and classification.

Sentiment analysis Sentiment analysis aims to identify subjective opinions based on text. We want to learn about sentiment analysis to associate keywords with product properties of aesthetic or utility.

5 Timeline

The project period is 8 weeks. Lists of tasks per stage:

- Week 1 & 2: Preparation
  - Learn about basic models for natural language
  - Learn about regression using text
  - Reproduce standard techniques to model data
- Week 3 & 4: Data processing and labeling
  - Develop proper definitions of status/utility for our data
  - Manually label the data using the definitions above
- Week 5 & 6: Modeling

  - Run experiments to try out different models and features
  - Identify those that are effective at predicting the labeled data
- Week 7 & 8: Analysis
  - High level analysis of our findings
  - Final report draft

References

[1] Amazon product data julian mcauley. `http://jmcauley.ucsd.edu/data/amazon/`. Accessed: 2015-07-12.

[2] Julie B Lovins. *Development of a stemming algorithm*. MIT Information Processing Group, Electronic Systems Laboratory, 1968.

[3] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.