Fine Food? We want to know!

**Jinhan Zhang**

**A53211930**

**jiz530@eng.ucsd.edu**

**Zhongjian Zhu**

**A53210727**

**zhz362@eng.ucsd.edu**

**Siqi Qin**

**A53203046**

**s7qin@eng.ucsd.edu**

**Abstract**

Two-dimensional[[1]](#footnote-2) arrays of bi-component structures made of cobalt and permalloy elliptical dots with thickness of 25 nm, length 1 m and width of 225 nm, have been prepared by a self-aligned shadow deposition technique. Brillouin light scattering has been exploited to study the frequency dependence of thermally excited magnetic eigenmodes on the intensity of the external magnetic field, applied along the easy axis of the elements.

CCS Concepts: • **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks →** Network reliability

General Terms: Design, Algorithms, Performance

Additional Key Words and Phrases: ACM proceedings, Word, text tagging

1 Introduction

Dining counts most for people, of which taste comes first. With the development of information technology, an increasing number of people purchase delicious food via online shopping platforms for convenience. But one problem with these platforms is that we can’t foretaste and directly perceived through the senses. Thus, the experiences from previous users become an important reference for us. Amazon, one of the most popular online shopping market in the world, provided a huge number of user reviews on all kinds of products. With these reviews, especially the review text and star rating, we can analysis the user’s sentiment towards a certain product and help us to make our own decision.

In this project, we use the dataset downloaded from Amazon Fine Food to predict the star rating based on the review information given by the users. We first do an exploratory analysis of the data and get some inspiration on the feature selection. Then, predictive task is determined and a baseline is set as a reference. We try seven different models and evaluate their performance on the dataset based on MSE. Finally, we find some recent literatures related to our project and put forward some future work. Based on the model given in this project, we can realize better prediction of the rating.

2 Dataset

2.1 Description

We use the Amazon Fine Foods Reviews dataset provided by Julian McAuley on the Stanford snap datasets[1]. This dataset consists of reviews of fine foods from Amazon within 1999 to 2012. There are 568462 reviews, in which 256059 users give their ratings towards 74258 products. And 568454 of the reviews have review text. In the review, product and user information, ratings and a plaintext review are given, which can be used for our prediction. We describe the features of each review in Table.1.

Table 1. Feature Description

|  |  |
| --- | --- |
| Feature | Description |
| productID | asin give for each product |
| userId  profileName  helpfulness  score  time  summary  text | ID of the user  name of the user  fraction of users who found the review useful  rating of the product  time of the review (unix time)  review summary  text of the review |

To select the features that best describe the prediction task we care about, we are interested in examining the structure of these dataset. So, we perform some exploratory analysis to better understanding these reviews. The analysis is based on the entire dataset.

2.2 Exploratory analysis

***2.2.1 Rating Distribution***

Rating is what we want to predict in this project, so we first take insight into the properties of the rating of the dataset. The rating distribution is shown as [Fig. 1](#fig1).

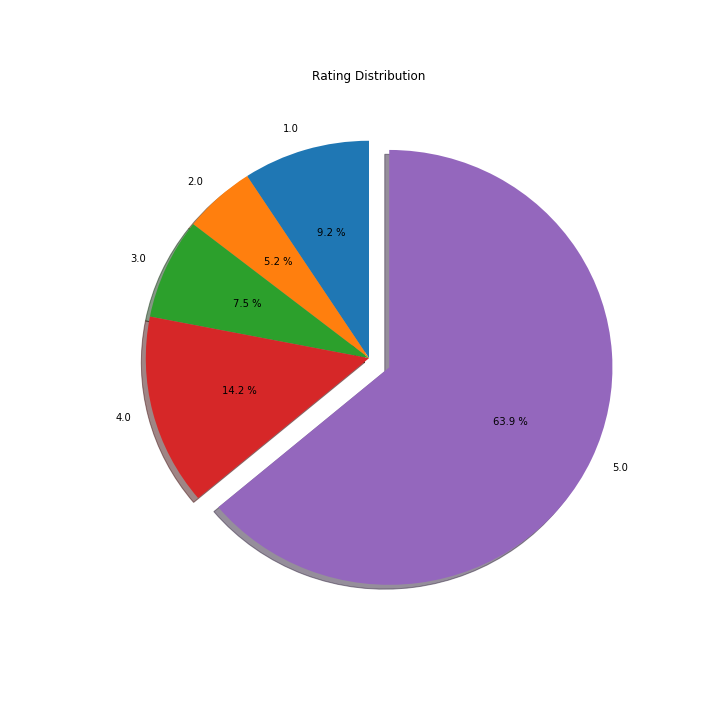


Fig. 1. Rating Prediction of the Dataset

As shown in [Fig. 1](#fig1), the 5-star reviews account for 63.9 % of all reviews. It seems that people are apt to give positive ratings nowadays.

***2.2.2 Length of review***

Longer review texts always provide more information compared to the shorter ones. So, we next analyze the relationship between the length of the review and the rating score. We count the length of review texts in rating score 1 to 5. Results are shown in [Fig. 2](#fig1).

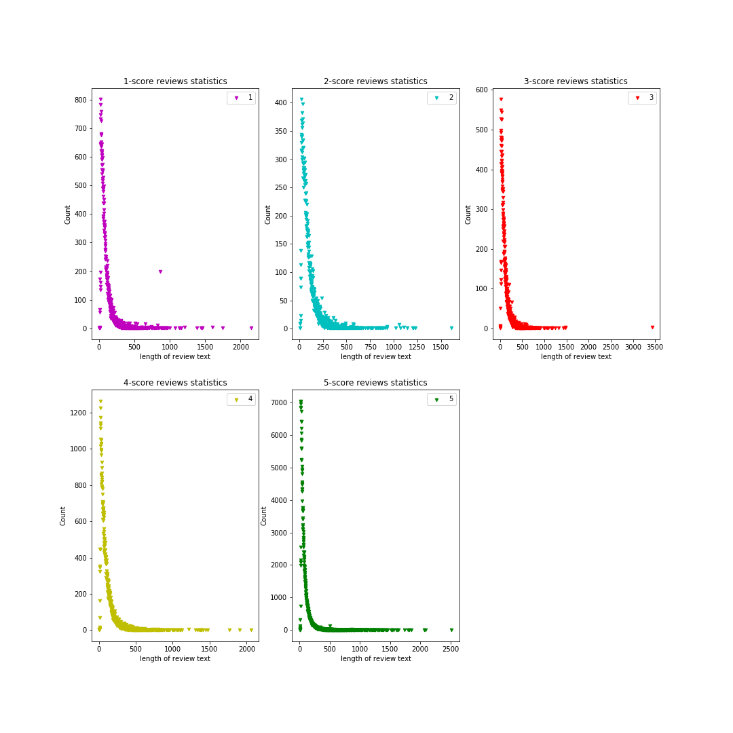


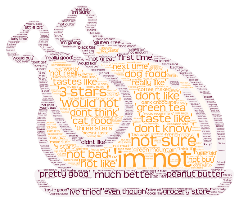
Fig. 2. Length of Review in Different Scores

In different scores, the length distribution all have the same trend. But we may find that the review text in 5-star rating is always shorted compared with the others.

***2.2.3 N-grams***

In the features description part, we find that the most informative feature is the review text. So here we analyze the reviews in a text mining viewpoint. We count the most popular unigram, bigram and trigram in 1-star rating, 3-star rating and 5-star rating. The results are as shown in [Fig.](#fig1) 3.





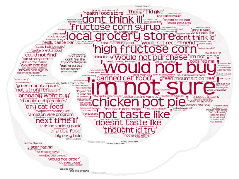


Fig. 3. Most Popular N-gram in Rating 1, 3, 5(add abcdefg)

From the most popular N-gram, we found that the bigram and the trigram are very representative for the score level they stand for. But the unigram is not varying very much among score 1, 3 and 5.

Considering that unigram is the most frequently used feature in text mining, we try to classify the unigrams based on the word class.

***2.2.4 Word Classification***

A word can play different role in a sentence. With natural language processing method, we can classify the words in the text into coordinating conjunction (CC), cardinal number (CD), determiner (DT), preposition (IN), adjective (JJ), noun (NN), Pronoun (PR), adverb (RB) and verb (VB). This classification is based on the NLTK library in python. The word class distribution of the dataset is shown as [Fig.](#fig1) 4.

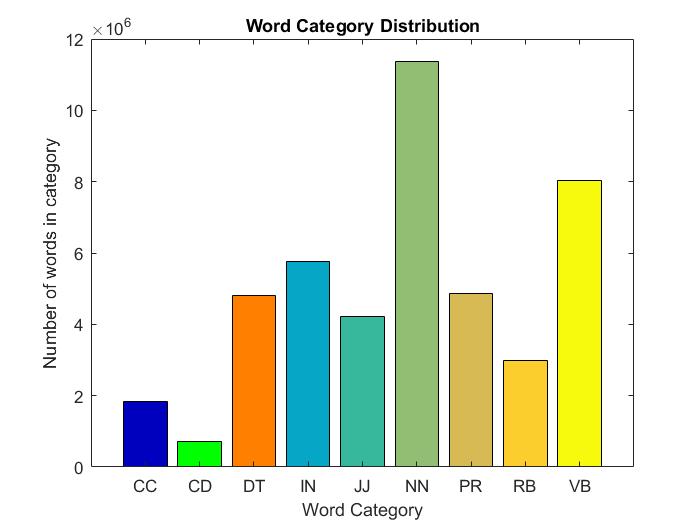


Fig. 4. Word Category Distribution

We can find here that some categories that contain little information about a user’s attitude, such as the preposition, are in huge amount. Maybe this is the reason that unigram can’t represent a certain star level well. So, in our project, we decide to try the adjective, noun and verb as feature, which we think contains more information. We finally count the most popular adjective, noun and verb in our dataset to help us better select the words we can use as features. The results are as shown in [Fig.](#fig1) 5.

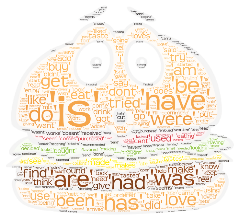
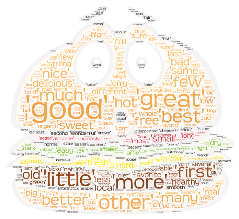


Fig. 5. Most Popular Adjective, Noun and Verb (add abcdefg)

Journal Name:Trans. on Graph.

Volume:9

Article:4

Year:2010

Month:March

Article Seq:39

DOI:0000001.0000001 2

1. Zhongjian Zhu, Jinhan Zhang, Siqi Qin [↑](#footnote-ref-2)