Mitigating Bias In Online Reviews

Anmol Singhal
IIIT Delhi
anmol17332@iiitd.ac.in

Tejas Oberoi
IIIT Delhi
tejas17367@iiitd.ac.in

Yashvardhan Singh
IIIT Delhi
yashvardhan17123@iiitd.ac.in

Abstract

A great deal of commercial activity that happens online on portals like Amazon depends on the reviews garnered by the products. The extent to which a product is endorsed significantly affects user purchasing behaviour online. The economic importance of reviews is massive for merchants within the platforms to ensure a good customer experience. However, the need to accumulate positive reviews for a product increases the chance of bias present in the reviews. Therefore, e-commerce portals must minimize the bias present, as it lies in the best interest of both their customers and merchants. In this work, we use statistical Bayesian models to determine the extent of bias present in user reviews on Amazon.

1. Introduction

The advent of e-commerce portals has transformed the world economy and has significantly shifted the usual market flow towards a virtual and technology driven environment. The essence of this noteworthy shift lies in the establishment of trust between the user and the merchants selling their goods through these portals. The idea of genuine online reviews for products sold on websites like Amazon.com is imperative to this notion of trust.

However, studies have shown that reviews provided by users that frequently shop online are susceptible to bias. This may be due to influence or personal inclination towards certain aspects of the product in question. Also, these reviews might not truly represent the underlying goodness and quality of the product. In other cases, ratings given to a particular product might be flawed and unable to capture the sentiment behind the review given to it.

Such discrepancy in reviews and ratings is of massive concern to these e-commerce portals. It also drastically affects user behaviour and is thus of utmost importance to merchants whose goods have to compete with other commodities sold online.

The goal of this paper is to determine the extent of bias present in online reviews. We use a rich dataset of Amazon reviews and draw upon Bayesian inference from random sampling to come up with an estimate of the probability distribution of bias of each user and goodness of each product reviewed. We use Markov Chain Monte Carlo (MCMC) methods and concretely the No-U-turn sampler, an extension of Hamiltonian Monte Carlo, to sample points estimating our target distribution.

We draw inspiration from Walker et. al and build upon their model by doing a sentiment analysis and calculating a sentiment score for each review. The results obtained show significant improvement over the baseline used.

2. Related Work

The primary motivation for our work is drawn from the work of L. Muchnik et al., who emphasises the influence of popular opinion in decision making. As part of their study, they analyse users who contribute news articles to a large news aggregation website. Their analysis reveals that users tend to upvote a comment that has received positive feedback more often than the ones which are new or have not received any positive responses yet, thus indicating positive herding and bias in user behaviour.

The first attempt to model bias mitigation, which is an unsupervised learning task, was made by Ge. et al, who used MCMC techniques to generate a probability distribution to detect bias in the papers accepted in NIPS conference. They use a Platt-Burges model to carry out their task, which is integral to our work as well.

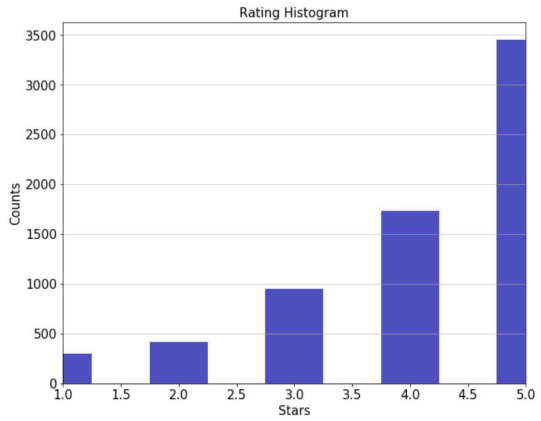
Walker et. al carry out the task of bias mitigation on the Amazon reviews dataset using the model proposed by Ge. et al. They train their models by assigning certain priors to the distribution of parameters and obtain the posterior by sampling from Amazon dataset. Our work is an extension of Walker et. al.

3. Methodology

3.1. Dataset

We use an Amazon product dataset compiled by Julian McAuley at UCSD. The data contains product metadata and review information for about 9 million reviews in 22 product categories, with dates spanning May 1996 to July 2014.

For each review in the dataset, we extract the user and product IDs, the rating, the review text and the helpfulness scores assigned by the user. We restrict our dataset to the Amazon Instant Video category, and consider only the users who have reviewed more than 10 products and products which have been reviewed by more than 10 users, to ensure sufficient richness in data. Our final dataset consists of about 7000 reviews, with approximately 500 users and 800 products. We use a 80-20 split to divide our data into a training and test set. Figure 1 shows the no. of reviews vs. the rating given.



For sentiment analysis, we retain the case of the textual reviews and do not remove punctuation marks. The sentiment scores are obtained for each sentence and are scaled from 1-5 to resemble product ratings.

3.2. Terminology

We formalise our problem by introducing some key notation that we use throughout this paper: index products with p and users with u. r_{pu} is the score reviewer u gave to paper p, g_p is the unobserved objective quality ("goodness") of product p, and b_u is the unobserved bias of reviewer u. We assume reviewer bias is constant for a given reviewer u, and the goodness is constant for a product p.

$$r_{pu} = g_p + b_u + \epsilon_{pu}$$
$$\epsilon_{pu} \sim \mathcal{N}(0, \sigma)$$

Platt and Burges solve the problem mentioned above by minimizing the following regularized least squares problem:

$$L = \frac{1}{2}\Sigma_p\Sigma_{u \in R_p}(r_{pu} - b_u - g_p)^2 + \frac{1}{2}\lambda\Sigma_u b_u^2$$

where λ is the regularization parameter, and R_u is the set of reviews for product p. Ge et al. propose a Bayesian reformulation of the Platt-Burges model which we consider as our baseline.

3.3. Baseline

We introduce the Bayesian Platt Burges used by Walker et. al by placing the following priors on the distributions of bias and goodness parameters. c and d are introduced as hyper-parameters, which also follow a Gaussian Distribution.

A_{pu} denotes the adjacency matrix of the data. It is defined by-

$$A_{pu} = \begin{cases} 1 & u \text{ reviewed } p \\ 0 & \text{otherwise} \end{cases}$$

The final rating distribution can be represented as-

$$r_{pu} \sim \mathcal{N}(g_p + b_u, 1/(dA_{pu}))$$
$$d \sim \mathcal{N}(a_d, b_d)$$

3.4. Model 1

The first model that we implement is by changing the priors of the goodness of the product according to the mean of product ratings r_{pu} on our dataset.

$$g_p \sim N(p_{mean}, p_{var})$$
$$r_{pu} \sim \mathcal{N}(g_p + b_u, 1/(dA_{pu}))$$
$$d \sim \mathcal{N}(a_d, b_d)$$

3.5. Model 2

The second model incorporates sentiment analysis using Vader sentiment analysis model. We use a pre-trained model of Vader and generate predict a polarity score ss_{pu} on each review. We further calculate the inverse review quality and represent it as I_{pu1}. We represent it in the prior distribution of the rating of the product in the following manner-

$$I_{pu1} = |R_{pu} - SS_{pu}|$$
$$r_{pu} = N(g_p + b_u \ , \ I_{pu1}/d)$$

Model	RMSE
Baseline	0.99873576695
Model 1	0.99324790267
Model 2	0.99857661991
Model 3	0.992470832753

Table 1. RMSE scores obtained on the Test Set

3.6. Model 3

The third model incorporates sentiment analysis using TextBlob model. We use a pre-trained model for TextBlob and again generate predict a polarity score on each review ss_{pu} .We further calculate the inverse review quality and represent it as I_{pu1}. We represent it in the prior distribution of the product rating in the following manner-

$$I_{pu2} = |R_{pu} - SS_{pu}|$$
$$r_{pu} \sim N(g_p + b_u, I_{pu2}/d)$$

After getting the priors from each model, the crux of our problem is estimating the parameters in our model, mainly the goodness g_p and bias b_u. We use Markov chain Monte Carlo (MCMC) to perform inference. Specifically, we work with STAN,which helps us for performing Bayesian inference. STAN is based on the No-U-Turn Sampler (NUTS), which provides several advantages over Hamiltonian Monte Carlo (HMC) and improves performance. This is an unsupervised learning task.

The posterior distribution obtained from STAN can be modeled by-

$$P(g, b|R_{train}) = \frac{P(R_{train}|g, b)P(g)P(b)}{\int P(g)P(b)P(R_{train}|g, b)dgdb}$$

4. Evaluation and Analysis

The key metric we use for the evaluation of our problem is by using RMSE, which we define using g_p and b_u as follows-

$$\text{RMSE} = \sqrt{\frac{1}{|R_{test}|} \sum_{r_{ij} \in R_{test}} (r_{pu} - \overline{g_p} - \overline{b_u})^2}$$

The average goodness and average bias is representative of the fact that multiple users are allowed to review a single product. We take the average over all such cases. If the assumptions of our model are correct, the value $b_u + g_p$ should be a close estimate for the product-reviewer pairing b U p in our test set.

The RMSE values obtained for our models are given in Table 1. The probability distributions obtained for each model for a particular product and user are shown in Figure 2.

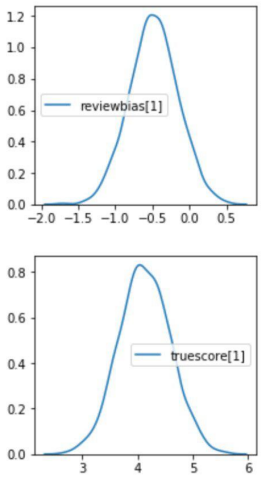


Figure 2- Baseline Probability Distribution for Product ID 1 and User ID 1

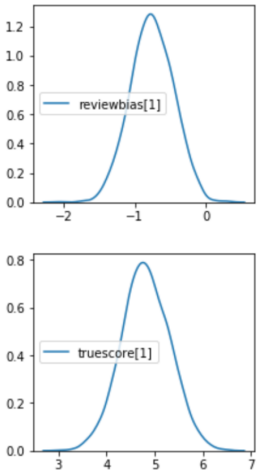


Figure 3-Model 1 Probability Distribution for Product ID 1 and User ID 1

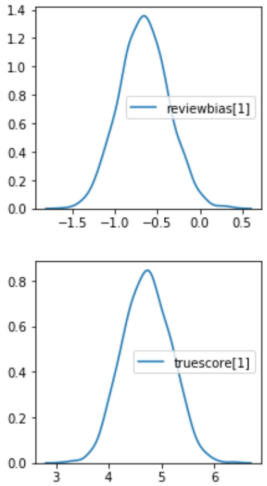


Figure 4-Model 2 Probability Distribution for Product ID 1 and User ID 1

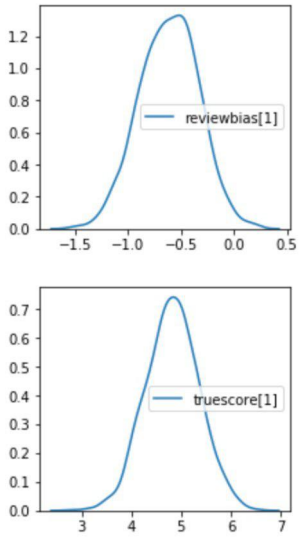


Figure 5- Model 3 Probability Distribution for Product ID 1 and User ID 1

5. Results

The results show that although the RMSE scores have a similar magnitude on various models trained, the error is the least when we incorporate sentiment analysis using TextBlob. We argue that this percentage of RMSE change is significant in terms of the probability distribution of reviews. Sentiment analysis using Vader, however, does not fetch the desired results.

6. Conclusion and Future Work

The paper aims to perform a Bayesian reformulation of reviewer calibration techniques introduced in Platt and Burges' NIPS paper. Moreover, we have attempted to apply extensions to the existing models and observed their results on a novel dataset. Gathered results clearly indicate that there are groups composed of reviewers whose ratings systematically deviate from the ones given by their peers. Bayesian models can be used to detect this latent bias.

A plausible extension of our work could be to perform behavioural research and examine if adjusted reviews are more helpful for users of e-commerce platforms. Platforms like Amazon could incorporate such models detecting reviewer bias and implement behavioural nudges in order to encourage their users to provide unbiased ratings. Alternatively, they could consider computing a weighted average of the ratings; assigning a higher weight to reviewers with a lower bias to get a better estimate of the goodness of the products.

7. References

1. C. Walker, B. Buttinger (2018)- Towards Mitigating Bias in Online Reviews (Stanford)

2. H. Gee, M. Welling, Z.Ghahramani (2017)- A Bayesian Model for Calibrating Reviewer Scores (NIPS)

3. Platt, B. and Burges, C. (2012). Regularized least squares to remove reviewer bias (NIPS)

4. Dai, W., G. Jin, J. Lee, M. Luca. 2014. Optimal Aggregation of Consumer Ratings: an Application to Yelp.com. NBER Working Paper