Examine the Top Reviewers on Amazon.com

Qianyun (Poppy) Zhang, Tyler Hendry

December 18, 2017

1 Introduction

Recent studies found that consumers rely on Amazon reviews before making purchasing decisions in both online and offline context [14]. Amazon has become one of the biggest information source that consumers reach to before purchasing. Consequently, reviews become strategically important to Amazon and sellers on Amazon[13].

Unsurprisingly, Amazon has been trying to provide incentives for members to actively contribute reviews or to entice users with low engagement levels whose contributions are expected to have high lifetime value to use a particular platform exclusively. To make the reviewing activity more engaging, Amazon recognizes its top contributor on the Amazon Hall-of-Fame[1]. On this page, Amazon lists top 10,000 reviewers out of total 44 million users based on the number of reviews they have contributed. Recognized reviewers will also get a badge attached to all the reviews they write, which will likely accumulate them more readers and helpful votes. One top reviewer, for instance, named iiiireader, has written as many as 4000 reviews, which accumulated more than 44,339 helpfulness votes by Oct 2017. It is not hard to imagine that being a top

[1] https://www.amazon.com/review/top-reviewers

reviewer at Amazon can also economically benefit the user. It is reported that some of the top reviewers at Amazon can even earn their livings by just writing and posting reviews[2].

However, becoming a top reviewer is not easy. An authentic and unbiased review actually takes a lot of efforts. According to our analysis, on average, top reviewers write around 200 words per review and around 90% of their readers find their reviews helpful. Being a top reviewers takes more than just writing: they need to know what information consumers need and they need be able to convince consumers of the pros and cons of the products. The top reviewers at Amazon should be very good at advocating the products and influencing the market, which potentially differentiate them from other consumers even in offline settings. We believe that these top-reviewer qualities could be inferred from the reviews they have written. In this study, we will focus on the examining the top reviewers on Amazon and discuss what takes to make a top reviewer. For this project, we first attempt to build a model to predict if a user will become a top reviewer based on the first several reviews. We use a random sample of users from Amazon for our analysis. We define a top reviewer as a user who has contributed at least 100 helpful reviews on the site. We first build a model with only numerical attributes, such as average review score and average number other users that mark the users reviews as being helpful. We then added textual features to our predictive model. We tested models built using the first one, three, and ten reviews of the user and in the end, we are able to correctly predict 90% of the reviewers with our best deep learning model. Using our prediction model, Amazon will be able to target and incentivize their top reviewers from early on.

We also examine the different types of successful reviewers and discuss

[2]https://www.npr.org/sections/money/2013/10/29/241372607/top-reviewers-on-amazon-get-tons-of-free-stuff

whether there are different routes to become the top reviewers on Amazon. For top reviewer who are extremely devoted to reviewing, which we defined as reviewers writing more than 500 reviews, we will segment them into different cohorts based on the products they review. These reviewers counts for less than 1% of the total reviewers on Amazon. This group is of particular interest to us because they represents the most committed and devoted community members on Amazon. Because the value of the products vary on Amazon, for some category, top reviewers can have a higher impact on profits compared to other category. By examining this grouping, we will learn the different types of reviewers and examine their characteristics.

2 Related Work

Given the importance of online consumer reviews, there have been fruitful work on how to identify helpful reviews and trustworthy reviews using review level characteristics. The underlying reason that accounts for why consumers adopt reviews in their decision-making processes is that reviews can provide diagnostic information to consumers about the products and services. This diagnostic value can be measured using the 'helpfulness vote' that is available from online platforms. For instance, on Amazon.com, after each customer review, there is a question: 'Was this review helpful to you?'. Individuals can vote on whether the review is helpful and can also see the summary statistics of the helpful vote of other users, such as '4 out of 5 people found this useful'. Helpfulness vote has been conceptualized as the peer-evaluation that facilitates a consumer's purchase decision-making([14]), and previous studies used the helpfulness vote as a proxy to measure the perceived diagnostic value of reviews to consumers. Long reviews with high star rating are also found to be more helpful to consumers([7]; [5]; [2];[16]). We will further discuss how we use this feature in the following

sections.

However, less attention has been paid to reviewers and only a couple of studies tried to examine how reviewers differ from each other. [12] examined the reviewer trustworthiness based on reviewers' posting frequency, reviewers' profile information completeness and number of badges reviewers acquired. They built a predictive model on distinguishing trustful reviewers from the general crowd. [6] found that reviewer information has a big impact on sales for unpopular products. [9] found that professional reviewers are more likely to add new perspectives to their reviews based on previous reviews while regular reviewers are more likely to repeat the same pros and cons as other reviews. Clearly, there are some differences in the motivation of writing reviews between very professional reviewers and regular ones: professional reviewers write more high-quality reviews. Since we do not observe the implicit motivation behind the review-writing behavior, we can only use the reviewer characteristics to infer the types of reviewer.

3 Part I: Predict who will become top reviewers

In this section, we aim to build a predictive model to predict who will become the top reviewers (write more than 100 reviews) on Amazon.

3.1 Data

Our data from a bigger project by Qianyun Zhang and Vishal Singh [3]. In the bigger project, they used a complete sample of reviewers and products on Amazon. The data covers all the reviews that have been posted on Amazon from 1990s to 2014. In the prediction part, we uses a random sample from this big dataset. To make our analysis manageable, we sampled 2000 reviewers,

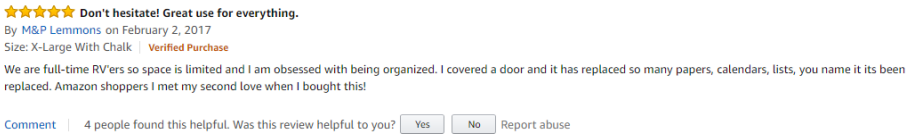
[3] www.vishalsingh.org

1000 top reviewers and 1000 non-top reviewers. We observe these information for each review that a reviewer writes:

- Star (from 1 to 5)
- Time-stamp
- Review Text
- Helpfulness (in percentage)

Figure 1 below shows an example of an Amazon review. Given these review level information, we can derive reviewer level characteristics such as the average star rating the reviewer has given, the average length and helpful of the reviews he or she has written.

Figure 1: Example of an Amazon Review



For our model, we computed 5 numerical attributes and 67 textual attributes. We vary the number of reviews (we will refer to as 'first n reviews' in the following paragraphs) to use in the model. We tried 3 conditions: just the first review, first 3 reviews, and first 10 reviews. For reviewers with less than 10 reviews, we assumed that if they continue writing, their average review characteristics will stay the same as their previous reviews. This is a strong assumption but given the data we have, this is the most practical assumption. In practice, at a given time, platforms can also use the time-stamp of the last review to infer whether the reviewer is still active. Since we are performing an ad-hoc analysis and our data was gathered before 2016, this information can be

of high predictive power but of little help to platforms in close to real-time predict. Thus, we will not use the time-stamp in our prediction. The 5 numerical attributes we generated are:

- Average review score: On Amazon, this score ranges from 1-5. For our test data, we used only the average scores from the first n reviews which we aimed to predict based on.

- Average length of review: This is the average total word count for the the users review. We considered this as we might expect more serious reviewers to put more efforts in writing reviews thus leave longer reviews than others.

- Average rating deviation from product mean: Each Amazon product has an aggregate review score ranging from 1-5 in half point increments. For each reviewer, we compute the difference between the users score and the aggregated product average score. Then we take average of this difference for each reviewer. Since the product rating is only in half point increments, we have some information loss. This measurement can tell us how deviant a reviewer is from general public opinion.

- Average helpfulness: Amazon allows users to tag a review as being helpful. For each review, we have the total number of times other users have tagged it as such. We then compute the ratio of the review being rated as helpful out of total votes. Our intuition for adding this was that a reviewer with more helpful marks, likely spent more time on the review and so may be more dedicated to adding future reviews.

- Female or not: Gender of the reviewer is inferred from the reviewer name using an algorithm from [15]. This is binary variable. If a reviewer is

identified as female, it takes 1. If the gender is unidentifiable or is male, it takes 0.

- Average review sequence: Using the time-stamp of a review, we can infer the sequence of a review given the total reviews of a product. If the sequence of a review is 1, it means this is the first review for a product. For each reviewer, we computed the average of the review sequence, weighted by the total number of reviews a product have by the time we scrape the data. This gives us a measurement on how eager or how innovative the reviewer is.

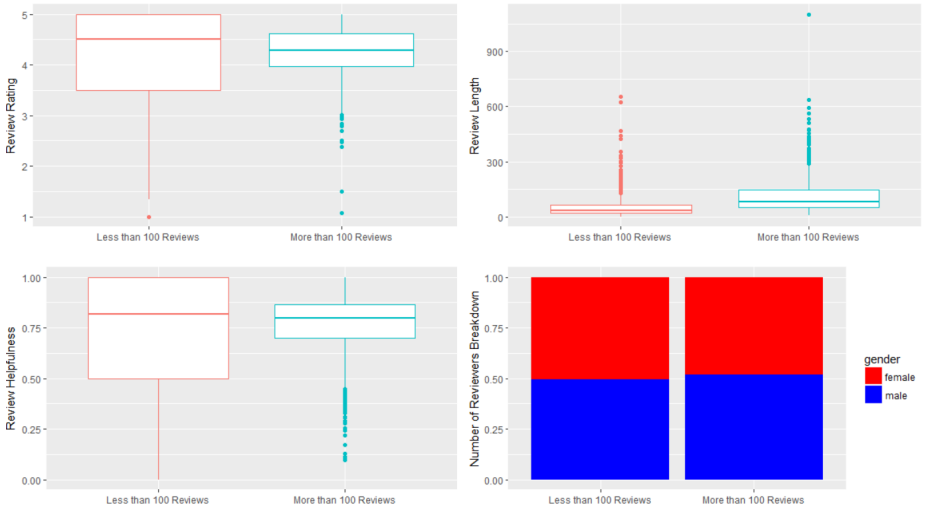
Figure 2 below plots the distribution of these characteristics for top and non-top reviewers. As can be seen from the figure, top reviewers on average do not give higher rating or have higher helpfulness votes compared to non-top reviewers. However, the distribution of average rating and helpfulness for top reviewers are very concentrated. This means that top reviewers are more similar to each other than non-top ones. We can also see that top reviewers write longer reviews compared to non-top reviewers. There are not significant gender differences between these two group of reviewers.

To generate the textual information embedded in the reviews, we used a standard language processing procedure from social-science field to derive the linguistic characteristics of the text[18]. Our text characteristics include but not limited to:

- POS Tagging: For each review we generated a count for each pos-tag. Examples include adjectives, nouns and verbs.

- Time tense: We compute how likely the review uses past tense, current tense or future tense.

Figure 2: Characteristics of Top vs Non-Top Reviewers



- Numbers: Numbers can help tell a story and makes the review more trustworthy because they offer support to textual information. So we compute how many times a number is used in the text.

- Pronouns: We pay special attention to pronouns because they can signal how objective or subjective a review is.

- Excessively strong language: We tagged certain swear words as being excessively strong. We expect that more serious reviewers will use such words more sparingly if at all.

Above are some characteristics that are of particular interest to us. A more complete lists of characteristics generated from this algorithm can be found at [17]. For the text classification models, we use only the text of reviews, to predict. For each reviewer, we stack the reviews they have written to a single vector and then perform word-embedding and classification with the word vectors.

To summarize, we have three different types of data: first is the numerical

value we generated from reviews such as review rating and review length; second is the numeric textual attributes such as weighted number of verbs and swear words; third is the word vectors generated from the text.

3.2 Model

We will use deep learning models to predict whether a reviewer is a top reviewer or not using the numerical characteristics and textual characteristics. The model and data combination we have tried are summarized below:

Model	Numerical Data	Numerical and Textual Attributes	Review Text
MLP	1st Reviews	1st Reviews	
MLP	3 Reviews	3 Reviews	
MLP	10 Reviews	10 Reviews	
LTSM		10 Reviews	
1-d CNN			10 Reviews
Fasttext			10 Reviews

Table 1: Model and Data

For the model, we used a hold-out sample to perform cross-validation and we tested our prediction accuracy on the test data. For MLP (multilayer perceptron), we used 2 RELU layers, each layer with a drop out rate of 0.1. The loss is computed using binary cross entropy. We used a batch method and the batch size is 150. We run the model for 50 epochs. For LTSM (Long short-term memory), we used a similar structure as MLP. For both MLP and LSTM, we tested their performances using both numerical attributes, such as average rating on reviewer level, and textual attributes, such as the pos-tags we generated, to predict.

We also tried using the review text to predict. For each reviewer, we stack their reviews to a single vector and we generated padded text vector with maximum length of 150 words and these vectors are filled with integers that represent the sequence of the words. To feed these information into the deep learning

models, we convert them into an embedding matrix. The maximum number of features in our model is 20,000 to maintain a reasonable time needed to train the model. For the 1-d CNN, we used 5 kernels and two RELU layers with 0.1 drop-rate. We used same binary cross entropy to calculate loss. We run the model for 20 epoches. In addition, we used a relatively new methods, Fasttext by Facebook [11]. This method added more information to the data by including the n-gram into the model input. This can help us to capture some information on the local word order[4].

3.3 Result

We first report the results of using MLP and the numerical attributes of reviews, namely the results for the model in second row of table 1, in below figure 3. We can compare the accuracy across the same model with different datasets. The x-axis represents the losses and accuracy calculated from train and test data over each epoch using the batch method. Clearly, the more information we have about reviewers, the more accurate is our prediction. The following figure 4 plots the results of models using both numerical and textual attributes. MLP performs less optimal compared to LSTM. Still, more information adds to more accurate prediction. In figure 5, we report the results using only review text and word embedding. We can see that by using the raw text, rather than summarized attributes, our model can learn more information about the two types of reviewers. The average accuracy from 1-d CNN is slightly lower than Fasttext but Fasttext performs more stable over each epoch. The highest accuracy we have in this task is around 90%. To make the comparison between models easier, we are reporting the overall accuracy of aforementioned models in the table below. Because MLP and Fasttext work with different data and have different build, this may not be a fair comparison between models. The easiest

[4]We thank Professor Joan Bruna for this suggestion.

model for platform managers to adopt is either MLP with numerical data or Fasttext with review text, depending on the capacity of the data-science team.

Figure 3: MLP with Numerical Attributes

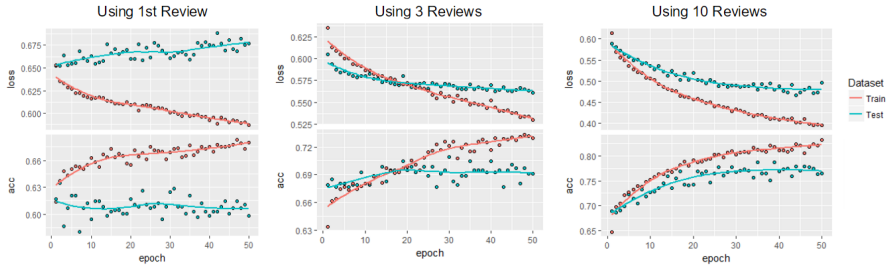


Figure 4: MLP and LTSM with Numerical and Textual Attributes

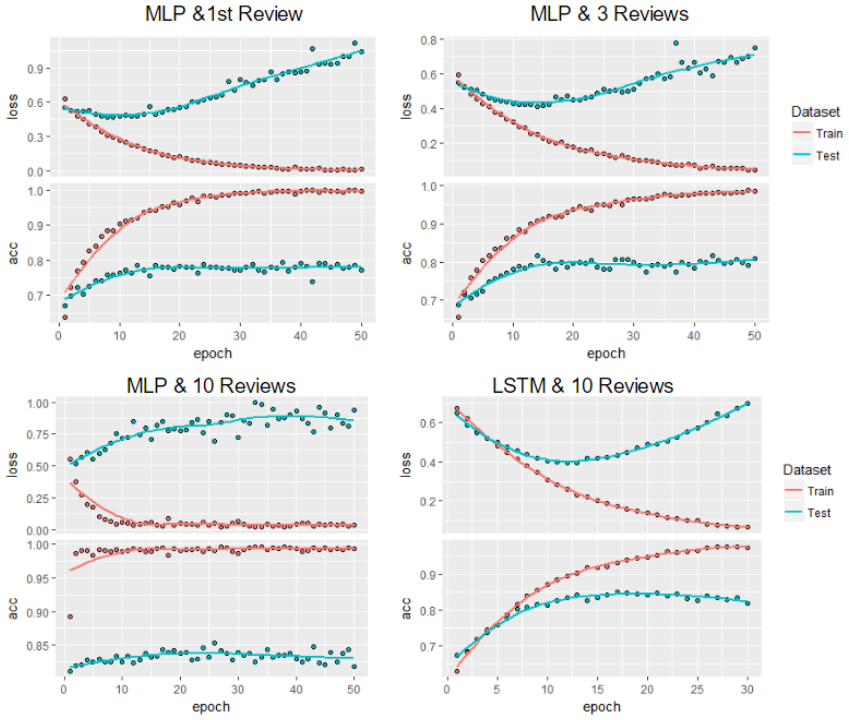
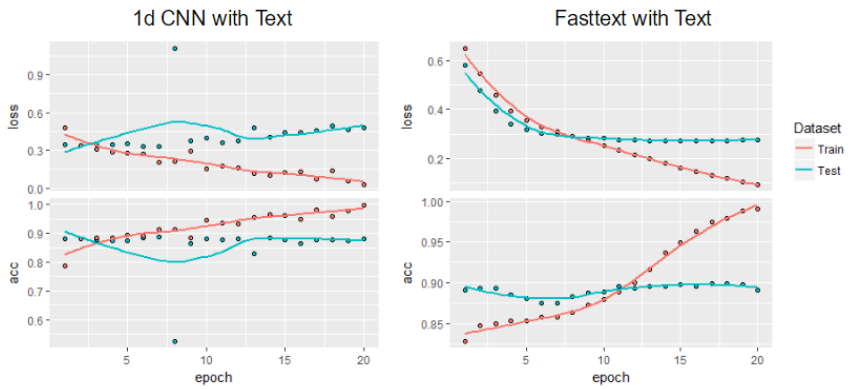


Figure 5: CNN and Fasttext with ReviewText



Model	Numerical Data	Numerical and Textual Attributes	Review Text
	Accuracy	Accuracy	Accuracy
MLP	1st Reviews: 61%	1st Reviews: 80%	
MLP	3 Reviews: 71%	3 Reviews: 81%	
MLP	10 Reviews: 78%	10 Reviews: 83%	
LTSM		10 Reviews: 84%	
1-d CNN			10 Reviews: 87%
Fasttext			10 Reviews: 90%

Table 2: Model and Data

3.4 Discussion

In this section, we have demonstrated how to use reviewer level attributes to predict whether a reviewer will write more than 100 reviews or not in the future. This model can help platforms such as Amazon to selectively target and motivate the reviewers. In this model, we only used public available data. We believe the accuracy of our model can be further improved once we integrate private user browsing data and purchase data. These data should be internally available to platform managers. In practice, platform managers can also constrain this prediction to certain time-frame. For example, they can predict who will write

more than 100 reviews in less than 2 months. The results they get from this type of analysis will be of high value

Besides being able to predict who will become the 'top' reviewer at Amazon, we are also interested in the different segments of reviewers. In the next section, we will select reviewers with more than 500 reviews to examine the product categories they review in.

4 Part II: Categorize the top reviewers

This section aims to gain a better understanding on the different types of top reviewers on Amazon based on the products they review.

4.1 Data

To examine the top reviewers who are truly committed to writing reviews, we choose to examine top reviewers with more than 500 reviews on Amazon. In total we have round 2000 reviewers from the full dataset that satisfies this criteria. This criteria also makes our analysis manageable. We have 24 product categories on Amazon. For each reviewer, we summarize how many reviews he or she has written for each product categories. This is the data we will feed into the topic model.

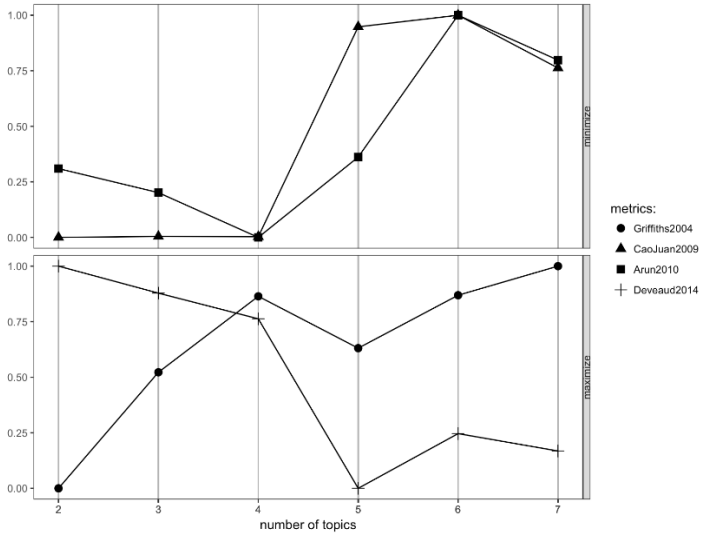
4.2 Model

We use LDA model [3] and treat each reviewer as a document and the product category as the word. The number of the reviews that a reviewer post in specific product category serves as the word frequency in our example.

There are 4 tests that are widely used to find out the optimal number of topics from LDA model: [1], [4], [8] and [10]. To find the best number of topics

in our data, we performed these 4 tests and plotted the results returned using these 4 metrics in the figure 6 for each number of topics. The x-axis represents the number of topics we have tried with the model. For the plot on the top, the lower value on y-axis, the better the assignment is. For the plot on the bottom, higher y-axis value represents better assignment. As can be seen, on average, number 4 gives the best fitted results. So we decided to work with 4 topics.

Figure 6: LDA Number of Topics



4.3 Result

The leading product categories (words in regular LDA models) for each reviewer type (topics in regular LDA models) are reported in table 3. In total, we have 4 types of reviewers and we labeled them as Books, Entertainment, Lifestyle, and Music based on the categories they review. For book reviewers, they primarily review book, kindle books and arts. For entertainment reviewers, they mainly review movie, shows and videos. These different reviewers have different purchases needs and also reviewing interests. Since we do not have the purchase

data, what we observe could either be a result of specialization of reviewers or could be that buying habits are very different between consumers on Amazon and reviewers tend to review some percentage of what they buy. In figure 7, we report how many reviewers we have in each type and the gender breakdown. The y-axis here represents the numer of reviewers. In figure 7, we are omitting reviewers that we cannot identify the gender for. As can be seen, most of the reviewers fall into the 'book reviewer' group. Entertainment and music reviewers are mostly male.

Book	**Entertainment**	**LifeStyle**	**Music**
Books	Movies and TV	Electronics	CDs and Vinyl
Kindle Store	Video Games	Health and Personal Care	Digital Music
Arts	Amazon Instant Video	Grocery and Gourmet Food	Musical Instruments

Table 3: Leading Categories for Reviewer Types

To understand how these reviewers differ from each other on other attributes, we summarized the average star rating, review length, helpfulness, readability(higher means less readable), and reviewer age(how long has been active). The results are plotted in the figure 8 below. We can see that entertainment reviewers give lower ratings and are less helpful compared to others. Even though their reviews are on average longer than others', their reviews are not as helpful as others. The music-loving reviewers has the longest history and write the most sophisticated reviews. The lifestyle reviewers give shorter but more helpful reviews. This is probably related to the fact that lifestyle related products, such as electronics and personal care products, are mostly utilitarian products so reviewers need to be concise and direct in their reviews. Books and music are experience products and reviewers need to articulate their experiences and their tastes with more words. However, because everyone has different preferences for experience products, one reviewers' experience may not help other reviewers very much. That is probably why we observe these data patterns.

Figure 7: Number of Reviewers in Each Group

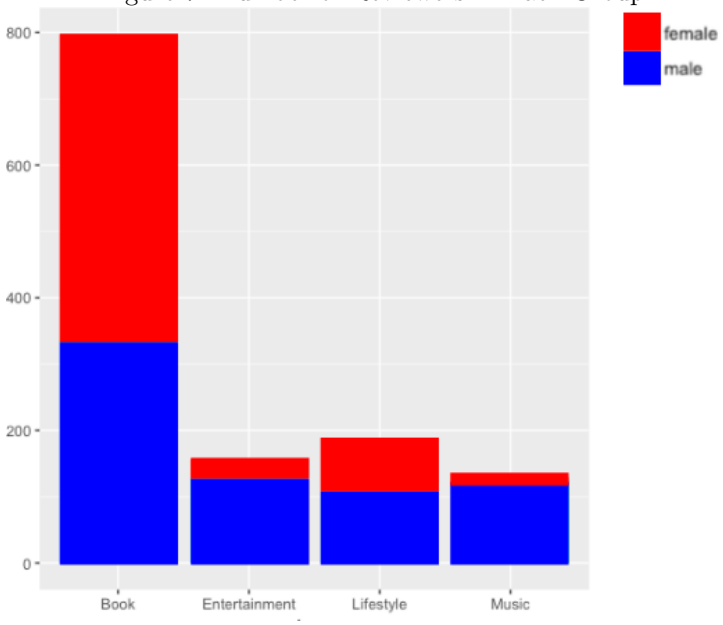
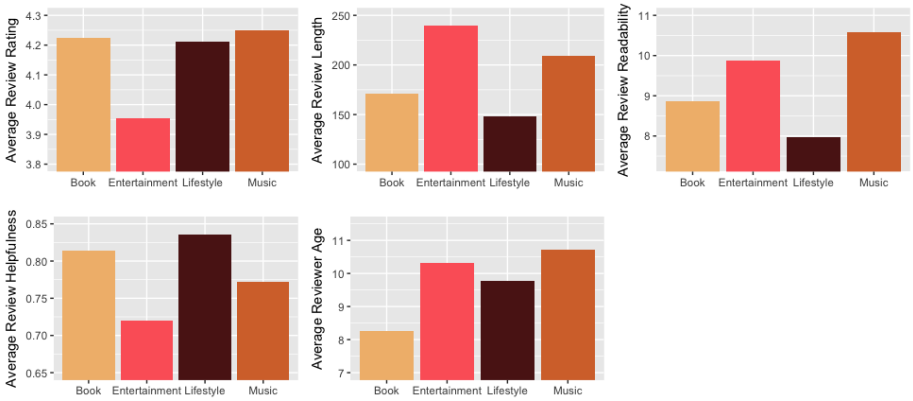


Figure 8: Reviewer Type and Characteristics



4.4 Discussion

Because Amazon started with selling books and rolled out new product offerings throughout its lifetime, the results of this categorization would be possibly biased towards older users reviewing mostly books. Surprisingly, we do not

observe this bias in our analysis. Our analysis demonstrates that there are 4 different groups of top reviewers and they mainly contribute to 4 different product categories. Of all types, the average age of the account of book reviewers was the lowest. If we have access to internal purchase data, we can also pin down these groups of reviewers in terms of their purchase habits and examine the portion of the products they review out of the total products they purchased. This could be a further extension to our analysis.

5 Conclusion

In this project, we first try to predict who will become the top reviewers on Amazon by using the first few reviews they have wrote. Our results offer business owners an opportunity to identify valuable users based on only a few reviews they observe. To make the results more useful, business owners can further integrate the monetary profits that reviews have brought into the model by examining the value of the products reviewed and predict who will become more valuable reviewers to the platform. The prediction accuracy can also be improved by further incorporate users browsing data and review-reading data. It is possible that top reviewers will spend more time browsing and reading reviews before they purchase.

To understand different types of top reviewers, we further categorized top reviewers into four groups based on the products they review, and we found reviewers have distinct review interests and behavior. For instance, electronic and grocery reviewers are more helpful than book reviewers. This suggests that different reviewers may help community in different ways. A further extension to our study can focus on understanding why reviewers demonstrate these different traits and what motivate them to be top reviewers.

References

[1] Rajkumar Arun et al. "On finding the natural number of topics with latent dirichlet allocation: Some observations". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2010, pp. 391–402.

[2] Hyunmi Baek, JoongHo Ahn, and Youngseok Choi. "Helpfulness of Online Consumer Reviews: Readers' Objectives and Review Cues". In: *International Journal of Electronic Commerce* 17.May 2015 (2012), pp. 99–126. ISSN: 1086-4415. DOI: 10.2753/JEC1086-4415170204. URL: http://mesharpe.metapress.com/openurl.asp?genre=article%7B%5C&%7Did=doi:10.2753/JEC1086-4415170204.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[4] Juan Cao et al. "A density-based method for adaptive LDA model selection". In: *Neurocomputing* 72.7 (2009), pp. 1775–1781.

[5] Pei-Yu Chen, Samita Dhanasobhon, and Michael D Smith. "All reviews are not created equal: The disaggregate impact of reviews and reviewers at amazon. com". In: (2008).

[6] Pei-yu Chen, Samita Dhanasobhon, and Michael D Smith. "Research Note : All Reviews are Not Created Equal : All Reviews are Not Created Equal :" in: May (2008).

[7] Judith a Chevalier and Dina Mayzlin. "The Effect of Word of Mouth on Sales: Online Book Reviews". In: *Journal of Marketing Research* 43.3 (2006), pp. 345–354. ISSN: 0022-2437. DOI: 10.1509/jmkr.43.3.345. arXiv: 0022-2437.

[8] Romain Deveaud, Eric SanJuan, and Patrice Bellot. "Accurate and effective latent concept modeling for ad hoc information retrieval". In: *Document numérique* 17.1 (2014), pp. 61–84.

[9] Eric Gilbert and Karrie Karahalios. "Understanding deja reviewers". In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM. 2010, pp. 225–228.

[10] Thomas L Griffiths and Mark Steyvers. "Finding scientific topics". In: *Proceedings of the National academy of Sciences* 101.suppl 1 (2004), pp. 5228–5235.

[11] Armand Joulin et al. "Bag of Tricks for Efficient Text Classification". In: *arXiv preprint arXiv:1607.01759* (2016).

[12] Marios Kokkodis. "Learning from positive and unlabeled Amazon reviews: Towards identifying trustworthy reviewers". In: *Proceedings of the 21st International Conference on World Wide Web*. ACM. 2012, pp. 545–546.

[13] Jumin Lee, Do-Hyung Park, and Ingoo Han. "The effect of negative online consumer reviews on product attitude: An information processing view". In: *Electronic commerce research and applications* 7.3 (2008), pp. 341–352.

[14] Susan M Mudambi and David Schuff. "What makes a helpful review? A study of customer reviews on Amazon. com". In: (2010).

[15] Lincoln Mullen. *gender: Predict Gender from Names Using Historical Data*. R package version 0.5.1. 2015. URL: `https://github.com/ropensci/gender`.

[16] Yue Pan and Jason Q Zhang. "Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews". In: *Journal of Retailing* 87.4 (2011), pp. 598–612. ISSN: 00224359. DOI: `10.1016/j.jretai.2011.05.002`. URL: `http://dx.doi.org/10.1016/j.jretai.2011.05.002`.

[17] James W Pennebaker, Roger J Booth, and Martha E Francis. “Linguistic inquiry and word count: LIWC [Computer software]”. In: *Austin, TX: liwc. net* (2007).

[18] James W Pennebaker, Martha E Francis, and Roger J Booth. “Linguistic inquiry and word count: LIWC 2001”. In: *Mahway: Lawrence Erlbaum Associates* 71.2001 (2001), p. 2001.