MGMT 190: Data and Programming for Analytics

Anna Zhang, Ying Jin, Sang Do, Anuj Shah, Milisha Merchant, Egemen Can Gök

Final Project Report: Insights from Amazon Product Reviews

Introduction:

Every year, more than 100 million Americans purchase goods from the online retail marketplace, which has become one of the fastest-growing sales channels in the United States. With the online retail industry on a rise, Amazon takes the lead in America and even some international markets.

Amazon is primarily a retailer, meaning that it exists to sell products to a wide customer pool at reasonable prices. It targets the middle and upper class individuals who value convenience when it comes to online shopping. One of the reasons Amazon dominates the retail industry is because of the product reviews on their website. User reviews have become proven sales drivers for Amazon; the majority of customers often rely on them before deciding to make a purchase. Research shows that online reviews not only help users make their purchasing decisions, but they also help ecommerce businesses obtain more customers. According to a market investigation done by Dimensional Research, an institution that helps technology companies succeed, 88% of the responding Americans who recently shopped online have been influenced by an online review when making a buying decision. This research shows the importance of product reviews for online shoppers. In addition, having reviews on an ecommerce website helps drive the volume of sales for those businesses. According to the Retail/E-Commerce Industry Report of 2011 by iPerceptions, a provider of web-focused customer analytics, based on visitor feedback, 63% of customers are more likely to make a purchase from an ecommerce website that has user-written reviews.

Amazon took a step forward by displaying customer reviews on each product and implementing a voting system that enables users to upvote a review when they believe that it is helpful and downvote a review when they feel like it is not helpful. Furthermore, Amazon ranks the reviews based on helpfulness, meaning that the reviews with the greatest amount of upvotes will appear at the top, while those with the least upvotes will be displayed on the bottom. This can be very helpful to the consumer because he or she will not have to scroll through a vast number of reviews in order to read the most informative ones. As the number of reviews can be quite large, consumers typically go through the reviews that are most visible to them before they stop reading. Therefore, by adding this feature, Amazon made it easier for their customers to locate the most helpful reviews on each product whether or not the reviewer gave high ratings for the product. According to an article published on the Business Insider in 2009, this voting feature contributes to

more than 2.7 billion dollars of revenue for Amazon every year. Amazon, recognizing and appreciating the profit brought by this feature as well as the value of their customers, continues to look for improvements on their review system by incorporating machine learning techniques to predict if a review is useful or not based on the data collected from past reviews and the number of upvotes that those reviews had.

Many research papers have been written on this subject to predict the helpfulness of each review based on previously collected data. Methods such as linear regression, sentiment analysis, and machine learning tools, such as Naive Bayes and Support Vector Machines (SVMs), were used in their research, and the results are very similar. Both researchers from UC San Diego and Stanford University were able to create models with 76.6% to 80.0% accuracy when predicting the helpfulness of each review. Further research was conducted on the subject, including the variation of this result across product types and the discussion of social impacts caused by this system of enabling opinions to be expressed on other individuals' views, but we chose to narrow down our focus.

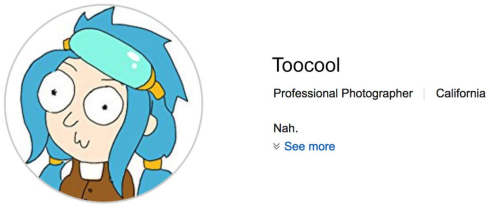
So, is there a relationship between the text and the sentiment of a review? What exactly makes a review helpful? As a group, we decided to observe the effect of the location of the customer, word count, types of words used, and rating on the helpfulness of a review. Moreover, we built a model to predict the rating of each review by analyzing the text. In this report, we will expand on each of these categories and further examine each of their correlations to helpfulness and overall ratings.

Data and Methodology:

After extensive research on the topic, we utilized Jupyter notebook to write the code. We narrowed down the large amount of data and focused our attention on the Cell Phones and Accessories category. From Julian McAuley's data collection, we were able to obtain a CSV file, which contains the user ID, product ID, rating, and timestamp of each review. Furthermore, we were able to gather additional data by opening and parsing JSON files. From these files, we were able to retrieve more information on the reviews, such as number of reviews, average rating, and the number of helpfulness upvotes.

We wanted to find the location of the raters, but this information was not readily available in the datasets that Julian provided, so we had to resort to web-scraping this information. By scraping the customer's profile, the location information of each rater was collected. Realizing that it would be impossible to web-scrape the customer profiles of all 3,447,249 users, we randomly selected 50 customers from the list of raters and web-scraped the location information on their profile pages. From Julian's data collections, we gathered the unique Amazon customer IDs of each user who rated products under the category of Cell

Phones and Accessories. To generate customer profile URLs, we simply appended each customer's ID number to this generic URL: https://www.amazon.com/gp/profile/. Then, we found the customer's location inside each web page using the BeautifulSoup library. A typical Amazon customer's profile is shown in Figure 1 below. Errors that arose during the web-scraping process were handled by the except clause.



[Figure 1: How a typical Amazon customer profile looks like online.]

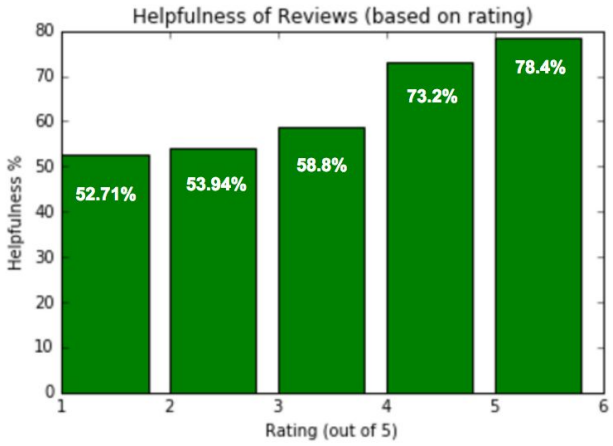
After we extracted the data, we moved on to exploring and graphically representing the data, which is where the Pandas and Matplotlib libraries came in handy. We used the Pandas library mainly to filter rows and group the data. For example, after obtaining the data frames through Pandas, we were able to uncover more specific information, such as the date of the first and last product rating.

Moreover, by using the Matplotlib library, we were able to produce a variety of aesthetic graphs to better visualize the data and the correlations to helpfulness. We used bar graphs, scatter plots, and pie charts. From these outputs, we were able to make some observations about the data, which will be discussed in the analysis section below.

Analysis:

We were able to draw many conclusions from the data that we obtained. Let's take a look at each of the graphs more closely. Keep in mind that we calculated helpfulness by dividing the number of upvotes by the total number of votes that the review received.

First, we analyzed the relationship between product rating and helpfulness as shown below. As we expected, the reviews with the highest ratings turned out to be the most helpful. Psychologically speaking, people tend to believe that reviews with higher ratings will provide more useful information and therefore, they are more likely to agree with them. This explains the increase in helpfulness as product rating increases. From a business standpoint, it would be a good idea to develop some type of award system for consumers who typically give high ratings so that they will be more likely to continue this behavior in the future. This will mutually benefit both the consumer and the seller who will be able to generate more profit.



[Figure 2: The percentage of helpfulness for each rating level.]

Second, we analyzed the reviews more extensively and looked at factors such as number of words, character count, and number of words in all caps as displayed in Figures 3.1, 3.2, and 3.3 below. We noticed that the scatterplots were all very similar. To summarize, the greater the number of words, characters, and number of words in all caps, the more helpful the review is. While the longer reviews were consistently more helpful, the shorter reviews varied greatly in terms of helpfulness. It would be strategic for businesses to encourage consumers to write longer reviews and provide some sort of incentive for them.

Third, as shown in Figure 3.4, we were able to analyze the average ratings over time. Around 2004, ratings have noticeably dropped, suggesting that people have become more critical of reviews as ecommerce becomes more popular. The decrease could also correspond to the status of the economy. After the financial crisis in 2008, there was a very steady increase in ratings. This could have occurred because consumers were more conscious of what they were buying and paid more attention to reviews and ratings. As consumers have become more wary of what they are buying, the importance of ratings has grown over time. People have started becoming more honest and discriminative in order to ensure that they are making a high return on investment in the products they are purchasing.

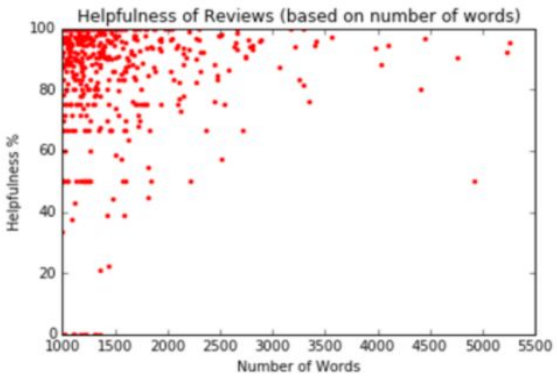


Figure 3.1

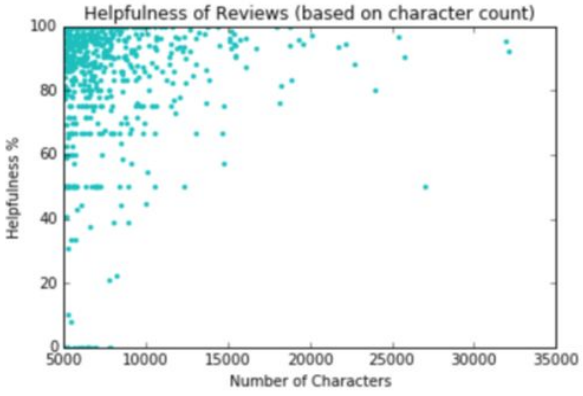
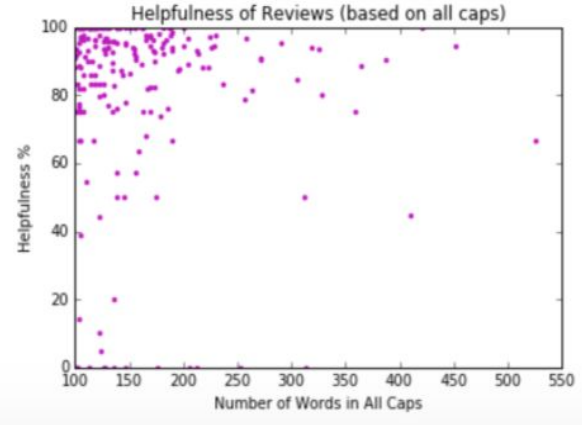


Figure 3.2



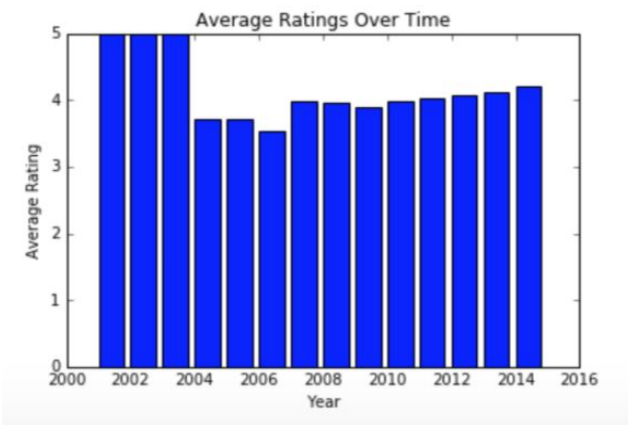
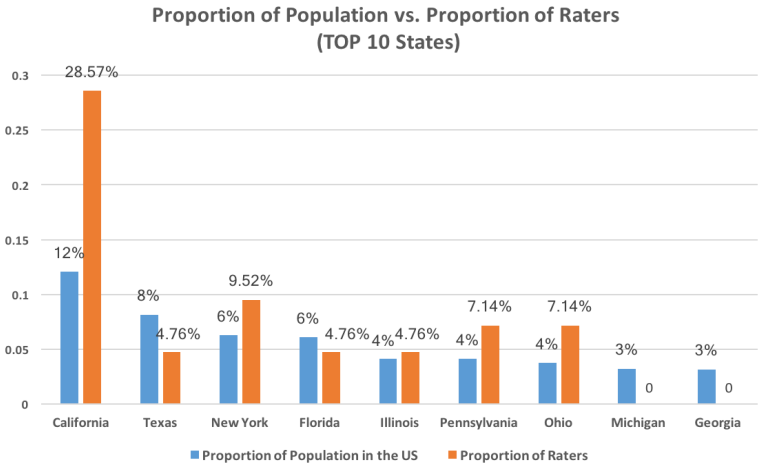


Figure 3.3 Figure 3.4

[Figure 3: The first 3 graphs depict the relationship between the length of the review and how helpful it is. The last graph illustrates the average ratings over time.]

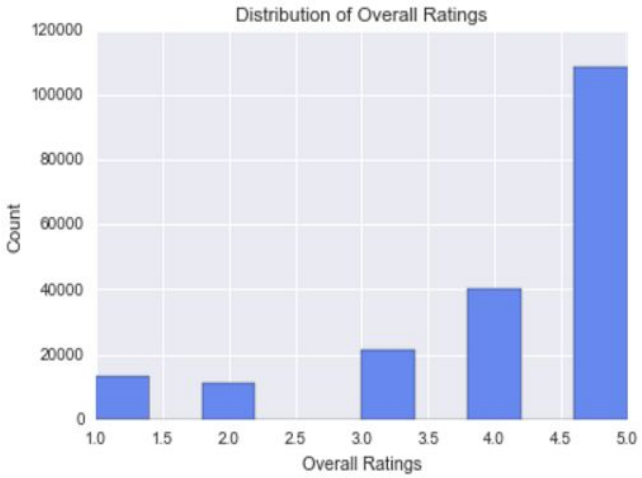
Fourth, we looked at location, which we had to obtain using web-scraping. After cleaning the data, we noticed that the largest portion of raters came from California. The second-largest amount of raters live in New York, while the next largest group of raters are from Pennsylvania and Ohio. Figure 4 shows a side-by-side comparison between the proportion of state population with the proportion of raters from that state. It was surprising to see that the proportion of raters from California is much higher than the proportion of Californians. This suggests that people in California love to share their shopping experiences, and that they are indeed passionate about technology-related items. The states with the biggest populations had the greatest number of reviews, meaning that sellers on Amazon should conduct marketing and testing in bigger states and cities in order to receive more feedback and bring in more revenue.



[Figure 4: A side-by-side comparison of the proportion of raters from a state versus the proportion of population of that state among the overall U.S. population.]

On another note, we attempted to draw a relationship between the text of each review with the sentiment of the review by using machine learning. We first took a look at the summaries of positive and negative reviews from the JSON file. Reviewers that gave the highest ratings said: “this [product] was just perfect!”, “Great replacement cable. Apple certified”, “Real quality”, and “I really like it because it works well with my…”. The people who gave high ratings to products in the Cell Phones and Accessories category wrote about quality, compatibility with their existing devices, and overqualification of their expectations. Some of the reviewers that gave the lowest ratings said: “not a good idea”, “horrible”, “don’t waste your money, pay more and buy one at Walgreen’s”, “be careful”, and “bad experience”. All of these sentiments are due to disappointments. These reviewers feel regretful about their purchases.

Knowing the general sentiments of customers who reviewed the products, we decided to take this a step further by looking at the frequency of each word in positive and negative reviews. Figure 5 shows the distribution of the overall ratings. We sorted our data based on the sentiment analysis: four or five star ratings with positive sentiment, one or two star ratings with negative sentiment, and three star ratings with neutral sentiment. Then, we dropped the data in the neutral sentiment category to build our predictive model because we didn’t want to feed our model the data with neutral sentiment.



[Figure 5: Distribution of Overall Ratings]

After using the Natural Language ToolKit to clean up the text data from the summary column, we were able to produce two different word clouds that display the word frequency data graphically. Our results are shown in Figure 6, where Figure 6.1 shows the most frequent words in positive reviews, while Figure 6.2 shows the most frequent words in negative reviews. This result shows that in the positive reviews, the reviewers often describe their emotional feelings towards the products by using words like “great”, “works”, and “cool”; whereas in the negative reviews, the reviewers often talk about the “money” and “work” they

have “wasted” on the products. Interestingly, we discovered that the words “good” and “work” appear frequently in both word clouds. When building the word cloud function, we used the “stopword” feature to remove stopwords like “not” when cleaning up the data and it is likely that the reviewers meant “not good” when leaving negative reviews. Sometimes, when it comes to writing product reviews on ecommerce sites, reviewers use small words like “not” when describing their experiences with the products (i.e. “not good” instead of “bad” or “not bad” instead of “good”).

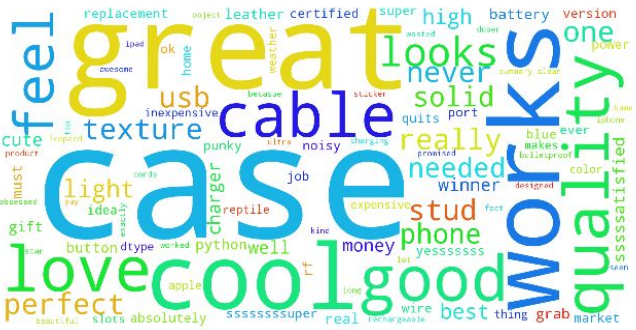


Figure 6.1

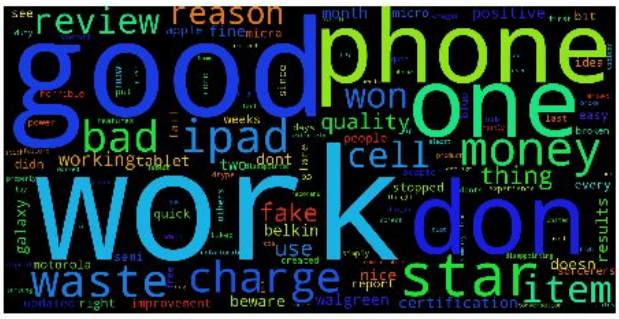


Figure 6.2

[Figure 6: A visual representation of the words frequencies in the positive and negative reviews. The frequency of a word is proportional to the size of the text in the word cloud. Figure 6.1 shows the most used words in positive reviews and Figure 6.2 shows the most used words in negative reviews.]

To build the machine learning model, we first used the *train_test_split* function in scikit-learn to split the data (70% of the data was used for training while 30% of the data was used for testing). Then, we used CountVectorizer and TfidfTransformer to convert the text data into numerical features usable for our machine learning model. After producing the word clouds, we realized that we should include unigrams and bigrams (1-2 word phrases) to improve our model’s predictive capability, since reviewers tend to use small phrases like “not good” and “not bad”. To incorporate unigrams and bigrams in our model, we set ngram_range equal to (1, 2). Each word acted as an independent variable, while the sentiment (whether the review is positive or negative) served as the dependent variable. By using these variables, we calculated a coefficient for each word or phrase. We constructed our machine learning model with Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Logistic Regression to find the best model for this particular analysis. In Figure 7.1, the receiver operating characteristic (ROC) curve determined that the logistic regression model had the most accurate result, since it has the largest area under the curve (AUC). The resulting accuracy of the logistic regression model is shown in Figure 7.2.

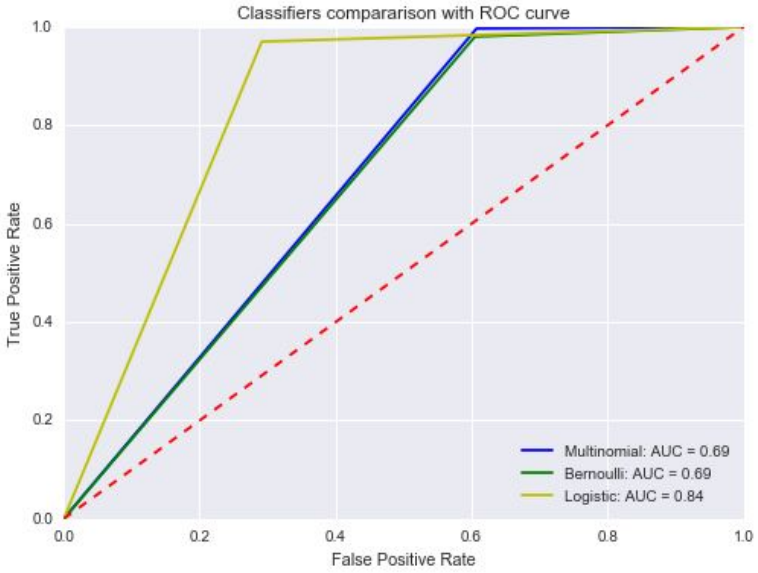


Figure 7.1

Accuracy Score 93.35 %
Reporting...

	precision	recall	f1-score	support
positive	0.79	0.71	0.75	7255
negative	0.95	0.97	0.96	44646
avg / total	0.93	0.93	0.93	51901

Figure 7.2

[Figure 7: Figure 7.1 shows the receiver operating characteristic (ROC) curve. Figure 7.2 shows the accuracy of the logistic regression model.]

	Feature	Coefficient
106529	worst	-31.972132
44379	horrible	-29.494287
90933	terrible	-28.790741
62903	not	-28.501653
106051	work yes	-28.107780
50617	junk	-28.078315
26335	doesn work	-27.622958
9458	be much	-26.944030
26576	dont buy	-26.923329
106740	wouldn	-25.591040
55116	love loved	-25.546070
72290	poor quality	-25.450832
91909	the armor	-25.125879
26338	doesnt	-25.124704
53980	little protection	-24.845425
613	actually high	-24.796086
106632	worthless	-24.609482
31317	fake	-24.234939
6886	at first	-24.210912
25310	died	-24.139358
26430	don buy	-24.107090
100262	useless	-24.032120
72319	poorly	-24.031802
103829	while it	-23.884018
6649	as strong	-23.660610
25285	didnt	-23.293560
7172	attractive easy	-23.188601
48280	is bright	-23.184678
54604	looking not	-23.056788
95602	time that	-23.047957
20507	compatible iphone	21.950595
25652	disappointed wanted	22.096687
42656	haven used	22.102689
102472	wasn sure	22.155938
63061	not disappointing	22.240283
18577	cheap nice	22.325880
78096	recommend white	22.387240
52926	light doesn	22.539056
105935	work case	22.554283
29312	even fit	22.662820
69661	perfect	22.769534
37306	galaxy this	23.224156
60376	my thing	23.424162
106362	works okay	23.487447
13534	bulky cheap	23.542177
62950	not awful	23.573633
2443	amazing	24.140004
62540	no problems	24.698794
63415	not rip	25.239919
63510	not terrible	25.409254
73491	pretty breaks	25.561533
96860	too week	25.735357
14054	but had	26.915642
106167	working cheaply	27.427650
31486	fantastic	27.862100
10262	best	28.704400
14546	but works	29.671053
101065	very cool	30.183100
62952	not bad	31.480662
10487	best not	41.060399

108250 rows × 2 columns

[Figure 8: This figure shows the coefficients of words in the unigram and bigram models.]

Examples of predictions made by the machine learning model are shown in Figure 9. For each test case, our machine learning model successfully predicted the actual sentiment of the reviews based on the given review text. Phone manufacturers in the industry can use this type of text analysis to improve the quality of their products by looking at the words or topics that customers complain about the most.

```
This is the first battery case I have had for my Galaxy S4. The S4 fits very well, is slim and doesn't add much weight to the Ga
laxy S4. It doubles the battery life. You can charge either the battery, the phone or both. There is a handy on-off switch with
 leds to indicate the level of charge.The battery case came on time and was packaged well. Well worth the price.

Sample estimated as POSITIVE: negative probability 0.01%, positive probability 99.99%

They look good and stick good! I just don't like the rounded shape because I was always bumping it and Siri kept popping up and
 it was irritating. I just won't buy a product like this again.

Sample estimated as POSITIVE: negative probability 0.04%, positive probability 99.96%

It worked for the first week then it only charge my phone to 20%. it is a waste of money.

Sample estimated as NEGATIVE: negative probability 99.97%, positive probability 0.03%

I am disappointed that the 1A didn't work with my iPad.  That's what I get for buying a cheap adapter.

Sample estimated as NEGATIVE: negative probability 56.50%, positive probability 43.50%
```

[Figure 9: Examples of predictions performed by the machine learning model.]

Conclusion:

With the information we uncovered by analyzing the review characteristics, Amazon can help users write more helpful reviews (they can advise users to write long, detailed reviews). Additionally, by using the analysis we did on the proportion of users in each state, businesses will be able to improve their relations with customers by catering a specific marketing strategy to each state in the United States. Lastly, our machine learning model shows that it is possible to predict the rating level of a review with high accuracy just by analyzing the text.

There are several interesting experiments that we can do in the future. We can use the code that we wrote for the Cell Phones and Accessories category on data collected from other categories to analyze the variance of these parameters across several product types. Hopefully, we can investigate the effect of word choice, grammar, and user sentiment on the helpfulness of each review for all product categories.

Works Cited

Ben Fox Rubin June 19, 2015 5:34 PM PDT @benfoxrubin: Amazon looks to improve customer-reviews system with machine learning, https://www.cnet.com/news/amazon-updates-customer-reviews-with-new-machine-learning-platform/.

Charlton, G.: Ecommerce consumer reviews: why you need them and how to use them, https://econsultancy.com/blog/9366-ecommerce-consumer-reviews-why-you-need-them-and-how-to-use-them/.

Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., Lee, L.: How opinions are received by online communities. Proceedings of the 18th international conference on World wide web - WWW '09. (2009).

Spool, J.: The Question That Makes Amazon $2.7 Billion Of Revenue, http://www.businessinsider.com/the-magic-behind-amazons-27-billion-dollar-question-2009-3.

What is the Impact of Customer Service on Lifetime Customer Value?, https://www.zendesk.com/resources/customer-service-and-lifetime-customer-value.

Zhang, Y., Lin, Z.: Predicting the Helpfulness of Online Product Reviews: A Multilingual Approach. SSRN Electronic Journal.

J. McAuley, C. Targett, J. Shi, A. van den Hengel: Image-based Recommendations on Styles and Substitutes. SIGIR, 2015.

J. McAuley, R. Pandey, J. Leskovec: Inferring Networks of Substitutable and Complementary Products. Knowledge Discovery and Data Mining, 2015.

B. Shitij: Predicting Amazon Review Helpfulness. https://cseweb.ucsd.edu/~jmcauley/cse255/reports/wi15/Shitij_Bhargava.pdf.