CS6220 Project Final Paper

Students:

Jiaxiang Zhu, Ryan Anderson, Siyao Cai, Sean Lu

Instructor:

Prof. Ling Liu

(Final Report of the BigData Project)

E-commerce Website Products Ratings vs. Reviews

1. Abstract / Introduction.

With E-Commerce websites such as *Amazon* and *Walmart* getting more attention and popularity, purchasing things online has become easier than ever in the present day thanks to quick processing times and shipping. An item purchased online in the morning could arrive as soon as the early afternoon. For items like groceries, people traditionally want to buy the product in person. However, things have changed as people nowadays might also purchase groceries online just for the convenience.

Regardless, there are still differences between shopping online and shopping in person. It is safer to purchase products in store that you can physically observe than to order online based on a few images from a website. In general, customers take big risks when shopping on E-Commerce websites. What happens if a customer buys clothes that turn out to be in bad quality or of the wrong size? The shopper then has to go through a long return and refund process. Normally, this scenario would not happen as often when shopping at a physical store.

In order to solve this issue, websites normally provide a star rating system for customers to check products beforehand. Normally, it works well as long as customers rate their purchases without biases. However, due to many factors, the system is not always accurate. A product with an overall high rating could still have issues and bad qualities. This may indicate that the product has inconsistent manufacturing, or perhaps that a lot of the ratings are illegitimate. In either case, customers would be confused about whether the product is actually good or not.

In order to combat the uncertainty of potential products, E-Commerce websites also use reviews to inform and justify user decisions and ratings. With so much reliance on product ratings and reviews, the user feedback needs to be as accurate as possible. Can the ratings adequately reflect the sentiment of the reviews? If not, what are some of the flaws with the current system? Also, can we possibly improve it for a better representation of the customers' sentiment? In order to answer these questions, we aim to analyze ratings and reviews of different types of products on different websites (*Amazon* and *Walmart*), with the goal of finding useful patterns among them.

2. Related Works.

People have spent some time and effort investigating the effects of online reviews and ratings. Previous researchers showed that customer ratings do not directly correlate with sales on some E-Commerce sites. **[1]**

Ganu et al. focus on improving rating predictions using the text contents of reviews. The paper discusses how the user experience could be improved by changing the structure of the reviews. The research involves using a sentiment-based text rating method to predict the overall sentiment expressed in the text review. This paper calculates the text rating with the number of positive and

negative sentences, and finds that the star rating generally matches the sentiment expressed in the text rating. **[2]**

Mudambi et al. mention that misalignment between the star rating and the text review can lead to the increased customer cognitive processing cost and suboptimal purchase decisions. The paper compares the rating system used by Amazon, CNET, and Yelp. This paper studies products of different categories. All the products selected were classified as experience goods and search goods. Search goods have more attributes that can be objectively measured and easily compared while experience goods are difficult to compare and the reviews are given based on users' sense. The Amazon data set was used as both training set and test set. The author used TagHelper as the sentiment analysis tool in this paper. Labeled text reviews from the Amazon dataset were used to train the model. After training, 1734 product reviews were used as test set to generate sentiment value and compare with the star ratings. Cohen's kappa was used to assess the alignment between text sentiment and the star rating. The results show that high reviews have higher misalignment compared to low reviews. Experience goods also showed higher misalignment rate than search goods. **[3]**

There are also studies on the implicit feedback dataset. The implicit feedback has some characteristics including no negative feedback, numerical value of implicit feedback, and evaluation of implicit feedback such as purchase history, watching habits, and browsing activity **[4].** The paper claims that it is hard to study the implicit feedback since we do not have enough input from the users regarding their preference, and there are so many hidden factors that cannot be determined.

Hutto et al. introduced a parsimonious rule-based sentiment analysis tool, VADER. The model is constructed based on generalized, valence-based, human-curated golden standard sentiment lexicon. No training set is required for this model. The author compared VADER to 11 other highly regarded sentiment analysis tools using social media text, Amazon product reviews, movie reviews, and NY Times editorials. VADER outperformed all other classifiers in overall precision and overall F1 score. **[5]**

3. Architecture / Methodology.

The project involves using the Amazon review data to build a baseline for the sentiment of a product review compared to the rating given to the product. As Amazon contains products from many different departments, such as Electronics or Clothing, we can compare how ratings, sentiment, and the relationship between the two differ by departments. Once we have this large standard of comparison, we can begin scraping Walmart website for review data. This data will

then be compared to the results from Amazon dataset, so that we can analyze patterns of review scoring and sentiment across platforms and item types.

Implementation Details:

- System: We are using a headless server running CentOS 7.7, with all the latest patches installed. The server hardware consists of a 4-core Intel Xeon E5-2680 CPU (with a base frequency of 2.7GHz and max turbo to 3.5GHz), 4GB memory, and 80GB of SSD storage. Since we are not planning to process image datasets, we have decided not to include a GPU in our setup, and will process our data solely on CPU.

- Datasets: We are taking our main dataset from the Amazon Product Data collected by Julian McAuley from UCSD (University of California San Diego) **[7][8][9]**. This collection of data is well-sorted with duplicate entries removed. We are also making a Walmart scraper to scrape Walmart review data for comparisons using Python.

- Tools & Frameworks: We are using *Valence Aware Dictionary and sEntiment Reasoner* (VADER) to generate sentiment values for the reviews **[5][6]**. It is a Lexicon and rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media, online reviews and other domains.

- Others: Most of our code will be written in scripting languages such as Shell or Python. We have built a MySQL database to store the data. We will use Excel to generate graphs for analysis. We also have a readme file in the Code Pack, please check there for all the explanations and usages of our code.

We have recorded a short video in the Code Pack. Please refer to that video for details of our system in action. Here is a screenshot of the terminal (See Figure 1).

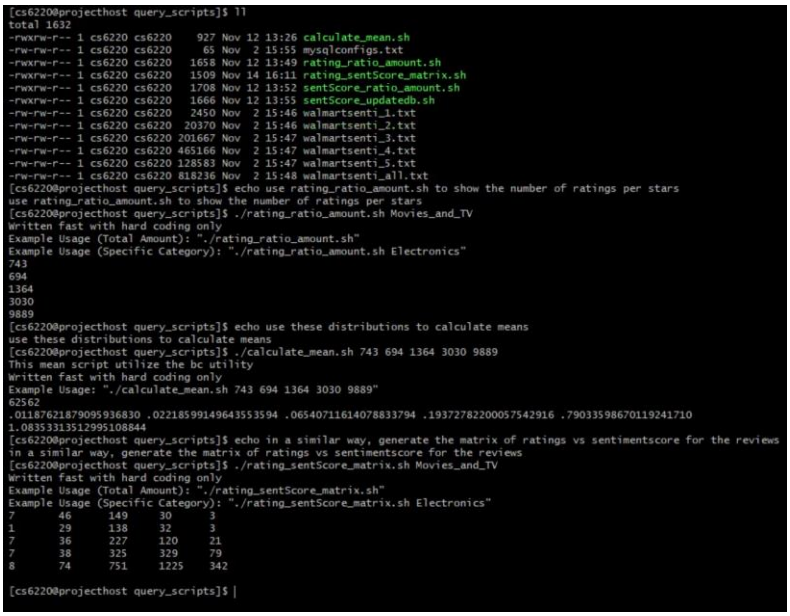

Figure 1. System Running Scripts to Output Tables and Calculate Means

4. Data Analysis and Experimental Results.

We selected 7 product categories that not only cover a wide range of product types, but also exist in meaningful ways across multiple E-commerce websites: Beauty; Clothing, Shoes, and Jewelry; Electronics; Movies and TV; Office Products; Sports and Outdoors; and Video Games.

The Amazon dataset originally had millions of reviews. For this project, we selected 100 products from each of the selected categories. In total, these 700 products had about 15,720 reviews associated with them. This data that we generated has a couple of potential issues. To start with, Amazon data only goes until 2014, whereas the Walmart data was from 2019. This means that comparisons across the two will have some differences, possibly due to how customer patterns of rating and reviewing products have changed in the last 5 years. Also, we drew our samples from the start of the Amazon dataset, meaning that they are all clustered in the early portions of the alphabet. We assumed that this would not have much of an impact on the distribution of ratings and reviews, but it is worth mentioning.

The Walmart dataset was generated from Walmart.com in October of 2019 by scraping several product pages within the seven categories listed above. The product pages tended to have around 30-40 products, so we scrapped around 3-4 for each product category. However, since some products did not have reviews, there are unequal number of products for each category, as seen in the table below.

Table 1. Walmart Product Amount across Different Categories

Beauty	Clothing, Shoes, Jewelry	Electronics	Movies and TV	Office Products	Sports and Outdoors	Video Games
99	106	100	109	80	81	106

The Walmart dataset had far more reviews per product than the Amazon dataset, with around 62,562 reviews across 681 products. This may be due to having scraped it directly from the Walmart.com website, where products are shown in order of best-selling. Therefore, the product data we obtained from Walmart may have different distributions of ratings / reviews, as the Walmart products are best-sellers while the Amazon products should be an indicative sample of the population.

As we can see in the below graphs, in addition to Walmart having far more reviews per product, the distribution of average numbers of review per product vary wildly across categories. For Amazon, they are all fairly even at around 20 reviews per product, with the exception of Movies and TV. However, for Walmart there is a much larger range, with around 15 reviews per product in Sports and Outdoors, and almost 400 reviews per product in Beauty.

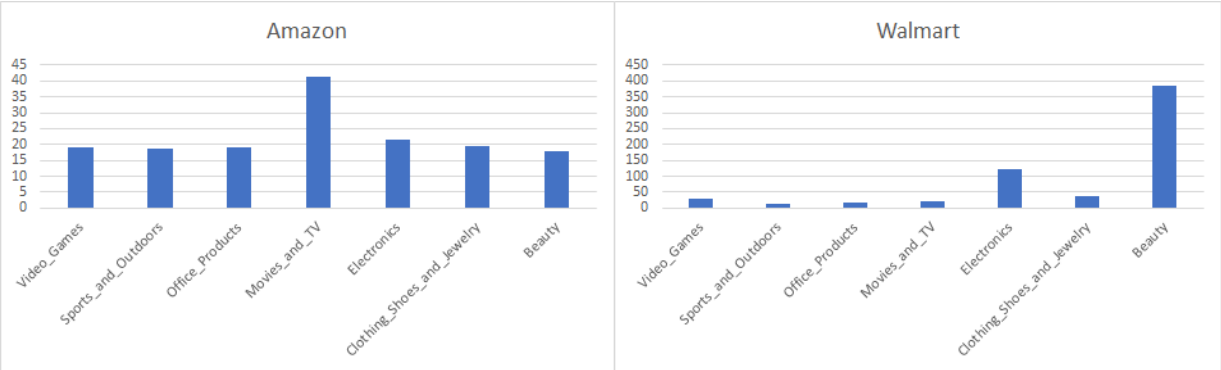

Figure 2. Amazon vs. Walmart Average Reviews Per Product Across Categories

We can easily perform data analysis and comparisons by utilizing our MySQL database. Some of the standard things that we look into for analysis are rating vs. sentiment across sites or product types / categories. The sentiment value is generated by analyzing the review texts with the VADER tool. The value output from this sentiment tool is from -1.0 to 1.0, where lower value represent negative sentiment and higher value represent positive sentiment. This range of value is then converted to an equally distributed scale to match the site ratings from 1 to 5 stars (we call this the sentiment score). In other words, a 1 star rating would map to -1.0 to -0.6, a 2 star rating would map to -0.6 to -0.2, etc. The details are shown in Figure 3.

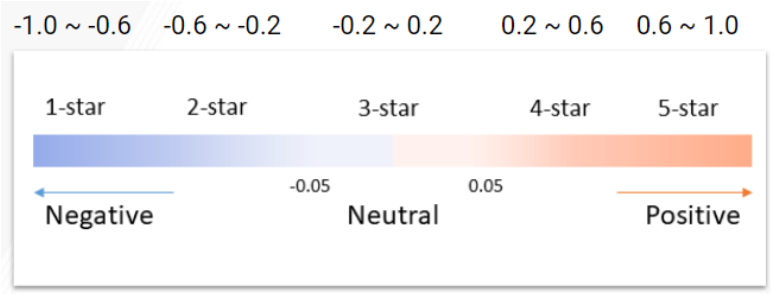

Figure 3. Sentiment Value Conversion to Star Rating

After converting the VADER sentiment value to a star rating, we create a 5x5 matrix to show the percentage of reviews that belong to each sector. Star rating is represented in the vertical axis, and sentiment score is in the horizontal axis.

To explain the graph, I will use the Amazon one as an example. We can see the percentage of 1 star rating and 1 sentiment score is only 1.75%. This means for all the 1 star product ratings, there is only 1.75% of them exactly match their reviews' sentiment score.

Table 2. Star Rating vs. Sentiment Score Matrix for Amazon and Walmart reviews

Amazon (% Match) ↓ Star Rating →Sentiment Score

	1	2	3	4	5
1	1.75%	16.3%	66.9%	10.6%	2.0%
2	0.9%	13.7%	67.0%	15.9%	1.2%
3	1.1%	5.1%	54.9%	31.3%	4.8%
4	0.3%	2.3%	38.1%	49.1%	8.4%
5	0.2%	1.4%	28.6%	54.8%	12.5%

Walmart (% Match) ↓ Star Rating →Sentiment Score

	1	2	3	4	5
1	2.1%	16.5%	64.5%	14.7%	2.2%
2	0.99%	8.99%	59.3%	28.2%	2.58%
3	0.46%	3.47%	38.1%	50.1%	7.26%
4	0.13%	1.04%	21.9%	63.5%	13.4%
5	0.07%	0.61%	16.2%	62.7%	20.4%

Using the MySQL database, we can query results to get the number of reviews that belong to each rating and sentiment score across sites. We use a range of SELECT statements, filtering sentiment score, rating score, and specific sites with WHERE. Below is the comparison between the default site ratings system between Amazon and Walmart. The graphs show similar distribution across both sites, where majority 60% of the ratings belong to 5 stars and around 20% belongs to 4 stars. The ratings from 1-3 stars are much less.

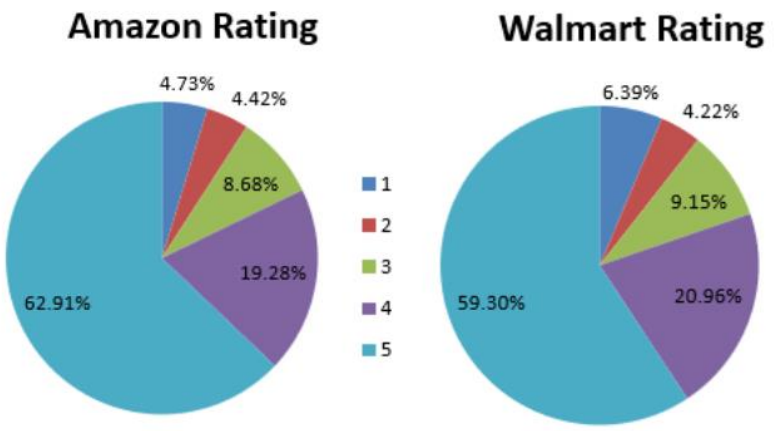

Figure 4. Rating Score Distribution for Amazon and Walmart Reviews

The sentiment matched star comparison between Amazon and Walmart using VADER however, shows more variable results. We can see that there are barely any reviews belong to the 1 and 2 stars. A majority belong to 3 and 4 stars for sentiment, whereas we saw the majority belonging to 4 and 5 stars for ratings. In addition, we can see that overall, Walmart has a more positive sentiment than Amazon. Walmart has more ratings in the higher values such as 4 stars, while Amazon has more ratings in the lower values such as 3 stars.

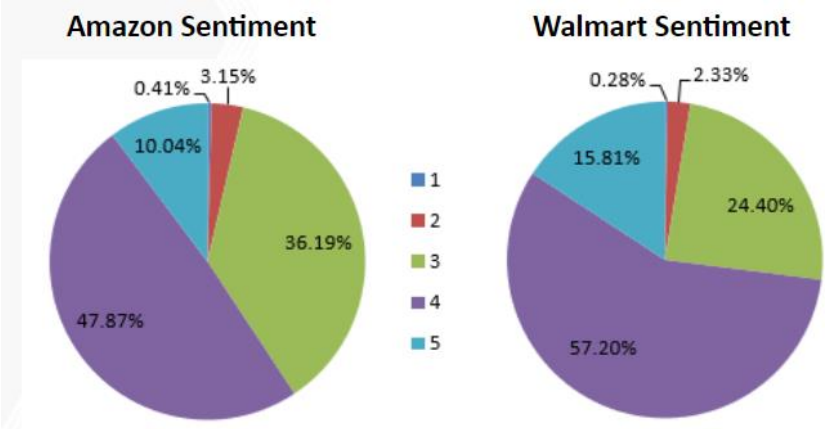

Figure 5. Sentiment Score Distribution for Amazon and Walmart Reviews

The stacked bar charts below represent the rating versus sentiment across the 7 different product categories for each website. Within each individual site, the results are very similar. For the rating, we can see that almost all of the scores are rated 5 stars. In sentiment, most of the scores are rated either 3 or 4 stars. The comparison between categories is less significant. Each category is relatively similar to each other from this large overview.

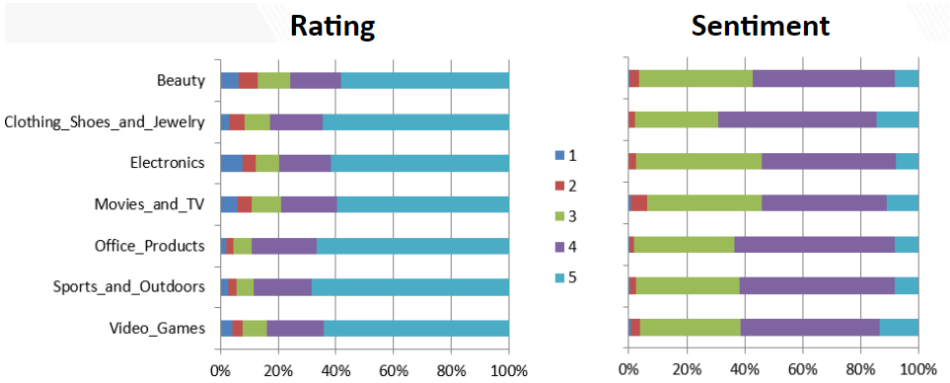

Figure 6. Amazon Star Rating vs. Sentiment Score of Different Product Categories

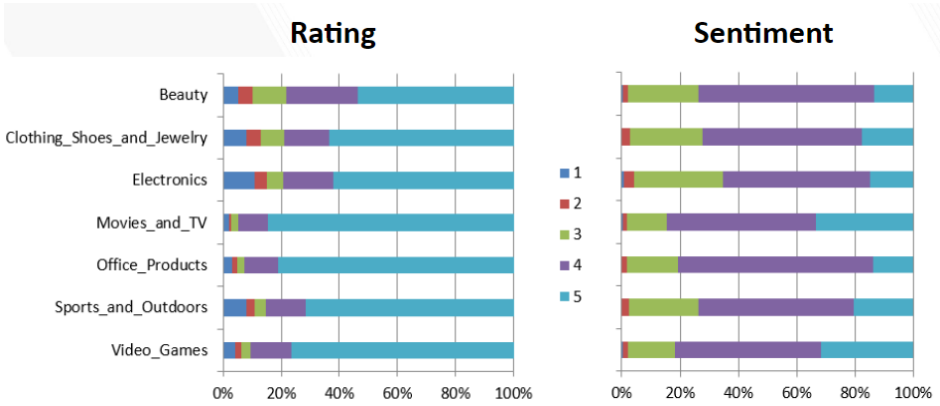

Figure 7. Walmart Star Rating vs. Sentiment Score of Different Product Categories

Looking more closely at specific categories however, we can actually see that there are some slight differences. Below, we show the rating and sentiment distribution for two specific categories. The first category is Sports and Outdoors, and the second category is Movies and TV. When looking at the Amazon sentiment chart, the two categories are very similar, almost identical. However, the Amazon review rating says differently. In the review ratings, Movies and TV has a much lower number of reviews, almost 10% lower, which are 5 stars. The Walmart review ratings surprisingly are actually the opposite of Amazon. For Walmart, the Movies and TV category has actually a higher number of positive 5 star ratings. The Walmart sentiment chart reflects this, as there is a very large number of 5 star sentiment scores compared to Amazon.

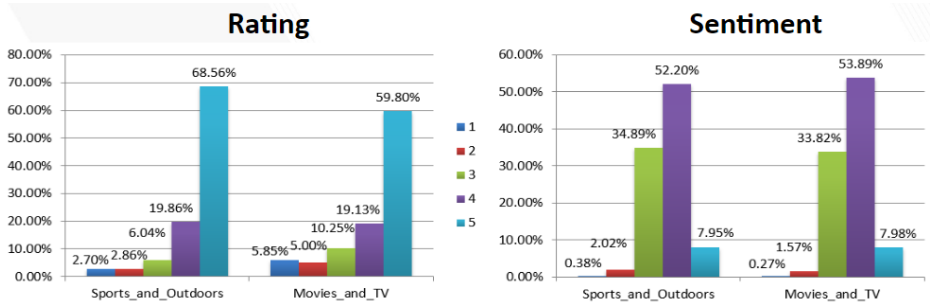

Figure 8. Amazon Star Rating vs. Sentiment Score on Specific Categories

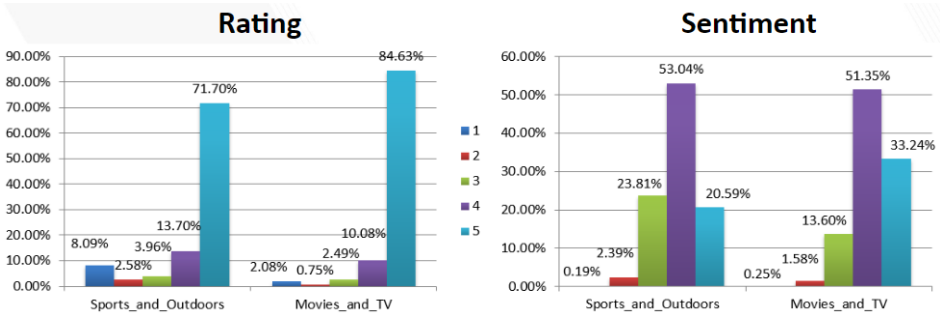

Figure 9. Walmart Star Rating vs. Sentiment Score on Specific Categories

We have processed more data and tables based on different categories. Please check the excel sheets we include in the analysis folder of our code pack.

5. Insights and Discussion.

Based on our results, sentiment-review misalignments exist in both Amazon and Walmart dataset. To explore the main reason for the misalignment, we analyzed customer reviews with large differences between the star rating and the sentiment score. The misaligned reviews can be classified into several categories. There are complaints about the shipping / packaging / price, subjective star rating, and misalignment due to the limitation of sentiment analysis tool. The current rating system in many E-Commerce websites only has one overall star rating, which is not sufficient to represent all the aspects of the purchase experience and the product. Customers can give low star ratings when one out of many parts of the purchase experience is disappointing, even if the product itself is flawless. One possible solution to this issue is dividing the overall star rating into product rating, shipping rating, packaging rating, and price rating. The star rating categories could also be different for experience goods and search goods. By splitting the overall

rating into different sub-categories, the star ratings can provide more detailed information about the whole purchase process.

Reading through the text reviews is very time consuming and it is infeasible to read all the customer reviews before making a purchase decision. As a result, the percentage distribution of the star ratings is a widely used indicator for all the users' opinions toward the product. Both Amazon and Walmart have an endorsement system that allows users to like or dislike other customer's review. Walmart even highlights the most liked and unliked ratings. However, even though the endorsement feature is helpful in minimizing the impact of unreasonable reviews and highlighting the good reviews, it cannot change the overall distribution of the star ratings. Therefore, we propose a new dynamic system to modify the distribution of the customer star rating. Based on the existing endorsement system, we can add different weights to each customer review in the overall distribution. In detail, all the reviews have an equal initial weight and the weight will increase if the review is endorsed by other customers. On the other hand, if the review is disliked by other customers, its weight in the overall star rating distribution will decrease. By applying this modification to the current system, we can further minimize the impact of the unreasonable and subjective reviews on the overall star rating distribution, and amplify the impact of objective reviews since they tell the truth about the product. The high and low bound of the weight and the increments can be determined by the E-Commerce websites.

Comparing our results to the previous work we referenced to, we found that rating-review misalignment mostly occurs in experience goods. Moreover, based on our results, rating-review misalignment occurs more often in reviews with low star ratings (1 or 2 stars). More than 60% of the reviews with low star ratings have sentiment scores within the neutral range based on their corresponding text reviews. This is caused by the difference between VADER and TagHelper, where Vader is based on the sentiment lexicon but TagHelper is a model trained with Amazon dataset. In this case, VADER is a better choice because it is based on a valence-based, human-curated sentiment lexicon and no training data is needed, so the sentiment value is relatively objective and independent from the star rating system. However, for TagHelper, the sentiment score is generated by a model trained with Amazon reviews with labeled star ratings. Since there are misalignments in the Amazon dataset, the training set is not perfect. Therefore, the sentiment score generated by TagHelper can also be biased. Since we have an overall much larger dataset (77000 reviews vs. 1734 reviews) and a much larger range of products (1400 products vs. 23 products) compared to the referenced paper, our results are more generalized and trustworthy.

The sentiment analysis tool we used also has some limitations and can be improved. The equation used to calculate the sentiment value of a customer review paragraph is shown in Eq. 1

$$S_{review} = \sum_{i=1}^{N} S_i \quad (1)$$

where S_i is the sentiment value of the i^{th} sentence within the paragraph. The sentiment value of the entire review is simply a summation of the sentiment values of all the sentences. This fails in some specific circumstances, where in the text review, the customers complained a lot about the product they previously used and only wrote a few positive sentences commenting on the new

product they bought. One possible solution is to add different weights to past tense sentence and present tense sentence when calculating the sentiment value of the entire text review.

Also, to map the sentiment value to sentiment scores, we used a linear cutoff between stars. However, the results may differ significantly with a different segmentation rule.

6. Conclusion.

In this project, we focused on the rating system on some major E-Commerce websites to study whether or not the star rating can adequately reflecting the sentiment of the text review. Customer review data from Amazon and Walmart were analyzed to investigate the correlation between star ratings and the corresponding text reviews. Seven categories of products were selected from Amazon dataset and Walmart website, which covers a total of 1400 products and around 77,000 reviews.

A sentiment lexicon-based sentiment analysis tool, VADER, was used to generate sentiment values for all the review texts. The output value of the sentiment analysis tool has a range from -1.0 to 1.0. The range was evenly segmented into 5 equal sections to map the sentiment value to the star rating system, which is called sentiment score in this report. Then, the sentiment scores were compared with the corresponding star ratings. The results showed that around 60% of the ratings on both Amazon and Walmart are 5-star ratings followed by 4-star ratings accounting for around 20% of all the star ratings. However, the sentiment distribution showed a different trend, where sentiment scores of 4 and 3 are the first two places (50% and 30% respectively). Both Amazon and Walmart data showed similar trends. Rating-review misalignment occurs in all seven categories of products but occurs more often in the experience good categories.

After looking into the special cases where the sentiment score is significantly different from the star rating, we found that shipping, packaging and price all have impacts on how people give the star ratings to products. Our data clearly showed that the current star rating system used by many E-Commerce websites cannot adequately represent the sentiment of the text reviews. Thus, we introduced several potential solutions to improve the current star rating system to guide customers to rate products objectively and minimize the impact of biased reviews on the overall distribution of the star ratings.

7. Lessons Learned and Possible Future Works.

We have learned a few lessons during our project:

- Since our Amazon data was old (up to 2014), we planned to scrape a batch of 2019 data for comparison. However, due to limited time, we decided to use the available data despite it being older. On the other hand, the Walmart dataset is newly scraped from the

official website (due to no publicly available data), so the dataset is fresh and up to date. We have some concerns when comparing 2014 Amazon data to 2019 Walmart data, but the final result still stands. (The 2018 Amazon data has just been published by UCSD in early November 2019. Although it is not in the scope of our current project, we plan to process it in the future.)

- The Amazon data from UCSD are clustered in the beginning of the alphabet, while our scraped Walmart data are sorted based on best seller due to its website design. Next time, we would like to pick specific products that are presented on both websites to keep the consistency.

- Due to the fact that Walmart data are sorted by best seller, there was a larger amount of reviews collected from Walmart compared to Amazon (62,562 vs. 15,720).

- The VADER tool has its limitations. For example, if a customer is comparing the product to the one he had, he might mention how bad the old one is and how good the new one is. This could confuse the sentiment tool and a wrong value might be given.

- As we can see in the figures, some of the categories have much more reviews than others. In that sense, 75,000 samples does not seem to be enough. We can definitely increase the total number of samples from both sites.

We would also like to expand our project further in the future:

- Process the entire dataset with neural network, and possibly create a prediction model.

- Add more categories with more products selected in the process.

- We have used an even distribution of sentiment values to calculate the sentiment scores. We could add some weights to past and present tense for more accuracy.

- Test different segmentation rules for mapping sentiment value to star rating and determine the optimal rule.

- UCSD has just published the 2018 Amazon dataset, and we would like to compare our current findings to the recent samples **[10]**.

- Adding bot detection to the reviews for noise filtering.

- Getting data samples from more E-commerce websites like Macy's and Newegg, for a more detailed and trustworthy comparison.

8. References.

[1].Chen, P. Y., & Wu, S. Y. (2007). Does collaborative filtering technology impact sales? Empirical evidence from Amazon. com. *Empirical Evidence from Amazon. Com (July 8, 2007)*.

[2]. Ganu, G., Elhadad, N., & Marian, A. (2009, June). Beyond the stars: improving rating predictions using review text content. In *WebDB* (Vol. 9, pp. 1-6).

[3] Mudambi, S. M., Schuff, D., & Zhang, Z. (2014, January). Why aren't the stars aligned? An analysis of online review content and star ratings. In 2014 47th Hawaii International Conference on System Sciences (pp. 3139-3147). IEEE.

[4]. Hu, Y., Koren, Y., & Volinsky, C. (2008, December). Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 263-272). Ieee.

[5] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[6] Vader Sentiment Analysis, GitHub Repository Webpage. https://github.com/cjhutto/vaderSentiment

[7]. Amazon product data, 1996-2014, Julian McAuley, UCSD. http://jmcauley.ucsd.edu/data/amazon/

[8]. R. He, J. McAuley. Modeling the visual evolution of fashion trends with one-class collaborative filtering. WWW, 2016

[9]. J. McAuley, C. Targett, J. Shi, A. van den Hengel. Image-based recommendations on styles and substitutes. SIGIR, 2015

[10]. Amazon product data, 2018, Jianmo Ni, UCSD. https://nijianmo.github.io/amazon/index.html