E Commerce Reviews: Sentiment Analysis

# Allyson Busch Winter 2019

**Abstract**

The explosion of online content and adoption of technology into almost all aspects of life have created a unique situation for businesses that are engaging in online and in person commerce. With information available online to users through the entire shopping process, it is critical to understand how electronic word of mouth affect the information gathering process. This study introduces the concept of e-commerce, emerging technologies in the field, and how electronic word of mouth is affecting consumers and businesses. With the evaluation of a public dataset containing information on reviews of an e-commerce clothing brand, this study was able to create a sentiment analysis model for evaluating reviews quickly and efficiently, finding that Naive Bayes is the best model for the dataset.

**Research Questions**

The project has research questions that build up to the main question of the study: what creates the best sentiment analysis model for the data? The research questions that build up to the sentiment analysis and help to understand the dataset in-depth include the following:

1. How often do common indicators of sentiment (i.e. love, hate, great, happy) appear in the review text variable in the dataset?
2. How often do the product class names appear in the reviews?
3. What are the common lengths of reviews? What is the range of reviews?
4. Is there any correlation between age and ratings?

Several of these questions exist to understand any underlying trends that may be present in the data that will only be understood with further investigation. These investigations can also help determine the quality of the data, whether that be if reviews are mostly one word reviews or if there are ratings above or below the one to five thresholds of ratings. Understanding the dataset can lead to a better model creation and a better understanding of the information to apply to other businesses and industries.

# Literature Review

Electronic commerce, commonly known as e-commerce, refers to the wide range of online business activities that center around products and services (Shahjee, 2016). Beginning in 1995, this emerging economy has transformed rapidly (Khan, 2016). This phenomena is altering the marketplace in respects to how consumers receive information and how firms structure their business models (Shahjee, 2016). The traditional physical storefronts associated with most businesses, including the employment of salespeople to inform guests and potential buyers, have transitioned to a combination of in person shopping (either through their brick and mortar or through a pop-up or partnership) as well as online stores (or e-commerce stores). E-commerce stores are associated with buying and selling goods over the Internet or conducting any transaction involving transferring ownership or rights to use goods, services, or other items through a computer-mediated network (Shahjee, 2016). With the growing use of the Internet, tablet devices, and smart phones as well as a larger consumer confidence, e-commerce is expected to continue to evolve and expand (Khan, 2016).

With the rise of e-commerce shopping as well as the increased time that users are spending online daily, there has been a shift in how consumers are obtaining their information. In the past, consumers would gain their information on goods they were intending to purchase either by paid sales people employed in brick and mortar stores or by word of mouth from their friends. The explosion and rapid adoption of the internet and expansion of online shopping has led the way to electronic word of mouth (or eWOM) to make a significant impact on behavior consequences (Phan & Pilik, 2017). With the exponential expansion of social media, businesses are able to open conversations with consumers that are more engaging and allow for easier transactional exchanges online (Khan, 2016). This has led to retailers striving to create better content and more realistic shopping experience with technologies such as augmented reality (Khan, 2016). The introduction of multi-channel services such as click & collect or return to store, as well as more upgraded technology solutions, have allowed a cohesive shopping experience that leverages both benefits of e-commerce and in-person shopping (Linzbach, Inman & Nikolova, 2019). With this transition, companies are beginning to turn to social media listening to understand customer sentiments (Agarap & Grafilon, 2018). Studies have shown that companies that inject big data analytics into their value chain for e-commerce experience a 5-6% higher productivity than their competitors (Akter & Wamba, 2015). A study by the BSA Software Alliance in the United States was able to isolate that 10% or more of the growth for 56% of firms could be contributed to big data analytics (Akter & Wamba, 2015).

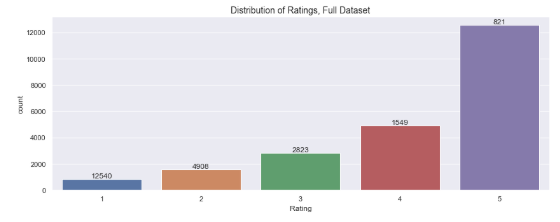
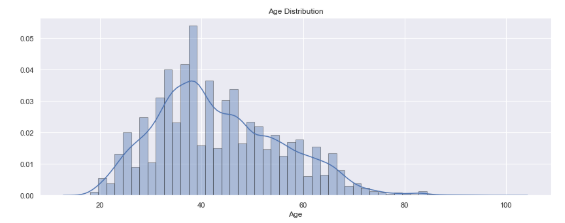
Part of this big data analytics push is to understand how online reviews affect word of mouth. Online consumers are increasingly coming to this important tool to evaluate and select products in their shopping experience (Roy, Datta, & Mukherjee, 2018). Scholars and practitioners have acknowledged that electronic word of mouth is a “dominant force in the marketplace” (Phan & Pilik, 2017). It is then not surprising that 90% of customers in the United States report that their buying decisions are influenced by online reviews (Manes & Tchetchik, 2018). However, electronic word of mouth information is perceived as less reliable than off-line word of mouth information to consumers, but it is considered more credible than information that is created by the sellers themselves (Manes & Tchetchik, 2018). Research has proven that reviews that disclose the reviewer’s identity, high reputation, and that contain elaborateness are perceived as more useful to consumers (Liu & Park, 2015). However, products that have more negative reviews in the first half of helpful reported reviews are likely to have lower product sales (Kaushik, Mishra, Rana, & Dwivedi, 2018).

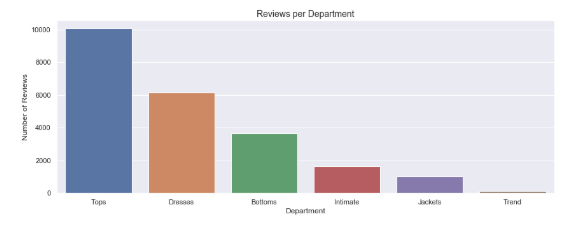
# Methodology

The overall project started with the procurement of data from Kaggle, a crowd-sourced platform with datasets that are open source, that contained reviews written by customers on a women’s clothing e-commerce brand (Brooks, 2018).. The initial stage of the project including importing necessary libraries into Python which include: pandas, numpy, matplotlib.pyplot, seaborn, datetime, sklearn for CountVectorizer, train\_test\_split, LogisticRegression,MultinomialNB, roc-curve/auc, MLPClassifier, SVC, and confusion\_matrix. After the appropriate libraries were imported, the dataset was read using pandas and contained variables of the following:

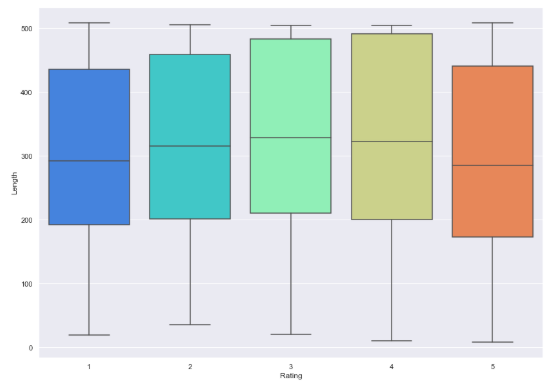
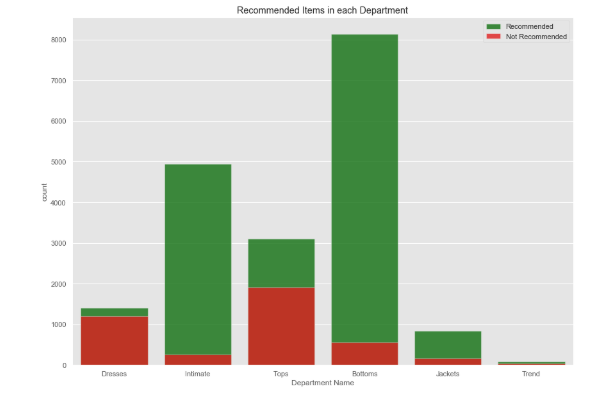
1. Clothing ID: This is an integer categorical variable that refers to the specific piece being reviewed.
2. Age: This variable is a positive integer of the reviewers age. It is unclear if it is self-reported or if it is pulled from something else.
3. Title: This is a string variable for the title of the review.
4. Review Text: This is a string variable for the review body text.
5. Rating: This variable is a positive ordinal integer for the product score granted by the customer from 1 (Worst) to 5 (Best).
6. Recommended IND: This variable is binary and states whether the customer recommends the product where 1 is equal to recommended, and 0 is not recommended.
7. Positive Feedback Count: This is the positive integer documenting the number of other customers who found this review positive.
8. Division Name: This is the categorical name of the product high level division.
9. Class name: This is the categorical name of the product class name.

After loading, and initially using .head() to look at the dataset to make sure the load was successful, .count() was applied. There are 22,641 entries, or reviews, in the dataset. After this, I dropped the unnecessary column that uploaded labeled Unnamed: 0. After this I looked at the count of rows again to see if there are any missing values. There were an inconsistent amount of each variable, particularly in the title, review text, and division name, department name, and class name variables. These are likely guests that left a rating but did not leave a text review, or ones that did not include a title for their review. Since we have such a large dataset and there was only a relatively small amount of null values, I decided to remove these null values but retained the original dataset to compare if there was any shift in the ratings. After the datasets were loaded, Exploratory Data Analysis began.

I started with the distribution of ratings between the initial dataset (reviews) and the dataset with the null values removed (ecom). Starting with the initial dataset, I created a plotly graph of the rating categories (1-5) with the counts on the y axis. The ratings were skewed to the right, meaning there were more positive ratings than negative ratings, following a linear relationship and coming close to even being exponential in growth from 1 to 5. I then created the same plotly graph, but with the ecom dataset that had the null values removed. This graph had the same overall shape and the ratings that were removed due to missing reviews seems to be uniformly distributed across the different ratings. I then looked at the summary statistics for both datasets, both of which were roughly even. The mean of the Rating variable was 4.18, and the dataset contained no outliers beyond the 1-5 rating hierarchy. The standard deviation was 1.12, meaning there is a relatively high variation from one user rating to the next. Moving forward, I will be using the ecom dataset that contains no null values as I do not believe losing this small sample has dramatically shifted any of the results we will get moving on. 

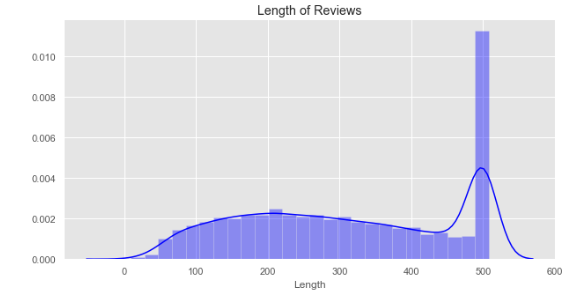
I then looked into the age of the guests in more detail. I started with creating an age distribution graph. This appeared to have a bell curve that was skewed to the left, with the peak of ages being at around 38. There appeared to be more users under the age of 50 than above, which could be explained in consumer behavior averages with younger demographics tending to shop online and review online whereas older demographics still tend to shop in brick and mortar stores, but there is not enough on this graph to make this assumption. Looking into the summary statistics of this variable, the average age of a reviewer on the website is 43, and the dataset contains a standard deviation of 12.33 which is relatively high. There did not appear to be any obvious outliers, as the data fell between the ages of 18 and 99, which are ages that could reasonably be reporting reviews online. While it is important to keep in mind that these are self reported ages by the reviewer, meaning that there could be errors in inputting them or that guests could potentially lie about their age, there are no factors that would be actively pushing them to lie other than if they were below 18. However, we do not see a large push of 18 year olds in the graph, so that does not appear to be a problem in the dataset. 

The next area that was evaluated in our Exploratory Data Analysis was the amount of reviews per product category. I started with the creation of a graph with the Department Name variable on the x axis and the count of reviews per department name on the y axis. Looking at the graph created, most of the reviews are for tops, with dresses and bottoms coming in second and third. This means that the data isn’t evenly distributed between department types, but this does not tell us if more people are reviewing tops because they are bad quality or because they enjoyed their purchase.

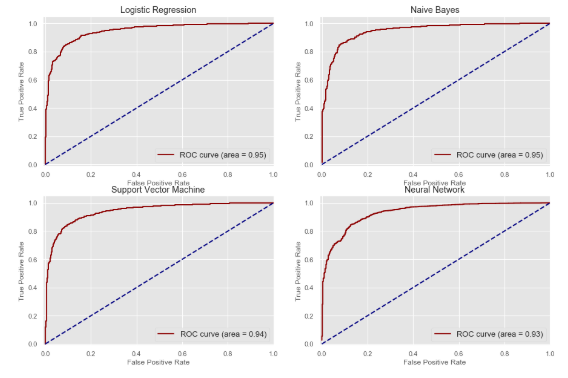
Moving forward, I added a length column by applying the len function to the variable “Review Text” in the ecom dataset. I then created graphs of the lengths of review by Rating category. These followed roughly the same distribution from Rating 1 to Rating 5, with there being far more long reviews than short reviews, and the reviews in the middle of the dataset following roughly a bell curve. I then created a boxplot to be able to easily compare averages and quartiles of the five rating categories with regards to length of reviews. The minimum and maximum of each rating were roughly the same, meaning the rating did not bear a huge affect on the length of the review. Ratings of 3 were more likely to be longer on average than other ratings, but the averages of all five were roughly the same. Reviews that bared a rating of 4 had a higher range of quartile length than the other categories, but there are no dramatic changes in the data found. After this evaluation, I moved into looking at the length of review by department. I created another boxplot of the review length by department name. This showed more variability in averages than when we divided the dataset by rating. Dresses and jackets tended to have longer reviews than items in the intimates or tops categories. There were no outliers in the dataset present in the boxplots. 

After this evaluation into length of review, I wanted to look into the recommended and not recommended reviews for each department. I created two separate data sets divided by the Recommended IND variable with 1 being recommended and 0 being not\_recommended. I then created a bar graph with this data to look at the recommended and not recommended items of each department. The intimates, jackets, and bottoms departments have a large percentage of positive recommendations versus the amount of not recommended reviews. However, tops and dresses have a large negative reaction, with dresses being particularly bad. 

Next, I want to look into possible correlations of the dataset variables. I did this with a simple corr function and then creating a heatmap with the variables. Through this I was able to see that not many of the variables are correlated, but there is a strong correlation between the Rating and the Recommendation IND variable. This is to be expected, as people who rate the item higher are more likely to recommend the item that they enjoyed using or owning. The next closest correlation is between the positive feedback count and the length, meaning that as length increases, the positive feedback count also increases. This correlation is low in significance, but it is interesting to keep in mind that longer reviews can garner more positive feedback.

I then looked into the length of reviews overall. I created a distribution plot of the length of reviews for the whole data set. This graph revealed the clear skew towards longer reviews, with the rest of the graph being a shallow bell curve. Looking into the descriptive statistics of the length variable, we are able to see that the mean of the dataset is 308 words, with the maximum being 508 words. This shows that most of the data is not consisting of small review texts that wouldn’t provide much insight to the study. For example, short reviews that just say “Great!” or “Love it!” might be positive feedback, but aren’t great feedback for the company overall and aren’t particularly useful to users who are looking at the reviews for information on the product. Having longer reviews can provide context to the rating, however it can lead to a more intensive process for marketing professionals to parse through reviews to gain insights.

After this evaluation, I moved into looking at how many times words were used in the dataset. I started with a CountVectorizer and added a word count column to the ecom dataset. I then selected words that were likely to hold sentiment as to what the user was feeling while writing. For example, this included words such as awesome, great, unhappy, wow, annoyed, upset, and hate. I then applied the function to the column and found that positive words such as love, great, super, and happy were used more often than negative words. The classes that people preferred to talk about were dresses and knits.

At this point it was time to move into the model creation, or creating the sentiment classifier. To start, I created a column in the dataset to show is the sentiment rated as negative or positive in the dataset. For the purpose of being able to classify reviews, I put reviews that were rated as 4 or 5 as positive (or True) and ratings that were 2 or lower as negative (or False). Reviews that contained a neutral ranking of 3 were excluded. I then split the dataset into a training and a test set. I ran the following analyses: Logistic Regression, Naive Bayes, Support Vector Machine, and Neural Network. I then added these results to the dataframe by first adding the prediction results to the training data. I then added the prediction probabilities of the models in a separate dataset so I would be able to evaluate those as well. After this I created AUC ROC curves to illustrate the diagnostic ability to check model performance. In addition, I created a confusion matrix to describe the performance of the classification model on the test data for which true values were known, which can be found in the appendix. It is after this evaluation that we could tell that Naive Bayes and Logistic Regression have the best results for the data used. As Naive Bayes took a shorter time to run, it can be argued that Naive Bayes is the better of the two models as it saves time for the amount of data being processed. 

**Results**

# The first research question that the study proposed was how often do common indicators of sentiment, such as love, great, happy, appear in the review text. We were able to determine that positive indicators of sentiment appear more often in the dataset with love appearing 8,951 times, great appearing 6,117 times, super appearing 1,726 times, and happy appearing 705 times. With most of the ratings being positive in the dataset, this is not much of a surprise as higher ratings tend to be associated with positive reviews which would contain positive indicator words.

The second question proposed in the study was how often do the product class names appear in the texts. The study was able to determine that dresses were mentioned 6,145 times, knits mentioned 4,626 times, blouses mentioned 2,983 times, and sweaters mentioned 1,380 times. When we combine this knowledge with the drop in recommendations and ratings for dresses, we can see this might be a trigger for negative word of mouth in the reviews.

The third proposed question is what are the common lengths of reviews as well as what is the range of reviews. The average length of review was 308.69 words, with a range from 9 words to 508 words. The review lengths were evenly distributed between the rating categories, meaning that rating did not have a correlation with the length of the review written. Knowing that the average is so long is helpful to know that the dataset will yield results, as most of the reviews have content beyond short reviews of “Loved it!” or “Hated it!” that wouldn’t yield much knowledge of why reviewers rated the way they did.

The fourth proposed research question is if there was any correlation between age and ratings. Our correlation analysis was able to determine the correlation between these two variables is only 0.03 and has no significant relationship. The only variables that had a significant relationship was between Rating and Recommendation Indicator, which is to be expected. Additionally, there was a low correlation between positive feedback count and the length of the review, meaning that more people found the review helpful as the review grew in length.

The research questions built up to the overall question of what is the best sentiment analysis model for the dataset, which we established was Naive Bayes, with Logistic Regression being a close second. Naive Bayes and Logistic Regression both had an ROC Curve of 0.95, however Naive Bayes was able to complete its run in 0.24 seconds, whereas Logistic Regression took 1.47 seconds. While this may not seem like a significant improvement, it could be a time saver when applied to a constantly updating dataset.

# Discussion & Conclusion

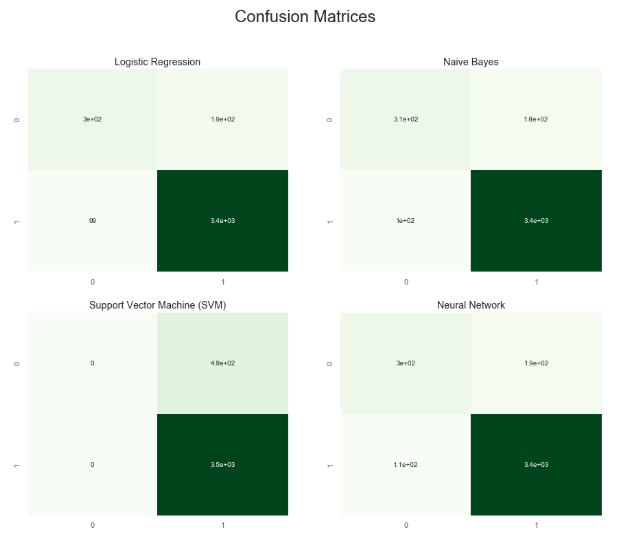
The aim of this study was twofold: I wanted a deeper understanding of the review dataset to see if there were any underlying trends and I wanted to determine the best model to determine sentiment in the reviews. These were successfully accomplished through the methodology and results above. Through this study, the e-commerce business has been given a functioning model as well as a better understanding of the information their reviews contain. The information discovered that offers the largest benefit to the company, which has been redacted from the dataset, is the application of the Naive Bayes sentiment analysis model to their reviews which can quickly identify negative reviews that would need feedback or monitored. These findings could help improve customer relations by giving opportunity to fix a negative guest experience through either offering a different item, replacing a damaged or broken product, or being able to answer questions or offer advice.

The models constructed, however, do have limitations in that we were only able to apply them to reviews that exist currently. To get the most benefit of this model, the company should apply it to incoming reviews on a timely basis, most likely being able to set a model that would trigger responses to marketing professionals. For further research, the dataset is well suited to other text mining predictions as well as text generation if there was an application that would benefit from auto-generating review texts.

# References

1. Agarap, A. F. M. & Grafilon P. M.. (2018, May 8). Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network. Retrieved from <https://arxiv.org/pdf/1805.03687.pdf>
2. Akter, S. & Wamba S. F.. (2016, Mar 16). Big Data Analytics in E-Commerce: A Systematic Review and Agenda for Future Research. *Electronic Markets, 26*(2), p.173-194. Retrieved from <https://link.springer.com/article/10.1007/s12525-016-0219-0>
3. Brooks, N. (2018). Women’s E-Commerce Clothing Reviews. Retrieved from <https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews>
4. Khan, A. G.. (2016). Electronic Commerce: A Study on Benefits and Challenges in an Emerging Economy. *Global Journal of Management and Business Research: B Economics and Commerce, 16*(1), np. Retrieved from <https://globaljournals.org/GJMBR_Volume16/3-Electronic-Commerce-A-Study.pdf>
5. Kaushik, K., Mishra, R., Rana, N. P., & Dwivedi, Y.K.. (2018, Nov). Exploring Reviews and Review Sequences on E-Commerce Platform: A Study of Helpful Reviews on Amazon.in. *Journal of Retailing and Consumer Services, 25,* 23-32. Retrieved from <https://www.sciencedirect.com/science/article/abs/pii/S0969698917307968>
6. Linzbach, P., Inman J J., & Nikolova, H.. (2019, Apr 24). E-Commerce in a Physical Store: Which Retailing Technologies Add Real Value? *NIM Marketing Intelligence Review, 11*(1), 42-47. Retrieved from <https://content.sciendo.com/configurable/contentpage/journals$002fnimmir$002f11$002f1$002farticle-p42.xml>
7. Liu, Z. & Park, S.. (2015, Apr). What Makes A Useful Online Review? Implication for Travel Product Websites. *Tourism Management, 47*, 140-151. Retrieved from <https://www.sciencedirect.com/science/article/abs/pii/S0261517714001903>
8. Manes, E. & Tchetchik, A. (2018, Apr). The Role of Electronic Word of Mouth in Reducing Information Asymmetry: An Empirical Investigation of Online Hotel Booking. *Journal of Business Research, 85,* 185-196. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0148296317305131>
9. Phan, Q. & Pilik, M.. (2017). Effect of Website Design on Positive Electronic Word of Mouth Intention: The Role Moderator of Gender. Retrieved from <http://m.publikace.k.utb.cz/handle/10563/1008706>
10. Roy, G., Datta, B., & Mukherjee, S.. (2018, July 24). Role of Electronic Word-of-Mouth Content and Valence in Influencing Online Purchase Behavior. *Journal of Marketing Communications, 25*(6), 661-684. Retrieved from <https://www.tandfonline.com/doi/full/10.1080/13527266.2018.1497681?scroll=top&needAccess=true>
11. Shahjee, R. (2016, Nov/Dec). The Impact of Electronic Commerce on Business Organization. *Scholarly Research Journal for Interdisciplinary Studies, 4*(27), 3130-3140. Retrieved from <http://oaji.net/articles/2017/1174-1484826380.pdf>

**Appendix: Confusion Matrices**

****