**Exploring plain-text reviews from Amazon with**

**natural language processing techniques**

Gerald Lee Jian Kai (gl2668)

6th May 2020

**Abstract**

Online reputation systems are especially important to digital marketplaces. They provide a formal channel for people to share and obtain information regarding products, services and other members from these platforms. Ratings and reviews help to reduce the information asymmetries that occur as a result of doing business over the internet and they facilitate a trust building mechanism between platform users. Nevertheless, several recent studies have found that numerical ratings from online reputation systems tend to be highly inflated and negatively skewed. This poses risks and challenges to the utility and reliability of the numerical ratings. Therefore, I hope to explore whether plain-text reviews can provide more accurate and less-inflated measures of customer reviews and experiences. I will be using plain-text review data from Amazon and applying natural language processing techniques, like sentiment analysis to analyze the reviews. Overall, the results demonstrate that sentiment scores of reviews produced with the AFINN lexicon can produce less inflated review scores but are limited in several ways. The results also demonstrate that it is possible to achieve an accuracy rate of 88% for classifying whether a product would be recommended or not based on the plain-text reviews. These findings are encouraging and suggest that there are potential benefits and opportunities for using plain-text reviews as a measure of customer satisfaction and product quality.

**Introduction**

Rating and review systems are indispensable to the digital economy because they offer a vital mechanism to mitigate the market failures of asymmetric information, which arise because of anonymity and unfamiliarity between transaction partners (Avery, Resnick, and Zeckhauser 1999; Bonsón Ponte, Carvajal-Trujillo, and Escobar-Rodríguez 2015; Ert, Fleischer, and Magen 2016; Nosko and Tadelis 2015; Teubner, Hawlitschek, and Dann 2017). Users of digital marketplaces do not have the opportunity to interact with one another in-person or engage directly, therefore, there are risks for doing business over these channels. Online reputation systems were designed and developed to reduce these risks by allowing platform users to provide feedback and share information about their experiences with products, services and other users. Other platform users can benefit from these ratings and reviews by screening and distinguishing between trustworthy and unscrupulous users (Nikhil and Ramesh 2019; Nosko and Tadelis 2015). This ultimately builds trust between users and for the platform itself. Platform users can also benefit from online reputation systems by using ratings and reviews to compare and assess the quality and level of satisfaction of products and services from these platforms (Nikhil and Ramesh 2019; Zervas, Proserpio, and Byers 2018). Overall, online reputation systems are highly valuable to digital marketplaces and they shape the consumer decisions of platform users (Filippas, Horton, and Golden 2018; Fradkin et al. 2015).

Despite their benefits, studies of online reputation systems have found that they tend to demonstrate issues with rating inflation, which undermine their reliability and utility (Cook 2015; Filippas, Horton, and Golden 2018; Hu, Zhang, and Pavlou 2009; Zervas, Proserpio, and Byers 2018). Studies of the largest digital marketplaces, like Uber, eBay and Airbnb, have found that numerical ratings tend to be highly negatively skewed and overwhelmingly positive across digital marketplaces. The studies assert that it is unlikely for these ratings to correspond with equally high frequencies of positive user experiences in reality. Hence, the inflated ratings misinform users and prevent them from reliably evaluating the true quality of the subjects being reviewed. Nosko and Tadelis (2015) suggest that this would lead to a breakdown of trust between platform users and those with a bad experience as a result of the inaccurate ratings would be less inclined to engage in future transactions at these platforms.

This study aims to contribute to the extant literature by studying the plain-text reviews from Amazon and investigating if they can provide less inflated and unreliable feedback. This study is also interested to see if plain-text reviews can be used to predict whether a product should be recommended or not recommended based on their 5-stars ratings. Amazon was chosen as the case study for this paper because the digital marketplace allows its users to provide text-based reviews and numerical 5-stars ratings for the products or services that are marketed over the platform. The extant literature has predominantly focused on numerical ratings so this research will be one of the few to analyze the distribution of plain-text review. I aim to apply sentiment analysis to the plain-text review to extract the opinions and sentiment of reviews so that they can be compared to the numerical ratings. Subsequently, I hope to explore the use of a logistic regression classifier to determine whether a product would be recommended or not based on its plain-text reviews. I also attempted to apply topic modelling to the plain-text reviews using Latent Dirichlet Allocation but achieved limited results.

**Methodology**

Natural language processing techniques, like sentiment analysis and topic modelling have been more frequently utilized in recent years across various domains to extract opinions and analyze large bodies of textual data (Bing, Minqing, and Junsheng 2005; Lawani et al. 2019). This study will be employing sentiment analysis to extract the opinions of customers from Amazon product reviews in order to derive a polarity score that denotes if a review was positive or negative. The method for doing so is to identify parts of the plain-text review that denote a sentiment or opinion based on a provided dictionary or lexicon of words or phrases (Luo 2018; Nasukawa and Yi 2003). Lexicons have been developed to specify whether certain words or phrases belong to a cluster of emotions or a positive, negative or neutral sentiment (Thet, Na, and Khoo 2010). These lexicons typically contain a list of positive and negative polar words with their associated polarity score or emotional category. In this study, I will be using the AFINN lexicon, which was developed by Nielsen (2011). The AFINN lexicon is a lexicon based on unigrams (single words) and has 2,477 terms. It offers a more granular representation of sentiment because it has a wider grading spectrum from -5 to +5, which would be more responsive in identifying the intensity of sentiment for each review. In the AFINN lexicon, the negative scores would indicate negative sentiments and vice versa for the positive scores. To perform the sentiment analysis in this study, the words used in each review would be assigned a polarity score based on the AFINN lexicon, and the total score of the review would be the sum of the scores of the words in that review.



To illustrate the sentiment analysis methodology used, I will use the following product review as a demonstration: “sewing poorly done sleeve tight body loose raise arm neck area shirt looked horrible bad sewing”. In this review, the text has already been pre-processed and cleaned (the full process is detailed in the data section of this paper). The words that are matched to the AFINN lexicon are ‘poorly’ (-2), ‘loose’ (-3), ‘horrible’ (-3) and ‘bad’ (-3). The sum of attributed positive scores (0) and attributed negative scores (-11) for this review is (-11), which denotes a negative review.

In the second part of this study, I will be using a logistic regression classifier to determine whether a product would be recommended (4-stars and above) or not recommend (3-stars and below) based on its plain-text reviews, which have been pre-processed and cleaned. To analyze the plain-text reviews, I carried out text feature extraction using a bag of words model, which transformed all the pre-processed and cleaned reviews into a dictionary of either uni-grams (singular words) or bi-grams (two words) that appear across the reviews. The output of this process is a vector format that has the numerical count of each word in the review. Subsequently, I split the dataset into a training and testing set (80% and 20%) and utilized a 5-fold cross validation method to further split the training dataset so that the effects of over or underfitting can be mitigated. The training datasets would be used to train the classifier while the testing dataset would be used to evaluate the trained classifier.

The classifier algorithm is based on a logistic regression model, which works by taking the text vector of the plain-text reviews and evaluating the coefficients or weights for each input variable, which in this case is either the uni-grams or bi-grams. The model then learns which input variables are most useful for discriminating whether a product is recommended or not recommended. A confusion matrix is produced to evaluate and visualize the performance of the model. The logistic regression classifier is evaluated on the testing data, which has not been used to train the model. The confusion matrix would compare the true positives, false positives, true negatives and false negatives. The confusion matrix provides parameters, like accuracy, precision and recall. Accuracy denotes the number of correctly classified reviews (true positives + true negatives) divided by the total number of reviews (true positives + false positives + true negatives + false negatives). Precision is the total number of reviews correctly identified as positive (true positives) divided by the total number of reviews identified as positive (true positives + false positives). Recall is the total number of true positives divided by the total number of positive predictions (true positives + false negatives).

In this study, I also attempted to incorporate a Latent Dirichlet Allocation (LDA) model to see if it is possible to categorize the plain-text reviews into different models without a supervised learning methodology. LDA is a generative probabilistic model of text corpus, which assumes that that the reviews are represented by random mixtures of latent topics where each topic can be characterized by a distribution of words (Blei, Ng, and Jordan 2003). A detailed explanation for this method has been provided by Blei, Ng, and Jordan (2003). The model that was built in this study used 10 topic models, set to 10,000 maximum features after ignoring features that appeared in more than 50% of the reviews. For the LDA model, it is generally recommended to remove common words because they would over influence the model and analysis. The initial results from this model were limited because it was difficult to differentiate between topics since the words within each topic clusters were very similar. As a result, this study will not focus on the results of the LDA analysis and their contributions to the extant research.

**Data**

The data used in this study are from product reviews on Amazon marketplace (United States). I decided to analyze Amazon’s online reputation system because it offers users the ability to provide both numerical ratings and text reviews for products and services. The dataset was obtained from an Amazon review data repository that was created and managed by Jianmo Ni. The reviews and metadata were web scrapped from the digital marketplace between May 1996 to October 2018 (Jianmo 2018). There are 233.1 million reviews on the repository and the dataset includes a number of variables, like ratings, text reviews, votes for helpfulness of review, price of object, reviewer name, etc.

I took a subset of the Amazon reviews because it would be too computationally intensive to analyze 233 million reviews, and also because I wanted to study ratings and reviews that were related to a homogenous product category. This was because, products and services can appeal very differently to different people, which would result in drastically different ratings and reviews. Customer expectations for a car and a pencil would be fundamentally distinct and attributes of the products would matter very differently to each individual. Thus, I decided to only look at reviews under the ‘Amazon Fashion’ category. Amazon produces and sells its own brand-name fashion products, like shirts, shorts, socks and leggings. They are budget alternatives for these products and are relatively homogeneous compared to other product categories. Due to their mass-production and budget image, consumers’ expectations of these products should be relatively consistent. The data subset that I retrieved contained the ratings and reviews for only the product category of Amazon fashion. There were 883,636 reviews and 6 variables:

* rating – five-stars rating
* reviewTime – time and date of reviews
* reviewerName – the name of the person who gave the review
* review – the plain-text review
* summary – the header of the review
* vote – the number of public votes for whether the review was helpful or not

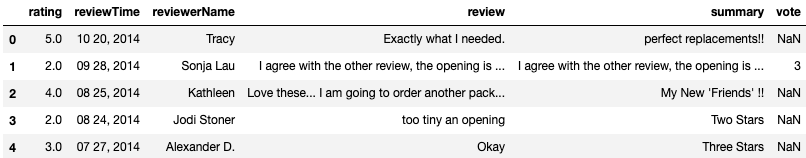


Table 1: Example of Dataset

I dropped four variables from the dataset because they had missing values or were irrelevant to the study. The vote variable was removed because it had a high percentage of missing data since only a small percentage of reviews were voted by other users as helpful. The reviewer name, review time and review summary were also irrelevant to this study and hence removed.

From the dataset, I visualized the univariate distribution of the numerical ratings and found that numerical ratings for Amazon fashion products followed the same pattern of rating inflation observed across other digital marketplaces, like Uber, Airbnb and eBay (Cook 2015; Filippas, Horton, and Golden 2018; Hu, Zhang, and Pavlou 2009; Zervas, Proserpio, and Byers 2018). In figure 1, the histogram highlights the negative skew (-1.01) in numerical ratings with a heavy concentration of ratings in the 5-stars category. 53 % of the ratings for Amazon fashion products from May 1996 to October 2018 were 5-stars and 16 % of the ratings were 4-stars. 1-star, 2-stars and 3-stars ratings cumulatively only made up 29 % of the overall ratings. The mean of the ratings was 3.91-stars and the median of the ratings was 5-stars. As suggested by the extant literature, it is possible but not realistic for every product and customer experience to be above the average. Rating inflation is the more likely explanation for why a majority of ratings are so high.

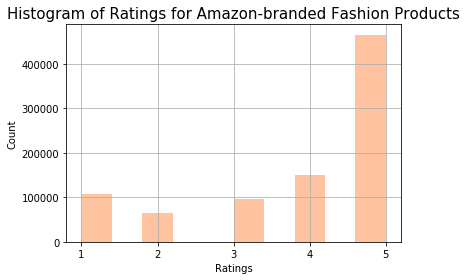


Figure 1: Histogram of Ratings for Amazon-branded Fashion Products

To analyze the plain-text feedback, it was necessary to clean and prepare the corpus of plain-text reviews. The following processes were taken to wrangle, pre-process and parse the text data for sentiment analysis:

* Removal of Punctuations and Digits
* Removal of English stop words
* Transformed all text to lowercase
* Split the sentences into tokens
* Stemmed the tokens with lemmatization

|  |  |
| --- | --- |
| Plain-text Feedback before Data Cleaning | Plain-text Feedback After Data Cleaning |
| Exactly what I needed. | exactly i needed |
| Too tiny an opening. | tiny opening |
| Love these… I am going to order another pack. | love i going order another pack |

Table 2: Sample of plain-text reviews before and after cleaning

**Results and Discussion**

The results from this study demonstrate that it is possible to obtain a less inflated review metrics by extracting sentiment scores from text-based reviews using the AFINN lexicon. The sentiment scores for Amazon-branded fashion products were less negatively skewed than the numerical ratings (as seen in figure 1) and appeared to follow a more normal distribution (as seen from the histogram in figure 2). The mean and median of the Sentiment Score were 4.66 and 3.0 respectively, which are both further away from the highest possible score. In general, the Sentiment Score for Amazon reviews tend to be closer to the neutral (0) to slightly positive (4) range.

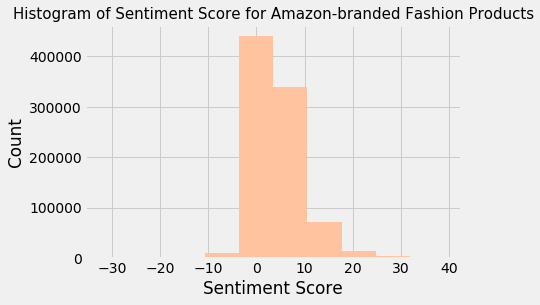
****

Figure 2: Histogram of AFINN Sentiment Score for Amazon Reviews

However, the skew and kurtosis of the sentiment score highlights that this measure of customer review is imperfect. The skew for sentiment score was 8.53 (highly positively skewed) and the kurtosis was 1118. This is the opposite from the 5-stars ratings, which was negatively skewed and suggests that the sentiment analysis algorithm is picking up a small percentage of extremely positive sentiment scores that are significantly above the mean. An explanation for this is that longer worded reviews could have more positive words and therefore contribute to much more extreme sentiment scores. To mitigate this in the future, it would be useful to divide the sentiment scores by the total number of words that were matched to the lexicon so that the word count of reviews can be controlled for.

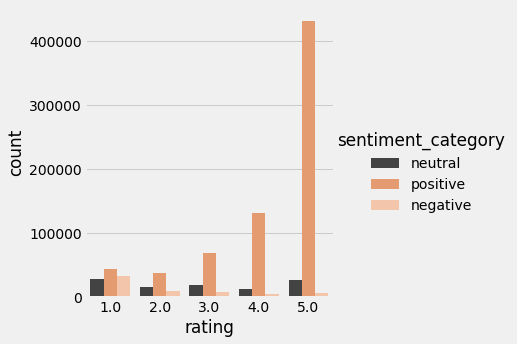
****

Figure 3: Visualizing AFINN Sentiment Score and Numerical Ratings for Amazon Reviews

Figure 3 and figure 4 also highlight that while the sentiment analysis algorithm is generally working properly, there are some limitations. From Figure 4, we can see that there is a generally positive relationship between the 5-stars numerical ratings and the sentiment scores from plain-text reviews. This is expected as people would tend to write more positively worded reviews for products that they have rated highly. However, in figure 3, we can see that there are instances where low-rated products (with 1-star) and positive sentiment. We can also see negative and neutral sentiment scores for products that were rated the highest rating (5-stars). In table 3, I have highlighted 2 examples of reviews that attributed a 5.0 and 6.0 sentiment score (2nd column from the right). The reviews were given a 1-star (2nd column from the left) and the plain-text reviews (5th column from the left) appear to be negative. In these situations, the bag of words model and AFINN lexicon cannot recognize complex linguistic structures and phrases, like “not big fan”, and thus attribute these reviews as positive. These examples highlight the inherent limitations of natural language processing techniques.

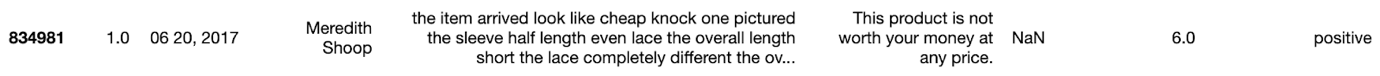


Table 3: Sample of Problematic Sentiment Scores

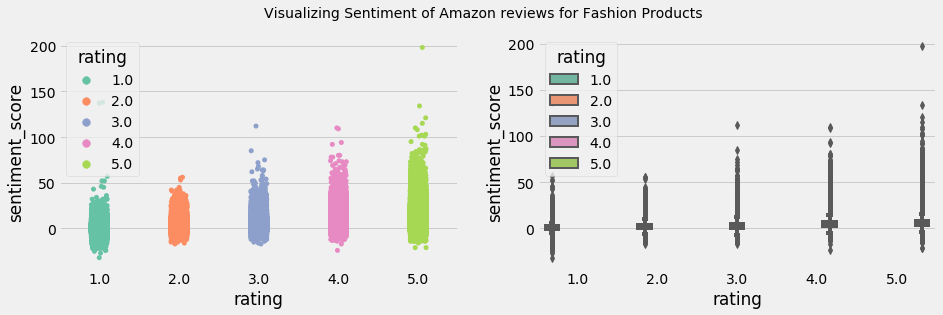
****

Figure 4: Visualizing AFINN Sentiment Score and Numerical Ratings for Amazon Reviews

In the second part of the study, I used a logistic regression classifier to distinguish whether a product would be recommended (4-stars and above) or not recommend (3-stars and below) based on the plain-text reviews. The results demonstrated that using the plain-text reviews from the training and testing data could produce a high accuracy score of 88% as seen from the confusion matrix in table 4. From the confusion matrix, we can also see that the model is better at determining Recommended products than Not Recommended products. The precision and recall rates are higher for the Recommended products than the Not Recommended products, suggesting that there are more false positives and false negatives in the Not Recommended predictions than Recommended predictions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| Not Recommended | 0.84 | 0.75 | 0.79 | 66,986 |
| Recommended | 0.90 | 0.94 | 0.92 | 152,923 |
| Accuracy Score |  |  | 0.88 | 220,909 |
| Macro Average | 0.87 | 0.84 | 0.85 | 220,909 |
| Weighted Average | 0.88 | 0.88 | 0.88 | 220,909 |

Table 4: Confusion Matrix of Logistic Regression Classifier Model

To understand how and what the logistic regression classifier model learned from the data, we can also look at the top 25 largest and smallest coefficients for the model. From figure 5, we can see the coefficients and their corresponding features that contribute to how the model would determine whether a product should be recommended or not. The words under the red bars highlight the uni-grams and bi-grams that signal whether a product that would be not recommended while the words under the blue bars signal a product that would be recommended. The result is as expected. Negative words like “poor”, “disappointing” and “garbage” denoted a product that would not be recommended and positive words like “excelent”, “awesome” and “perfecto” denoted a product that would be recommended. Overall, the results from the logistic regression classifier were encouraging and highlight that it is possible to use a similar model to predict future plain-text reviews. In future studies, it would be recommended to improve the model performance by applying different text pre-processing functions to the raw plain-text reviews, using tf-idf features or using other machine learning models.

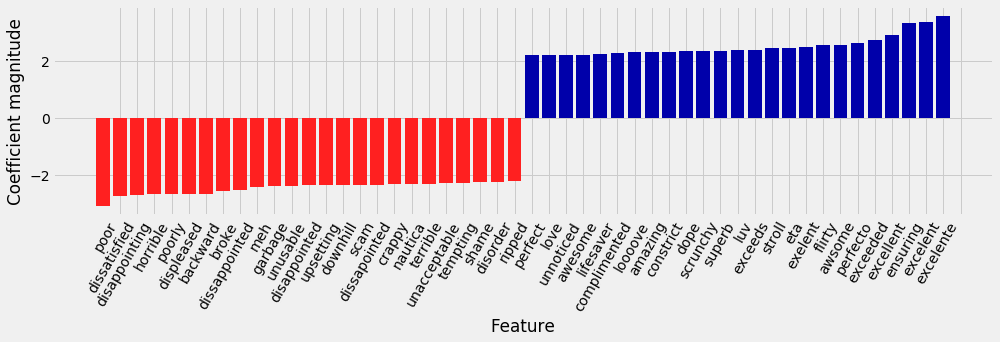
****

Figure 5: Largest and Smallest Coefficients for Logistic Regression Classifier of Amazon Reviews

This study also explored the use of the Latent Dirichlet Allocation model to find groups of topics that cluster together frequently. The study’s hypothesis was that it would be possible to find clusters of topics from the plain-text reviews based on different sentiments, opinions or subjects related to the product reviews. For instance, there might be a cluster related to poor sizing and fit, and another cluster related to materials. The LDA results turned out to be limited because the features in the different topic models were very broad and similar, which made it difficult to differentiate between distinct topics. In table 4, you can see the top 10 words for each topic. There are words that are repeated across topics, such as how size and fit are present in topics 0, 1, 2, 4, 6, 8 and 9. Across the 10 topics, we can see that comfort, material, size and fit tend to be the main broad topics across the reviews although it is difficult to differentiate between them. 10 topics might have been too little to capture the different types of reviews in the dataset, which explains why the words in each topic are broad and repetitive. For future analysis, it is recommended to test the LDA model on larger sizes of topics to see if the model can extract topic models that have more distinct features. LDA is an unsupervised learning method and its findings are ultimately limited due to the lack of category or topic labels in the dataset. However, they can be useful to digital marketplaces that have more distinct topic categories, like different product categories.

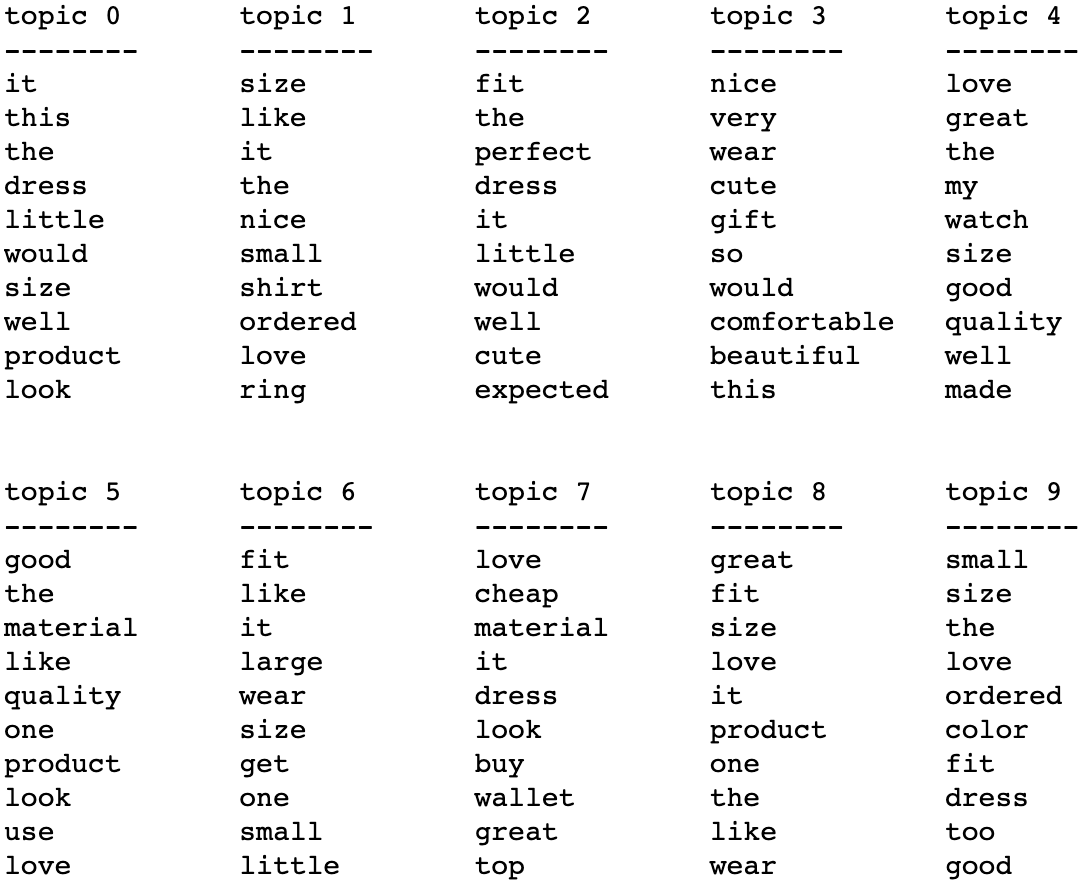
****

Table 4: Top 10 words for each LDA topic model (10 components)

**Conclusion**

In conclusion, this study has highlighted that there are numerous interesting and useful insights that can be obtained by analyzing plain-text reviews from online reputation systems. This study contributes to the extant literature by demonstrating that sentiment analysis based on the AFINN lexicon can extract less-inflated sentiment scores for Amazon-branded fashion products. This can be a useful measure of product quality and customer satisfaction that would work alongside numerical ratings. However, there are limitations to the use of sentiment analysis due to its inability to capture more complex linguistic patterns. This study also highlights that it is possible to use a logistic regression classifier on the plain-text reviews to distinguish between recommend and not recommend Amazon-branded fashion products at an accuracy rate of 88%. Future work can look into refining and improving the model and data to produce a more accurate model, which could have multiple uses in digital marketplaces. This study’s exploration of using an LDA model to distinguish topic models amongst plain-text reviews, highlight the challenges of using unsupervised learning methods on text data. Ultimately, these findings add to the literature of online reputation systems and suggest that there are potential benefits and opportunities for using plain-text reviews as a measure of customer satisfaction and product quality.

**Bibliography**

Avery, Christopher, Paul Resnick, and Richard Zeckhauser. 1999. “The Market for Evaluations.” *American Economic Review*.

Bing, Liu, Hu Minqing, and Cheng Junsheng. 2005. “Opinion Observer : Analyzing and Comparing Opinions on the Web.” *Proceedings of the 14th international conference on World Wide Web*.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research*.

Bonsón Ponte, Enrique, Elena Carvajal-Trujillo, and Tomás Escobar-Rodríguez. 2015. “Influence of Trust and Perceived Value on the Intention to Purchase Travel Online: Integrating the Effects of Assurance on Trust Antecedents.” *Tourism Management*.

Cook, James. 2015. “Uber ’ s Internal Charts Show How Its Driver ­ Rating System Actually Works.” *Business Insider*.

Ert, Eyal, Aliza Fleischer, and Nathan Magen. 2016. “Trust and Reputation in the Sharing Economy: The Role of Personal Photos in Airbnb.” *Tourism Management*.

Filippas, Apostolos, John Joseph Horton, and Joseph Golden. 2018. “Reputation Inflation.” In *ACM EC 2018 - Proceedings of the 2018 ACM Conference on Economics and Computation*,.

Fradkin, Andrey, Elena Grewal, David Holtz, and Matthew Pearson. 2015. “Bias and Reciprocity in Online Reviews: Evidence From Field Experiments on Airbnb.” In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*,.

Hu, Nan, Jie Zhang, and Paul A. Pavlou. 2009. “Overcoming the J-Shaped Distribution of Product Reviews.” *Communications of the ACM*.

Jianmo, NI. 2018. “Amazon Review Data.” https://nijianmo.github.io/amazon/index.html.

Lawani, Abdelaziz, Michael R. Reed, Tyler Mark, and Yuqing Zheng. 2019. “Reviews and Price on Online Platforms: Evidence from Sentiment Analysis of Airbnb Reviews in Boston.” *Regional Science and Urban Economics*.

Luo, Yi. 2018. “What Airbnb Reviews Can Tell Us? An Advanced Latent Aspect Rating Analysis Approach.” *Graduate Theses and Dissertations*.

Nasukawa, Tetsuya, and Jeonghee Yi. 2003. “Sentiment Analysis: Capturing Favorability Using Natural Language Processing.” In *Proceedings of the 2nd International Conference on Knowledge Capture, K-CAP 2003*,.

Nielsen, Finn Årup. 2011. “A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs.” In *CEUR Workshop Proceedings*,.

Nikhil, Garg, and Johari Ramesh. 2019. “Designing Informative Rating Systems: Evidence from an Online Labor Market.” : 36.

Nosko, C., and S. Tadelis. 2015. “The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment.” *NBER Working Paper Series*.

Teubner, Timm, Florian Hawlitschek, and David Dann. 2017. “Price Determinants on Airbnb: How Reputation Pays off in the Sharing Economy.” *Journal of Self-Governance and Management Economics*.

Thet, Tun Thura, Jin Cheon Na, and Christopher S.G. Khoo. 2010. “Aspect-Based Sentiment Analysis of Movie Reviews on Discussion Boards.” *Journal of Information Science*.

Zervas, Georgios, Davide Proserpio, and John Byers. 2018. “A First Look at Online Reputation on Airbnb, Where Every Stay Is Above Average.” *SSRN Electronic Journal*.