Final Report

Team 16 ATL Go Jackets

Fall 2018

Identification of Constructive Sentiment from Potentially Biased Reviewer Remarks

April Draper
aprildraper@gatech.edu

Divya Paduvalli
dpaduvalli3@gatech.edu

Jake Rocker
jacob.rocker@gatech.edu

Priya Sampathkumar
priyad11@gatech.edu

Table of Contents

Introduction-Motivation

Product reviews often include purchasing experience information in addition to product satisfaction. It is difficult to truly evaluate if a review is accurate. Consumers must read through multiple reviews to ascertain whether a product's rating is an accurate reflection of the reviews. One cannot tell whether the reviewer is equally responsive for both positive and negative experiences. Some reviewers may have a positive or negative bias and may post reviews more frequently in such a manner rather than posting equally in both situations (16). Identifying a reviewer's bias would be cumbersome without automated analysis. To aid consumers assessment of online reviews, creation of a web application to find bias of reviewers on products is proposed. The user will select a product and our application gathers reviews of that product. An analysis of the general sentiment of the reviews will be completed. The user will be presented with a summary indicating the positive and negative reviews with the potential bias of the reviewers. A word cloud of most frequently occuring tokenized words will also be presented to give users a quick overview of the commonly occuring themes that reflect emotions.

Problem Definition

Retail sites gather customer experience reviews but do not share insights on reviewers. Some reviewers are inherently biased towards positivity or negativity (16). Consumers rely on honest reviews to evaluate the potential usefulness of a product. If these reviewers are negatively biased, their reviews will be overly negative and may skew the results of the overall rating for a product. Based on their bias, their review may hinder rather than help consumers properly evaluate a product.

Survey

Sentiment analysis, a subfield of natural language processing (NLP), refers to the automated extraction of feeling or opinion from a text document. It is most popularly used for analyzing online reviews. Techniques for sentiment analysis can be broadly classified as lexicon based or machine learning based (12,13). Alternatively, a hybrid of the two methods can be used, which will be our approach. The learned part of the model will use a recurrent neural network (RNN).

Some challenges in sentiment analysis lie in the training data used in models. There is a disparity between a reviewer's rating and the overall review's sentiment (6). We plan to use R to write suitable algorithms to identify positive/negative sentiments (1). Sentiment analysis has been explored to research whether product reviews reflect the product sales, but usually just from one manufacturer (2). Our project will compare the reviews from multiple manufacturers.

To study the effect of online reviews on sales, sentiment analysis performed on books from Amazon.com and BarnesandNoble.com found that Amazon had a more sophisticated, user friendly interface for customers to express their opinions, which impacts sales (10). Sentiment analysis along with ARIMA was applied on a car sales platform, the results showed a high accuracy for predicting car sales (8). Online reviews are part of electronic Word of Mouth (eWOM) which greatly impacts consumer purchasing decisions (9). Research shows that positive reviews positively affect sales (3). The manufacturer will gain an understanding of the overall customer sentiment for their products on which they can base their decision of whether to manufacture the product. Furthermore, manufacturers may choose to charge more for a guaranteed positive customer experience.

Proposed Method

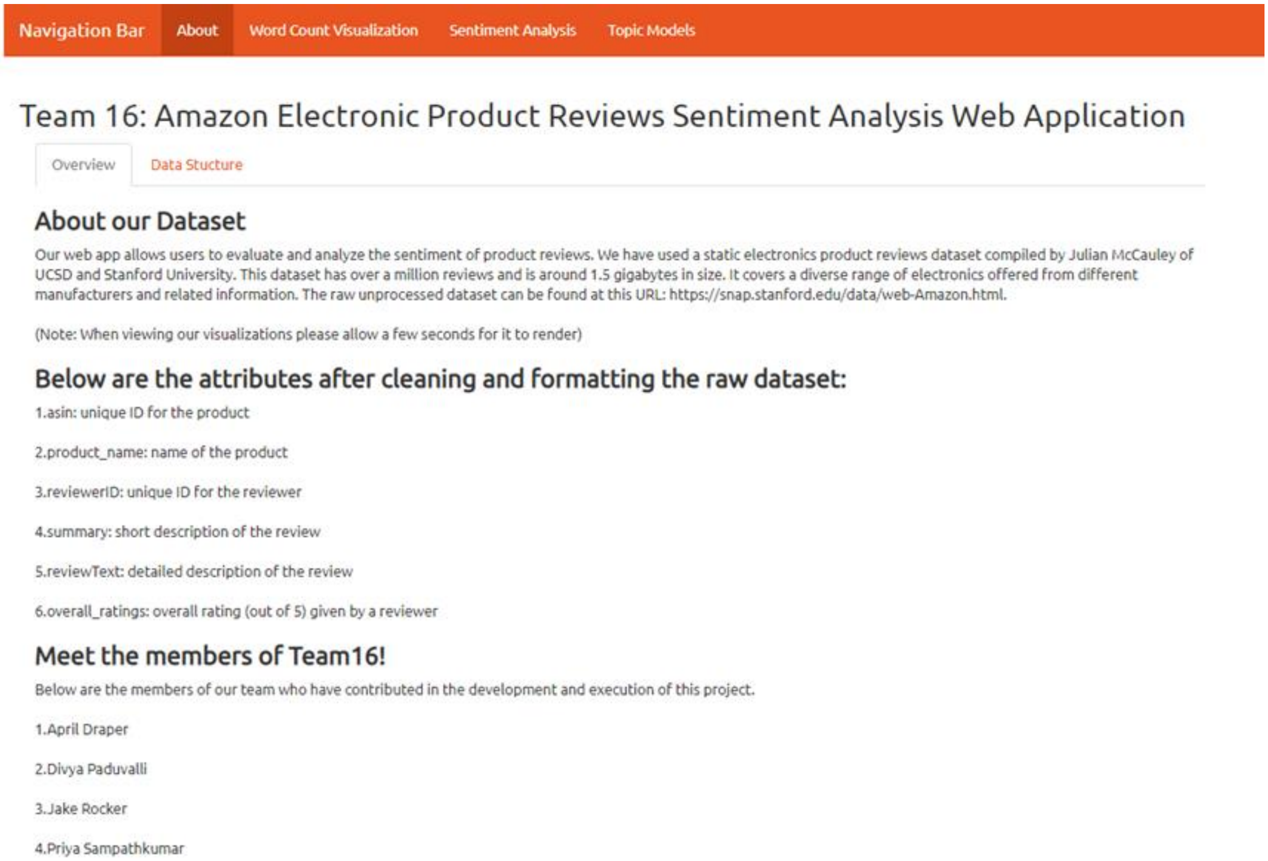
Our initial goal was to scrape all the product reviews under the category of laptop from the websites of Amazon, Walmart, and Target. Unfortunately, neither of these websites provide an API to do this. There are around 30,000 search results and more than 400 pages for this category. Scraping using Python is not feasible due to potentially getting blocked for sending multiple HTTP requests. We instead used a static electronics product reviews dataset compiled by Julian McCauley of University of California, San Diego and Stanford University (17). This dataset has over a million reviews and is around 1.5 gigabytes in size. It covers a diverse range of electronics offered from different manufacturers and related information. We performed our analysis on multiple electronic products and retailers.

Although sentiment analysis has been done many times, our innovations are in how we are grouping the analysis and presenting it to the user. One innovation in our project is identifying potentially biased reviewers. We will be allowing users to evaluate the validity of a review by revealing the potential bias of the reviewer. Another innovation is identifying products that negative and positive reviewers are more likely to purchase. This could allow manufacturers to identify products they may want to use as models for future innovations. Allowing users to confirm the validity of a review by checking the reviewer for potential bias is a new idea. This allows the user to make decisions based not only on the average reviews, but also whether the reviewers have a potential bias.

Tokenization is one of the methods we employed. A token is a meaningful unit of text, most often a word, that we are interested in using for further analysis, and tokenization is the process of splitting text into tokens (15). We used this approach to create a bag of words from our reviews. Long strings were split into smaller tokens and the order of the words was ignored. We kept tokens that were useful for our analysis by removing punctuation, converting words to lowercase, and remove insignificant words (ex: a, an, and etc). We also stemmed words since families of related words (ex: like, liked, likeful etc) with similar meaning can be considered as a single unit by reducing words to their stem base or root form. In order to visually describe the overall theme of the reviews and also to understand what the major key words are, we have included word count analysis, word count chart, word cloud etc.

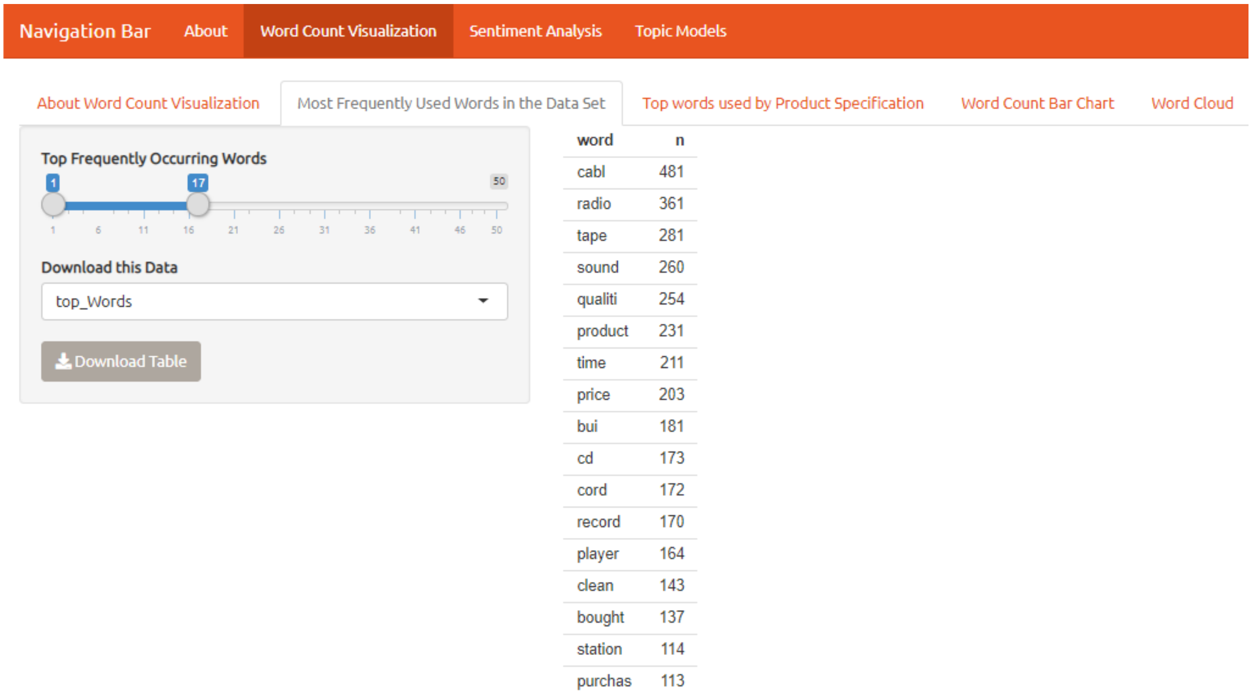
In order to quantitatively predict the sentiment of each token, we used the pre-built R dictionaries, mainly AFINN, Bing, and NRC. For each token, AFINN provides the sentiment score (-1,-2), Bing provides the binary sentiments (positive/negative), and the NRC provides the emotion (fear, happy). Since every document is a mixture of topics and every topic is a mixture of words, we used Latent Dirichlet Allocation (LDA) model which uses groups we have identified to classify commonly occuring topics in the reviews.

We used Rstudio to analyze our cleaned dataset and performed sentiment analysis and used Shiny to build our user interface (UI).

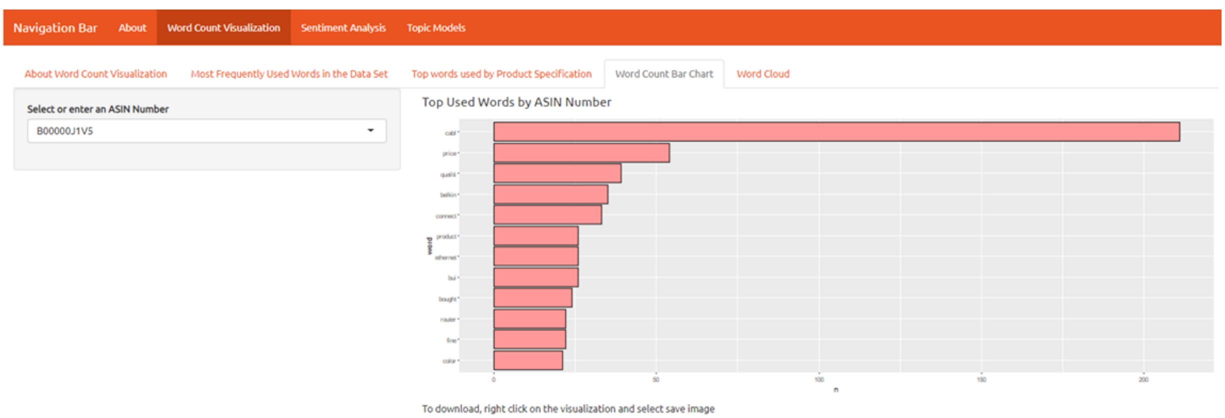


Our UI is a website that users can navigate through to explore the sentiment analysis of the reviews and view the overall bias (if present) of reviewers. There are multiple tabs to explore, giving the user more control to view the results in a way that helps the individual user easily understand the sentiment of the review and reviewer. The users can also download visualizations and tables that they find useful.

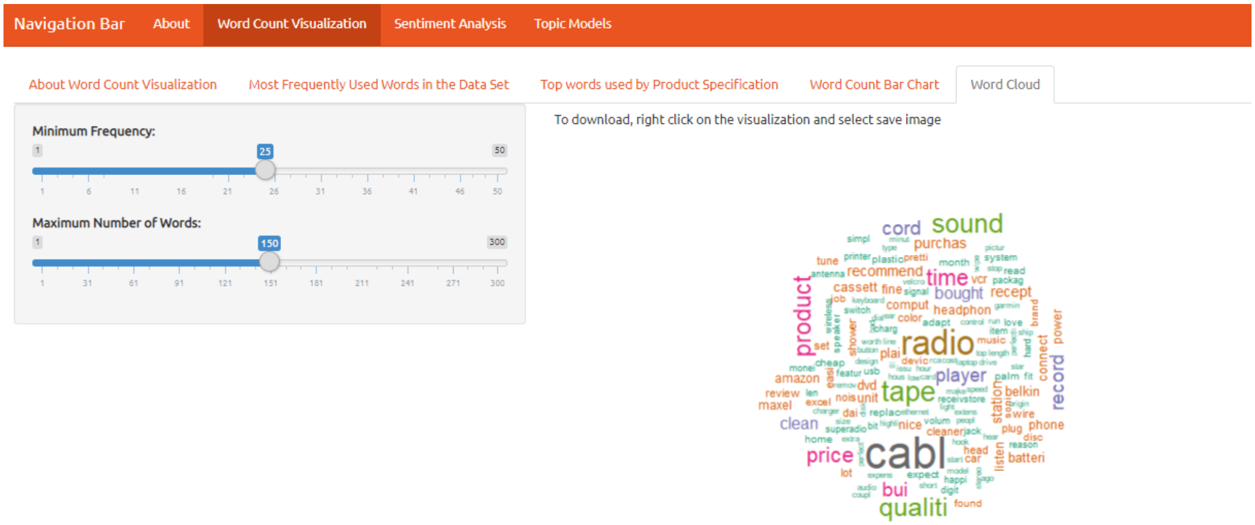
The first tab, Word Count Visualization, allows users to view most frequently used words in the reviews. They can select how many of these words to include in their list.



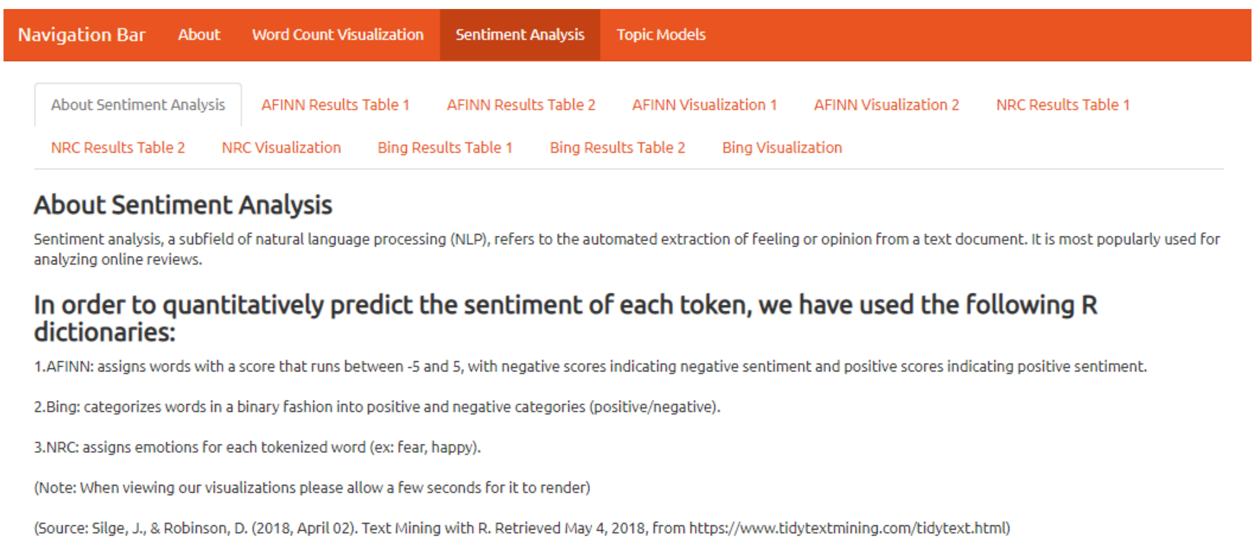
Users can also look at the most frequently used words by asin number.



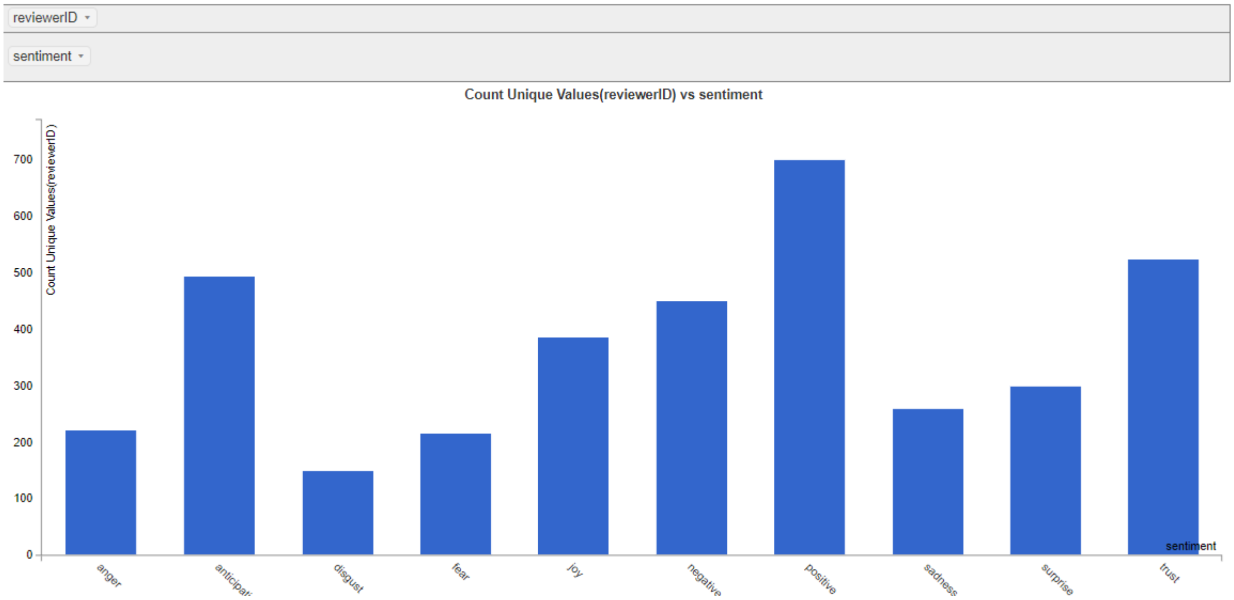
Finally, under the Word Count Visualization tab, users can create a word cloud. This visual representation of the words used in the reviews gives users a quick view of the most frequently used words in a review. Larger words appear more often and smaller words are used less often. The user has the freedom to choose how many words to use in the word cloud as well as the frequency of the words included. For example, the user can select twenty-five for the minimum frequency, meaning the words included in the word cloud must be used in the reviews at least twenty-five times. They could select one hundred fifty for the maximum words included in the word cloud. This would create a word cloud of the top one hundred fifty words that occured at least twenty-five times in the reviews.



The next tab, Sentiment Analysis, has multiple interactive tables and visualizations for users to explore to view the sentiment analysis for certain products. If they are looking for reviews on a particular product, they can select that product to view the analysis in a graph, bar chart, or table.



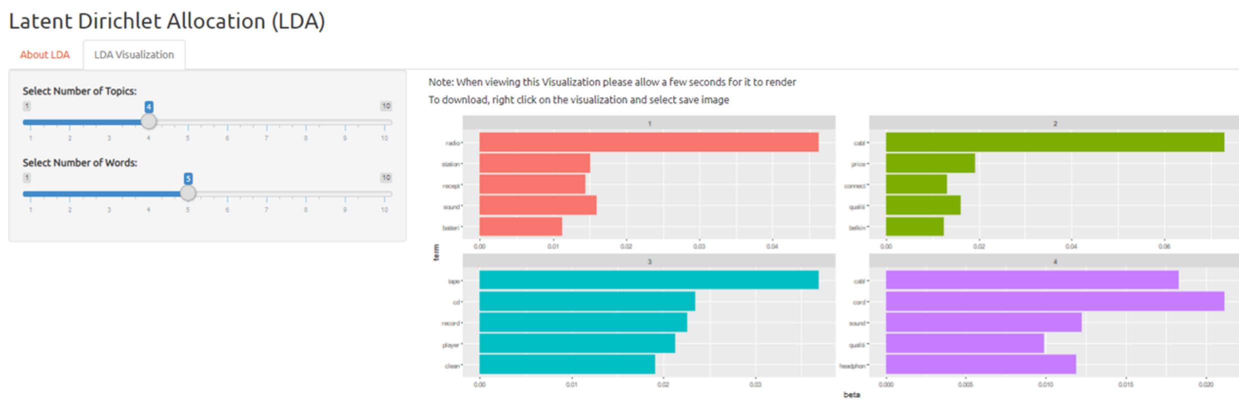
A bar chart can be created to give a sense of reviewer bias. The chart shows the different sentiments expressed and how often they occur in the review. Feelings such as anger, disgust, fear and sadness express a negative view while anticipation, joy, surprise and trust are positive. These feelings are concluded from the sentiment analysis and used to give the review an overall positive or negative bias. The chart below is an example of what the user could create and view. This chart shows an overall positive review, but does not appear to be extremely biased in that there is negative emotion expressed as well. The user could conclude that this reviewer gives a balanced review of products.



Another chart users could create is a heatmap. Heatmaps allow for quick interpretation of results because the intensity of their colors correspond to information in the data. For example, the higher a number in the data, the darker, more intense the color on the heatmap and vice versa.

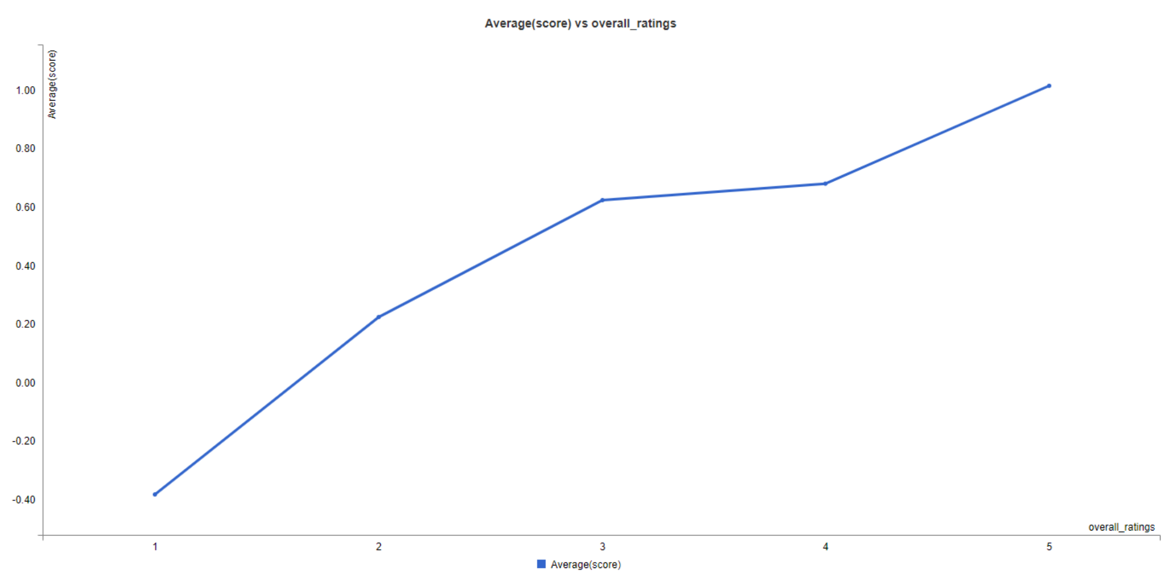
asin \ sentiment	negative	positive	Totals
B00000J061	115	133	248
B00000J1V5	90	126	216
B00000DM9W	69	101	170
B00000J1SC	46	70	116
B000001ON6	23	85	108
B00000J1QR	36	64	100
B00000J1QK	33	63	96
B000001OL6	43	38	81
B000001OM5	31	50	81
B000001OM4	41	37	78
B00000J1EJ	44	26	70

The Topic Models tab presents the Latent Dirichlet Allocation or LDA of the reviews. Users can select the number of topics (up to ten) to split the review into and how many of the most frequently used words to include in the graph (up to ten). A horizontal bar graph is created for each topic the user selects with one bar for each word in the topic. LDA finds the mixture of words that is associated with each topic while determining the mixture of topics in each review. Users can view these results to help them understand the different topics within a review along with the most frequently used words in each topic. This is another tool to give the user a bigger picture of the overall meaning and evaluation of the reviews.



Experiments and Evaluation

We visualized the distribution of the resulting sentiment classes (positive, negative etc) of our model to identify if there was extreme skewness. Since there was an imbalance of the sentiment classes, we bootstrapped, a resampling method, to equalize the baseline accuracy. To measure the accuracy of our model, we split ~70% of our data for training and ~30% for testing. We used the training set to fit and tune our model and a testing set to evaluate model accuracy. We used k-fold cross validation, where the training set was further split into k = 10 (we decided to use 10 folds since this is a commonly used approach in Analytics) subsets where we trained on 9 subsets and tested on one subset. This approach yielded us 10 scores which was averaged (average score/accuracy = 91%). In addition, we also used regression to understand if there was a correlation between the resulting sentiment scores and review ratings. There was a strong correlation between the sentiment scores and review ratings (correlation coefficient R = .86).



The graph above shows the results of comparing the predicted sentiment from our model with the actual ratings of the review. The line suggest a high correlation between our prediction and the actual rating, giving us confidence in our model at predicting the sentiment of a review along with any reviewer bias.

Experiments performed on our model include searching for specific keywords or products and evaluating the analysis returned in the interactive dashboard. The recommendation returned by the dashboard suggests the bias of the reviewer. Once we built our model, we compared its performance with the built in dictionaries in R. We made adjustments to the rules until we were satisfied. When comparing the results, we found the built in R dictionaries were more accurate than our own and decided to use them in our analysis.

When building our model, we tried grouping similar products together in order to simplify the interface for the end user. This proved to be counter-productive. From feedback we received, the user would like information on a particular product. Having other products grouped together with the product concerned changed the results. This did not give the user the actual results they were seeking. After reviewing the feedback and results from grouping, we decided to eliminate it from our model to improve accuracy and user experience.

Once the rules were established from our experiment results, we developed our final web application, as well as our rules, to calculate the sentiment of the reviews. We had individuals test our app thoroughly and give us direct feedback on suggested improvements. The idea is for the app to be self explanatory and enjoyable to use, while providing interesting and useful information. We used these suggestions to fine tune our UI.

Conclusions and Discussion

Through our analysis, it was beneficial to use sentiment analysis to not only compare the sentiment of the text of a review to the actual rating, but also to detect a bias of the reviewer, if present. By reviewing and comparing the interactive tables and visualizations available in our app, users are able to gain a greater understanding of the message conveyed by a review. Our app is useful in aiding consumers to make better purchase decisions.

Distribution of Team Member Effort

All team members have contributed equally.

Appendix

(1) Matloff, Norman. *The Art of R Programming: a Tour of Statistical Software Design*. No Starch Press, 2011.
(2) Li, Xiaolin, et al. "The Effect of Online Reviews on Product Sales: A Joint Sentiment-Topic Analysis." Information & Management, 2018, doi:10.1016/j.im.2018.04.007.
(3) Eslami, Seyed Pouyan, and Maryam Ghasemaghaei. "Effects of Online Review Positiveness and Review Score Inconsistency on Sales: A Comparison by Product Involvement." Journal of Retailing and Consumer Services, vol. 45, 2018, pp. 74–80., doi:10.1016/j.jretconser.2018.08.003.
(4) Yu, Xiaohui, et al. "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain." Knowledge and Data Engineering, IEEE Transactions On, vol. 24, no. 4, 2012, pp. 720–734., doi:10.1109/TKDE.2010.269.
(5) Lak, Parisa, et al. "Star Ratings versus Sentiment Analysis - A Comparison of Explicit and Implicit Measures of Opinions"." An Introduction to Biometric Recognition - IEEE Journals & Magazine, Wiley-IEEE Press.
(6) Maks, Isa et al. "Sentiment Analysis of Reviews: Should we analyze writer intentions or reader perceptions?" Vu University, Faculty of Arts.
(7) Mali D. "Sentiment Analysis of Product Reviews for ECommerce Recommendation." International Journal of Management and Applied Science, ISSN: 2394-7926.
(8)Ren, Yuyang, et al. "Analysis of Social Media Sentiment to Predict Product Sales" ttps://Pdfs.semanticscholar.org, 2005, pdfs.semanticscholar.org/bb31/48a29e9a9543845fb45851548fd8d3f7e13c.pdf.
(9) Bickart, Barbara & M. Schindler, Robert. (2002). Internet forums as influential sources of consumer information. Journal of Interactive Marketing. 15. 31. 10.1002/dir.1014.abs.
(10)Judith A. Chevalier and Dina Mayzlin (2006) The Effect of Word of Mouth on Sales: Online Book Reviews. Journal of Marketing Research: August 2006, Vol. 43, No. 3, pp. 345-354.
(11) Collomb, Ana¨ıs, et al. A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation. University of Lyon, INSA-Lyon, F-69621 Villeurbanne, France, 2018, liris.cnrs.fr/Documents/Liris-6508.pdf.
(12) Medhat, Walaa, et al. "Sentiment Analysis Algorithms and Applications: A Survey." NeuroImage, Academic Press, 27 May 2014, www.sciencedirect.com/science/article/pii/S2090447914000550.
(13) Jeevanandam Jotheeswaran, Dr S. Koteeswaran. "Sentiment Analysis: A Survey of Current Research and Techniques."
(14) Shelke, Nilesh, et al. "Domain Independent Approach for Aspect Oriented Sentiment Analysis for Product Reviews." SpringerLink, Humana Press, 1 Jan. 1970, link.springer.com/chapter/10.1007/978-981-10-3156-4_69.
(15) Silge, J., & Robinson, D. (2018, April 02). Text Mining with R. Retrieved May 4, 2018, from https://www.tidytextmining.com/tidytext.html
(16) Kanouse, D. E., & Hanson, L. (1972). Negativity in evaluations. In E. E. Jones, D. E. Kanouse, S. Valins, H. H. Kelley, R. E. Nisbett, & B. Weiner (Eds.), Attribution: Perceiving the causes of behavior. Morristown, NJ: General Learning Press.

(17) Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. R. He, J. McAuley. http://jmcauley.ucsd.edu/data/amazon/, 2016