A Multifaceted Product Recommendation System

Introduction

The US consumer pet industry generates about $100 billion[20] in annual revenue where nearly half of pet owners indicate Amazon as their preferred option for online pet goods purchases.[21] Amazon generates product recommendations based on seller-sponsored listings, customer views and past purchases[9] but does not factor in consumer reviews. Our approach leverages consumer reviews, synthesized with multiple Natural Language Processing (NLP) algorithms and visualizes the results with an interactive display to create a buyer-focused experience.

Problem Definition

The approach used by Amazon to recommend products highlights the conflict of their role as both a platform provider and a competitor to their sellers[10] as well as it being a seller-focused model. Our project aims to address this issue by using an innovative, user-driven recommendation system which combines Latent Dirichlet Allocation (LDA) topic analysis, sentiment analysis and normalized ranking of product reviews along with a new algorithm developed by the team that separates products into categories based on review text.

Survey

Reviews and their helpfulness have been studied using a variety of techniques. For instance, Chua observed higher Amazon review ratings are correlated with the Amazon helpfulness metric when performing regression analysis.[15] Jerziorski studied sponsored search ads and their impact on the user experience which we attempt to neutralize by using customer reviews from all products.[11] Also, sentiment analysis has been used to analyze review data using techniques such as a boolean measure of sentiment[1] as well as applying hybrid classification methods that use The General Inquirer Based Classifier (GIBC) and rule-based classifiers.[2] Kaur and Sharma utilized TextBlob sentiment along with supervised learning techniques to analyze scraped Twitter data relating to the Metoo movement.[13] Tripathy uses supervised learning techniques such as Naive Bayes, Stochastic gradient descent, and Support vector machines along with an n-gram model to predict sentiment of movie reviews[3]. Li provides an interesting addition by grouping comments based on product features such as price, cost, and payment allowing consumers to segment sentiment based on these groupings known as aspects[4].

To better understand consumer's interests, Sutherland and Kiatkawsin used LDA to find topics using Airbnb reviews[8]. The AirBNB topics were able to represent couch surfing customers to mansion seekers. This is analogous to the wide range of individuals looking for products that are being explored in our project. Examples of the AirBNB topics include "Residential Pets", "Arrival and Departure Convenience", and "Sleep Disturbance" which would not show up in a traditional number of bedroom and bathroom search on AirBnB. Our project's topic search creates a demand driven search in addition to sentiment analysis on normalized ratings which Sutherland and Kiatawsin did not provide. Additionally, Yaakobi provides helpful context by explaining traditional frequency based approaches vs. LDA which uses a Bayesian technique; one application was the ability to rank startups using LDA alongside similarity scores.[7]

Online product ratings have been analyzed by Sun who noted that a product's average rating and standard deviation of ratings impact future demand[5]. Fan, Xi and Liu assessed ranking products based on ratings and product attributes such as average rating, popularity and participant count using percentage distributions, stochastic dominance degree and the PROMETHEE-II ranking method[12]. The research of both Wang[14] and Kauffmann[22] informed our decision to use the combined methods discussed in detail in this report and also raised the risk of review credibility which the team determined is outside the scope of this project. The algorithms discussed in those papers would be an interesting addition to our methodology.

Team BoxCoxRox (Team 17): N. Abramson, M. Kunnen, K. Matisko, K. McCanless, M. Porter and S. Tay

Our approach in ranking involves combining techniques from both research by using standard deviation of the ratings and attributes such as average ratings, rating counts, review text length to create a ranking for the visualization component. However, these approaches are frequently used in isolation. Schafer discusses several recommender systems, none of which use a multi-pronged approach as we aim to do.[18]

Proposed Method
Intuition / List of Innovations

1. **Increase in user satisfaction -** Analysis of customer reviews to identify satisfaction and sentiment helps identify the strengths and weaknesses of products, allowing future recommendation to be driven by user feedback.
2. **Automated review analysis** - Manual parsing of reviews is expensive and time-consuming. A strong algorithm and visualization platform can deliver valuable insights in much less time. An integrated python script was developed to allow a front-to-end execution of the entire data procurement and algorithm pipeline (See Approach and Data sections below for more information) to produce the information needed for the visualization application. This script, boxcoxrox_pipeline.py, is located within our CODE folder.
3. **Ensemble NLP Algorithm -** Topic identification, sentiment analysis, ranking of review scores and LDA topic analysis are combined in a way that has not been explored in the literature we surveyed. The combination of these four analytical techniques provided a more enhanced and in-depth analysis of the review dataset.

Description of Approaches - Algorithm and Interactive Visualization
Algorithms - Many NLP approaches are frequently used in isolation. Our approach uses a continuous sentiment score and leverages not only sentiment analysis, but also normalized ratings, and topic extraction in order to recommend products that may interest consumers. Our algorithm includes a four-step process as discussed below.

Direct Frequency Topic Identification (DFTI) - An algorithm developed by the team whereby a set of predefined categories and associated key words are used to group products based on reviews. The algorithm works by aggregating all of the review text for a given product. Next, the number of occurrences of the key words for each predefined category are counted. The predefined category with the highest key word count is then used to classify each product. Unlike the product taxonomy which is predefined by Amazon and is potentially biased in favor of sellers, our approach focuses specifically on the text of product reviews as a means to group products into categories. Unlike LDA, this approach is a supervised learning algorithm.

LDA Topic Analysis - LDA is used to identify topics from the combined reviews of each product. LDA uses a sparse Dirichlet prior which reduces overfitting as the size of the corpus increases[16], which makes it a more appropriate algorithm for our large data set compared to pLSA. We chose to use the Python scikit-learn implementation of LDA. We found our initial approach of using Spark based LDA unnecessary once we first separated the dataset into individual categories (see DFTI above.) To further enhance the LDA model, we applied GridSearchCV[22] to determine the optimal number of topics in each category.

Sentiment Analysis - A sentiment score for each product is calculated using Vader (Valence Aware Dictionary and Sentiment Reasoner) Sentiment, a lexicon trained by 10 human graders. Vader yielded a high correlation coefficient when utilized for analysis on social media text, editorials, and reviews[17]. Vader's robustness made it a more appropriate choice for our project compared to TextBlob sentiment that was used in Kaur and Sharma's research[13]. Additionally, it

contains logic for emoticons and acronyms one would typically find in a review. Each review is scored and aggregated by ASIN (Amazon Stock Identification Number) then summarized by mean, median, mode, and standard deviation. Values range from -1 to 1 with sentiment thresholds as follows: less than -.05 for negative, greater than .05 for a positive, and the remainder neutral.

Ranking - When building a probabilistic model using ratings to create a reliable product quality indicator, Xie and Lui noted that the average score produces a more reliable and robust quality assessment than the majority and median rule.[6] Thus, we use average in our approach. A rank is created by ASIN using parameters which include average rating by product, number of reviews, and average length of reviews. These parameters are used to assess usefulness of the reviews data for each individual product in order to produce a rank for the entire dataset, as well as a rank within the product category and topic. A percentile ranking is generated based on the result to normalize the ranking.

Final Dataset Assembly - Each of the above processing tasks generate an intermediate csv file which is a result of an inner join on the product ASINs. Then the following steps were performed to prepare the data for usage:

1. **Product link validation -** We used Selenium to validate Amazon product page links to only display links with a valid page on Amazon.com.
2. **Text data cleansing -** Several text processing steps such as cleaning up any HTML tags present in the product title and description, and shortening and formating the description for easy display in the user interface were performed.
3. **Removal of irrelevant products** - A number of ASINs that existed in the raw dataset have been reused by Amazon to tag a non-pet product, or were otherwise misclassified as pet products. The last step in the final data assembly removed these products based on a blocklist that were developed through manual testing.

Integrated pipeline - A single script was created to consolidate data preparation, the four modelling techniques noted above (Direct Frequency Topic Identification, Sentiment Analysis, Ranking and LDA Topic Analysis) and the generation of the final dataset that is used for the interactive visualization component detailed below. This file is called boxcoxrox_pipeline.py and is located in the CODE folder.

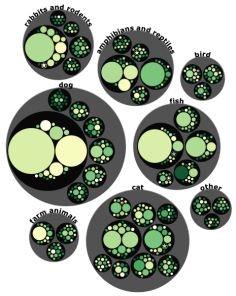
Interactive Visualization / User Interface - Our visualization consists of an interactive multi-layer bubble chart to visualize product clusters containing product recommendations for each topic. Using LDA topic analysis, we identify topic clusters to be used to create individual bubbles in the chart. Results from sentiment analysis and ranking are used to fill the topic bubbles with associated products. To ensure the visualization provides meaningful information to users, we include the top one hundred highest ranked products in each topic bubble. The color and size of the product bubble highlights the magnitude of the sentiment analysis and the number of reviews for that product respectively. Two sliders are present to allow users to limit the number of products displayed per LDA topic and the minimum sentiment of products. We allow users to drill-in to each topic by highlighting the product information for each product via a tooltip. The highest-ranked product in each topic cluster is indicated with a star icon.

We originally planned to use Tableau to drive our visualization, however Tableau was unable to provide the level of customization and zoomable drill-through functionality we needed. For this reason, we switched our visualization rendering to D3.js.

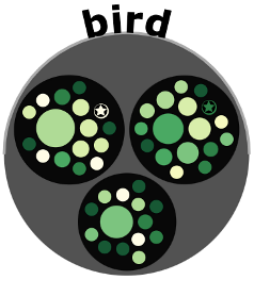
A Multifaceted Product Recommendation System

We hosted the site on gitlab.io so it is publicly available. We have embedded the following features in our visualization application:

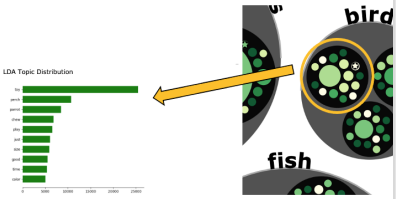
Global View
Display of the bubble chart which shows all the pet categories



Zoomed-in View
User clicks and view is centered and zoomed to topic

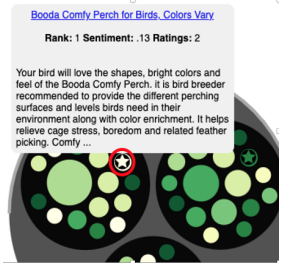


LDA Topic Distribution
Hover on the topic bubbles to display the topic distribution bar chart on the left.

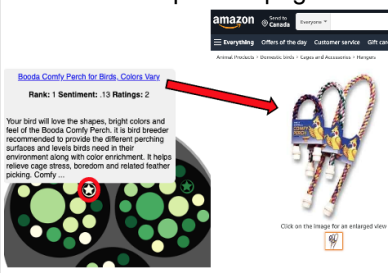


Personalized Filters
Filters located at the top allow users to adjust the visualization by filtering products by rank and average sentiment

Tooltip Information
Hover on product bubbles to display product information. Starred products are first-ranked products in each topic bubble.



Links to Amazon
Products with valid Amazon pages will have an embedded hyperlink that redirects to its Amazon product page



Data

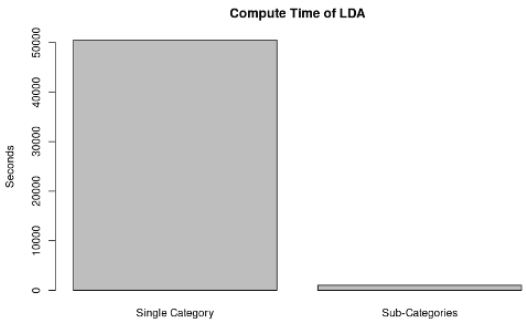
Our Amazon review dataset is sourced from Ni, Li and McAuley's research comparing reference-based and aspect conditional models in product reviews using relevance scores[19]. There are two files associated with pets which contain a list of pet products and individual product reviews. Further data cleaning is required given that the files are not structured as a singular JSON document but rather are made up of individual lines of JSON objects.

The downloader component of our data pipeline downloads and decompresses the raw datafiles from the website mentioned above. After the data has been downloaded and decompressed into ".json" files, it is converted from the non-standard JSON used in those files and inserted into a structured SQLite3 database. This database is used as input to the different models described above. The final dataset consists of 6,542,483 product reviews and 205,999 products. The size of the SQLite3 database itself is 4.7Gb.

Design of Experiments (DoE) and Evaluation

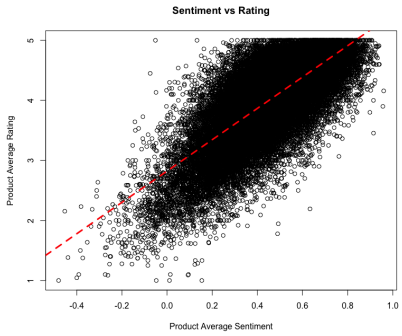
LDA Topic Analysis Performance - Is it more efficient to process all documents using LDA at once, or group the documents into a high level predefined clustering? Our initial modeling attempts were completed with Spark in a single node cluster on a multi-core computer. Computation time for all documents (reviews) was 50,492 seconds (~14 hours). This approach,

unfortunately, had several negative consequences. First, the long compute time made it difficult to rapidly iterate on our design and visualizations. Secondly, with such a large range of products, we found that the topics were not very intuitive. To address this, we conducted an experiment using DFTI whereby we predetermined the first layer of categorization of the products (dog, cat, bird, fish, etc.) before performing LDA on the resulting classifications.



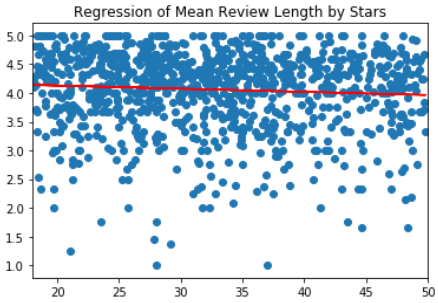
Results: After categorizing the products prior to performing LDA, not only did the resulting LDA topics make much more sense, the compute time was reduced by 98% to 1034 seconds (~17 minutes). The "Compute Time of LDA" bar chart shows the performance improvement of LDA on pre-categorized product reviews vs a single category of reviews.

Sentiment Analysis - Does Vader produce accurate sentiment? Per Hutto,[17] Vader was developed and tested on social media posts. We wanted to validate if Vader also works on product reviews. To confirm our approach, we performed a linear regression of mean product rating (stars) onto mean product review sentiment as measured by Vader. In the graph on the right, each dot represents a product. The x and y axis correspond to average sentiment and average rating for each product dot.



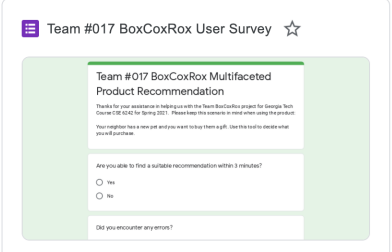
Results: When regressing mean rating onto mean sentiment, we observed an R-squared value of 55%. We also observed that the coefficient of 2.61 was significant at the .05 confidence level. This provides confirmation that Vader is an appropriate tool to use on product reviews and that we can use the sentiment score to filter product recommendations for our visualization component.

Product Rank - Do parameters used in the ranking process correlate with the final result? We performed a regression of average rating onto sentiment to explore the relationship between the average length of review text and the average rating (stars) for a random sample of 1000 products.



Results: The R-squared value for this regression was extremely low < (0.01). The test, therefore, failed to prove that longer reviews are associated with higher customer rankings. However, we were still faced with how to rank products that are tied. We decided that products with longer review text should be ranked higher despite the evidence above simply because products that have longer average reviews should contain more information. We felt that in the case of a tie, the product with the most information should be listed first.

Final Product & Visualization Usability Testing - Can users quickly understand how to navigate the tool? Usability Testing Plan - We will observe at least 10 users via a shared screen and ask them to perform the following task: Your neighbor has a new



pet and you want to buy them a gift. Use this tool to decide what you will purchase. A Google survey will tabulate the survey results.

Results: Total number of users surveyed is 18

Survey Metrics	**Results**	**Goal**
Did the user find a suitable recommendation within 3 minutes?	**67%**	60%
Did the user encounter any errors?	**94%**	90%
Rate the following on scale of 1 to 5:		
Helpfulness	**3.6**	>= 3.5
Usability	**3.7**	>= 3.5
Topics Relevance	**3.9**	>= 3.5
Does it provide enough product details to make a purchase decision?	**56%**	60%

Conclusion and Discussion This project resulted in the development of an integrated algorithmic model that presents an alternative approach in recommending pet products to users by leveraging customer reviews, which provide a buyer-driven approach in the purchasing experience.

During the course of this project, a few key observations were noted that can be considered for future research on customer reviews from online shopping platforms:

1. **Discontinued products yielding high rank and strong positive sentiments** - A few products that were highly ranked and yielded strong positive sentiments have been discontinued on Amazon. This may be a valuable insight from the seller's perspective on potential revenue lost after removing these products from the platform.
2. **Expansion of similar analysis on other product types** - This project was performed specifically on pet products. It would be worthwhile to expand this research in the future on other product types such as food, musical equipment and household products.
3. **Enhancement of user interface** - Based on survey results, users found the application useful in finding potential products to buy but it lacked sufficient information to make a final purchase. Potential improvements could include embedding more detailed product and pricing information.
4. **Combined approach of using seller metrics and customer reviews to facilitate a balanced recommendation system** - Leveraging insights from both sellers (product sale counts, repeat buys etc) as well as customers may lead to a more effective recommendation system, higher customer satisfaction and better sales yield.
5. **Dynamic data requires frequent modeling updates** - Due to products changes, frequently updating the models and results may ensure the best results for consumers.

Contributions - All team members have contributed a similar amount of effort toward the completion of this project.

Reference

1. Fang, Xing, and Zhan, Justin. “Sentiment Analysis Using Product Review Data.” *Journal of Big Data*, vol. 2, no. 1, 2015, pp. 1–14.

2. Prabowo, Rudy, and Thelwall, Mike. “Sentiment Analysis: A Combined Approach.” *Journal of Informetrics*, vol. 3, no. 2, 2009, pp. 143–157.

3. Tripathy, Abinash, et al. “Classification of Sentiment Reviews Using n-Gram Machine Learning Approach.” *Expert Systems with Applications*, vol. 57, 2016, pp. 117–126.

4. Li, Yan, Li, Yan, Qin, Zhen, Qin, Zhen, Xu, Weiran, Xu, Weiran, Guo, Jun, and Guo, Jun. "A Holistic Model of Mining Product Aspects and Associated Sentiments from Online Reviews." *Multimedia Tools and Applications* 74.23 (2015): 10177-0194. Web.

5. Monic Sun. “How Does the Variance of Product Ratings Matter?” Management Science, vol. 58, no. 4, INFORMS, 2012, pp. 696–707, doi:10.1287/mnsc.1110.1458.

6. Xie, Hong, and Lui, John C. S. "Mathematical Modeling and Analysis of Product Rating with Partial Information." *ACM Transactions on Knowledge Discovery from Data* 9.4 (2015): 1-33. Web.

7. Yaakobi, Alon, Goresh, Moshe, Reychav, Iris, McHaney, Roger, Zhu, Lin, Sapoznikov, Hanoch, and Lib, Yuval. "Organisational Project Evaluation via Machine Learning Techniques: An Exploration." *Journal of Business Analytics* 2.2 (2019): 147-59. Web.

8. Sutherland, Ian, and Kiatkawsin, Kiattipoom. "Determinants of Guest Experience in Airbnb: A Topic Modeling Approach Using LDA." *Sustainability* (Basel, Switzerland) 12.8 (2020): 3402. Web.

9. Galit Shmueli, Peter C. Bruce, and Nitin R. Patel. “Data Mining for Business Analytics. 3rd ed.” Hoboken: Wiley, 2016. Web. Chapter 14.

10. Edward J. Janger & Aaron D. Twerski, “THE HEAVY HAND OF AMAZON: A SELLER NOT A NEUTRAL PLATFORM”, 14 Brook. J. Corp. Fin. & Com. L. 259 (2020).

11. Jeziorski, Przemyslaw, and Ilya Segal. 2015. "What Makes Them Click: Empirical Analysis of Consumer Demand for Search Advertising." *American Economic Journal: Microeconomics*, 7 (3): 24-53.

12. Fan, Zhi-Ping, et al. “Supporting Consumer’s Purchase Decision: a Method for Ranking Products Based on Online Multi-Attribute Product Ratings.” Soft Computing (Berlin, Germany), vol. 22, no. 16, Springer Berlin Heidelberg, 2018, pp. 5247–61, doi:10.1007/s00500-017-2961-4.

13. Kaur, Chhinder, and Sharma, Anand. "Social Issues Sentiment Analysis Using Python." 2020 5th International Conference on Computing, Communication and Security (ICCCS) (2020): 1-6. Web.

14. WANG, Xuan, et al. “Improved LDA Model for Credibility Evaluation of Online Product Reviews.” IEICE Transactions on Information and Systems, vol. E102.D, no. 11, The Institute of Electronics, Information and Communication Engineers, 2019, pp. 2148–58, doi:10.1587/transinf.2018EDP7243.

15. Chua, Alton Y.K, & Banerjee, Snehasish. (2016). “Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality.” *Computers in Human Behavior*, *54*, 547–554

16. “Girolami, Mark; Kaban, A. (2003). On an Equivalence between PLSI and LDA. Proceedings of SIGIR 2003. New York: Association for Computing Machinery. ISBN 1-58113-646-3”

17. 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.

18. Schafer, J Ben, Konstan, Joseph A, and Riedl, John. "E-Commerce Recommendation Applications." *Data Mining and Knowledge Discovery* 5.1 (2001): 115-53. Web.

19. Ni, Jianmo, et al. “Justifying Recommendations Using Distantly-Labeled Reviews and Fine-Grained Aspects.” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019. Crossref, doi:10.18653/v1/d19-1018.

20. Bedform, Emma. “U.S. Pet Industry Expenditure 1994–2020.” Https://www.Statista.Com/, Statista, 10 Feb. 2021, www.statista.com/statistics/253976/pet-food-industry-expenditure-in-the-us.

21. Statista Research Department. “Leading Online Marketplaces Visited to Buy Pet Products in the United States in 2020.” Https://www.Statista.Com/, Statista, 22 Mar. 2021,www.statista.com/forecasts/1223382/leading-online-marketplaces-for-pet-products.

22. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.