Cheering for the Underdog: A Case Study of Generosity in Online Consumer Behavior

Matthew Adner

GATI, University of California, Irvine

Polygence Summer Program

September 7, 2021

Abstract

This paper focuses on the social behavior characterized as generosity, demonstrated by verified purchasers leaving mobile phone reviews and product ratings on Amazon. We hypothesize that consumers show a bias in favor of companies with weaker market positioning, as opposed to companies who dominant the market of certain products. This paper frames an analysis of the data set of consumer reviews of mobile phones purchased online. Our analysis employs natural language processing (NLP) to evaluate the hypothesis of consumer bias, in cheering for the underdog brand.

Introduction

The social dimensionality of online consumer behavior is an area of research expanding alongside the rapid growth of e-commerce. This paper is motivated by the potential for consumers to display generosity to certain companies in their online transaction reviews. Specifically, this paper evaluates the proposition that consumers favor companies in a weaker brand position, as compared to companies in a premium brand position, thus demonstrating a consumer bias in cheering for the underdog. Using a data set of ratings and customer reviews collected from Amazon.com transactions in consumer electronics, we may evaluate proposed consumer bias with tools from the fields of data science and natural language processing (NLP). The idea of the underdog refers to a narrative in popular culture where the weaker competitor overcomes a seemingly dominant competition. Our results from the analysis of online reviews may **[1] validate the main proposition that consumers show bias, cheering for the underdog; or [2] contradict the proposition by suggesting a similar bias in the opposite direction; or [3] invalidates the proposition by showing no significant bias in either direction.]**

In e-commerce retail transactions, big data has become a key to business profitability. In particular, online ratings and reviews have become the standard source of insight into how consumers think about their purchases. Given the increased availability of data collected through consumer transactions on digital platforms, our research sits at the crossroads of both marketing, and data science – offering an exploration of consumer bias with the use of NLP, and a statistical analysis that evaluates the evidence of consumer bias.

Organized in three parts, this paper begins with a literature review of Marketing research on social bias of consumers that might surface in e-commerce. To evaluate the proposition that consumers show social bias in their online transactions, we then draw upon data science, to frame a study of a data set. Our data set is a collection of reviews of consumer electronics transactions, including star ratings (1 to 5) and text-based reviews for which data pre-processing is applied. Finally, after framing the data analysis, we evaluate the evidence regarding proposed social bias by examining the text-based reviews. In this final discussion, we seek to contribute towards the growing field that merges marketing and data science, with a reflection on this case-study of potential social bias in consumer behavior online.

Literature Review

An important area of study in Marketing is how brands are meaningful to consumers, as brands play a heavy role in how consumers identify with companies and their products (Keller, 2020). With the new online ecosystem curated by digital platforms like Amazon.com, marketing research as a field has extended to understand consumers' social behavior in this new environment of e-commerce (Rapert, Thyroff, and Grace, 2021). Online platforms may offer new dimensions of social behavior to explore, insofar as overlapping with social media. Consumers' increased reliance on the feedback of their friends and fellow consumers may lead to an increased importance of reviewers' reputation (Zhang, Zhao, Cheung, and Lee, *Decision Support Systems*, 2014). Consumers exhibit pro-social values, for example, favoring brands that promote social missions, such as fair-trade or sustainability (Kotler and Lee, 2008). In particular, Saccard et al. (2021) finds that language used in an online transaction can impact consumers' pro-social behavior, "nudging" them to support a social cause. Companies use narratives in their brands to make a connection to customers (Cooper, Schembri and Miller, 2020). Paharia et al. (2011) reference four "identity mechanisms" that consumers feel towards a brand and propose that customers who themselves identify as an underdog, may favor the brand with a perceived "humble origin[s], lack of resources, and determined struggle against the odds."

In age-old cultural stories of the underdog, such as "David and Goliath," where the weaker player wins, is present too in markets. Historically, it is not uncommon for when competition is uneven that consumers cheer for the weaker party: local vs. global, main street vs. big box stores, or the expected party vs. against-all-odds. Underdog products are products that maintain a weaker market position in a specific product space. For example, a bargain brand in a headphone market dominated by big brand luxury names like Apple-branded Air Pods or Google-branded Pixel Buds. Consumers may show social bias in markets when they cheer for the underdog, sending more potential customers their way. This paper furthers this research on how consumer social bias shows up in e-commerce.

The Model and Data Set: Marketing ‘Big Data’ Meets Data Science

With the growth of e-commerce and the increase of digital information, or big data, new areas of research and new methodologies such as machine learning and natural language processing have opened up at the intersection of marketing and data science (Jain, Pamula, Srivastava, 2021). The question driving this study is if consumers are rooting for the underdog, referring to the weaker brand in the marketplace. To frame this analysis, the market position of the brand is simplified to two positions: dominant and weak. The premium band holds the dominant market or the Goliath position, while the bargain brand holds the weak brand position or the underdog position. For this study, products that come from a company with a market cap above 1 billion (USD) were considered premium, and those below were considered bargain brand. Dataset: https://www.kaggle.com/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones. Reviews include a simple rating on a scale of 1 to 5 stars, and a text-based section. Customers may read these reviews before a purchase, to look for evidence of product features or performance. Reviewers may earn a reputation as a reviewer. While reputation was part of the dataset, it was not included in this analysis. Our study furthers the growing connection between the study of consumer behavior and data science, as we use data science techniques and methodology to analyze our dataset.

The Data Set and Framing the Analysis

Our approach automates the analysis of ratings and reviews via a Python script, NLP libraries, and machine learning (ML) models. We chose the Python language for its dynamic flexibility in conducting analysis on large datasets via dataframes, and the specifics of this analysis are discussed below. Throughout this study, coding was executed within Jupyter notebooks (ending .ipynb). The libraries used were: fuzzywuzzy, numpy, pandas, csv, nltk, sklearn, warnings, and matplotlib. The models examined were TFIDF, count vectorization, N-grams, naïve Bayes, and logistic regression. We use these models to predict the star rating from the text review in both the big and small company data, with models trained on an equal number of reviews from each. We experimented with multiple ML model types before selecting the most accurate of the bunch for our final test set to ensure due diligence, as there is no one-size-fits-all model in ML for NLP. TFIDF examines the number of times a word occurs in a text, generating a virtual keyword from the text. TFIDF avoids common words like 'the' and 'as', which would be unhelpful in determining a review's sentiment (Steccanella, 2019). We looked at this model because seeing what kinds of words tend to come with positive/negative reviews could be helpful for predicting unbiased star values. Count vectorization turns all the words it encounters into vectors that have the same number of dimensions as the total number of words it encounters (Heidenreich, 2018). We looked at this model because this kind of model is a successful general approach that works well across a variety of NLP problem types. N-grams chunks together different sets of words of different length, and checks how frequently they occur (Kumar, 2017). We examined at this model because we believe that it is likely that there are chunks of a few most commonly used words that tend to be found across negative reviews. Naïve Bayes uses Bayes' theorem to predict the probability of something being true about the text given it contains pre-found attributes (Saxena, 2017). We checked this model as being able to predict the probability of what an attribute of the text is, is well-aligned with our goal of accurately predicting sentiment. Logistic regression works similarly to linear regression, but instead of outputting continuous values (slope) it outputs an effectively binary value (going to ∞ or -∞) (Swaminathan, 2018). We looked at this model because outputting an effectively binary value (1 or 0 as positive or negative), is more suited to our purposes than having to set thresholds for a continuous value, which is what would have to be done with linear regression. In reviewing how different approaches to the analysis impacted model efficacy, we hoped to find a most-suited approach for our final analysis.

Challenges with the Data Set

There are many challenges that come with working with big data, which can be summed up as the challenge of “cleaning the data.” Cleaning the data, in general, refers to the process of identifying and removing inaccurate, corrupt, or incomplete records from a set, which may be misleading (Bird et al., 2019). One problem was that there were differing brand name spellings in the (e.g. “Samsung” as “Samsung Korea LTD”). To counter this, we replaced all effectively identical names of brands with the same name. Another problem was that many of the reviews did not have a listed brand. We handle this by searching the actual text of the review for brand names, and if only one brand was listed, then that brand was assumed to be the brand of the product that the review was about. Unfortunately, this search-strategy for addressing the listed brand problem only worked for about 10 percent of the reviews without listed brand, and so the rest of the reviews without listed brands were dropped from our dataset. For companies with very few reviews, say fewer than 500, we dropped the rows from our data, as statistically-speaking we would not have enough to work with. A final issue was that there is no preexisting list of big and small companies. To solve this, we research the market caps of phone companies online. If a market cap was above 1 billion dollars (USD), the company was considered to be large, if less or not able to be found, then the company was considered small. There were many general concerns for challenges presented by the data set for machine learning analysis. Some problems were handled by adjustments in the code. Other problems were handled by just removing that subset of reviews. In summary, the data set provides an opportunity for exploring the consumer behavior of generosity, but there are challenges and problems that must first be anticipated and addressed in setting up the analysis of the data.

Possible Sources of Error

There were additional factors that were possible sources of error. One of these is the fact that bargain brand companies' products were lower quality, which is a difficult variable to control for via programming alone. Another possible source of error is that there may be a correlation between the type of customer and the type of phone purchased, reflecting different quality standards customers have. For example, a person who buys a premium brand phone, with a higher price, might also be more likely to be highly sensitive to the product details. Another possible source of error is the distribution and quantities of data: the majority of the reviews were from large companies. While there was enough data to perform analysis, more data can give greater confidence in our analysis. In summary, big data analysis comes with possible sources of error, but the hope is that the large amount of information which one can examine in a study, using reliable statistical tools gives us faith in results that outweigh the smaller concerns with error.

Summary Analysis and Conclusions

The aforementioned models were run, with the most accurate found to be n-grams, after comparison via ROC (receiver operating characteristic), which shows the classification model performance across threshold. Because of this, n-grams was used for our final analysis. Our metric for assessing the hypothesis was a generosity score, calculated by finding the percentile difference between the number of customer-reported negative ratings, and the number of negative ratings predicted from the review text by the model. The generosity score for the larger companies was -6.19. The generosity score for the smaller companies was -2.41. This aligns with our hypothesis that greater generosity would be shown towards the underdog company, and means that our model (trained on both large and small company data) predicts lower star ratings for the associated reviews of small companies, implying some level of benefit being provided by the consumer in their actual rating.

What this tell us is that, by a factor of over 2.5, the ratings for underdog companies are still being given a push, and warrants additional research. Similar studies should be conducted outside of the smartphone-specific market, and beyond even consumer electronics as a whole. It will be beneficial for companies and marketing groups to research how these trends display across different platforms, other than Amazon, and even perhaps the e-commerce sales of used goods. This form of research may also be relevant for trends of environmentally conscious brands like Patagonia, or Fjällräven's, in terms of how consumers may provide a bump based on their perception of a company's branding. Similarly, quantifiable insights towards luxury brands, like Gucci and Louis Vuitton, would create value – as we seek to further understand how consumer's standards come into play when analyzing beyond typical wealth consumption lines.

References

Bird, Sarah, Krishnaram Kenthapadi, Emre Kiciman, and Margaret Mitchell. (2019). Fairness-aware machine learning: Practical challenges and lessons learned. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM),* 834–835.

Chang, T. e. a. (2021). "Accelerating mixed methods research with natural language processing of big text data." *Journal of Mixed Methods Research, vol. 15, no. 3*, 398–412. Retrieved from doi:10.1177/15586898211021196

Cooper, H., Schembri, S., & Miller, D. (2010). Brand-self identity narratives in the james bond movies. *Psychology & Marketing, 27*(6), 557-567.

Gladwell, M. (2013). *David and goliath: Underdogs, misfits, and the art of battling giants* Little, Brown.

Heidenreich, H. (2018, August 24). Count Vectorization with Scikit-learn-e. *Towards Data Science Blog.* https://towardsdatascience.com/natural-language-processing-count-vectorization-with-scikit-learn-e7804269bb5e

Jain, P. K., Pamula, R., & Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Computer Science Review, 41*, 100413. doi:https://doi.org/10.1016/j.cosrev.2021.100413

Keller, K. L. (2020). Consumer research insights on brands and branding: A JCR curation. *Journal of Consumer Research, 46*(5), 995-1001.

Koh, N. S., Hu, N., & Clemons, E. K. (2010). Do online reviews reflect a product's true perceived quality? an investigation of online movie reviews across cultures. *Electronic Commerce Research and Applications, 9*(5), 374-385. doi:https://doi.org/10.1016/j.elerap.2010.04.001

Kotler, P., & Lee, N. (2008). *Social marketing: Influencing behaviors for good*. Sage.

Kumar, P. (2017, October 21). An Introduction to N-grams: What They Are and Why Do We Need Them. *XRDS Crossroads: the AMC Magazine for Students.*

Paharia, N., Keinan, A., Avery, J., & Schor, J. B. (2011). The underdog effect: The marketing of disadvantage and determination through brand biography. *Journal of Consumer Research, 37*(5), 775-790.

Phang, C. W., Zhang, C., & Sutanto, J. (2013). The influence of user interaction and participation in social media on the consumption intention of niche products. *Information & Management, 50*(8), 661-672. doi:https://doi.org/10.1016/j.im.2013.07.001

Rapert, M. I., Thyroff, A., & Grace, S. C. (2021). The generous consumer: Interpersonal generosity and pro-social dispositions as antecedents to cause-related purchase intentions. *Journal of Business Research, 132*, 838-847.

Saccardo, S., Li, C. X., Samek, A., & Gneezy, A. (2021). Nudging generosity in consumer elective pricing. *Organizational Behavior and Human Decision Processes, 163*, 91-104.

Saxena, R. (2017, Feb 6). How the Naïve Bayes Classifier Works in Machine Learning. *Dataaspirant*. https://dataaspirant.com/naive-bayes-classifier-machine-learning/

Stecanella, B. (2019). What is TF-IDF?: A Simple Introduction. *Monkey Learn Blog.* https://monkeylearn.com/blog/what-is-tf-idf/

Swaminathan, S. (2018, 15 March). Logistic Regression: Detailed Overview. *Towards Data Science*. https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

Zhang, K. Z., Zhao, S. J., Cheung, C. M., & Lee, M. K. (2014). Examining the influence of online reviews on consumers' decision-making: A heuristic–systematic model. *Decision Support Systems, 67*, 78-89.

Zhu, F., & Zhang, X. (. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing, 74*(2), 133-148. doi:10.1509/jm.74.2.133

https://www.kaggle.com/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones