Helpfulness in Amazon Reviews

Shruti Hegde
University of Maryland
College Park
shegde17@umd.edu

Kalpita Raut
University of Maryland
College Park
kraut@umd.edu

Himanshi Manglunia
University of Maryland
College Park
himanshi@umd.edu

ABSTRACT

Understanding how users perceive reviews' quality is crucial since it can reveal insights on what drives a user's purchasing decisions[1]. Through this paper, we examine the Amazon reviews data and its helpfulness rating to understand which features in a review are the most significant to determine its helpfulness. Semantic features, linguistic features, emotions computed through LIWC and attributes like reviewer expertise are computed for the analysis. Further, a model is built using Multiple linear regression to determine the significant variables that impact helpfulness. The results from the model show reviewer count to be having the strongest effect followed by personal pronoun 'I'. The linguistic feature - number of words - is significant towards helpfulness. Also, emotions such as anger, sadness , anxiety, positivity have an impact on review helpfulness.

Author Keywords

Helpfulness, reviews, bias, emotions, social analytics, LIWC

INTRODUCTION

The increasing impact of the Internet has dramatically changed the way that people shop for goods. More and more people are now gravitating to reading products reviews prior to making purchasing decisions. Online customer reviews have become a significant source of product-related information for consumers. Such reviews have become an indispensable component of e-commerce Websites such as Amazon. As a result of the growing number of customer reviews, determining the helpfulness of a review is important. The increasing number of reviews available for various products has created information overload for consumers (2). Pinpointing which reviews are most helpful is critical in reducing this information overload. A review diagnosticity theory is defined as the piece of information that is helpful in making informed purchase decisions and is linked to the assumption of information diagnosticity, which includes the question of whether a certain text of information is helpful during the processes of decision-making [2].

MOTIVATION

Online Product reviews have become significant source of information for customers. We have reviews for everything nowadays. Hence resulting in an abundance of information. Although such profound information might seem like a good thing, it can also be characterized as information overload. Information overload (also known as infobesity, infoxication, information anxiety, and information explosion) is the difficulty in understanding an issue and effectively making decisions when one has too much information about that issue. When it comes to reviews, there's a possibility that due to a lot of reviews, certain review that is actually helpful might be buried or a new review might get buried. Therefore, it is necessary to process this abundance of information in such a way that we can extract what's important and provide that to the customer/user. In one prominent example that explains why helpfulness of a review is important, it is estimated that Amazon.com added $2.7 billion to annual revenue by asking the simple question "Was this review helpful to you?"

APPROACH

Deciding the product purchase relies on how helpful the product review is. However, several research papers identified a wide range of biases and each of them influences the helpfulness rating given by a customer for the product differently. Some of the biases stated are as follows :

- Underreporting Bias[3] : Among people who purchased a product, those with extreme ratings (5-star or 1-star) are more likely to express their views to "brag or rant" than those with moderate views (underreporting bias)
- Social Influence Bias [4] : Social influence bias refers to a tendency that one's opinion is influenced by the opinions expressed in other reviews.
- Selection Bias [4] : Selection bias, roughly speaking, happens when the set of users that submit a review is not representative of the entire purchasing population.

Additionally, there are other issues that further question whether the customer reviews are trustworthy (5).

- For newly posted reviews, most likely no vote or only a few votes would be cast, and therefore, identifying their helpfulness is difficult.
- Presenting the reviews ranked by their user-voted helpfulness scores may create situations of "monopoly" in that only the highest ranked reviews get viewed, leaving no opportunities for the newly published yet unvoted reviews to show up on users' radar.
- In some cases, reviews can be incorrectly labeled as helpful or not helpful due to spam voting.

Therefore, in this study, we aim to identify the features which are significantly related to helpfulness of a product review and build a model to predict the helpfulness score of the reviews. Based on the scores, the model will provide the most helpful review related to a product in order to reduce information overload and to avoid customer biases.

Research Questions

In this study, we aim to answer two main questions -

1. What are the most significant features to predict the helpfulness score of a review?

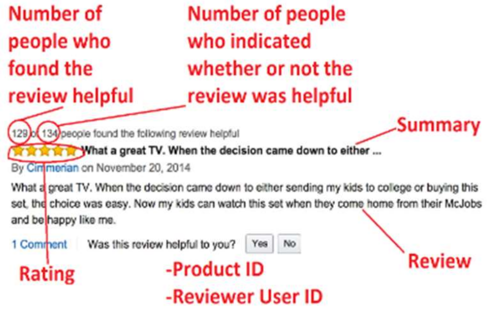
- The sheer amount of reviews with varying content and quality makes it impossible for customers to evaluate and comprehend all the information before making a purchase decision The possibility of sorting reviews based on helpfulness enables potential consumers to shorten their information search, evaluate alternatives more efficiently, and make better purchase decisions

2. Impact of emotions on Review helpfulness.

- Seller seek to utilize product reviews to boost trust and increase sales. When writing a text review, an unsatisfied customer often express negative emotions, which have the potential to influence the attitude and behaviors of future customers to a greater extent than positive emotions. However, this does not mean that the more emotional a negative review is, the more helpful. If the most helpful reviews of an online retailer are mostly positive, the retailer can expect benefits in terms of reputation, trust, and sales. In contrast, if the most helpful reviews are largely critical, the retailer is likely to suffer. We want to find the differential effects of emotions like anxiety, anger on helpfulness.

About the Dataset

This dataset [6] contains product reviews and metadata from Amazon, including features 142.8 million reviews spanning May 1996 - July 2014.This dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).


Number of people who found the review helpful
Number of people who indicated whether or not the review was helpful
Summary
Review
Rating
-Product ID
-Reviewer User ID
Yes
No

129 of 134 people found the following review helpful

★★★★★ What a great TV. When the decision came down to either ...

By Cimmerian on November 20, 2014

What a great TV. When the decision came down to either sending my kids to college or buying this set, the choice was easy. Now my kids can watch this set when they come home from their McJobs and be happy like me.

1 Comment | Was this review helpful to you?

Fig 1. A snppet of an Amazon Review

The dataset contains the following features:

- asin – ID of the product
- helpful – helpfulness rating of the review
- overall – rating of the product
- reviewText – text of the review
- reviewTime – time of the review (raw)
- reviewerID – ID of the reviewer
- reviewerName – name of the reviewer
- summary – summary of the review
- unixReviewTime – unix time of review

Scholars investigating review helpfulness have focused primarily on a number of determinants that are easily observable, such as ratings and reviewer characteristics and a few studies have investigated the content and substance of reviews themselves. All the research projects that we considered for finding the factors that impacts helpfulness have used factors such as reviewer's information, review_score, review_date, etc. A paper (5) proposed that helpfulness could be made more effective by taking reviewer expertise into account. Extending this to our study, we shall also extract reviewer expertise. The dataset that we have does not contain reviewer expertise as a separate entity. We propose to compute the reviewer expertise by counting the number of times a reviewer has posted their reviews. This is based on the finding by the paper [7] that states that more number of reviews by a reviewer is associated with greater reviewer expertise. We also plan to use linguistic features such as adjective and semantic features while building the model. The linguistic and semantic features have been elaborated below.

Features computed for model:

- Reviewer expertise by getting a count of the number of reviews posted by a reviewer
- Number of words in a review
- Number of words in summary
- Number of characters in a review
- Number of sentences in a review
- Helpfulness percent

Data Cleaning and Preparation

We performed simple cleaning processes like removing all the missing values. We could afford to do this since our dataset has around 190000 observations and we didn’t lose much data even after removing all NA values.

The data for helpfulness column was in the form of a string as [“upvotes”,”downvotes”]. This column was further processed by splitting these values into two different columns, upvotes and downvotes, to calculate the helpfulness percent. The helpfulness percent was computed as percentage of ratio of upvotes and total votes. This helpfulness percent was the dependent variable for this study.

We also removed rows where sum of upvotes & downvotes is less than 5.

Since our entire study was based on the textual content of our reviews, we cleaned our textual data as well by performing the following methods-

- Stopword removal & Unwanted characters removal : “Stop words” are the most common words in a language like “the”, “a”, “on”, “is”, “all”. These words do not carry important meaning and are usually removed from texts
- Tokenizing : Tokenization is the process of splitting the given text into smaller pieces called tokens.
- Lemmatizing : This is a process of reducing words to their word stem, base or root form (for example, books—book, looked—look).

METHODOLOGY

For research question 1, we will try to analyze the above features and see which one is most strongly related with helpfulness by constructing a Multiple Linear Regression model. Multiple linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. In our study, our dependent variable is helpfulness score

For research question 2, we have used the LIWC dictionary to compute linguistic as well as emotion based features from the review texts obtained after cleaning the data. LIWC maintains a dictionary composed of almost 4500 words and word stems, each of which defines one or more categories. As each word in the review is processed, LIWC searches its dictionary file for a match with the review word. If a match occurred, the appropriate category scale(s) for that word would be incremented. At the end of this procedure, a final score is assigned for each category, representing the percentage of words in the review that matched that category. There are 63 categories in LIWC. Some of those categories are:

Psychological emotions - positive or negative like sad, anger, anxiety, affection, positive, negative,
Personal concerns - money, achieve, leisure
Perceptual processes - feel, hear, see and more.

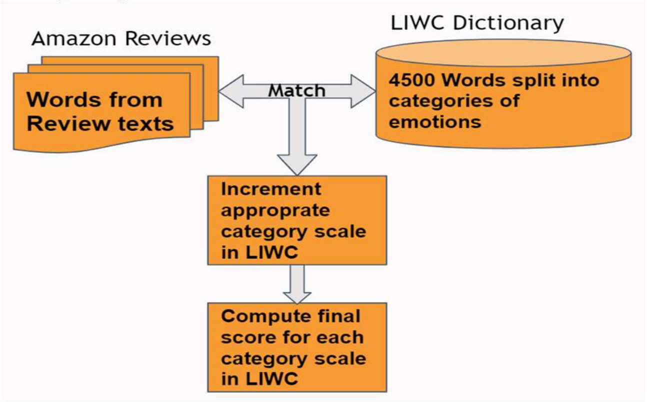
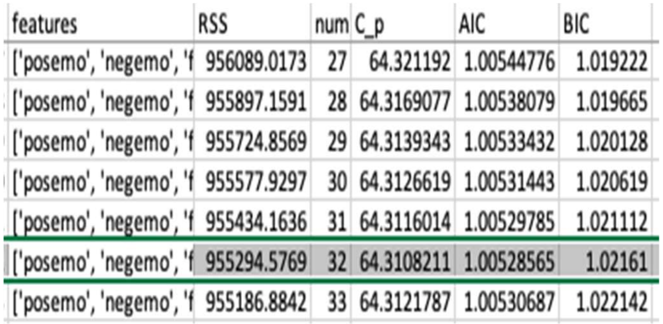

Amazon Reviews
LIWC Dictionary
Words from Review texts
Match
4500 Words split into categories of emotions
Increment appropriate category scale in LIWC
Compute final score for each category scale in LIWC

Fig 2. Flowchart of how LIWC works

The score for each review, for each category was stored in a dataframe, and combined with the other features computed for research question 1. A Multiple linear regression model was created on all of these features together. The dependent variable was helpfulness score for both the models.

RESULTS

We combined all the linguistic features and features from LIWC as our independent variables and helpfulness score as the dependent variable. First, all the variables were standardized. In other words, variables in the regression were all converted to z-scores before running the analysis to get standardized coefficients. Standardized coefficients help us compare the variables easily to each other after building the model. We then did a best model selection using forward selection, which involves starting with no variables in the model, adding the variable (if any) whose inclusion gives the most statistically significant improvement of the fit, and repeating this process until none improves the model to a statistically significant extent. The best model is the one with the smallest AIC score. Akaike information criterion (AIC) estimates the quality of the model.

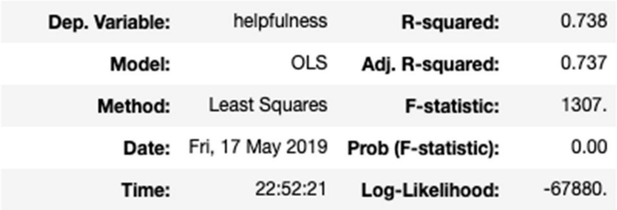


features	RSS	num	C_p	AIC	BIC
['posemo', 'negemo', 'f	956089.0173	27	64.321192	1.00544776	1.019222
['posemo', 'negemo', 'f	955897.1591	28	64.3169077	1.00538079	1.019665
['posemo', 'negemo', 'f	955724.8569	29	64.3139343	1.00533432	1.020128
['posemo', 'negemo', 'f	955577.9297	30	64.3126619	1.00531443	1.020619
['posemo', 'negemo', 'f	955434.1636	31	64.3116014	1.00529785	1.021112
['posemo', 'negemo', 'f	955294.5769	32	64.3108211	1.00528565	1.02161
['posemo', 'negemo', 'f	955186.8842	33	64.3121787	1.00530687	1.022142

Fig 3. Results of best model selection

The best model consisted of the independent variables 'posemo', 'negemo', 'feel', 'negate', 'insight', 'reviewer_count','work','inhib','article','summary_word_count', 'we', 'family', 'they', 'anx', 'number', 'anger', 'sad', 'i', 'tentat', 'excl', 'preps', 'ingest', 'bio', 'percept', 'conj', 'adverb', 'humans', 'nonfl' and 'quant'.

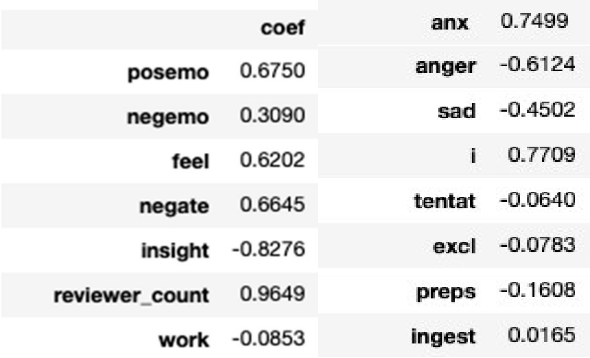
After getting the best features, we ran multiple linear regression.



Dep. Variable:	helpfulness	**R-squared:**	0.738
Model:	OLS	**Adj. R-squared:**	0.737
Method:	Least Squares	**F-statistic:**	1307.
Date:	Fri, 17 May 2019	**Prob (F-statistic):**	0.00
Time:	22:52:21	**Log-Likelihood:**	-67880.

Fig 4. Results of the MLR model

Figure above shows the summary statistics of the best model. Adjusted R^2 for the best regression model is 0.737 which means that 73.7% variability in helpfulness score is explained by the model we built with the best features.



	coef
posemo	0.6750
negemo	0.3090
feel	0.6202
negate	0.6645
insight	-0.8276
reviewer_count	0.9649
work	-0.0853
anx	0.7499
anger	-0.6124
sad	-0.4502
i	0.7709
tentat	-0.0640
excl	-0.0783
preps	-0.1608
ingest	0.0165

Fig 5. Coefficients of all significant features from the model

Figure above shows some of the variables that are significant in our model. To find the relative importance of variable and compare which measures have the strongest influence on helpfulness, standardized coefficients have been used. According to our model, reviewer count has the strongest effect, followed by insight (includes words like think, know, consider) and personal pronoun I. Emotions like anger, sadness, positive emotions and negative emotions also play a role in helpfulness. The sign of the standardized coefficient describes the direction of the relationship and the magnitude shows the effect size of the variable. For example, we can see that the reviewer expert count has a positive relationship with helpfulness which would mean that the expert reviewers write reviews that are helpful to the users. The negative coefficient of insight suggests that increasing use of insightful words does not seem to be helping users which seems kind of strange. Also, Anxiety, Positive and negative emotions are positively related to helpfulness. However, anxiety and positive emotions seem to be more strongly related than the negative emotions. Anger and sadness are negatively correlated to helpfulness which reveals that users do not find such reviews filled with anger or sadness helpful. Since sadness had a p-value greater 0.05, it is not significant independent variable. Anger is negatively correlated to helpfulness which reveals that users do not find such reviews filled with anger helpful. The magnitude of the coefficient also shows that anger is quite strongly negatively associated with helpfulness. Therefore, results of our model suggest that anxiety filled reviews like "I had some doubts about the item I purchased, never got an answer" were perceived more helpful to the users than anger filled reviews like "These people are TERRIBLE. They stalled the order for days". Also, the presence of the variable 'I' in our model, which has a positive coefficient, in the model suggests that more personal anecdotes are helpful to the users. The model also included word counts as one of the significant predictors positively associated with helpfulness which indicated that users found elaborate reviews are helpful.

CONCLUSION

The results from the model show reviewer count to be having the strongest effect followed by personal pronoun 'I'. Reviewer count isn't really visible to the user. But our model shows that expert reviewers seem to be giving reviews that are likely to be helpful to the users. The linguistic feature - number of words - is significant towards helpfulness. Also, emotions such as anger, sadness , anxiety, positivity have an impact on review helpfulness.

LIMITATIONS

As much as we tried to address the limitations and biases, there were certain limitations that we could not address.

1. On public platforms like Amazon where most people often check the reviews before buying products, there are sellers who post fake reviews on the products to influence the decision of the prospective buyers. We have performed our analysis without accounting for those fake reviews.
2. The dataset on which the analysis has been performed dates back to 2014. There are a lot of changes in the dynamics of such platforms which could not be accounted for because of the data.

Future Scope

We suspect that if the fake reviews are filtered from the dataset, the results would be more accurate. This can be a scope for the future of this study.

REFERENCES

1. Bobby Nguy. Evaluate Helpfulness in Amazon Reviews Using Deep Learning. 2015. Retrieved April 9, 2019 from https://cs224d.stanford.edu/reports/Nguy.pdf
2. Yin et al. Dreading And Ranting: The Distinct Effects Of Anxiety And Anger In Online Seller Reviews. 2010. Retrieved April 9, 2019 from https://pdfs.semanticscholar.org/3196/4bd6a54d3cee40182e475d81e4fa1aef381f.pdf?_ga=2.131368975.1002297413.1552342105-112313596.1549820713
3. Hu et al. Overcoming the J-Shaped Distribution of Product Reviews. 2013. Retrieved April 9, 2019 from

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2369332

4. Askalidis et al. Understanding and overcoming biases in online review systems. 2017. Retrieved April 9, 2019 from https://www.sciencedirect.com/science/article/pii/S0167923617300428?via%3Dihub
5. Liu et al. Modeling and Predicting the Helpfulness of Online Reviews. 2008. Retrieved April 9, 2019 from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4781139&tag=1
6. Jmcauley. Amazon product data. 2014. Retrieved April 9, 2019 from http://jmcauley.ucsd.edu/data/amazon/
7. Connors et al. Is It the Review or the Reviewer? a Multi-Method Approach to Determine the Antecedents of Online Review Helpfulness .2011. Retrieved April 9, 2019 from https://ieeexplore.ieee.org/document/5718695