What NLP Elements Make Amazon Reviews Helpful?

Anonymous ACL-IJCNLP submission

Group member:

- Shichao Han: *schan21@berkeley.edu*
- Ophelia Wang: *wangyumeng2017@berkeley.edu*.
- Han Liu: *han_liu@berkeley.edu*

Abstract

Natural language processing provides useful information when it comes to customer feedback or product review. The main objective of this research is to identify linguistic features that make reviews helpful. Prediction models are trained to reveal features that predict helpful contexts. Causal inference is conducted to quantify whether the features have causal relationship to the perceived helpfulness, as well as providing featurization guidance for model improvement. The selected features with significant causality are put into an even simpler classification model, leading to a better predicting performance. The current study identified several linguistic features that have justified causal relationship to the helpfulness of customer review through an integration of text-based classification and causal inference.

1 Introduction

The Internet has been an established venue for sharing ideas and reviews - an important element in the E-commerce ecosystem, as customers seek information from reviews in their purchasing decision and merchants gain market insights from users' feedback(Malik and Hussain, 2018)(Zhu et al., 2020). Identifying factors that contribute to the helpfulness of reviews help both customers and sellers.

In the literature of review helpfulness prediction, there are determinants that are identified by previous researchers: length of review, cumulative helpfulness of reviewers, social network structure of the reviewers and some linguistic characteristics(Huang et al., 2015)(Mudambi and Schuff, 2010)(Zhou and Guo, 2017). The current study further examines the effect of linguistic characteristics, as on the users side, texts are their primary source of information in reviewing online contents.

Researchers discovered those linguistic features and structural features that are useful and can be incorporated in the classification models in the current study, to improve model accuracy(Krishnamoorthy, 2015)(Ngo-Ye and Sinha, 2014)(Qazi et al., 2016)(Zhou and Guo, 2017). However, many of the existing studies focus on features and models' performance based on metrics such as accuracy, Area Under Curve(AUC), correlation coefficient, and sensitivity analysis, and leave the causal relationship unjustified(Egami et al., 2018)(Qazi et al., 2016)(Zhou and Guo, 2017). In light of some recent development of text-based causal inference frameworks, causation of the features on the helpfulness is measured using average treatment effect estimates.

1.1 Project Overivew

Therefore, the driving research question in the project is to discover the natural language ele-

ments that make Amazon reviews to be considered helpful, and quantify their causal effects on the perceived helpfulness. There are two major components of the project.

1.2 Component1-Classification Models

Classification models are developed not only to filter the features that work the best for different models, but also to put the data into groups that are correctly labeled and falsely labeled. They should also unveil the characteristics of the language features that confuse the classification models, as well as human readers.

Three models are considered in the current report: (1) BERT for text classification, (2) convolutional neural network(CNN), and (3) few-shot learning with GPT-2(Xu et al., 2020). Beyond their presence in INFO256 class of UC Berkeley with detailed discussions and implementations, they are representative of a wide range of models - traditional deep learning models, pre-trained models with embeddings, and few-shot learning models.

We further examined linguistic features in the sentences that are wrongly classified by our models in the futures. We managed to extract the feature representations in the trained models, e.g., the filter weights in CNN and the embedding in the BERT model. Detailed discussions can be found in section 2 and 3.

1.3 Component2-Causal Analysis on Linguistic Features

After extracting features from the modeling outcomes, causal inference is conducted to analyze the causal relationship between selected linguistic features and the number of "helpful" that the reviews have received. Highlights of our causal analysis for usefulness are: (1) we used the pre-trained BERT model to generate a sentence embedding for each review and treated the sentence vectors as features that are potentially confounding to our outcome of interest, which is the usefulness of reviews; (2) we have conducted causal analysis on testing whether some linguistic features (e.g. the length of the reviews) have causal relationship to the perceived helpfulness.

1.4 Data Description

A pre-labeled dataset contains product reviews and metadata from Amazon mainly in the Electronics section, including 142.8 million reviews spanning May 1996 - July 2014. Data includes reviews (ratings, text, helpfulness votes) and product metadata (descriptions, category information, price, brand, and image features). For each review, the number of "helpful" and "not helpful" counts are also recorded in the dataset.

To define "helpfulness" for classification, we followed previous researchers to first calculate $\frac{\text{helpful Counts}}{\text{helpful Counts + unhelpful Counts}}$, and then discretize the value at the threshold of 0.7(Ghose and Ipeirotis, 2010). The binary response "helpful" and "not helpful" are denoted by 1 and 0. The original dataset is unbalanced, having 80% helpful reviews. Hence, we sampled $10,000$ helpful reviews and $10,000$ non-helpful reviews for future analysis such as usefulness predictions and causal inference to prevent to problem of class imbalance. Also we removed all the duplicates in the dataset.

2 Prediction Models

Using the overall information from the data, which consists of the helpfulness votes of the reviews, we could infer how satisfied consumers are towards the product reviews, with 1 demonstrating users find reviews not useful and 0 otherwise. Therefore, we could conduct a supervised classifier with the actual review text as the core predictor variable, with a usefulness dummy variable as the response variable.

2.1 Helpfulness Votes and Features

The dataset contains "helpful" for each data, which is a list with two elements: the number of users that voted helpful, and the total number of users that voted on the review (including the not helpful votes). In any predictive framework, featurization can be important. From each review, the predictive model should be able to extract core information and determine if a review is assigned as "helpful". To derive the features for our purposes, we will examine samples of Amazon product reviews and their corresponding helpfulness ratings and use the product ratings, content, and presentation of the reviews to assess their helpfulness with readers (Rodak et al., 2014). The groups of features in our main approaches including:

(a) Words count
(b) Number of exclamation marks
(c) Number of question marks
(d) Number of all-capitalized words
(e) Occurrences of words that are highly correlated with helpful reviews
(f) Automated Readability Index

Our motivation for these structural features were to capture token-based and syntatic-based analysis of the text of the reviews. There are also non text-based features that are found to be predictive(Ghose and Ipeirotis, 2010). After doing words counts analysis within helpful reviews group and unhelpful reviews group separately, we found a group of words that show up in helpful votes much more frequently than in unhelpful votes, including like, quality, price, well, very, definitely and great. For each of these words highly correlated with helpful votes, we create a new word-count feature. Intuitively exploring other ways to measure how useful one review is, we found that the readability of review text is also an important factor that affects customer experiences. The Automated Readability Index, constructed based on two ratios, one representing word difficulty(number of letters per word) and the other one representing sentence difficulty(number of words per sentence)(Senter, 2012), can well quantify the readability of review texts.

$$ARI = 4.71\frac{characters}{words} + 0.5\frac{words}{sentences} - 21.43$$

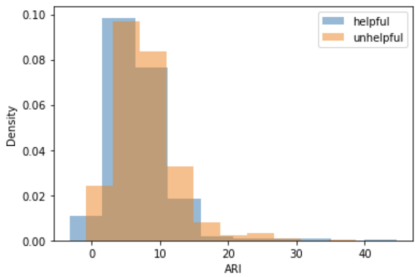

Figure 1: ARI Distribution.

A high-valued ARI indicates a high level of difficulty for readers to digest the text data. We expect reviews with high ARI to be less helpful than those of low ARI. The above distribution verifies our initial guess as we can see the the distribution of ARI within the unhelpful review group is on the right of that within the helpful review group.

2.2 BERT

The first model we consider implementing is the BERT algorithm, Bidirectional Encoder Representations from Transformers, designed by Google. BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers(Devlin et al., 2019). We fitted a BERT-BASE model with 12 layers, 768 embeddings and 10 epochs on the training dataset, achieved an accuracy of 0.65.

2.3 CNN

Convolutional Neural Network(CNN), given its complexity and flexibility in representing

underlying features of the texts, is widely used in many NLP tasks and in helpfulness prediction (Chen et al., 2018) (**?**). CNN incorporates lexical and semantic features, which are found to produce high accuracy in classification of helpfulness of reviews (Mitra and Jenamani, 2021). After hyperparameter tuning, a CNN model with window of size 2 and 96 filters is selected and produces an accuracy to be 66.6%.

2.4 Few-shot learning

In few-shot learning, we are trying to recognize the labels of data using only a small amount of labeled data. The goal of the model is to generalize unseen examples quickly and computationally-efficient. Here, the labeled data in the training set has been used to generate the prompt to classify the test data - reviews, into ”positive” and ”negative” categories. After preliminary test, a few-shot learning model can only achieve around 53% accuracy. The corresponding confusion matrix is shown in Figure. 2.

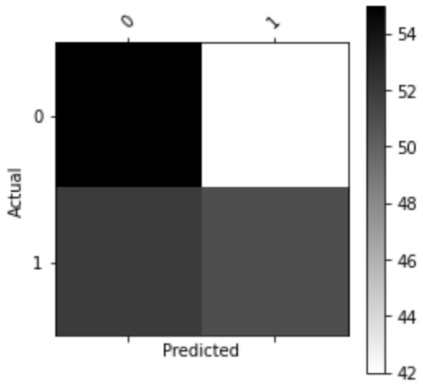

Figure 2: Confusion Matrix.

2.5 Misclassified Data Points

Reviews that are misclassified by our models have been further investigated. Here are a few false positive examples:

Table 1: Prediction Accuracy

Model	Accuracy
BERT	65.45%
CNN	66.60%
Few-shot learning	53%

- I got this to use with a Linux based XBMC media center on a ZBOX as the DVD player. It was recognized out of the box. It plays DVDs great, until it just quits. I’ll have to buy another drive as this one is not working for me.

- I have no complaints about the sound quality of these headphones. It is a pity though that they were not made to last - after only about 3 years of light office use the plastic shielding on the cable shows signs of cracking and the leatherette cushion on the headband has fallen off. The headphones still sound good though.

Above are two reviews having helpfulness feature equal to 0, indicating that the majority of reviewers found this review unhelpful, but our algorithm classified them as helpful reviews. The tone is very ambiguous in these reviews: reviewers first listed out a few product benefits, but then commented on the downsides right after. Hence, intuitively it makes sense that other users think these reviews are unhelpful. However, these reviews contain some strong positive phrases, such as ”plays DVDs great” and ”have no complaints about”. These words could result in false positive prediction in our models.

2.6 Features from Misclassified Data

In order to featurize the characteristics of the reviews that are often wrongly classified, KMeans clustering for text is conducted to divide the wrongly classified data into several clusters(Li and Wu, 2010). From the frequency

summarization of the tokens in different clusters. Tokens with highest occurrences in each cluster are included as features for examination in the causal inference part. Model features, along With these new linguistic features from mis-classified data, are passed to causal inference with the helpfulness indicator. Further analysis on how to improve the modeling through better featurization is also included in the next section.

3 Causal Analysis

This section covers the extraction of candidate features, as well as the qualification of their causality to the perceived helpfulness.

3.1 Causal Inference Analysis

Causal inference aims to quantify the causality of selected linguistic features, which are found to be useful in predicting models. Consider the data to be $\{\mathbf{X}_i, y_i\}$, where $\mathbf{X}_i \in \mathbb{R}^d$ is the sentence vector representation of each review, and $y_i \in \{0, 1\}$ to be the indicator of whether the review is perceived as helpful or not. In the current study, we used a BERT-base-uncased model to extract the last layer of embedding of the tokenized reviews(Devlin et al., 2019). Let $\mathbf{z}$ denote a feature that might potentially "causes" the text to be of a specific class, and $\mathbf{z}_i$ is the value of feature $\mathbf{z}$ of unit i. Our causal inference question is thus formed to study the causal effect of $\mathbf{z}$ on $\mathbf{y}$ with the presence of some possible observed covariates $\mathbf{X}$ for the reviews.

The primary challenge in the causal inference in the current study is the presence of confounders. Other linguistic features, including word relations and some sentiments, known to be related to our outcome of interest, are encoded in the sentence embedding using BERT features $\mathbf{X}$(Reimers and Gurevych, 2019)(Ghose and Ipeirotis, 2010). We also studied the causal effect of the length of words count for each review on the perceived level of helpfulness. Here, z_i would be the length for each review. However, since the distribution of $\mathbf{X}_i$ is not controlled and might distribute different among different levels of $\mathbf{z}$, the current study can be classified as an observational study in causal inference(Stuart, 2010)(Rosenbaum, 2002b).

3.2 Experiment on Data

If X_i is in high dimension, however, another problem of high dimensionality might occur for text data, and therefore, we need some dimensionality reduction strategy for the embedded data(Egami et al., 2018). A propensity score is defined as $e(X_i) = P(\mathbf{z}_i|\mathbf{X}_i)$ and can be used as a 1-dimensional representation of the $\mathbf{X}_i$(Rosenbaum and Rubin, 1983)(Paul, 2017). We chose propensity score as a dimensionality reduction strategy. To quantify the causal effect and accounting for the confounding from X_i, we employed the average treatment effect(ATE) of z_i obtained in linear regression with adjustment for the possible confounders in their propensity score$e(X_i)$(Rosenbaum, 2002a). Result of regression adjustment for $\mathbf{y}$ on different $\mathbf{z}$ are presented in Table 3. Note: the coefficients are the fitted coefficient for regressor $\mathbf{z}$ in regression $\mathbf{y}$ on $\mathbf{z}$ and $\mathbf{X}$. If the p-value is smaller or equal to 0.05, we can conclude that the feature $\mathbf{z}$ has causal relation to the perceived helpfulness.

Based on adjustment for 1-D representation of sentence embedding, the estimated average treatment effect for features length of the reviews, number of "definitely", "number of exclamation marks" and readability index are still significant at $\alpha = 0.05$ level, while "number of question marks" and "number of all-capitalized words" are not significant. The conciseness of the sentence, the less ambiguity in tone, and strong emotion in the contents are features that

Feature	Coefficient	SE
Length	0.05	0.01
# of !	0.77	0.39
# of ?	0.74	0.89
# of caped words	0.03	0.05
# of very	0.23	0.64
# of definitely	10.08	2.33
# of great	-0.38	0.77
# of use	0.44	0.73
# of like	-0.22	0.74
# of quality	1.34	0.96
# of price	1.11	1.18
# of well	-0.63	1.31
# of any	0.007	1.06
Readability Index	0.35	0.09

Table 2: ATE Estimate in Regression Adjustment

Feature	P-value
Length	4.7e-12
# of !	0.05
# of ?	0.40
# of caped words	0.56
# of very	0.71
# of definitely	1.74e-5
# of great	0.62
# of use	0.51
# of like	0.97
# of quality	0.16
# of price	0.35
# of well	0.63
# of any	0.99
Readability Index	7.27e-5

Table 3: ATE Estimate in Regression Adjustment Continued

cause the reviews to be perceived helpful.

3.3 Model Based on Causal Features

After identifying the linguistic features that have a significant causality to the perceived helpfulness, a simpler classification model - logistic regression is built on the same set of data. With a 0.5 cut-off value, the logistic regression achieved an prediction accuracy of 70.7%, which outperforms all the classifiers based on the overall structure of the review text, including BERT, CNN and Few-shot Learning. The selected features are not only shown to have causality to the helpfulness, but are predictive, improving prediction model efficiency.

4 Conclusion and Future Work

The current study has identified linguistic features that have causality to the perceived helpfulness of the reviews, as shown in section 3. Readability of the text, usage of punctuation and words of strong emotion, and conciseness of the text are all important for helpful customer reviews. The features with strong causal relationship also improve the classification modeling performance, using even a simpler model. Business decision can be much better leveraged with a carefully justified causal inference of the natural language processing results.

The same method can be applied to more linguistic and non-linguistic features as mentioned in section 1, such as the number of times that the user has posted in history, retail price of the product, and the category of the product. Although the study combined NLP and causal inference for identifying features that make reviews helpful, other facets of users' online behaviors are not included in the modeling, nor being adjusted for in the causal inference. Here is where the limitations, as well as the opportunities lie.

References

Cen Chen, Minghui Qiu, Yinfei Yang, Jun Zhou, Jun Huang, Xiaolong Li, and Forrest Bao. 2018. Review helpfulness prediction with embedding-gated cnn. *arXiv preprint arXiv:1808.09896*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2018. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*.

Anindya Ghose and Panagiotis G Ipeirotis. 2010. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE transactions on knowledge and data engineering*, 23(10):1498–1512.

Albert H Huang, Kuanchin Chen, David C Yen, and Trang P Tran. 2015. A study of factors that contribute to online review helpfulness. *Computers in Human Behavior*, 48:17–27.

Srikumar Krishnamoorthy. 2015. Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, 42(7):3751–3759.

Nan Li and Desheng Dash Wu. 2010. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision support systems*, 48(2):354–368.

MSI Malik and Ayyaz Hussain. 2018. An analysis of review content and reviewer variables that contribute to review helpfulness. *Information Processing & Management*, 54(1):88–104.

Satanik Mitra and Mamata Jenamani. 2021. Helpfulness of online consumer reviews: A multi-perspective approach. *Information Processing & Management*, 58(3):102538.

SM Mudambi and D Schuff. 2010. What makes a helpful review? a study of customer reviews onamazon. com (ssrn scholarly paper no. id 2175066). *Social Science Research Network, Rochester, NY*.

Thomas L Ngo-Ye and Atish P Sinha. 2014. The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Systems*, 61:47–58.

Michael Paul. 2017. Feature selection as causal inference: Experiments with text classification. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 163–172.

Aika Qazi, Karim Bux Shah Syed, Ram Gopal Raj, Erik Cambria, Muhammad Tahir, and Daniyal Alghazzawi. 2016. A concept-level approach to the analysis of online review helpfulness. *Computers in Human Behavior*, 58:75–81.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Jordan Rodak, Minna Xiao, and Steven Longoria. 2014. Predicting helpfulness ratings of amazon product reviews.

Paul R Rosenbaum. 2002a. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327.

Paul R Rosenbaum. 2002b. Overt bias in observational studies. In *Observational studies*, pages 71–104. Springer.

Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

E.A. Senter, R.J.; Smith. 2012. Automated readability index. In *Wright-Patterson Air Force Base*, pages 1–14.

Elizabeth A Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

Shuzhe Xu, Salvador E Barbosa, and Don Hong. 2020. Bert feature based model for predicting the helpfulness scores of online customers reviews. In *Future of Information and Communication Conference*, pages 270–281. Springer.

Shasha Zhou and Bin Guo. 2017. The order effect on online review helpfulness: A social influence perspective. *Decision Support Systems*, 93:77–87.

Yongmin Zhu, Miaomiao Liu, Xiaohua Zeng, and Pei Huang. 2020. The effects of prior reviews

on perceived review helpfulness: A configuration perspective. *Journal of Business Research*, 110:484–494.