Adapting Online Review Helpfulness Learnings to Online Business Reviews

Kevin Chow, Gia Schutzer, Soumik Mukherjee

Abstract

The ever-growing volume of online business reviews poses a challenge for customers seeking to identify helpful reviews. This project aims to address the issue by adapting insights from adjacent domains to develop a predictive model to rank the helpfulness of business reviews within the Yelp Open Dataset. Key factors influencing review helpfulness include review length, readability, star rating, business nature, review aspects, and the review type. Our approach involves a combination of feature engineering, extractive summarization, regression analysis, and fine-tuning pre-trained transformer models. The final model integrates these components into an ensemble framework via model stacking, combining all the learnings from adjacent review domains, to predict and rank review helpfulness. Business nature (search or experience) plays a key role in predicting review helpfulness, with review length, readability, star rating, review aspects (e.g. ambiance or service), and review type (regular, comparative, or suggestive) having slight influences on helpfulness.

1 Introduction

Customers increasingly rely on online reviews to make informed purchasing decisions. The exponential growth of user-generated reviews, however, has created a challenge. While more information should facilitate better decision-making, the sheer volume of reviews often leads to information overload. Business review platforms like Yelp now host hundreds of millions of reviews, which make it difficult for users to discern which reviews are helpful. Additionally, researchers have noted that progress in online review helpfulness prediction has been fragmented, with limited systematic building upon previous successes. Our research addresses these issues by developing a review helpfulness prediction model that can determine and rank helpful reviews for Yelp to surface.We achieve this by incorporating current NLP methodologies and building on past research in adjacent review domains, namely Amazon and TripAdvisor. Ultimately, we aim to change how online business platforms surface their reviews.

In this project, we looked at multiple factors that are important in determining review helpfulness. We examined basic structural features of the text such as the length in characters, words, and sentences. We also assessed how easy the review text is to understand through a range of readability metrics that are calculated based on factors like the number of words in a sentence, spelling errors, the number of letters or syllables per word, and word frequency. The individual star ratings given by each review, which play the role of providing a succinct and helpful way for customers to be better informed, was also studied. Another key factor is the service provided. This can be divided into two natures: search or experience. Search nature businesses generally provide services that customers can obtain information or quality about prior to purchasing, i.e. objective information about the businesses' services. Experience nature businesses have subjective qualities about their services, such as if the food is delicious at a restaurant. Review texts were also analyzed for specific aspects regarding the reviewer's experience with a business. For business reviews, these aspects can span elements such as ambiance, brand, contextual information, price, product / food, service, and other. The type of review (regular, comparative, or suggestive) combined with the review length also influences helpfulness. Comparative reviews tend to be more helpful when longer, as they provide more detailed contrasts between businesses, while suggestive reviews are more effective when concise.

We utilized pre-trained transformer models to extract key features and then created a system for ranking reviews by helpfulness. By surfacing the most helpful reviews, online business review platforms can enhance user trust, increase platform engagement, and provide more meaningful connections between customers and businesses.

2 Background

Prior work in review helpfulness has largely been focused in the product space. Our research explores and builds upon the survey of methodologies to predict online product review helpfulness by Diaz and Ng by applying it to business reviews [1]. Their audit yielded a set of key influential features. We focus on the adoption of a few specific features for the business review space: review length, readability, star rating, review aspects, business nature, and review type.

Liu et al. show that review length features can commonly be used to predict helpfulness, based on the intuition that longer reviews contain more information and are thus more valuable to readers [2]. This straightforward but effective feature captures the amount of detail provided in a review, influencing the likelihood of the review being helpful. Ghose and Ipeirotis further explore readability as a function of various structural features including length (of characters, words, and sentences), spelling errors, and a range of readability assessment metrics. Their work indicates that reviews with higher readability scores are associated with more product sales [3].

Hypothesizing why extreme product reviews, particularly positive reviews with high star ratings, are often preferred by review readers, Huang et al. studied the relationship between a reviewer's product rating and review helpfulness [4]. It was found that for top reviewers, product rating is shown to be a significant predictor of review helpfulness.

Further exploring the relationship between product review extremity and helpfulness, Mudambi and Schuff examined reviews based on the nature of the product [5]. Specifically, how product nature interacts with star ratings, word count, and total votes. For experience natured goods, where satisfaction with the product is subjective to each individual's experience, their research supports that reviews with extreme ratings are less helpful than reviews with moderate ratings. For search nature goods, where product satisfaction is largely based on objective truths about the product, review length had a greater impact on review helpfulness.

Research has also been done on Amazon data using a topic modeling technique that identifies top words across review texts and several product categories. These words were then grouped based on similarities into a set of higher-level aspects, which were used to predict helpfulness scores. Yang et al.'s approach points to a 7% improvement in predicting helpfulness [6].

Additionally, Qazi et al. investigated how different types of reviews (regular, comparative, and suggestive) affect review helpfulness, finding that review type moderates the relationship between review length and perceived helpfulness [7]. Their study of 1,500 TripAdvisor reviews revealed that comparative reviews tend to be more helpful when longer, while suggestive reviews are more useful when concise. They demonstrated that review type classification can be a valuable predictor of helpfulness.

Building upon these findings, we translate and adapt these learnings for the online business review domain. Taking into consideration the unique characteristics and expectations of the business review domain, we further extend their findings by incorporating the use of pre-built transformer models to process and analyze the content of each review's text, allowing us to more efficiently extract deeper, more nuanced features for predicting review helpfulness. Further, we combine the models explored in the research above in an ensemble framework, creating a novel helpfulness prediction model.

3 Methods

3.1 Dataset

Our dataset is sourced from Yelp.com and includes 6.99 million rows of review data for businesses across the United States. Yelp allows readers to provide useful votes to every review, displaying the number of previous votes below each review text for readers to see. We focus our research on businesses that had at least 350 total reviews, with a final dataset composed of 722,652 helpful reviews (num. of helpful votes > 0) and 722,652 not helpful reviews (num. of helpful votes $= 0$). This data was selected to balance the ability to capture meaningful review patterns with computational resource constraints.

3.2 Feature Engineering

Review Helpfulness: We operationalized helpfulness as the amount of useful votes a review received divided by the total number of useful votes across all reviews for that business. The helpfulness percentage normalizes data by accounting for variations in the total number of votes across businesses. This allows us to better quantify the impact of individual reviews within the context of each business's overall helpfulness, ensuring fair and meaningful comparisons regardless of differing review activity levels. Helpfulness is heavily skewed to the right, where the majority of reviews have a helpfulness score of less than 1%, as indicated in Table 1.

Mean	Std. Dev.	Min	25%	50%	75%	Max
0.1904	0.4748	0.0000	0.0000	0.0075	0.2288	38.0846

Table 1: Helpfulness percentage summary statistics

Review Length: For our baseline Tobit model, we measured review length as the raw count of words in each review. Reviews showed substantial variation in length, ranging from extremely brief 1-word reviews to detailed reviews containing up to 1,032 words. The median was 74 words with an average review length of 104 words. This disparity in length suggests that word count could be a meaningful predictor of review helpfulness.

Readability: Our work adapting readability closely replicated that of Ghose and Ipeirotis [3]. We measured structural elements, namely number of characters, words, sentences, and spelling errors in a text. Next, we calculated several metrics that score how readable a text is. These include: Automated Readability Index, Coleman-Liau Index, Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning fog index, and SMOG.

Star Rating: We adapted Huang et al.'s findings on review extremity for top reviewers in a slightly modified manner for the business review domain. Unlike Amazon product review data, our data fails to capture `unhelpful` votes. Instead, we operationalize `reviewer_cumulative_helpfulness` by determining the average helpfulness percentage of a reviewer's past reviews. The resulting Tobit regression model captured the relationship between business review extremity and helpfulness for top reviewers and their reviews [4].

Business Nature: The first application of Mudambi and Schuff's product review work to business reviews consisted of an initial model that mirrored their final model. While they focused on six specific products with binary labels 0 (search) and 1 (experience), our dataset had 2,753 businesses to label. As such, we first operationalized business nature based on each business's categories. Each of the 418 unique categories assigned to businesses were hand labeled with their primary nature, such as "Laser Hair Removal" (0 or search) and "Restaurants" (1 or experience). The majority class was used to assign each business a final binary business nature. As acknowledged by Mudambi and Schuff, "products can be described as existing along a continuum from pure search goods to pure experience goods [5]. Applying this logic to businesses and their reviews, we extracted weights from each review's text to quantify where along the continuum each business resided. Extractive summarization, first utilizing term frequency before improving with TF-IDF, was utilized to shorten each review. Summarization length was adjusted until a final maximum of three sentences was reached, balancing computational resources and preserving review content. Extractive was chosen over abstractive summarization to preserve the original review content as much as possible. The removal of stop words was also utilized to improve efficiency and remove noise. Scikit-learn's `CountVectorizer` was used to quantify each review summary's text before being compared to each business category. Based on each category's binary label, a final weighting for both search and experience was calculated.

To improve upon the business nature model, the weights calculated with `CountVectorizer` were replaced with calculations of the cosine similarity between review summary embeddings and the business's category embeddings. Both were extracted using Google's `bert-base-uncased`, allowing for contextualized word embeddings to be generated and compared. According to each category's binary classification, these new similarity weights were calculated before being averaged over the total number of categories in each nature. This averaging accounted for the varying number of categories per business, some with as little as one and others with more than 10. Finally, `reviewer_cumulative_helpfulness` was added to assess how Huang et

al.'s findings might further improve the business nature model. The resulting business nature model is:

$$\begin{aligned}\text{Helpfulness \%} = {} & \beta_1 \text{Rating} + \beta_2 \text{Rating}^2 + \beta_3 \text{Business Nature} + \beta_4 \text{Word Count} \\ & + \beta_5 \text{Total Votes} + \beta_6 \text{Rating} \times \text{Business Nature} + \beta_7 \text{Rating}^2 \times \text{Business Nature} \\ & + \beta_8 \text{Word Count} \times \text{Business Nature} + \beta_9 \text{Search Similarity} + \beta_{10} \text{Rating} \times \text{Search Similarity} \\ & + \beta_{11} \text{Rating}^2 \times \text{Search Similarity} + \beta_{12} \text{Experience Similarity} \\ & + \beta_{13} \text{Experience Similarity} \times \text{Word Count} + \beta_{14} \text{Reviewer Cum. Helpfulness} + \epsilon\end{aligned}$$

Review Aspects: We define 7 high-level aspects by which we categorize reviews:

1. **Ambiance / Environment:** The overall atmosphere and setting of the place
2. **Brand:** The identity and reputation of the business
3. **Contextual Information:** Relevant details related to the experience with the business
4. **Price:** The cost of the products or experiences, including value for money
5. **Product / Food:** The quality, taste, and presentation of the items offered
6. **Service:** The quality of customer service
7. **Other:** Any reviews that did not fit the above criteria

While our methodology follows a similar approach to Yang et al., rather than utilizing an LDA-type technique we first implement a zero-shot classification technique leveraging Facebook's `bart-large-mnli` model. This model was chosen because of its high inference ability on natural language tasks. Reviews are assessed relative to a prompt: "This review is about [aspect]". Next, we implemented a few-shot classification approach, manually categorizing 266 sentences in the data. The model was trained using 10 examples per label, with the remaining examples used for validation and testing.

The model is applied only to reviews that have 5 or more useful votes, an approach replicated from Yang et al. Classification is applied at the sentence level for each approach, given that one review may cover several or all of these aspects. We then test two variations of assessing the aspects. The first involves assigning review text the aspect that occurs in the majority of sentences and utilizing a new aspect "Multiple" if there is no majority. In the second, we calculate the total number of times an aspect occurs and divide it by the number of sentences in the text to yield a percentage of times an aspect is mentioned. Helpfulness score predictions were computed using an SVM regression model with a RBF kernel, as consistent with Yang et al.'s approach [6].

Review Type: Following Qazi et al.'s research on review types, we classified Yelp reviews into three categories: regular (standard descriptive reviews), comparative (reviews that compare multiple businesses), and suggestive (reviews offering specific recommendations) [7]. To perform this classification at scale, we again used Facebook's `bart-large-mnli`. We initially classified reviews using zero-shot classification and then fine-tuned the model on a subset (100 reviews) of manually labeled Yelp reviews. The classified reviews based on the fine-tuned model were then combined with review length, to create interaction terms that capture how different types of reviews might influence helpfulness. To predict helpfulness using review type and these interaction terms, we employed a random forest model. Random forest was chosen for its ability to model non-linear relationships and capture complex interactions between features, such as the differential effects of review length across categories (e.g., comparative reviews benefiting from longer length while overly verbose suggestive reviews detract from perceived helpfulness).

Final Stacked Model: Our work involves a two-step modeling approach. First, we individually modeled helpfulness as a function of each of the features above. This step served as an initial assessment of the relationship between each feature and the outcome. Unless otherwise specified, Tobit regression was utilized in each model. We make use of Tobit regression analysis in our models to account for the nature of our dependent variable (helpfulness) and the nature of the sample. A review's helpfulness never becomes negative (i.e. the data is left-censored), as useful votes cast only indicate a positive helpfulness and does not capture unhelpfulness [4]. Furthermore, the use of Tobit regression helped address the potential selection bias inherent in this type of data. The dataset does not provide the number of persons who read the review, but only the

number of total useful votes cast. Since it is unlikely that every reader who found a review helpful voted on helpfulness, there is a potential selection problem [5].

Next, we created a stacked model that ensembles the predictions from the first step. We omit the baseline of review length given that feature is incorporated into most of the first step models. Additionally, the stars model is not included as its own feature, but is instead incorporated as part of the business nature model. We held out a validation dataset to build and refine step 1 models and used the testing set to evaluate the final ensemble model.

The final resulting stacked model is defined by the following function:

$$\text{Helpfulness } \% = \beta_1 \text{Readability} + \beta_2 \text{Review Type} + \beta_3 \text{Business Nature} + \beta_4 \text{Review Aspects} + \epsilon$$

Evaluation Metrics: Models are evaluated using two key types of metrics highlighted by Diaz and Ng [1]. We used Normalized Discounted Cumulative Gain (NDCG) to assess the quality of the ranking that our models yield. NDCG values range from 0 to 1, where the highest possible score is achieved when the true scores are perfectly ranked. All models produced NDCG scores (k = 5, 25, 50, 100, 1000, and all reviews) quantifying their ability to appropriately predict and rank the top k reviews' helpfulness. We also assessed the strength of our helpfulness predictions by comparing root mean squared error (RMSE) and mean average error (MAE). RMSE and MAE allowed us to understand how far our predictions are from their true values in an easily interpretable manner. These two key types of metrics allowed us to both assess model performance individually, as well as compare model performances against each other.

4 Results and Discussion

The review length baseline model served as a starting point for predicting review helpfulness. With an RMSE of 0.4644 and an MAE of 0.2507, the model captured the relationship between review length (measured in number of words) and helpfulness. However, its NDCG (k = all) of 0.8138 indicates limited ability to rank reviews effectively compared to other models. Despite its limitations in ranking, the baseline model's simplicity provides a useful benchmark for evaluating the effectiveness of other models we explored.

Model	**RMSE**	**MAE**	**NDCG (k = all)**
Review Length (Baseline)	0.4644	0.2507	0.8138
Readability	0.4926	0.2274	0.8904
Stars (Top Reviewers Only)	1.4530	0.9070	0.8268
Business Nature	0.3373	0.2386	0.9673
Review Aspects	1.3990	0.7307	0.8309
Review Type	0.4396	0.2050	0.8259
Ensemble Model	**0.3753**	**0.2800**	**0.9595**

Table 2: Model performance metrics (RMSE, MAE, NDCG)

We see a 9.3% improvement in MAE and 9.4% increase in NDCG with the implementation of the readability model. This model incorporates review length along with several other text readability metrics, offering a more precise prediction.

The final stars model, emulating Huang et al., demonstrated the ability to predict and rank helpfulness for only top reviewers' reviews. Its findings were incorporated into the final business nature model.

Direct implementation of Madumbi and Schuff's business nature model yielded a RMSE of 0.3514 and a MAE of 0.2444. Compared to our baseline model, it demonstrated a 24.3% and 2.5% reduction in errors, respectively. The NDCG (k = all) showed an 18.3% improvement. Through utilization of BERT's contextual embeddings to quantify nature on a continuum, combined with `reviewer_cumulative_helpfulness`, the final business nature model demonstrated a slight improvement over the model based on Madumbi and Schuff. RMSE and MAE dropped by 4% and 2.5%, respectively. NDCG scores also increased across all values of k. Compared to our baseline, this model's NDCG scores proved business nature to be a key component of determining review helpfulness (Fig. 1).

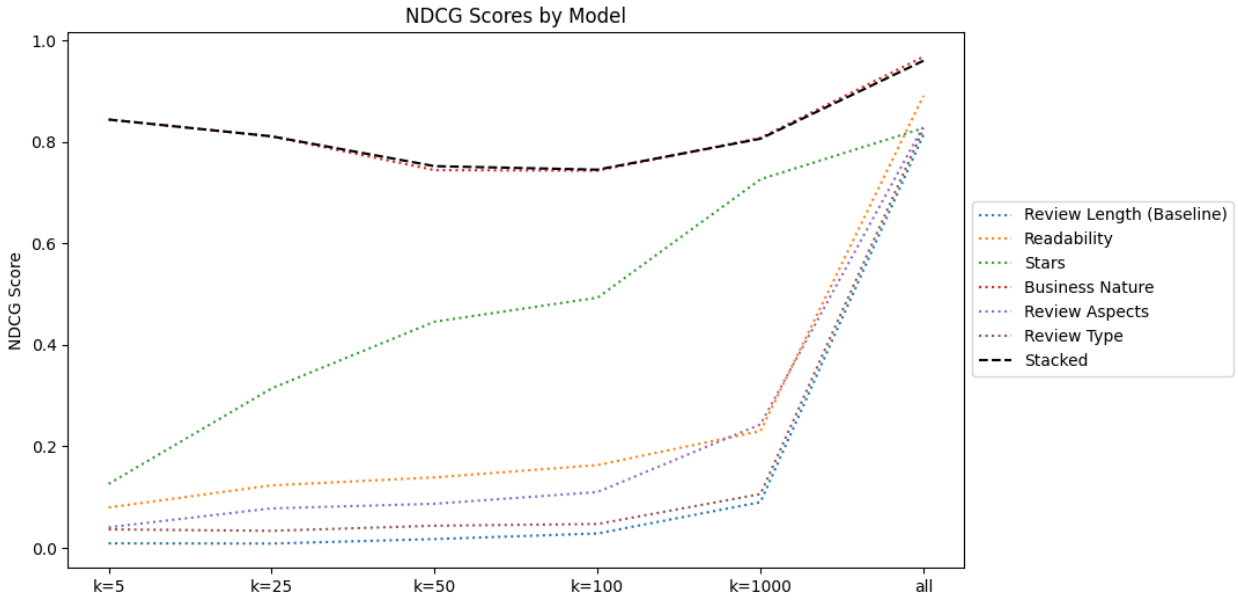

Figure 1: NGCG scores for all models

We tested the four key review aspect models outlined above. When first attempting the review aspect using a zero-shot classification technique and majority aspect approach, it yields an NDCG of 0.8283 for k = all. Defining the aspects as a percentage and utilizing a few-shot approach leads to a stronger NDCG of 0.8309, an 0.3% improvement over the initial model and a 2.1% improvement over baseline. Although the few-shot model led to a slightly higher NDCG, it is still lower than ideal. We hypothesize that this is because we largely adapted the aspects from the product review category, and business reviews tend to contain different information – often involving more storytelling of the experience.

The review type model, which incorporated classifications of reviews into regular, comparative, and suggestive categories, demonstrated a meaningful improvement over the baseline model (Table 1). This model achieved a reduction in MAE by 17.0% compared to the baseline. The model had the lowest MAE, as it was predicting the majority of reviews as not helpful. We believe this model did not perform as well due to the differences in review types found in business versus product reviews. With a majority of reviews classified as regular, it is possible that the types represented in Qazi et al.'s work do not translate evenly to business reviews. While the model performed slightly below other models like the readability model in terms of NDCG, its results contributed to the ensemble model.

The final stacked model showed considerable improvement over the review length, readability, review aspects, and review type models, with NDCG (k = all) improving by 17.9%, 7.8%, 15.5%, and 16.2%. When compared to the business nature model, the stacked model shows slight increases in both types of error. Despite this, NDCG scores are notably similar across the board (Fig. 1). The inclusion of the other models, even with the business nature model's individual performance, was decided upon to leverage the diversity of features and increase model robustness.

Across all models, a notable increase in NDCG was observed when k was set to all reviews. This increase is likely attributable to the imbalance and heavy skew in the helpfulness distribution, reflecting real world review helpfulness. With half the data having no helpfulness, their predictions and rankings provide an artificial inflation to the NDCG (k = all) score. Some error in our models could also be attributed to the limited scope (i.e. specific features) through which we examine reviews.

5 Conclusion

In this research, we aimed to develop a model that predicts and ranks the helpfulness of Yelp business reviews, addressing the challenge of information overload in online reviews. We successfully created an ensemble model

by adapting insights from adjacent domains, incorporating factors like review length, readability, business nature, review aspects, and review type. Our findings show that business nature plays a significant role in predicting review helpfulness, with the other factors playing moderating roles. Future work should explore model refinements, such as investigating the impact of more business-centric review aspects and review types. Future work could also utilize the prediction and ranking of review helpfulness to create business level summaries to further simplify customer decisions.

References

[1] G. Diaz and V. Ng, "Modeling and Prediction of Online Product Review Helpfulness: A Survey," 2018. Accessed: Oct. 04, 2024. [Online]. Available: https://aclanthology.org/P18-1065.pdf.

[2] H. Liu, Y. Gao, P. Lv, M. Li, S. Geng, M. Li, and H. Wang, "Using Argument-based Features to Predict and Analyse Review Helpfulness," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017, pp. 1358–1363, Association for Computational Linguistics.

[3] A. Ghose and P. G. Ipeirotis, "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 10, pp. 1498–1512, Oct. 2011, doi: https://doi.org/10.1109/tkde.2010.188.

[4] A. H. Huang, K. Chen, D. C. Yen, and T. P. Tran, "A study of factors that contribute to online review helpfulness," *Computers in Human Behavior*, vol. 48, pp. 17–27, Jul. 2015, doi: https://doi.org/10.1016/j.chb.2015.01.010.

[5] S. Mudambi and D. Schuff, "Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com," *MIS Quarterly*, vol. 34, no. 1, pp. 185–200, 2010, doi: https://doi.org/10.2307/20721420.

[6] Y. Yang, C. Chen, and Forrest Sheng Bao, "Aspect-Based Helpfulness Prediction for Online Product Reviews," Nov. 2016, doi: https://doi.org/10.1109/ictai.2016.0130.

[7] A. Qazi, K. B. Shah Syed, R. G. Raj, E. Cambria, M. Tahir, and D. Alghazzawi, "A concept-level approach to the analysis of online review helpfulness," *Computers in Human Behavior*, vol. 58, pp. 75–81, May 2016, doi: https://doi.org/10.1016/j.chb.2015.12.028.

[8] jamesdj, "tobit." GitHub. https://github.com/jamesdj/tobit, Accessed: Nov. 10, 2024.

[9] Hugging Face, "Setfit." GitHub. https://github.com/huggingface/setfit, Accessed: Dec. 1, 2024.