Predicting Restaurant Review Helpfulness and Evaluating Cuisine Bias

Jia Lu and Suzy Choi
University of California, Berkeley

Abstract

Customers rely on online reviews to make purchase decisions. However not all reviews are helpful and what is helpful may differ depending on product types and individual users. The purpose of this paper is to predict helpfulness of reviews and explore differences in review helpfulness by restaurant cuisine type. This study is based on Yelp reviews of American, Asian, and Mexican restaurants in Arizona, and explores several variations of models using Logistic Regression, LSTM, and BERT. The results show BERT as the best performing model and no consistent difference in predicting review helpfulness by cuisine.

1. Introduction

With the rise of applications like Yelp or Google My Business, consumers increasingly rely on reviews to make purchasing decisions. Therefore being able to identify potential useful reviews could offer potential value to consumers. However unlike sentiment analysis, predicting the usefulness of reviews is less straightforward.

Predicting review usefulness is more challenging when the reviews are of experience goods. Experience goods are products or services where attributes prior to consumption are difficult to observe. One major example would be restaurants--before dining in a restaurant it is difficult to obtain information on attributes like food taste, ambience, or service quality. This is why prior to going to restaurants, consumers rely heavily on reviews on applications like Yelp. However, reviews on experience goods tend to be subjective and influenced by the reviewer's bias on attributes like cuisine.

In this project, we use natural language processing and deep learning to predict helpfulness in yelp reviews and detect any culture bias in restaurants of different cuisine types.

2. Related Work

Previous research showed that a variety of factors including linguistic factors (such as review length), sentiment polarity (positive/ negative), and business/user characteristics (user reputation, business rating) affect review helpfulness [6] [19].

Some studies created models to predict the review helpfulness percentage in which case most used linear regression [6] or support vector regression (SVR). Other studies have used logistic regression, support vector classification (SVC), random forest [3] to predict whether a review is helpful or not. Deep learning methods such as Convolutional Neural Network [13], various forms of Recurrent Neural Network including Long Short Term Modeling [5] have also been applied to the problem of predicting review helpfulness.

Most review helpfulness studies used task-agnostic embedding layers, such as Word2Vec or GloVe, which could not capture context related word features. The Bidirectional Encoder Representations from Transformers (BERT) has been introduced recently and has achieved state-of-the-art results in various NLP tasks [1]. Xu et al. [18] applied the BERT feature based model to predict the helpfulness of Amazon product review and achieved high accuracy. Song et al. [14] used BERT as the embedding layer for end-to-end aspect-based sentiment analysis and outperformed existing models for both laptop and restaurant data.

Studies on cross-cultural difference or cultural bias on reviews have been less studied, and even those studies focus more on sentiment differences. Wenyi Tay [15] used aspect-sentiment classification and opinion clustering. Nakayama and Wan [9**]** compared the Japanese and American yelp reviews to identify the difference in sentiment for aspects such as service quality, food quality, and price fairness.

3. Methods

Previous attempts have reached an accuracy rate of 70-80% using handcrafted features and feeding them into support vector machines or RNN/LSTM models [5][16]. We aim to reach an accuracy rate beyond 80%. Accuracy will be our evaluation metric for ease of comparison with related studies.

This study focuses on how people perceive review helpfulness differently depending on cuisine type. The hypothesis is that there is a difference in what people look for in reviews that they consider helpful depending on cuisine. We explore the application of BERT to predict restaurant review helpfulness across different cuisine types. We will train BERT alongside other logistic regression models and neural network models. We expect the contextualized word embedding layer of BERT to increase predictive power.

3.1 Dataset details

We use the Yelp review dataset (https://www.kaggle.com/yelp-dataset/yelp-dataset) as the primary data. The review data contains over 8 millions reviews across various industries and countries. We limit our analysis to restaurants in Arizona because it is the state with most businesses in the Yelp data. The business category field is used to determine the cuisine type. Chinese, Korean, Japanese, Thai, Sushi Bar, and Vietnamese restaurants are grouped together to form the Asian cuisine group. American, Pizza, Burgers, Barbeque, Sandwiches, Chicken Wings restaurants are grouped to form the American cuisine group.

The review helpfulness is determined by the count of “useful” votes. The unhelpful reviews include reviews with 0 useful votes. The review data is skewed with 58% of the reviews having no useful votes, 20% having 2 or more votes, and 1% having over 10 votes. In order to evaluate the impact of different useful vote thresholds, we create two helpful review categories: the helpful review (votes > 1) group contains reviews with 2 or more useful votes and the helpful review (votes > 9) group contains reviews with 10 or more useful votes. To make sure the class populations are balanced, we randomly select 10,000 reviews from each category.

3.2 Baseline Model

For the review helpfulness classifier, we created several models and compared the model performance based on accuracy. We first gauged a good baseline by applying Bag-of-Words and TF-IDF with logistic regression.

The bag-of-words (BoW) model keeps track of the frequency of words while the Term Frequency-inverse document frequency (TF-IDF) measures how frequently a term appears in the document multiplied by the inverse document frequency.

3.3 Model Architecture

For our deep learning models, we compared BERT with several Long Short Term Memory (LSTM) architectures. We trained generic classifiers across all cuisine types, as well as cuisine specific classifiers. Long Short Term Memory (LSTM) models were tested with various hyperparameters, use of attention, and pre-trained GloVe embeddings [12].

As mentioned in the literature review section, previous studies have pointed to review extremity (polarity), review length, and peripheral factors (product/user/business-specific) as factors that affect review helpfulness [19]. To capture these features we added review length, star deviation, and total reviews per business into the LSTM model. We fed the LSTM output and above-mentioned review meta-data into fully connected layers (Figure 1). Activation function RELU was used for the fully connected layers and sigmoid was used for the output layer. Review length was calculated by the number of words in the review, and pulled the number of reviews per business from our original dataset.

For the BERT models, we used BERT Base, which has 12 encoder layers, 768 hidden dimensions, and 12 self-attention heads. For the first BERT model, we fed the pretrained BERT layer into the classifier. For the second BERT model, we retrained the last six layers of BERT. We then fed the BERT output into a fully connected neural network (Figure 2). The activation function RELU was used for the fully connected layers. (Please refer to the Appendix for model details)

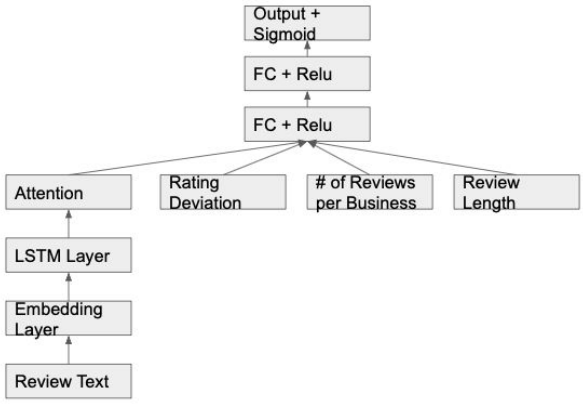


Figure 1. The LSTM model with review meta data

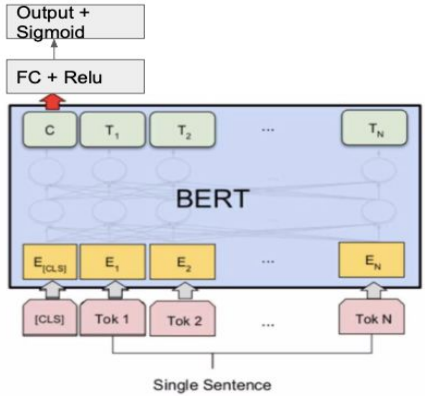


Figure 2. BERT model

3.4 Training Details

Logistic Regression

The NLTK tokenizer class was used to tokenize the text after removing stopwords and capitalization. CountVectorizer and TfidfVectorizer were used to extract the features which were input into the logistic regression model.

LSTM

Prior to running the deep learning classification models, we performed tokenization on the text to make the text easier to featurize. The keras tokenizer class was used to turn the text into a sequence of integers. The vocabulary size of 10000 was used. Reviews longer than 500 words were truncated because most of the reviews contained less than 500 words.

The embedding layer sits between the input and the LSTM layer, and is used to create word vectors for incoming words. The embedding layer can be initialized randomly or with word embeddings like Glove or word2vec. For our models, we created one with random initialization and another with GloVe embeddings. We use word embedding vectors of 200 dimensions for our models.

BERT

For BERT models, the BERT uncased tokenizer was used to preprocess review text. Consistent with the LSTM models, we padded the reviews that are shorter than 500 words and truncated the reviews that are longer than 500 words. We tested various hyperparameters and landed with: Adam as the optimizer, binary cross entropy as the loss function, 0.0005 learning rate, and maximum of 5 epochs. Out of the various learning rates we tried, 0.0005 showed the best performance. In the case of the number of epochs, we found that having an epoch number higher than 5 did not improve the performance on the testing data. Since previous research showed that dropout could help reduce overfitting problems, dropout of 0.1 for dense layers was applied. We split the data so that we could use 20% of the data as the testing data.

4. Results

The results are summarized in Table 1. It shows that the LSTM model with additional features performed better than logistic regression models and other LSTM models across cuisine types. Additionally, the LSTM with pre-trained GloVe embeddings generally outperformed the LSTM with trained embeddings, which is consistent with previous studies.

Table 1. Accuracy Rates for Predicting Helpful Reviews

	Helpful Reviews (Votes > 9)				Helpful Reviews (Votes > 1)			
Models	**Overall**	**Asian**	**Amer.**	**Mex.**	**Overall**	**Asian**	**Amer.**	**Mex.**
Logistic Regression w/ BoW	78.6	79.5	77.9	79.2	63.8	66.2	64.1	63.3
Logistic Regression w/ TF-IDF	80.5	80.6	79.4	78.0	66.9	67.9	66.0	64.9
LSTM w/ attention	78.2	76.8	79.1	80.0	66.5	67.2	66.1	63.4
LSTM w/ attention and GloVe	79.6	76.7	79.2	81.0	68.5	69.9	68.2	65.1
LSTM w/ additional features	79.5	79.0	80.1	82.1	**69.3**	69.9	69.7	67.0
Bert Fine Tuning	**83.0**	**82.9**	**82.4**	**84.6**	68.9	**71.2**	**70.6**	**70.4**

The BERT models consistently outperformed logistic regression models and LSTM models for all cuisine types. In particular, the BERT model for Mexican restaurant reviews achieved the highest accuracy rate of 84.6%. The accuracy rates for Asian cuisine and American cuisine are similar to the accuracy rate of overall cuisine. Our experiments showed retraining the last six layers of BERT performed better than using pre-trained BERT layers. Therefore we retrained the top six BERT layers for all the BERT models. We also obtained better results with a two-layer neural network compared with a single-layer neural network as the classifier.

For the task of predicting helpful reviews (votes > 1), BERT is still the best model for individual cuisine datasets, but not for the overall dataset. It is also interesting to see that the BERT models trained on individual cuisine datasets generate higher accuracy than the BERT model trained on the overall dataset. This is consistent with our hypothesis that there are cultural biases in the review context, and BERT models for individual cuisines, compared with overall models, are better at capturing the context nuances.

5. Conclusion and Future Work

This study shows that the fine tuning BERT works the best in predicting restaurants review helpfulness. The average accuracy rate of the individual cuisine models, however, did not consistently perform better than the overall model. It could be that unlike what we initially hypothesized, the factors that predict review helpfulness are similar across the cuisines in the study (American, Mexican, and Asian). Alternatively, there may exist a difference across the cuisines but our models were unable to capture the subtleties consistently due to the small sample size. Future research could test this with larger datasets.

Future studies in understanding the subtle cultural differences between reviews could consider reviews from different countries. Nakayama and Wan's study compared Yelp reviews of Japanese restaurants in Japan and Western countries (which the study categorized as US, UK, Canada, and Germany). Nakayama and Wan [9] chose 10 entree items and extracted sentiment phrases associated with each entree. After categorizing the phrases into one of four aspects (food, service, physical environment, and price fairness),

they tabulated the correlation values (review proportion of an entree item given a sentiment over the review proportion of an entree item over all reviews) by sentiment and helpfulness. Results showed that when it came to usefulness, Yelp reviews from Western countries focused more on the negative price fairness aspect than the ones from Japan, whereas reviews from Japan focused more on the positive price fairness aspect.

It may also be interesting to see if such a difference can also be captured within the same country/city when the user profile is not as different as it is in Nakayama and Wan's study.

Reference

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805. https://arxiv.org/abs/1810.04805

[2] Ying Ding and Jing Jiang. 2015. A Joint Model of Product Properties, Aspects, and Ratings for Online Reviews. Proceedings of Recent Advances in Natural Language Processing, 131-137. https://www.aclweb.org/anthology/R15-1019.pdf

[3] Srikumar Krishnamoorthy. 2015. Linguistic features for review helpfulness prediction. Expert Syst. Appl. 42, 3751–3759.

[4] Xin Li and Lidong Bing and Wenxuan Zhang and Wai Lam. 2019. Exploiting BERT for End-to-End Aspect-based Sentiment Analysis. arXiv:1910.00883. https://arxiv.org/abs/1910.00883

[5] David Liu. 2018. Understanding and Predicting the Usefulness of Yelp Reviews. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2760995.pdf

[6] Susan M. Mudambi and David Schuff. 2010. What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. MIS Quarterly, Vol. 34, No. 1, pp. 185-200. https://www.jstor.org/stable/20721420

[7] Makoto Nakayama and Yun Wan. 2017. Is culture of origin associated with more expressions? An analysis of Yelp reviews on Japanese restaurants. Tourism Management, 66, 329-338. https://doi.org/10.1016/j.tourman.2017.10.019

[8] Makoto Nakayama and Yun Wan. 2019. Same sushi, different impressions: a cross-cultural analysis of Yelp reviews. Inf Technol Tourism 21, 181–207.

[9] Makoto Nakayama and Yun Wan. 2019. Cross-Cultural Examination on Content Bias and Helpfulness of Online Reviews: Sentiment Balance at the Aspect Level for a Subjective Good.

[10] Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and Prediction of Online Product Review Helpfulness: A Survey. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), 698–708. https://www.aclweb.org/anthology/P18-1065.pdf

[11] Marco Passon, Marco Lippi, Giuseppe Serra, Carlo Tasso. 2018. Predicting the Usefulness of Amazon Reviews Using Off-The-Shelf Argumentation Mining. Proceedings of the 5th Workshop on Argument Mining, 35–39. https://www.aclweb.org/anthology/W18-5205.pdf

[12] Jeffrey Pennington, Richard Socher, Christopher Manning. 2014. GloVe: Global Vectors for Word Representation.

[13] Sunil Saumya. 2020. Predicting the helpfulness score of online reviews using convolutional neural network. Soft Computing.

[14] Youwei Song, Jiahai Wang, Zhiwei Liang, Zhiyue Liu and Tao Jiang. 2020. Utilizing BERT Intermediate Layers for Aspect Based Sentiment Analysis and Natural Language Inference. arXiv:

2002.04815. https://arxiv.org/abs/2002.04815

[15] Wenyi Tay. 2019. Not All Reviews are Equal: Towards Addressing Reviewer Biases for Opinion Summarization. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 34–42. https://www.aclweb.org/anthology/P19-2005.pdf

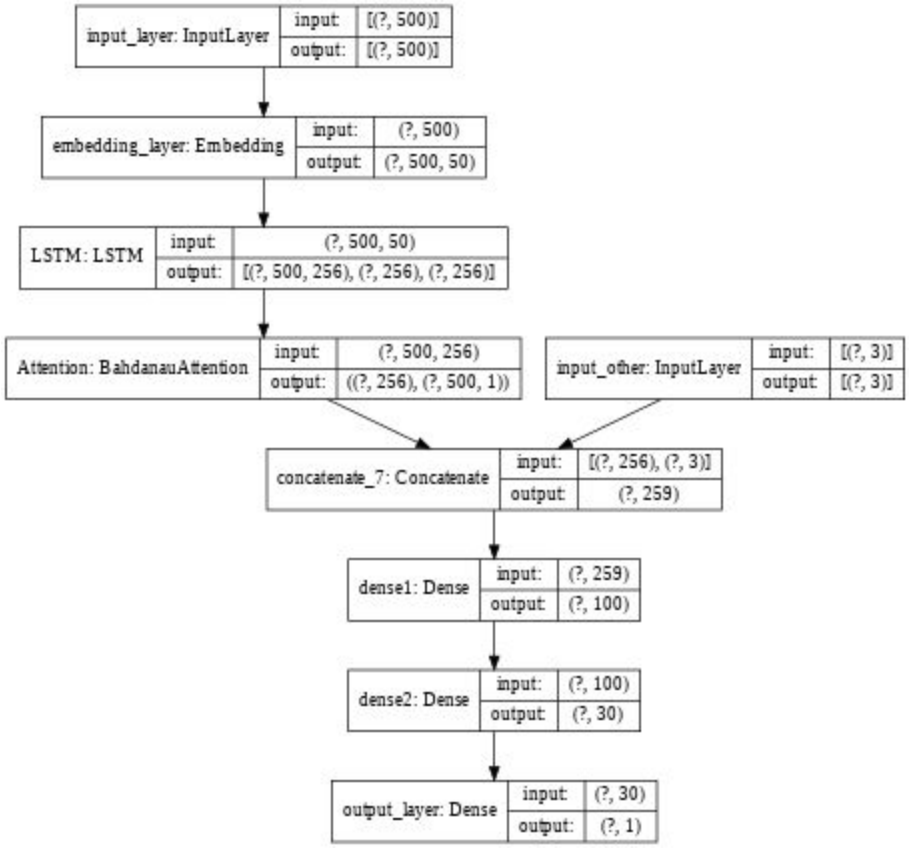
[16] James Wei, Jessica Ko, Jay Patel. 2016. Predicting Amazon Product Review Helpfulness.

[17] Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. arXiv:1805.04601. https://arxiv.org/abs/1805.04601

[18] Shuzhe Xu, Salvador E. Barbosa, Don Hong. 2020. BERT Feature Based Model for Predicting the Helpfulness Scores of Online Customers Reviews. In: Arai K., Kapoor S., Bhatia R. (eds) Advances in Information and Communication. FICC 2020. Advances in Intelligent Systems and Computing, vol 1130. Springer, Cham. https://doi.org/10.1007/978-3-030-39442-4_21

[19] Guopeng Yin, Li Wei, Wei Xu, and Minder Chen. 2014. Exploring Heuristic Cues For Consumer Perceptions Of Online Reviews Helpfulness: The Case Of Yelp.Com. PACIS 2014 Proceedings. 52. http://aisel.aisnet.org/pacis2014/52

Appendix I Model Architecture for LSTM with additional Features



Appendix II Model Architecture for BERT