Yelp Food and Service Aspect-Based Sentiment Analysis

Problem Definition

The goal of the project is to use aspect-based sentiment analysis to classify the polarity of food and service reviews on Yelp. This project implemented Natural Language Processing (NLP) techniques and machine learning classification algorithms to classify "positive" and "negative" reviews.

A recurring business concern is connecting consumers to products that are a good fit for them. This project uses machine learning and NLP to help Yelp users more accurately predict if they will have a positive experience based on their preferences. Business owners on Yelp will also benefit from having target customers seek them out for goods and services. The visualization and interactivity portion of this project will allow Yelp users to write reviews as usual, with the advantage of having a real-time display of a binary scoring for food and/or service. Similarly, Yelp business owners are able to see an analysis of their business through reviews and use the web-based tools to drill-down the data further.

Introduction

Social commerce platforms have always aimed to connect customers to relevant goods and services by transforming review data into business insights. With increasing amounts of review data, an ideal solution not only processes reviews in a time-efficient manner, but also ensures that the information extracted is still meaningful. Sentiment analysis is one area of natural language processing (NLP) that can be utilized to derive business insights from text-based reviews.

Processing text-based reviews data into a form appropriate for an NLP algorithm is a hot topic of research. Nakayama and Wan's paper [5] states that based on sentiment analysis of Yelp reviews from different language versions of Yelp, culture does impact reviews and ratings on Yelp (2019). The paper describes an algorithm to label reviews as "positive" or "negative", which we used for our project. However, the strategy the authors came up with to evaluate businesses in multiple categories leads to a smaller dataset focused on 10 keywords, which does not apply to our project.

Artificial neural networks are popular choices of algorithms for sentiment analysis. In [1], the authors attempt to create a Review Rating Prediction model. Through their analysis they identify that convolution Long Short Term Memory (CLSTM) emerges as the preferred model for generating accurate rating predictions based on their reviews (Rafay et al., 2020). In addition, the authors in [11] and [13] apply a neural network model to classify the document sentiments. Wang and Gang (2018) also discuss using Convolutional Neural Network (CNN) for emotion analysis. Tang et al. (2015) finds that their model outperforms a number of state of the art algorithms and that gated Recurrent Neural Networks (RNN) performs better than standard RNN. In contrast, Wen et al. (2018) uses a combination of RNN and CNN frameworks. In Chuttur and Pokhun's paper, two neural network algorithms tested, CNN and LSTM, did not perform well in classifying Yelp reviews into one of several different emotions (2021). These papers are helpful because they describe what neural network and optimizer combinations are most accurate in sentiment analysis of Yelp reviews.

Research has also found that supervised learning algorithms can be effective for sentiment analysis. In [14], the authors discuss a method of predicting star ratings using a combination of text, reviewer history, and restaurant statistics, and the random forest tree algorithm yielded the best results in their experiment (Yu et al., 2015). Similarly, in [7], the authors tested various methods for feature extraction and modeling. Another paper identified that high-performing algorithms include Linear support vector classification and stochastic gradient descent (Salinca, 2015). Whereas in Faisol et al.'s paper, a combination of combined unigram and bigram feature extraction and a Gaussian Naive Bayes training algorithm is most effective at classifying Yelp reviews as "positive" or "negative" (2020). This paper gives us potential preprocessing methods and training algorithms for our project. However, the paper's training dataset only had 1000 rows. We plan to train with a larger dataset with 10000 rows.

Some of the risks and drawbacks include false, biased, or outlying data. In [4], Luo et al. (2019) uses LDA to find key categories in their Yelp reviews. They also conclude that reviews in expensive metropolitan areas speak more harshly about value and its relation to those factors. Although, a sparsely

written review could also affect the performance. Using LDA for short texts could be misleading because the algorithm might not find related topics and ignore the review completely. We plan to use aspect-based sentiment analysis which has limitations. Current, popular techniques do not properly represent aspect-terms and aspect-categories, thus, impacting the prediction performance. Tan et al. (2020) attempts to correct this issue by using aligning aspect embedding. That being said, the payoff of this project is that users can quickly assess whether the food and service were positive or negative regardless of the 5-star rating system. Visually showcasing the polarity will cut time and help customers more because they are able to make rounded decisions about consuming goods or services from the restaurant and business owners are able to get an automated analysis of their strengths and weaknesses.

Proposed Method

From a product standpoint, our approach explores feature additions to an existing service i.e. Yelp. Our first innovation is to offer real-time sentiment analysis of food and/or service after a review is written. The interactive visualization we made displays the review, along with the given star, and a predicted "thumbs up" or "thumbs down" if the review contains an opinion on the business's food/service. This innovation not only gives users another layer of information to base consumption decisions off of, but also automates the process of evaluating a review. The process of opinion mining a review can retroactively be used on older reviews that contain sentiments about food/service. Our second innovation is to create an interactive tool that allows Yelp users, both business owners and consumers, to drill-down the review information given in the form of data visualizations. The success of our project could be measured as an increase in positive reviews, since with the aid of our approach, users can find target restaurants that better fit their preferences. From a research standpoint, we implemented various NLP models as mentioned above, including aspect extraction methods that can detect sentiment such as BERT. Nandwani and Verma review various NLP techniques for sentiment analysis, including an overview of emotion detection (2021). We also took high-performing algorithms for sentiment analysis of Yelp reviews described in papers above, conducted an experiment that compares prediction metrics and training time of the algorithms, and selected the overall best-performing algorithm to classify reviews for our interactive visualization to ensure that our product is better than the state of the art. By exploring various NLP and classification algorithms, our project is successful in developing a model that correctly classifies reviews with high accuracy, while mitigating as many drawbacks as possible, such as algorithm runtime.

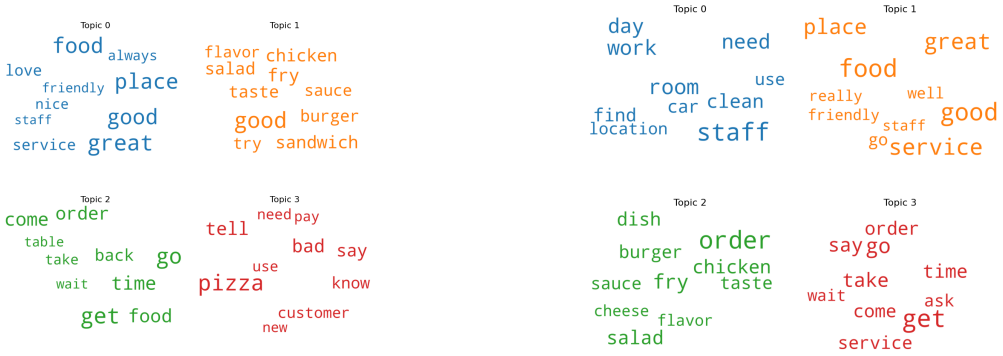
This project incurs no costs. Our dataset is a free-to-download dataset from Yelp. Python is compiled through Google Collab and Jupyter Notebooks. Our web application is hosted on AWS EC2 using the streamlit package. S. and Ramathmika investigate whether NLP and classification algorithms can effectively classify Yelp reviews into a positive or negative sentiment, and they also describe the hardware used for this research, helping us further evaluate project costs (2019).

After downloading the data from Yelp's website, we kept these files contained in its extracted zipped folder: review.json and business.json. We used python's Pandas library to read the json file into a dataframe and removed columns irrelevant to our analysis and experiment. The final dataset, saved as a csv file, is 1.2 GB in size.

All the text reviews were pre-processed by two methods: stemming and stopwords removal. Firstly, all the textual data were transformed into lower-case and sentences were split into lists of single words. Punctuation, emoticons, numbers, and extra spaces were removed. The Natural Language Toolkit (Nltk) modules in python were imported to remove stopwords including ("you", "I", "we", "she", "the", etc.) [15] and they could not provide any information for meaningful analysis. To further simplify the dataset, stemming was used to reduce all forms of words to their roots, e.g. tasting, tasted, taste, tastes; these words are reduced to "taste" to ease grouping and processing.

We implemented LDA (similar to Luo et. al, 2019), Latent Dirichlet Allocation, on the 1,000,000 row dataset as a clustering technique for finding words related to food and service; although, topic modeling can be subjective, as it is up to the researcher to understand the domain of the dataset. Results of our LDA model are shown below. Four topics were chosen as the output; fewer topics created

over-generalized groupings and more topics resulted in some overlap between the dominant words. The word cloud on the left is for food words and the right word cloud is for service words. It can be observed that the food word cloud is dominated by types of food (nouns) and other adjectives to describe the quality of the food. Topics 1 and 3 in the service word cloud are clearly focused on the quality of service and it can be argued that these two topics represent good and bad service topics, respectively.



Reviews were subsequently labeled as “positive” (= 1) or “negative” (= 0). If the review’s star rating was 1 or 2 stars, it was labeled as “negative”. If the star rating was 4 or 5 stars, it was labeled as “positive”, else for a rating of 3 stars, the review is dropped from the dataset to prevent sentiment ambiguity. After labeling the reviews, we shuffled and reduced the size of the dataset to 200,000 rows, and balanced the ratio of the positive and negative reviews to a 1:1 ratio. The review dataset was then merged with the business dataset using a left join on business_id so it could be used for data visualization.

We filtered the merged dataset down to reviews that contain food and service related words as identified by the corpus produced from topic modeling. Since our objective is to identify food and service sentiments, we trained separate models for food and service for prediction. In order to do so, we created two similar functions (get_food_sent, get_service_sent) that returns food and service related sentences of each review for both models; for sentences containing words in both food and service corpus, we part-of-speech tagged target-words along with its associated adjective using python’s NLTK package. The latter attempts to address the problem of aspect-based sentiment analysis. While it is common for positive reviews to have both positive sentiments for food and service and vice versa, we did not manually look into reviews that have differing sentiments. This could cause a decrease in reported accuracy when tested.

Feature extraction and feature selection techniques were used to vectorize the dataset before model training. We used a combined unigram and bigram feature extraction method as part of the data pre-processing and TF-IDF for data vectorization. Our reasoning for the use was because Faisol et al. and Ruchi, S. and Jongwook, W.’s papers used the same feature extraction and data vectorization methods.

After vectorization, the dataset then was trained on 6 different models: Multinomial Naive Bayes, BERT, CLSTM, bidirectional LSTM, Random Forest, and linear Support Vector Machine. Rather than combine models, we compared these proposed models to see which model works best with our preprocessing techniques, dataset, and imposed time constraints. Our exploratory data analysis made with Seaborns, MatplotLib, and WordCloud was incorporated with our web application using Streamlit.

Exploratory Data Analysis

During the project, statistics such as histograms and boxplots were implemented to explore data insights and distribution of review ratings. A. Rafay (2020) confirmed the distribution could be skewed to 4 and 5 stars rating and Yelp reviews become biased.

The distribution of star-ratings in the Yelp dataset after filtering for reviews that are food- or service-related is 119,711 (11.71%) one-star reviews, 87,007 (8.51%) two-star reviews, 117,362 (11.48%) three-star reviews, 246,836 (24.13%) four-star reviews, and 451,817 (44.18%) five-star reviews.

Figure 1: Star Rating Distribution

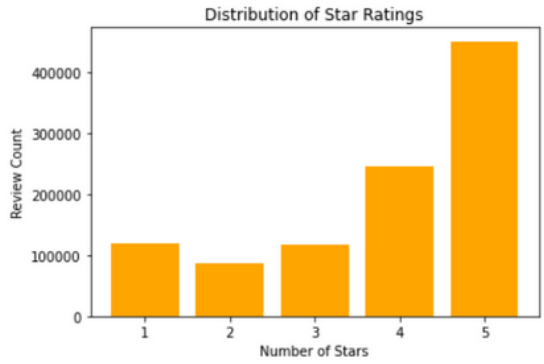
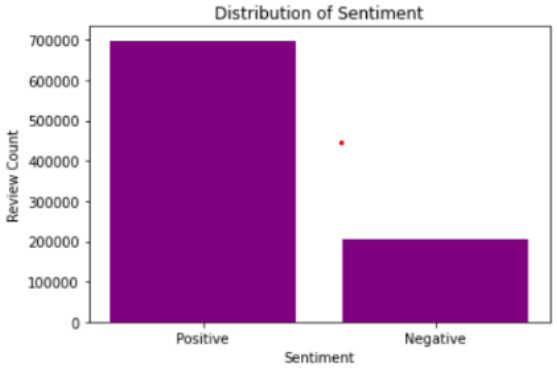
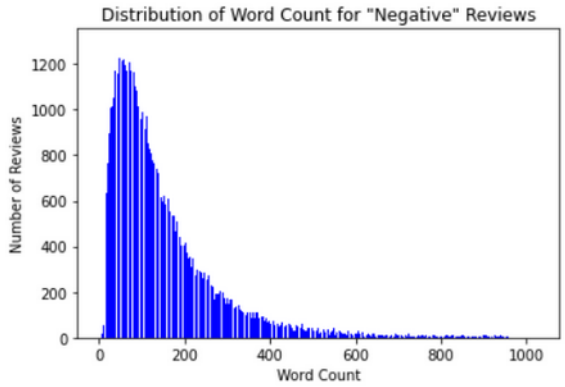


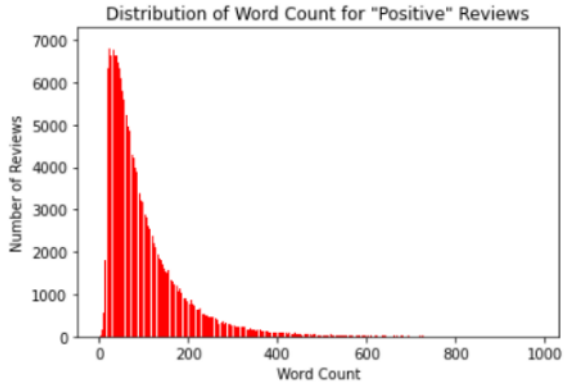
Figure 2: Sentiment Distribution



After removing three-star reviews and labeling data, but before balancing the dataset, 698653 (68.31%) of the reviews in the remaining Yelp dataset were labeled as "positive", while 206718 (20.21%) of the reviews were labeled "negative." This distribution is visualized in Figure 2, above.

We also explored the relationship between the word count reviews and their corresponding sentiment. "Negative" reviews, on average, have a higher word count (148 words) than that of "positive" reviews (100 words), and have a higher standard deviation of word counts (128 words) than that of "positive" reviews (91 words). These relationships are visualized in the figures below.





Experiments/Evaluation

Our project seeks to identify a classification algorithm that classifies the sentiment Yelp reviews to a high degree of accuracy and in a timely manner, so that we can incorporate the optimal algorithm into our product. Consumers and business owners active on Yelp are the main beneficiaries of this project since their user experience will be impacted favorably. Our approach will add another level of granularity for the user, leading to more informed decisions. One potential impact is a decrease in negative reviews on Yelp, as people will likely take into account the additional sentiment analysis and topic modeling in their decision making. This would be done by continuously updating and tracking restaurant reviews using our models. If our research could be implemented into the Yelp product, another metric could include accuracy of models developed through distributed federated learning, as discussed by Si et. al (2022), which is more relevant today as virtually everyone owns a smartphone for processing.

We conducted experiments that compared different classification algorithms' ability to accurately classify the sentiment of reviews in a timely manner. The questions our experiments were designed to answer were as follows: what classification algorithm yields the highest quality model for either food-related reviews or service-related reviews? Are the differences in runtime for each classification algorithm substantial enough to be a valid trade-off for accuracy in the final product?

We used a processed version of the Yelp dataset for the experiments, starting with the preprocessing steps described in the "Proposed Method" section, and after the step where we created two

distinct datasets for food and service-related reviews, we further reduced the size of each dataset to 10,000 rows, 5,000 positive-sentiment rows and 5,000 negative-sentiment. We then trained the classification algorithms we were investigating on each of these two 10,000 row datasets. After training, for each of the six models for each of the two datasets, we recorded the accuracy (Salinca, A. (2015)), the precision, recall and f1-score for negative sentiment and positive sentiment, and the approximate training time for the models.

Multinomial Naive Bayes, Random Forests, and Linear SVM had model hyperparameters tuned through 5-fold cross-validation of each training set of food or service using GridSearchCV from sklearn. Then, using their corresponding optimized hyperparameters, the food training set and the service training set each had a model trained on it, and the trained models were then used to predict labels for their respective food or service test set, and model quality metrics were computed from the results.

BERT	**Accuracy**		**Precision**	**Recall**	**f1-score**
Food Data	0.83	**0 (negative)**	0.90	0.76	0.83
		1 (positive)	0.76	0.90	0.83
Service Data	0.81	**0 (negative)**	0.93	0.75	0.83
		1 (positive)	0.68	0.90	0.78

Multinomial Naive Bayes	**Accuracy**		**Precision**	**Recall**	**f1-score**
Food Data	0.82	**0 (negative)**	0.79	0.80	0.79
$\alpha = 1$		**1 (positive)**	0.85	0.83	0.84
Service Data	0.80	**0 (negative)**	0.78	0.81	0.80
$\alpha = 0.1$		**1 (positive)**	0.82	0.79	0.80

Random Forests	**Accuracy**		**Precision**	**Recall**	**f1-score**
Food Data	0.80	**0 (negative)**	0.76	0.78	0.77
200 estimators Max depth = 75		**1 (positive)**	0.83	0.82	0.83
Service Data	0.78	**0 (negative)**	0.74	0.80	0.77
200 estimators Max depth = None		**1 (positive)**	0.81	0.75	0.78

Linear SVM	**Accuracy**		**Precision**	**Recall**	**f1-score**
Food Data	0.81	**0 (negative)**	0.82	0.77	0.79
$C = 1$		**1 (positive)**	0.81	0.85	0.83
Service Data	0.80	**0 (negative)**	0.78	0.82	0.80
$C = 1$		**1 (positive)**	0.83	0.79	0.81

C-LSTM	Accuracy		Precision	Recall	f1-score
Food Data	0.80	**0 (negative)**	0.80	0.80	0.80
		1 (positive)	0.77	0.77	0.77
Service Data	0.81	**0 (negative)**	0.85	0.82	0.83
		1 (positive)	0.72	0.72	0.72

Bidirectional LSTM	Accuracy		Precision	Recall	f1-score
Food Data	0.76	**0 (negative)**	0.79	0.79	0.79
		1 (positive)	0.76	0.81	0.78
Service Data	0.74	**0 (negative)**	0.73	0.73	0.73
		1 (positive)	0.71	0.75	0.73

BERT does fairly well in predicting food and service sentiments, however, it has a train time of two hours given 10,000 rows of data. As the model scales to size to retroactively predict older reviews of Yelp, BERT is not an ideal model to choose. Multinomial Naive Bayes also does fairly well in predicting food and service sentiments, and it has a train time of half a minute given 10,000 rows of data. Random forest does relatively poorly in predicting food and service sentiments compared to the BERT and multinomial Naive Bayes, and it has a train time of about 45 minutes for 10,000 rows of data. Linear SVM does better than random forest in predicting food and service sentiments, but worse than Multinomial Naive Bayes. Since it takes 10 minutes to train on 10,000 rows of data, linear SVM is an inferior choice of algorithm to multinomial Naive Bayes for the purposes of our project.

For both food and service data, BERT has a pattern of having better precision for negative reviews than that for positive reviews, and vice versa for recall. Multinomial Naïve Bayes, Random Forest, and Linear SVM generally share the same pattern of having worse metrics (i.e. precision, recall, f1-score) for predicting the sentiment of negative reviews than that of positive reviews for the food data. For service data, precision tends to be worse for negative reviews than positive, and vice versa for positive, which evens out to an about equal f1-score for both sentiments

Convolution Long Short Term Memory (C-LSTM) performed better with 80% accuracy for food data and 81% for service data, it has a train time of an hour given a small size rows of data. Long Short Term Memory (LSTM) performed lower among all other models and achieved 76% for predicting food and 74% for predicting service data. LSTM required three to four hours of train time given 10,000 rows of data using epoch =1 which consumed a huge amount of hardware resources.. Increasing the number of epochs would help the model generalize better to perform well in the real world. Both deep learning algorithms have better precision, recall, and f1-score for negative reviews.

Conclusions and Discussion

All team members have contributed a similar amount of effort.

References

1. A. Rafay, M. Suleman and A. Alim, Robust Review Rating Prediction Model based on Machine and Deep Learning: Yelp Dataset. *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, 2020, pp. 8138-8143, doi: 10.1109/ICETST49965.2020.9080713.
2. Chuttur, Y., Pokhun, L. (2021). An Evaluation of Deep Learning Networks to Extract Emotions from Yelp Reviews. In: Panigrahi, C.R., Pati, B., Pattanayak, B.K., Amic, S., Li, KC. (eds) Progress in Advanced Computing and Intelligent Engineering. Advances in Intelligent Systems and Computing, vol 1299. Springer, Singapore. https://doi.org/10.1007/978-981-33-4299-6_5
3. Faisol, H., Djajadinata, K., & Muljono, M. (2020). Sentiment Analysis of Yelp Review. 2020 International Seminar on Application for Technology of Information and Communication (iSemantic), 179-184.
4. Luo, & Xu, X. (2019). Predicting the Helpfulness of Online Restaurant Reviews Using Different Machine Learning Algorithms: A Case Study of Yelp. Sustainability (Basel, Switzerland), 11(19), 5254–. https://doi.org/10.3390/su11195254
5. Makoto Nakayama, Yun Wan, The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews, Information & Management, Volume 56, Issue 2, 2019, Pages 271-279, ISSN 0378-7206, https://doi.org/10.1016/j.im.2018.09.004
6. Nandwani, P., Verma, R. A review on sentiment analysis and emotion detection from text. *Soc. Netw. Anal. Min.* **11**, 81 (2021). https://doi.org/10.1007/s13278-021-00776-6
7. Ruchi, S., Jongwook, W. (2019). Applications of Machine Learning Models on Yelp Data, Volume 29, No 1. https://doi.org/10.14329/apjis.2019.29.1.35
8. Salinca, A. (2015, September). Business reviews classification using sentiment analysis. In *2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)* (pp. 247-250). IEEE.
9. S., H., & Ramathmika, R. (2019). Sentiment Analysis of Yelp Reviews by Machine Learning. 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 700-704.
10. S. Si, J. Wang, R. Zhang, Q. Su and J. Xiao, "Federated Non-negative Matrix Factorization for Short Texts Topic Modeling with Mutual Information," 2022 International Joint Conference on Neural Networks (IJCNN), 2022, pp. 1-7, doi: 10.1109/IJCNN55064.2022.9892602.
11. Tan, Cai, Y., Xu, J., Leung, H.-F., Chen, W., & Li, Q. (2020). Improving aspect-based sentiment analysis via aligning aspect embedding. Neurocomputing (Amsterdam), 383, 336–347. https://doi.org/10.1016/j.neucom.2019.12.035
12. Tang, D., Qin, B., & Liu, T. (2015, September). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1422-1432).
13. W. Wang and J. Gang, "Application of Convolutional Neural Network in Natural Language Processing," 2018 International Conference on Information Systems and Computer Aided Education (ICISCAE), 2018, pp. 64-70, doi: 10.1109/ICISCAE.2018.8666928.
14. Wen, S., & Li, J. (2018, December). Recurrent convolutional neural network with attention for twitter and yelp sentiment classification: ARC model for sentiment classification. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence* (pp. 1-7).
15. Yi, L., Xiaowei X.,(2021, April). Comparative Study of Deep Learning Models for Analyzing Online Restaurant reviews in the era of COVID-19 Pandemic. *International Journal of Hospitality Management, Volume 94*. https://doi.org/10.1016%2Fj.ijhm.2020.102849
16. Yu, M., Xue, M., & Ouyang, W. (2015). *Restaurants Review Star Prediction for Yelp Dataset*. Technical Report 17. UCSD.