To Yelp or Not to Yelp

Brendan Foo
Data W266: Natural Language Processing
UC Berkeley School of Information
brendanfoo@ischool.berkeley.edu

Abstract

This study explores the impact of customer reviews, particularly on Yelp, in shaping the reputations of restaurants. It investigates the efficacies of using large language models for summarization tasks and regression models to predict restaurant ratings. The goal is to provide a concise yet accurate summary of the restaurant's reputation based on given customers' reviews of said restaurant. This will help customers in their decision-making process in determining which restaurant best fits their preferences while also enhancing a restaurant-owner's understanding of their reputation and perception based on this summary.

Introduction

Reputation determines our initial impressions of some entity as they are developed through the opinions of the larger community. This social evaluation metric is what makes us choose one thing over another, one restaurant over another. Restaurants build their reputation through the customers' experiences. In the past, that would be done through word of mouth or the words of some infamous food critic. Now, every individual has the authority to leave a review on the internet, on some platform. These platforms are now databases that collect these reviews of customers' experiences, whether it be raving over the French onion soup at renown 3 Michelin starred restaurant or complaining about the rude customer experience at their local deli. The most prominent platform that collects these restaurant reviews used in North America is Yelp.

Yelp has become the go-to for many in deciding whether to go to Restaurant A or Restaurant B. Their first reaction would be to look at their ratings. These ratings are the average ratings of all the customers who decided to leave a review and it would be out of 5: (1 – poorest rating of the business; 5 – highest rating of the business). However, every customer knows ratings do not tell the full story as the ratings are subjected to every customer's interpretation of this ordinal scale. For example, some may see a 4 as excellent while others see 4 as average. For customers to build a better understanding of the restaurant's reputation, they would investigate these reviews.

Currently, Yelp would show the *top* reviews of a restaurant. This is primarily determined by if other customers find their reviews helpful or if the writer has a history of writing credible reviews. You would be able to develop a much better grasp of the restaurant's reviews through reading their Yelp reviews. However, many times, you may believe the reviews past customers had written do not accurately reflect the ratings they had given that restaurant. Or another scenario where you are stumped whether a 5 star review carries more weight in your decision-making than a 1 star review that had detailed the shortcomings the restaurant meticulously.

Hence, the goal of the project is to be able to summarize all the Yelp reviews for each restaurant and business that still contains sufficient information and sentiment, which should then accurately reflect the restaurant's ratings and reputation. The project would be using machine learning model techniques for text summarization: three of them being large language models, along with regression models for making restaurant rating predictions. Its focus is for a customer to be able to quickly grasp the general reputation of the restaurant and improve the business owners' awareness about their performance.

Background

A summary can be defined as "text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significant less than that. (Mani, 2002)" What is different about summarizing Yelp reviews is each review is sentimentally different from each other (i.e. a 1 star Yelp review will have more usage of negative words or words with negative connotations whereas a 5 star Yelp review will have more usage of positive words). This means capturing the overall sentiment with our text summarization approaches is one of the target of this project.

There are two types of text summarization: extractive summarization and abstractive summarization. These methods are essential in natural language processing (NLP) in distilling important information from large text corpora.

Extractive summarization selects the most informative sentences and/or phrases verbatim from the source text and concatenates them to form the summary. By identifying and extracting these key segments of the source text, we are able to retain the most salient information while omitting less relevant details. The models employed for extractive summarization in this paper are BERT (Bidirectional Encoder Representations from Transformers) and LSA (Latent Semantic Analysis).

Abstractive summarization, on the other hand, generates new sentences and/or phrases to capture the essence and sentiment of the source text. This approach is the closest to resembling human-like summarization, which typically involves text interpretation, generalization, and paraphrasing/rephrasing. The models employed for abstractive summarization in this paper are T5 (Text-to-Text Transfer Transformer) and Pegasus (Pre-training with Extracted Gap-sentences for Abstractive Summarization Sequence-to-Sequence models).

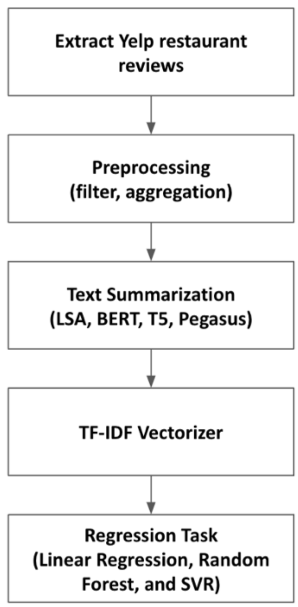
A related work used Bayesian scoring to improve the global star rating of every business and they used TF-IDF (Term Frequency-Inverse Document Frequency) and ExpandRank (an algorithm derived from PageRank) to extract keyphrases (Bechon et al., 2011). Another related work uses extractive summarization method to summarize and generate tips from Yelp reviews (Meng et al.). There was a related work whose approach is most similar to this project as they also first summarize the reviews before proceeding with their Yelp rating predictions. However, the key difference would be their summarization is limited to only extracting three keywords that best represents the overall sentiment of a single Yelp review (Suresha).

Approach

For our choice of dataset, we used the Yelp Open dataset. It is a subset of Yelp's businesses, reviews, and user data for use in connection with academic research. It contains 6,990,280 reviews from 150,346 different businesses in 11 metropolitan areas.

In Figure 1, it shows the model approach used in this project, starting from extracting the Yelp restaurant reviews.

Figure 1: Summary of Proposed Approach Involving Text Summarization and Regression Phases



As the dataset is already mostly cleaned, we only need to perform some data preprocessing, filtering, and aggregation to find a sufficient sample size for modeling purposes. We first only looked 'restaurants' under the 'categories' section, then filtered out any restaurants with fewer than 30 reviews and each review must have at least 50 words. However, this filtered dataset was still large for computational purposes. So we randomly sampled 0.1% of the filtered dataset to achieve a sample of 2,849 reviews. We aggregated these reviews by their associated business ID and performed one last filtering step by only considering the businesses with more than 3 reviews. This reduced our final sample size to 63 rows.

After this was completed, we implemented our text summarization models. Each model required different hyperparameter tuning (i.e. number of beams and repetition of n-grams). To determine the most suitable number of beams and size for n-grams repetition, we ran our abstractive summarization models (T5 and Pegasus) through all combinations of 1 through 10 for number of beams as well as 'no_repeat_ngram_size' for one source text. These hyperparameters were tuned based on a human perception of which summary output "summarized the best" while referencing the source text.

These summaries were then passed through a TF-IDF vectorizer. This feature extraction method transforms these summaries into a matrix of TF-IDF features. The values in the matrix represent the TF-IDF scores for each term. Through this vectorization technique, the most important terms can be identified while downplaying stopwords. These features were then randomly split our training and test sets (80-20 split). I would then feed the training set into three different regression models: Linear Regression, Random Forest Regressor, and Support Vector Regressor to give us our results in the section below.

Results

Figure 2: RMSE Values for Regression Models in conjunction with Extractive and Abstractive Summarization Methods

	Extractive		***Abstractive***	
Model Type	**LSA**	**BERT**	**T5**	**Pegasus**
Linear Regression	0.65	0.82	0.66	0.92
Random Forest Regressor	0.70	0.75	0.81	0.88
Support Vector Regressor	0.66	0.79	0.65	0.88

Figure 2 displays the Root Mean Square Error (RMSE) values for different combinations of summarization models and regression algorithms. RMSE is a standard measure of magnitude of error between predicted values and actual values. Given the context of Yelp rating predictions from 1 through 5, a lower RMSE value would indicate a higher predictive accuracy as predictions would be closer to the actual value.

The results are as follows:

- **Linear Regression:** LSA extractive summarization and T5 abstractive summarization exhibited the highest accuracy, with RMSE values of 0.65 and 0.66 respectively. Pegasus performed the worst with an RMSE value of 0.92.
- **Random Forest Regressor:** LSA and BERT extractive summarization models performed the best here but among all three regression models, this produced the highest RMSE values
- **Support Vector Regressor:** LSA extractive summarization and T5 abstractive summarization again led to the most accurate predictions, with RMSE values of 0.66 and 0.65 respectively. But the RMSE values for the SVR overall are lower than the Linear Regression model.

To give a sense of what the generated summaries look like across different models as well as its rating prediction, below is an example.

Source Text: "This has become a place I visit at least once while visiting Clearwater FL. I had the salmon and it was great before. They have a special menu item added every day and today was the roasted duck. It was so good and the sauce that came with it was sweet but had a spicy kick to it! The breakfast looks like it would be awesome, I just can't get up that early! Try the Gelato shop next to them after-wards for a good treat!This place was recommended to us by valet at our hotel. We said we wanted someplace with great food that was lively. We ate "outside" under cover and not long after we arrived it started raining, then pouring. Some of the rain dripped through, which was not a big deal, but there was a situation with a gentleman and his dog that had come in the entryway to take cover from the rain. It ended up with the police being called and a fist fight. We did want lively. Our waitress was Irish and was really wonderful. The food was good, I didn't think it was great. I had the chicken with lemon sauce. We had the chicken satay for an app and it was good. We ordered top shelf margaritas straight up in martini glasses. The first was very good, then next two were just ok, sort of disappointing. I would go here again. Also, the outside seating is very crowded, barely room to pull out your chair. They could do with a few less tables out there.This was the best restaurant that we stopped at on our trip to Clearwater. The lobster ravioli had huge (2-3inch) pieces of lobster and the dual colored pasta. My wife had the crab cakes which were overly breaded but still good. Her rice and vegetables we're spectacular especially after a few days of fried food. Highly recommend this establishment. We saw some local police officers eating dinner there which is always a good sign."

LSA Summary Output: "The breakfast looks like it would be awesome, I just can't get up that early! They could do with a few less tables out there.This was the best restaurant that we stopped at on our trip to Clearwater. We saw some local police officers eating dinner there which is always a good sign."

BERT Summary Output: "This has become a place I visit at least once while visiting Clearwater FL. Her rice and vegetables we're spectacular especially after a few days of fried food."

T5 Summary Output: "the lobster ravioli had huge (2-3inch) pieces of lobster and the dual colored pasta . the crab cakes were overly breaded but still good ."

Pegasus Summary Output: "We had a great time in Clearwater, Florida, and the food was just what we were looking for, but the service wasn't as good as we would have liked."

Predicted Rating: *3.7097*
Actual Rating: *3.6667*

Conclusions

The project demonstrates the potential of using machine learning models to summarize restaurant reviews and predicting their ratings to determine their reputation. Summarization models employed in this project, which are BERT and LSA for extractive summarization, T5 and Pegasus for abstractive summarization, show varying degrees of success in capturing the essence and sentiment of the customers' perception of the restaurant. Future research could combine both text summarization with sentiment analysis to more accurately capture and effectively summarize these reviews. Other steps that could be taken include exploring other regression models as well as classification ones for greater accuracy.

References

[1] Bechon, A., Grimaldi, M., & Merouchi, R. (2011). Improving Yelp Reviews. Retrieved from https://cs229.stanford.edu/proj2011/BechonGrimaldiMerouchi-ImprovingYelpReviews.pdf

[2] Suresha. (n.d.). Summarizing Reviews and Predicting Rating for Yelp Dataset. Retrieved from https://cs224d.stanford.edu/reports/Suresha.pdf

[3] Mani, I. (2002). Introduction to the Special Issue on Summarization. *Computational Linguistics*, 28(4), 399. Retrieved from

https://direct.mit.edu/coli/article/28/4/399/1774/Introduction-to-the-Special-Issue-on-Summarization

[4] Meng, R., Fu, J., & Wang, M. (n.d.). TL;DR? Tips Generation for Yelp Reviews. University of Pittsburgh. Retrieved from https://memray.me/uploads/tl-dr-tips.pdf

[5] Kumar, K. S., Priyanka, M., Rishitha, M., Teja, D. D., & Madhuri, N. (2021). Text Summarization with Sentimental Analysis. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)*, 9(4), 18–21. https://doi.org/10.21276/ijircst.2021.9.4.4

[6] Yang, S., Cui, L., Xie, J., & Zhang, Y. (2020). Making the Best Use of Review Summary for Sentiment Analysis. School of Engineering, Westlake University; Institute of Advanced Technology, Westlake Institute for Advanced Study; Zhejiang University; Tencent SPPD. Retrieved from https://aclanthology.org/2020.coling-main.15.pdf

[7] Abuka, G. (n.d.). Text Summarization and Sentiment Analysis of Drug Reviews: A Transfer Learning Approach. Retrieved from https://jewlscholar.mtsu.edu/server/api/core/bitstreams/ecc3bc6b-b018-4b65-a1d5-e71ced19c724/content