Comprehensive Rank Strategy for Yelp Customer Reviews
SI671 - Data Mining - Term Project Report

Xiaoyang Sheng (shengxy@umich.edu)
Yulin Gao (yulingao@umich.edu)
Zicong Xiao (zicongx@umich.edu)

Dec 2023

Abstract

E-commerce platforms like Yelp, faces challenges in managing customer reviews' impact on sales and reputation. Positive reviews boost sales, while deceptive ones can cause financial losses and erode trust. Our goal is to develop a strategic approach for effective review ranking and presentation, aiming to enhance credibility and maximize positive feedback impact. We've developed a sequential algorithm incorporating three models: an NLP model for usefulness prediction (acc.= 0.79), another NLP model for rating star prediction (MAE = 0.91), and an XGB model for usefulness regression (MAE = 5.40). The outcomes are displayed through the sorting of reviews based on both their actual rating and the model scores.

1 Introduction

E-commerce platforms like Yelp, encounter formidable challenges when it comes to navigating the influence of customer reviews on both sales and reputation. The impact of reviews in this digital era is profound, with the potential to either substantially enhance or detrimentally affect a business's bottom line and customer perception.

The significance of relevant reviews cannot be overstated, as they play a pivotal role in positively influencing sales. Genuine and insightful feedback from customers serves as a valuable endorsement, aiding potential buyers in their decision-making process. These reviews act as a form of social proof, instilling confidence in consumers and encouraging them to make informed purchasing decisions. Conversely, deceptive or misleading reviews pose a significant threat to businesses. Not only can they result in financial losses by steering potential customers away, but they also have the potential to erode the hard-earned trust of existing clientele.

Our overarching goal is to develop and implement an effective strategy for ranking and presenting reviews. This involves the creation of a system that not only prioritizes relevant and genuine reviews but also employs mechanisms to identify and mitigate the impact of deceptive or misleading feedback.

2 Data

We will use the open Yelp dataset, accessible at `https://www.yelp.com/dataset`[1]. Yelp is a big company renowned for curating crowd-sourced reviews on various businesses. The dataset aligns seamlessly with our research objectives, offering comprehensive data that encompasses the core aspects we require. Yelp's dataset encompasses a wealth of information, including business details such as location and attributes, full review text data with corresponding user_id and business_id, as well as user-related data that includes friend mappings and various user-associated metadata.

Here are some examples for the review dataset: Columns for the user dataset:

user_id	name	review_count	yelping_since
elite	friends	fans	average_stars
compliment_cute	compliment_list	compliment_note	compliment_plain
useful	funny	cool	compliment_cool
compliment_hot	compliment_more	compliment_profile	compliment_funny
compliment_writer	compliment_photo		

review_id	user_id	business_id	stars	useful	funny	cool	text	date
KU_O5udG6z pxOg-VcAEodg	mh_-eMZ6K5R LWhZyISBhwA	XQfwVwDr-v0 ZS3_CbbE5Xw	3	0	0	0	If you decide to eat here, just be aware it is going to take about 2 hours from beginning to end. We have tried it multiple times, because I want to like it! ...	2018-07-07 22:09:11
BiTunyQ73aT9 WBnpR9DZGw	OyoGAe7OKp v6SyGZT5g77Q	7ATYjTIgM3j Ult4UM3IypQ	5	1	0	1	I've taken a lot of spin classes over the years, and nothing compares to the classes at Body Cycle. From the nice, clean space and amazing ...	2012-01-03 15:28:18
AqPFMleE6 RsU23_auESxiA	_7bHUi9Uuf 5__HHc_Q8guQ	kxX2SOes4o- D3ZQBkiMRfA	5	1	0	1	Wow!	2012-01-03 15:28:18

3 Research Question

The proper assessment and arrangement of customer reviews have consistently posed significant challenges in the realm of e-commerce. This is due to the fact that displaying relevant reviews can significantly impact consumer psychology, leading to increased sales and a favorable reputation. Conversely, the prominence of deceptive or misleading feedback can potentially result in financial losses and the erosion of customer trust. Hence, the development of a method to effectively rank and present reviews can be a game-changer. Such an approach would empower sellers and e-commerce platforms to boost their sales while ensuring that customers enjoy a seamless and trustworthy shopping experience.

4 Methodology

4.1 NLP Model for Usefulness Prediction

No matter for platform, providers and customers, valuable comments should be prominently featured in the top rankings of reviews. Such comments offer detailed insights and suggestions for users, enabling them to gain a better understanding of the product and service. This not only enhances user knowledge but also has the potential to boost sales, particularly when the products and services are of high quality.

Within our dataset, there exists a single indicator that quantifies the usefulness of each review—the useful votes assigned by platform users. However, upon the initial posting of a new review, it may not garner a substantial number of useful votes, rendering this indicator inadequate for promptly reflecting its usefulness. To address this limitation, we have developed a model aimed at predicting whether a review would potentially be deemed useful or not.

4.1.1 EDA

To predict whether a review would potentially be deemed useful or not, we utilize the power of natural language processing, which accepts sentences as the input and output the predicted usefulness. Before we build and train the model with the review texts, it's necessary to examine the possibility or feasibility of the NLP approach, and whether the review texts provide meaningful indication for the usefulness of the review.

First, we visualize the distribution of the length of review texts and the distribution of the usefulness of reviews, and check whether either of the both distribution exhibits heterogeneity among different

groups. The distribution of the length of reviews (number of words) for different levels of usefulness is shown below in 1.

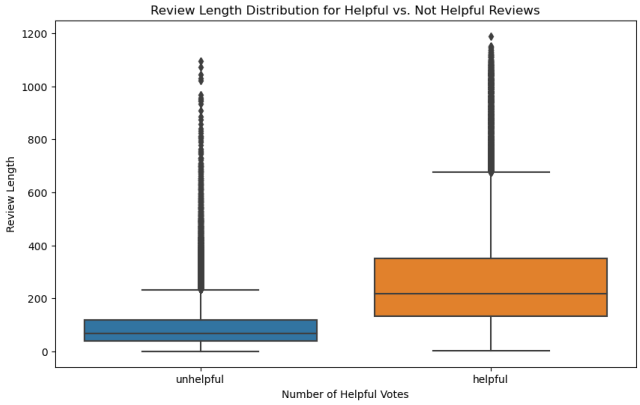

Figure 1: Distribution of the length of reviews for useful and not useful reviews

To further prove the potential of the review texts as the input data for NLP approach, the distribution of the number of useful votes for long and short reviews is also plotted below in 2.

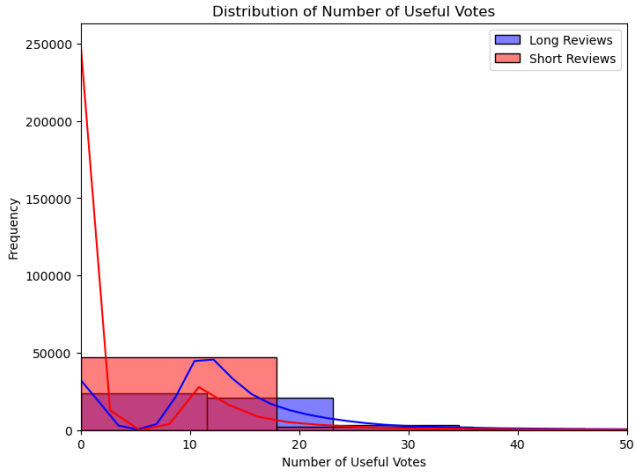

Figure 2: Distribution of the number of useful votes for long and short reviews

We've proved that the review texts can bring meaningful information towards the prediction of usefulness. However, the length of review texts could be effectively utilized by the traditional regression model, like the tree models. Before we start to apply the MLP model, we use sentiment analysis to understand the overall sentiment of reviews and check if sentiment correlates with helpful votes:

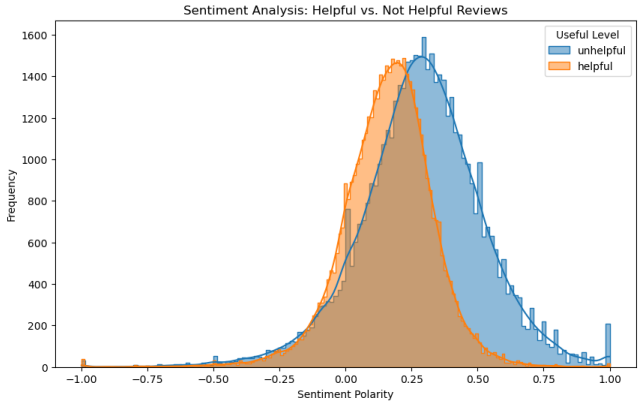

Figure 3: Distribution of sentiment polarity for useful and not useful reviews

For the sentiment analysis, we notice the heterogeneity of the distribution of sentiment polarity between groups with different useful level. For reviews that receive more "useful" votes, the sentiment polarity is greater, which means useful reviews tend to be less emotional.

4.1.2 Model Structure

Tokenization and Encoding

Tokenization serves as the initial stage in any NLP (Natural Language Processing) pipeline, exerting a profound impact on subsequent processes. A tokenizer functions by breaking down unstructured data and natural language text into discernible chunks of information, each considered a discrete element. The occurrences of these tokens within a document can be directly employed as a vector, effectively transforming an unstructured text document into a numerical data structure suitable for machine learning.

This transformation enables the utilization of tokens directly by computers to prompt useful actions and responses. Moreover, they can be seamlessly integrated into a machine learning pipeline as features that trigger more intricate decisions or behaviors.

Text encoding is a method used to transform meaningful text into a numerical or vector representation. This conversion preserves the contextual and relational aspects between words and sentences, allowing machines to comprehend patterns within the text and discern the context of sentences. Typical encoding

In this model, we employ the Hugging Face tokenizers library for preprocessing tasks. This library encompasses key methods for tokenization, including converting token strings to IDs and vice versa, as well as encoding/decoding (combining tokenization with integer conversion). The library also facilitates the addition of new tokens to the vocabulary in a manner independent of the underlying structure. It adeptly manages special tokens such as mask, beginning-of-sentence, and other tokens. Moreover, we use the tokenizer customized for our used Bert model mentioned below, and it will exactly perform the tasks to get the appropriate input for the model.

Model

We use the distilbert-base-multilingual-cased model on hugging face. It is a open-source model on the famous ML community hugging face, and developed by Developed by Victor Sanh, Lysandre Debut, Julien Chaumond and Thomas Wolf[2]. It is distilled version of the BERT base multilingual model[3] and the overall pre-training and fine-tuning procedures for BERT is shown in figure 4.

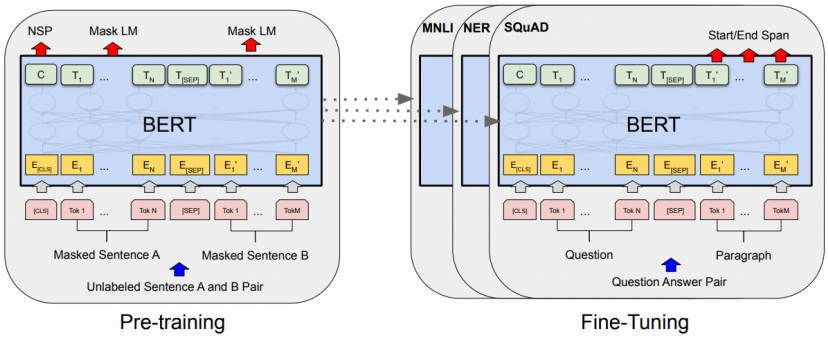

Figure 4: Overall pre-training and fine-tuning procedures for BERT[3]

The model is trained on the concatenation of Wikipedia in 104 different languages. The model has 6 layers, 768 dimension and 12 heads, totalizing 134M parameters. This model use leverage knowledge distillation during the pre-training phase and it is possible to reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster. A triple

loss combining language modeling, distillation and cosine-distance losses are also introduced to make it smaller, faster and lighter model is cheaper to pre-train[2].

After customized the Adam optimizer and the binary cross-entropy loss function to the model and directly output the result for the binary classification, the summary of the model structure is shown below:

Layer (type)	Output Shape	Param #
input_word_ids (InputLayer)	[(None, 129)]	0
tf_distil_bert_model (TFDistilBertModel)	TFBaseModelOutput (last_hidden_state=(None, 129, 768), hidden_states=None, attentions=None)	134734080
tf.__operators__.getitem(SlicingOpLambda)	(None, 768)	0
dense (Dense)	(None, 1)	769

Training

The data we use to train this model is sampled from the total review dataset. For training dataset, we select 37500 historical reviews with useful votes greater than or equal to 10 as useful reviews, and 37500 historical reviews with useful votes equal to 0 as useless reviews. As for the test dataset, we select 12500 historical reviews with useful votes greater than or equal to 10 as useful reviews, select 12500 historical reviews with useful vote equal to 0 and 10000 historical reviews with useful vote between 0 and 10 as useless reviews. The reason why we do not include reviews with useful votes between 0 and 10 in the training dataset because they are less typical then those with 0 vote and therefore the model could study the features in more distinguishable way. But anyway, we still test those of votes between 0 and 10 in our test dataset.

We set the hyper-parameters as following:

Hyper-param	value
Epochs	3
Batch size	12
Max length of truncation and padding	129

We train the model using tensorflow GPU version and CUDA package, together with a NVIDIA GeForce RTX 4050 Laptop GPU.

4.2 NLP Model for Rating Star Prediction

4.2.1 EDA

Since we already have shown the EDA for the text itself in the previous section, we focus on some particular on the rating.

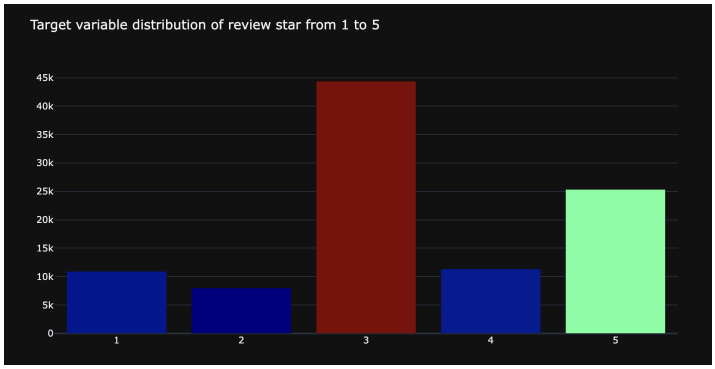

Figure 5: Distribution on rating.

The depicted figure illustrates the distribution of ratings ranging from 1 to 5 within the dataset. Notably, there is a substantial number of records with ratings of 3 and 5, aligning with the common observation that customers often lean towards being generous or moderate in their ratings. However, when taking into account the sample size, we find that the training size for each rating is sufficient for the model, and as such, no further adjustments are made.

4.2.2 Model Structure

As is often observed, individuals may express their reviews or comments with intense personal emotions, diminishing the objectivity and meaningfulness of their feedback for other customers. Additionally, the presence of malicious reviews, intentionally assigning low or high ratings to business holders, further complicates the reliability of user ratings. Hence, it becomes essential to develop a model capable of predicting ratings based solely on the text content. By comparing this predicted rating with the user-assigned rating, we can identify and exclude these outliers, ensuring that the top rankings remain both meaningful and useful.

Tokenization and Encoding

We use the same tokenization, encoding and other pre-processing as the previous model, since the text data are from the same dataset.

Model

We use the same basic pre-trained Bert model as the previous one, but we customize the Adam optimizer as well as the categorical cross-entropy loss function, and add another extra layer to the end, computing a weighted average of each rating possibility. This model will finally give a predicted regression result in the range of [0,5].

The summary of this model is shown below:

Layer (type)	Output Shape	Param #
input_word_ids (InputLayer)	[(None, 192)]	0
tf_distil_bert_model (TFDistilBertModel)	TFBaseModelOutput (last_hidden_state=(None, 192, 768), hidden_states=None, attentions=None)	134734080
tf.__operators__.getitem(SlicingOpLambda)	(None, 768)	0
dense (Dense)	(None, 5)	3845
regression (weighted average) layer	(None, 1)	0

Training

The data we use to train this model is sampled from the total review dataset. For training dataset, we select 75000 historical reviews. As for the test dataset, we select 25000 historical reviews.

We set the hyper-parameters as following:

Hyper-param	value
Epochs	10
Batch size	8
Max length of truncation and padding	192

We train the model using tensorflow GPU version and CUDA package, together with a NVIDIA GeForce RTX 4050 Laptop GPU.

4.3 XGB Model for Usefulness Regression

Acquiring a significant number of useful votes for a review is a gradual process, often resulting in newly published reviews receiving minimal votes despite their actual utility. Having employed the NLP

model to classify forthcoming reviews into "useful" and "useless" categories, we now seek to enhance the capability by introducing a new model. The subsequent model should produce a quantitative prediction on the potential usefulness of those "useful" reviews in future scenarios.

Given that we possess the dataset containing reviewers' (user) profile information, we operate under the assumption that individuals consistently providing "useful" reviews are more likely to continue producing valuable content. In other words, we assert that the usefulness of a review can be dependent on user information. The task can be translated into the form of a regression problem, with **useful** in review dataset as Y, and **corresponding attributes** in user dataset as X.

4.3.1 EDA

Within our curated dataset of 25,000 filtered "useful" comments, the "useful" score exhibits a range spanning from 11 to 539. The distribution reveals a slight right-skew, with three quantiles at 12, 14, and 19. The mean is 18.5 and the standard deviation is 15. According to the correlation matrix between **useful** and **other attributes** in Fig. 6, the top correlated user attributes with usefulness are **history_useful** and **history_cool** (corr = 0.23), which separately refers to the history count of useful votes and cool votes to that user. These features would hold potential for predicting the review usefulness. As each user record owns only 20 attributes, we decide to utilize all of them. The feature selection procedure will be integrated into the training stage to explore the most influential factors in predicting usefulness.

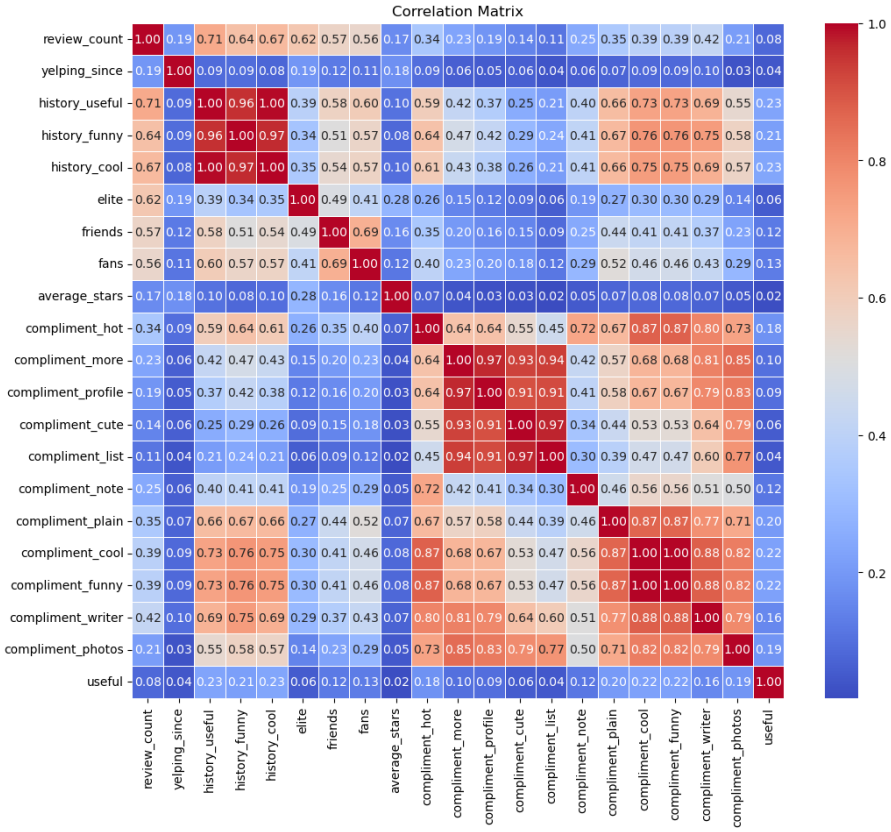

Figure 6: Usefulness - User Attributes Correlation Matrix

4.3.2 XGBoost

Tree-based model XGBoost is employed to tackle the regression task, considering its robust performance and the ability to handle complex relationships within dataset. The mean absolute error (MAE) is adopted as the loss function, and hyperparameters such as max depth, learning rate are

tuned through a comprehensive GridSearch within the context of five-fold cross-validation setup.

Following the optimization of hyperparameters through GridSearch, an importance plot is generated from the resulting model to identify the features strongly associated with the response variable:

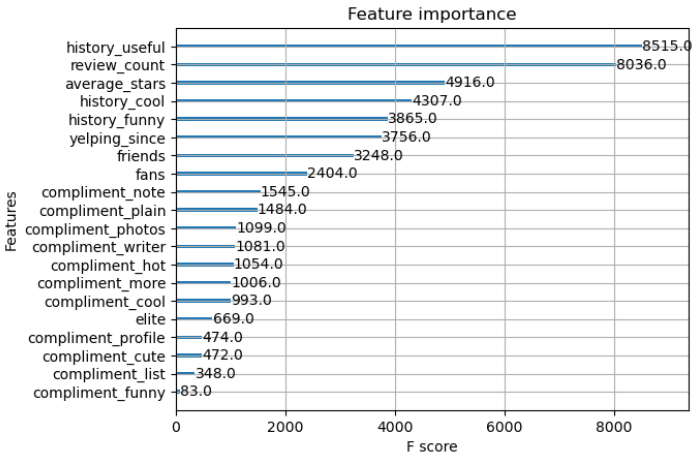

Figure 7: XGB Feature Importance Plot

The features are eliminated from the training set in ascending order of their importance. At each iteration, a new model is fitted using the refined X matrix and the previously determined hyperparameters. The ultimate feature selection is determined when it achieves the lowest MAE, specifically achieving 5.40 when the feature set is reduced to the top 16s, from **history_useful** to **elite**.

4.4 Final Ranking Strategy

Finally, after training and tuning the three models above, we come up with the aggregated strategy to calculate the scores for ranking.

In the concluding ranking strategy, three distinct models tailored for different sub-tasks are stacked together using a series-connected approach in Fig. 8. When a new review surfaces, incorporating the review rating, review text, and user background information as inputs, we employ BERT model 1 (from Section 4.1) to qualitatively assess its usefulness. If the BERT model classifies the case into the "useless" category, it will be assigned 0 points for model score. Subsequently, we adopt BERT model 2 (from Section 4.2), designed to predict review rating completely based on the review text. If the predicted rating exhibits an absolute difference greater than 1.0 from the ground truth provided by the user, we consider the "real" review rating as an outlier, indicating a mismatch with the text. Likewise, we assign 0 points to reviews with outlier ratings. The remaining reviews receive potential usefulness scores predicted by XGB model 3 (from Section 4.3). The comprehensive ranking is established by sorting all reviews in descending order based on the rating, model score, and publication date.

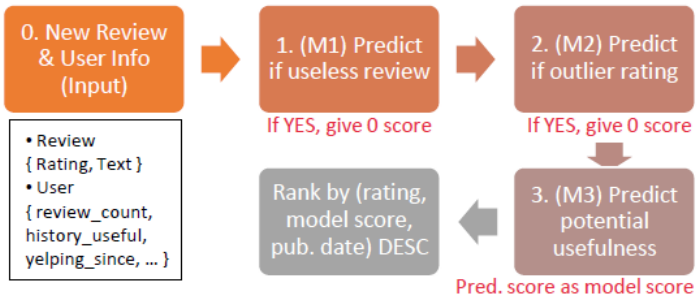

Figure 8: Ranking Strategy

5 Result

5.1 Model Performance

5.1.1 NLP Model for Usefulness Prediction

On the test dataset described in the methodology section, this binary classifier has the **test loss** of **0.929**, and the **test accuracy** of **0.79**.

5.1.2 NLP Model for Rating Star Prediction

On the test dataset described in the methodology section, this NLP regression model has the **test MAE** of **0.91**.

5.1.3 XGB Model for Usefulness Regression

On the test dataset described in the methodology section, the XGB regression model has the **test MAE** of **5.40**, compared to standard deviation of **14** for response variable.

5.2 Final Ranking

According to our strategy mentioned above, we finally apply it to a unseen dataset, with 50 records for each rating 1 to 5, totally 250 records. Because it is a ranking algorithm and the original dataset does not have the ranks, here we just put some ranking results in **5 star rating** for result presentation, and we could tell the algorithm performance by comparing the reviews with high scores and low scores:

review text	model score
This was my first time at Lure and I can't wait to go back! We were there on a Saturday night and for such a busy restaurant they really know about great service, not to mention the food was amazing, and the restaurant itself has a beautiful look. There were four of us dining, each had something different and all enjoyed our meal. Oysters, fish, clam chowder and lobster ravioli were delicious, and the staff were so helpful, sometimes at a busy restaurant you don't get much attention, but not here, we had a great experience and will return again.	11.458033
They have everything here. It astonishes me time and time again how this deli carries things that usually require a trip to the grocery shop. They have typical deli stuff and then all the random crap I need at the last minute like baking pans, batteries, greek yogurt, good cheese, Green & Black's chocolate, etc. They make their own KimChi too. I hope they start making more Korean food like bibimbap. The staff are incredibly friendly. Their flower selection is good too. Oh- we got a lottery ticket here that won us $12 so it's lucky.	10.954467
...	...
Love this truck and the delicious desserts they are putting out. The macaroons are really worth the trip. The flavors are really creative and spot on. I've had macs in Paris at laduree and it may just be the American in me but I like these better for the taste, size and value. The cookies have the texture and flavor that makes macs delicious and unique and the fillings are decadent and tasty as they should be. The milk and honey was one of my favorites as many others have also said.	9.745833
...	...
Called regarding a broken garage door spring. Tech came out next day, was very friendly, and quickly completed the repair. Total repair took under 30 minutes. Very pleased!	0
Had the $39 Sunday special. Wonderful. Salad was scrumptious. Steak with Prawns was perfect.Desert yummy. What more can s man want?	0
Great Pho and Ramen! Also the bahn mi was the best I ever had. We will be back soon.	0

Although all the reviews mentioned above are given along with 5-star ratings, it is evident that they vary significantly in terms of details and information provided. Utilizing our ranking algorithm, reviews featuring more extensive details and sincere comments accumulate higher scores, securing a position in the top ranks. Conversely, shorter and perfunctory reviews receive lower scores, potentially even zero, relegating them to the bottom of the list.

The outcomes of our models and ranking algorithms demonstrate a clear trend: more useful and detailed reviews are consistently positioned higher in the rankings. This substantiates that our approach effectively addresses the research problem introduced at the outset. We are confident that our algorithm provides a potential solution in the realm of online reviews ranking.

6 Discussion

In discussing the existing limitations, conventional methods for sentiment analysis primarily rely on natural language processing (NLP) models. These models collect information only from textual content within reviews. With the the evolving dynamics of virtual communication, in contrast, many users, especially the younger generation, often integrate non-textual elements, such as emojis to convey sentiments. Such preference leads to the challenge for current models, which can have poor performance in understanding non-textual contexts.

In addition, while each individual model has demonstrated advisable performance in the given task, it's worth noting that the series connection within the ranking strategy can introduce some level of effectiveness loss. To assess the extent of this potential loss, possible approach involves conducting A/B testing, comparing our proposed method against other ranking strategy, which demands a substantial number of user judges. An alternative method for evaluating such loss is to compute the sequence similarity between our output ranking and the ground truth obtained from Yelp. The approach offers a quantitative measure to assess the alignment and accuracy. Consequently, future work will be gathering additional ranking information from Yelp, and exploring the optimal weights for each model within the strategy.

7 Conclusion

In this project, we come up with a ranking algorithm for the customer reviews on the platform Yelp, in order to display useful and relevant reviews and further gain positive impact on consumer psychology, leading to potential increased sales and a favorable reputation. We developed the algorithm by three models, one NLP model for usefulness prediction, one NLP model for rating star prediction and one XGB model for usefulness regression. All three models perform well in the test dataset, and finally we pick some samples to present the ranking results on 5-star reviews, and the outcome shows the desired effect and therefore gives a potential solution to our research problem.

8 Reference

[1] Yelp. Dataset. Accessible at www.yelp.com/dataset. (2023).
[2] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.
[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Appendix: Code

For the code implementation, please refer to:

`github.com/xiaoyang-sheng/NLP-Ranking-of-Yelp-Customer-Reviews`