A Deep Dive into Hotel Reviews: What Matters Most When It Comes to Ratings?

Zhuohao Yin
Dept. of Computer Science and Engineering
HKUST
Hong Kong, China
zyinad@connect.ust.hk

I Chieh Chen
Dept. of Computer Science and Engineering
HKUST
Hong Kong, China
icchen@connect.ust.hk

***Abstract*—In this era of information, human's lives have been closely relative with data and the data technology has been affecting people's daily living in every aspects, including food, cloths, accommodation and transportation. From the perspective of hotel managers, the collection of hotel service rating and reviews has provided them with a way to review on their current operation strategies, but how to effectively extract information from such data remains a question. In this project, we investigated on two information extracting techniques: distinct feature engineering on texts and text encoding. We further build regression models on the two extracting technique and discuss about the results.**

***Index Terms*—Data mining, Exploratory Data Analysis, Natural Language Processing**

I. Introduction

How to examine current operation strategies has always been an important topic to different business operators, as it is critical for them to identify their current deficiencies and make improvements accordingly. Thanks to the rapid evolution of technology, customers nowadays are able to provide comments to merchants through the internet. This provides business managers to collect review on their own products and services efficiently and effortless, and the next important topic to the technique to extract important information from the great amount of data. As different customers may convey their own ideas in different ways and normally they would have their own idea own different aspects of the same product, their isn't a distinct format that different people would narrate their comments, and it remain a complex problem on how to retrieve the main ideas of different comments.

In this project, we try to inspect on a dataset containing numbers of hotels' service review and service rating pairs, and we try to explore on different ways to extract information from text reviews. By text information extraction techniques, we want to inspect on how different slices of text are correlated with the final rating provide by each users. In particular, we apply language models to extract features from text reviews, and we further make prediction on the rating with the features we derived. We will further discuss the results from using different text informaton extraction techniques with their advantages and disadvantages respectively.

II. Related Works

In this section we discuss several related works that have been commonly applied in the natural language processing field.

A. BERT

BERT is an abbreviation of Bidirectional Encoder Representations from Transformers. BERT is a designed to break the limitation of previous word that normally make use of text unidirectional, which is not optimal for sentence-level tasks [1]. Different from previous works, BERT adopts a multi-layer bidirectional Transformer based architecture which considers both left and right context, and the architecture can be further extend to perform a variety of tasks such as question answering and next sentence prediction.

To allow the BERT model to be further adapted to different tasks, [1] highlighted the importance of BERT pre-training. BERT models is pre-trained on BooksCorpus and English Wikipedia, and the main focus is put on the document-level corpus so that the data can contain more long contiguous sequence. The first task of BERT pre-training is the procedure of "Masked Language Model" , which focus more on word prediction in a sentence. The input corpus is randomly masked for a percentage of tokens. Then, the masked corpus is fed into BERT and BERT is trained to predicted the masked words over the know vocabulary.

The second task focuses more on the relationship between sentences. In this task, each time two sentences are sample for pre-training, and with a probability the second sentence is either the actual sentence that follows the first sentence in the corpus or a random sentence sampled from the corpus. [1] has demonstrated that learning such relationship between sentences can be beneficial to downstream tasks such as Question Answering and Natural Language Inference.

With the two pre-training tasks mentioned above, the Transformer-based model's self-attention mechanism has allow us to push the pretrained model to more downstream tasks. To finetune the BERT for a task, we can simply plug the specific and output into BERT, and we finetune all the model parameter end-to-end. Thanks to the previous work in pre-training, finetuning actually take less time and resources to the finish, and we do not need to retrain the whole model for each tasks to have satisfying results.

B. BERTopic

Topic models have been proposed to find the latent topics across different sentence in corpus, the multiple works have been proposing the feasibility of the clustering technique in solving such task. [2] presented BERTopic, which is a model that can generate coherent topic by clustering embeddings that is generated from pre-trained language models.

The most preliminary task of BERTopic is the conversion of text into embeddings, and we need to assure that different sentences with the same topic should be semantically similar. [2] adopted the state-of-art framework [3] that allows users to convert sentences and paragraphs to latent vectors, and one of the most important features of [3] is its use to cluster semantically similar documents, which supports [2] to perform documents clustering.

However, the concept of spatial locality is not well-defined in high dimensional space, and [3] adopts UMAP [4],which can preserve more feature while projecting high dimensionality into low dimensionality, to reduce the dimensionality of the generated latent vectors mention above. Then, the dense vectors with reduced dimensionality are clustered with HDBSCAN [5].

After clustering with the embedding vectors, we need to generate topic for each of the clusters. [2] propose a variant of TF-IDF, which is measurement of the importance of each word to a corpus. The adapted variant is a class-based TF-IDF that can model the importance of word in clusters, and with such technique we can generated numbers of topics for each of our cluster.

By the above mentioned procedures, BERTopics is suggested to be capable in learning coherent coherent patterns and perform stably in different tasks.

C. RoBERTa

Following the work of BERT [1], [6] continues on the development of BERT models' training and claimed that the original BERT is significantly undertrained. [6] proposed RoBERTa, which is the abbreviation of "A Robustly Optimized BERT Pretraining Approach", and the work highlight that the training process is highlt sensitive to hyperparameters tuning, and with proper hyperparameters setups the BERT models can have better performers than those work published after BERT.

The core task of RoBERTa is to experiment and evaluate the impact of each hyperparameter's impact on the BERT models' training result. For example, [6] claims that BERT model can be very sensitive to the Adam epsilon term, and training on large batch size can improve perplexity perplexity for the masked language modeling objective and also increase accuracy on the task. After investigating on different aspects, RoBERTa is a work that aggregate the improvements mentioned, and RoBERTa is able to achieve state-of-the-art results on datasets such as GLUE and RACE at the time.

III. Method

Although our ultimate goal is not the accurate prediction of hotel ratings but rather to extract key factors that influence the ratings, building a classification model is a feasible way to do it. We first adopt a topic model, BERTopic [2], to summarize frequently mentioned topics in the dataset. Next, we formulate them as multiple choice questions and utilize a MC-Question-Answering (MCQA) model, Roberta [6], to answer the questions based on the content of each review. The categorical answers are regarded as discrete features for a logistic regression model. We refer to this approach as the **QA approach**. As a comparison, we also train BERT [1] with a classifier head to predict the ratings, taking as input the raw reviews. This approach is referred to as the **BERT approach**.

A. Exploratory Data Analysis

As we would like to explore on the key factor that contributes the most to hotel service rating, we make use of the "Trip Advisor Hotel Reviews" from Kaggle [7]. Our dataset contains 20492 pairs of hotel ratings with their corresponding English text reviews, and the hotel ratings range from 1 to 5 with the distribution shown in Figure 1.

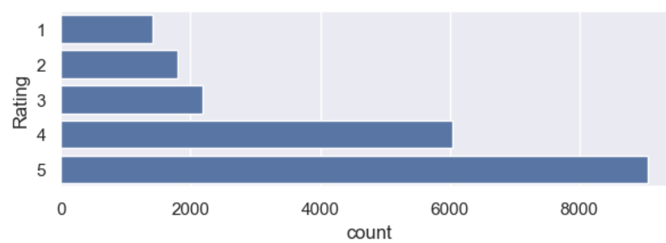

Fig. 1. The count of reviews in each ratings.

For the text review in this data, the minimum text length is 9, the maximum text length is 1933, the mean text length is 106.375 and the standard deviation is 100.655. And by prelimarary classifying review with rating bigger than 3 as positive rating, and reviews with rating smaller than 3 as negative rating, we can get the word frequency as Table I (Excluding words that occur in both sides).

Positive Word	Count	Negative Word	Count
location	7494	told	1361
clean	5998	asked	818
friendly	4764	called	652
little	4668	star	638
walk	4546	money	526
excellent	4438	manager	521
best	4054	pay	508
recommend	3536	came	498
area	3470	air	476
restaurant	3389	towels	474

TABLE I
TOP 10 WORDS EXISTING IN POSITIVE AND NEGATIVE EXAMPLES.

B. Data Preprocessing

In the QA approach, the first step of data preprocessing begins with extracting frequent topics that are discussed by users. Specifically, the total 20491 reviews are concatenated together to form a whole document. Then BERTopic [2] examines the document to produce topics that are frequently discussed, where each topic is represented by several keywords. In total, the model produced 118 topics. Table II shows the top 10 extracted topics. The subsequent step requires manual examination of each extracted topic since some are related to specific tourist attractions instead of generic hotel attributes. After manually filtering the topics, we summarize aspects about hotels that the extracted topics mentioned. Each aspect is then formulated into a multiple choice question, with 2 or 3 choices to choose from. A summary of the total 38 aspects is shown in Table III. For each aspect, a question is formulated, as well as several answer choices. For instance, for aspect "has good service", the question and answers are:

Q: Does the hotel have good service?
A_1: The hotel has good service.
A_2: The hotel doesn't have good service.
A_3: Not sure if the hotel has good service.

Subsequently, the MCQA model takes each review and iterate through each question to fill in the answers. Due to the categorical nature of the answers, they can be treated as features which are later used for classification.

For the BERT approach, there is not much preprocessing to be done. The reviews are simply tokenized and feed into the BERT model.

C. Model Training

The model used in the QA approach is logistic regression. The answers generated in the data preprocessing step are in the range of $\{0, 1, 2\}$, which are one-hot encoded to be inputs to the logistic regression model. The logistic regression is trained until convergence, setting the max interation to be 500. The χ^2 test is used for feature selection.

The BERT model is loaded with pretrained weights and it is assembled with a classifier head using the embedding of the `CLS` token as the input to the classifier. The classifier is a composition of a dropout layer of 0.1 probability and a linear layer, followed by a softmax function to produce classification probabilities. The assembled BERT-for-classification model is then finetuned for 10 epochs.

IV. Results and Discussion

The test set is formed by randomly sampling 10% from the whole dataset. The rating prediction is evaluated by classification accuracy and the macro F1 score, as shown in Table IV. As one can observe, the BERT approach outperforms the QA approach on both accuracy and F1 score. This is not hard to explain because BERT implicitly learns a high dimensional embedding for each review, which possesses much richer representation power than explicit categorical features. On the other hand, in the QA approach, features are manually determined and constructed.

Approach	Accuracy	Macro F1
BERT	0.66	0.60
QA	0.59	0.49

TABLE IV
Test results of the 2 approaches.

However, the ultimate goal of this project goes beyond the mere prediction of hotel ratings. From the business perspective, being able to accurately predict the customers' ratings given their review is barely meaningful. By contrast, if we can derive acute insights by solving the rating prediction as a surrogate task, it will bring business values to hotel runners as they improve their service.

For the purpose of data mining, the QA approach enables us to evaluate different features through statistical hypothesis testing. Specifically, we perform the χ^2 test to rank all the features. As displayed in Figure 2, some features are much more crucial than the other. In other words, there are certain factors that are what customers care the most and thus have the largest impact on the ratings. Table V shows the top 10 ranked features.

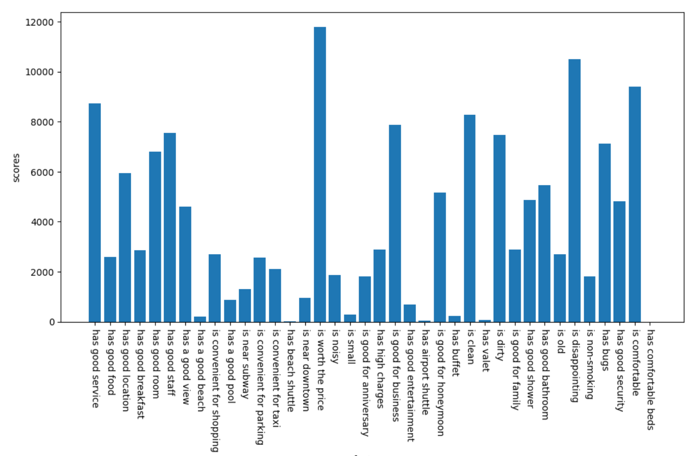

Fig. 2. A barplot showing the χ^2 score for all features.

Rank	Feature
1	is worth the price
2	is disappointing
3	is comfortable
4	has good service
5	is clean
6	is good for business
7	has good staff
8	is dirty
9	has bugs
10	has good room

TABLE V
Top 10 features that are the most impactful on ratings.

Feature "is worth the price" wins the first place, meaning that people do care if a hotel provides good value for money.

Topic	Keywords
1	['punta', 'cana', 'resort', 'beach', 'food', 'people', 'did', 'not', 'nt', 'vacation']
2	['barcelona', 'ramblas', 'metro', 'hotel', 'city', 'catalunya', 'euros', 'location', 'rambla', 'las']
3	['paris', 'metro', 'eiffel', 'hotel', 'tower', 'location', 'small', 'rue', 'staff', 'louvre']
4	['amsterdam', 'hotel', 'canal', 'tram', 'room', 'dam', 'location', 'station', 'central', 'breakfast']
5	['florence', 'duomo', 'italy', 'hotel', 'train', 'breakfast', 'station', 'location', 'walk', 'ponte']
6	['seattle', 'downtown', 'pike', 'needle', 'parking', 'market', 'space', 'hotel', 'stay', 'center']
7	['york', 'nyc', 'square', 'westin', 'new', 'times', 'ny', 'manhattan', 'hotel', 'room']
8	['juan', 'san', 'puerto', 'rico', 'old', 'el', 'beach', 'condado', 'pool', 'area']
9	['location', 'great', 'hotel', 'excellent', 'nice', 'staff', 'clean', 'good', 'stay', 'friendly']
10	['waikiki', 'hawaii', 'beach', 'honolulu', 'hawaiian', 'ocean', 'view', 'oahu', 'aqua', 'outrigger']

TABLE II
TOP 10 EXTRACTED TOPICS, EACH WITH KEYWORDS AND THE CORRESPONDING REPRESENTATIVE DOC.

Aspect
has good service
has good food
has good location
has good breakfast
has good room
has good staff
has a good view
has a good beach
is convenient for shopping
has a good pool
is near subway
is convenient for parking
is convenient for taxi
has beach shuttle
is near downtown
is worth the price
is noisy
is small
is good for anniversary
has high charges
is good for business
has good entertainment
has airport shuttle
is good for honeymoon
has buffet
is clean
has valet
is dirty
is good for family
has good shower
has good bathroom
is old
is disappointing
is non-smoking
has bugs
has good security
is comfortable
has comfortable beds

TABLE III
A SUMMARY OF THE ASPECTS ABOUT THE HOTELS.

Among the ranking, negative attributes are as prominent as positive ones, indicating that people are extremely averse to several things, such as disappointment, tidiness and bugs. Whether such conclusions are valid or not, we propose a novel framework that can be applied on similar datasets in other industries.

One of the reasons that the QA approach achieves poor classification results is that the MCQA model did not yield the optimal answers to the questions. A considerable amount of the answers are A_3 in the aforementioned QA example, which means the model is uncertain to draw a positive or negative conclusion about the question given the review. This can hinder classification performance as a number of the answers stand for uncertainty. However, a possible solution is to prompt users to complete questionnaires after they post reviews and allow users to answer the questions. In this way, we are able to acquire QA data of high quality that can be used to further finetune the MCQA model for our domain.

V. CONCLUSION

In this project, we have performed data mining on the TripAdvisor dataset with a Question Answering approach. Our main objective is to reveal which aspects about hotels most significantly influence the rating. We have accomplished this objective by conducting feature selection on the logistic regression with categorical features construsted according to popular topics. Our findings provide insights to hotel runners on potential improvements they can achieve in order to raise customer ratings. Moreover, our data mining framework is transferrable to any other industry with similar user feedback data, such as restaurants.

REFERENCES

[1] Jacob Devlin, undefined, et al. 2019."BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in North American Chapter of the Association for Computational Linguistics,

[2] M. Grootendorst. 2022. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," in arXiv preprint arXiv:2203.05794,

[3] Nils Reimers, Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". arXiv:1908.10084, 2019

[4] L. McInnes, J. Healy, and J. Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426.

[5] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. The Journal of Open Source Software, 2(11):205.

[6] Yinhan Liu, undefined., et al. 2019."RoBERTa: A Robustly Optimized BERT Pretraining Approach,". arXiv:1907.11692.

[7] Alam, M. H., Ryu, W.-J., Lee, S., 2016. Joint multi-grain topic sentiment: modeling semantic aspects for online reviews. Information Sciences 339, 206–223.