Understanding Product Quality with Unstructured Data: An Application of LLMs and Embeddings at Airbnb

Jikun Zhu
Airbnb.com
jacob.zhu@airbnb.com

Zhiying Gu
Airbnb.com
zhiying.gu@airbnb.com

Brad Li
Airbnb.com
brad.li@airbnb.com

Linsha Chen
Airbnb.com
linsha.chen@airbnb.com

ABSTRACT

Ensuring exceptional service and product quality is essential for a business's sustained success. Traditionally, product quality has been assessed using signals from structured data, such as customer review ratings, types of customer support issues, and resolution times. Alternatively, some studies in the economics literature estimate quality through structural choice models. We discuss some of the limitations of those methodologies. With the rise of cutting-edge Large Language Models (LLMs), our research introduces an innovative approach that harnesses LLMs and embeddings to extract valuable insights from unstructured data, such as product descriptions and review texts, enhancing the conventional methods of quality measurement.

Our approach unfolds in three steps. First, we generate text embeddings using state-of-the-art pre-trained embedding models. Next, we employ these embeddings to cluster products, pinpointing clusters that are most indicative of customer retention patterns. Finally, we assign quality scores based on the presence of features highly predictive of retention that we extracted from clusters, thereby crafting a novel metric for intrinsic product quality. We evaluate the predictive power of this new quality measurement on customer retention.

We show in comparison with quality measurement purely based on structured data, this method is immune to cold start problem and offers a richer, more nuanced understanding of product quality by seamlessly integrating structured and unstructured data, paving the way for more informed and strategic business decisions.

KEYWORDS

LLM, Large Language Model, Text Embedding, Clustering, Retention, Quality, Review, Survey

1 INTRODUCTION

In today's competitive market landscape, the quality of products and services is a critical determinant of a business's long-term success. High product quality not only attracts customers but also fosters loyalty, driving sustained revenue growth. Consequently, accurate measurement and continuous improvement of product quality have become paramount for businesses seeking to gain a competitive edge.

Historically, product quality has been assessed using structured data. Common indicators include customer review ratings, types of customer support issues, and resolution times. These methods provide quantifiable and easily interpretable measures of product quality, but have their own limitations. For instance, [3] discussed online platforms' feedback mechanisms for the word-of-mouth effect. In the paper, the summary of eBay feedback data points out limitations of using such structured data. About half of the buyers do not leave feedback, and reviews tend to be overwhelmingly positive. Additionally, the format of review ratings, ranging from positive, neutral, negative to a 1-5 star scale, results in partial information about product or service quality. This signal can be biased because potential reviewers may omit relevant information. [4] offers a detailed discussion on this bias and shows users who do not leave reviews often had worse experiences, highlighting this bias further. Another common way to understand quality with structured data through surveys [2], which suffers from similar limitations.

Another stream of research has focused on the estimating unobserved quality with variants of fixed effects through structural choice modeling following [1]. For instance, [7] estimates quality as a combination of product fixed effects, time fixed effects, and product-time fixed effects, leveraging cross-country variations in an international trade setting. This approach captures unobserved quality components beyond just observed signals like reviews. However, such models rely on strong assumptions of the model structure and provide limited insights into the drivers of quality.

The advent of Large Language Models (LLMs), has opened new avenues for extracting meaningful signals from unstructured data. LLMs have demonstrated remarkable capabilities in understanding and generating texts in natural languages, making them ideal for analyzing product descriptions, customer reviews, and other textual data. Recent advances in natural language processing (NLP) have shown that text embeddings can capture intricate semantic relationships and contextual nuances. [8] discussed how recent develop in the field empowered the product review understanding. [6] provides a good overview on the performance of embedding clustering based on different algorithms. The work closest to ours to our knowledge is [13]. The paper determines the services quality of e-commerce site using Tokopedia users' review on the Trusted Company, an online review site. We believe that richer information can be extracted from customer reviews such as appraise or complaints of product features.

In this paper, we propose an approach to estimating product quality by integrating structured and unstructured data using LLMs and embeddings. Our methodology involves three key steps: (1) generating text embeddings using pre-trained models, (2) clustering products based on these embeddings to identify patterns related to customer retention and using TF-IDF to understand the drivers of quality perception, and (3) assigning quality scores based on the presence of features that are predictive of retention. One unique aspect of our approach is that our source of unstructured data signals includes not just customer feedback such as reviews, issue reports, and surveys, which can have partial coverage or be biased, but also product description data, which is always available and relatively objective - limiting idiosyncrasy of individual customer preferences. Another advantage of incorporating product description data is to overcome the cold start problem - we have no signal on quality of newly launched products/services if quality purely relies on customer feedback/reviews. This approach not only enhances traditional quality measurement methods but also provides a more comprehensive and nuanced understanding of product quality.

By leveraging the strengths of LLMs and embeddings, our paper aims to bridge the gap between structured and unstructured data, offering a robust framework for product quality estimation. This paper contributes to the growing body of literature that seeks to harness the power of AI and machine learning to enhance business decision-making processes.

We illustrate our approach using Airbnb data, demonstrating the practical application and effectiveness of our proposed methodology.

2 WHAT IS PERCEIVED QUALITY?

Quality definition can often be vague and dimensions of quality may change based on services and products that are being transacted. [14] defines four general service quality dimensions for online platforms: efficiency, system availability, fulfillment and privacy. This approach relies on insights from the extant literature and a comprehensive qualitative study. For some products that have clear feature specs, it is relatively easy to define quality. For example, laptop quality can be measured by its CPU capability, memory size etc. However, in case of service offerings that are heterogeneous and diverse, like Airbnb stays, quality measures can be challenging, especially for services/products that we don't see any customer feedback.

Fortunately, large literature has shown the linkage between quality and customer retention such as [5]. We propose to identify quality features using this intrinsic linkage between quality and retention. Whether customer i comes back to the Airbnb platform depends on the quality of the previous trip experience. Formally,

$$retention_i = f(Q(features))$$

$Q()$ is a function that transforms product/service features into the quality measure. $retention_i$ is a binary variable becomes 0 when the customer churns. In the case of this Airbnb exercise, we define churn as customers do not come back in 12 months after their previous trip.

The goal of this paper is to identify features inside the function $Q()$ that predicts customer retention and assign a quality score to each listing based on the presence of the identified features.

3 DATA

We mainly use 3 sources data for signals of quality:

(1) Reviews customers left for listings on Airbnb
(2) Detailed description page for Airbnb listings
(3) Listing photos

3.1 Reviews

Not all Airbnb guests leave a review after their stay. Figure 1 shows that reviews are skewed towards 5-star ratings. Consequently, we cannot purely rely on the reviews stars for each trip as a measure of trip quality. Several products have been developed at Airbnb to leverage the review information. One successful work is using text sentiments on reviews to measure trip quality. However, we want to explore more signals from reviews beyond sentiment. In this work, we select most recent 10 reviews of a listing as the underlying data for embeddings (Figure 2.), then extract the word features from the reviews to infer listing quality. We will compare different ways of quality measurement in the last validation section.

Figure 1: Airbnb Review Ratings and Quality Ratings

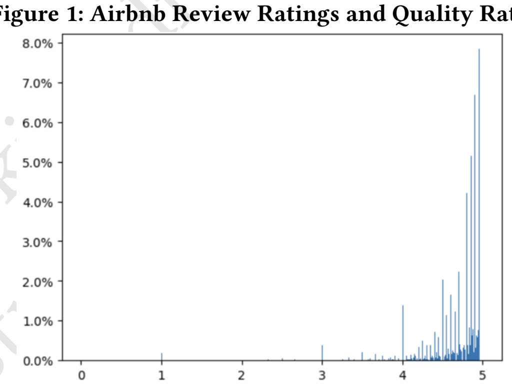
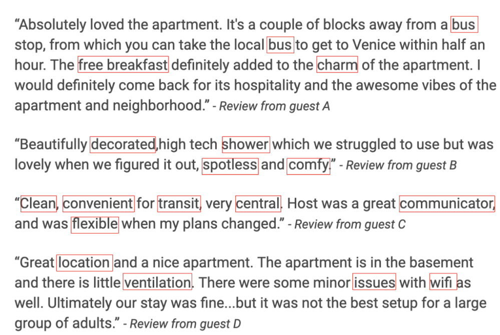


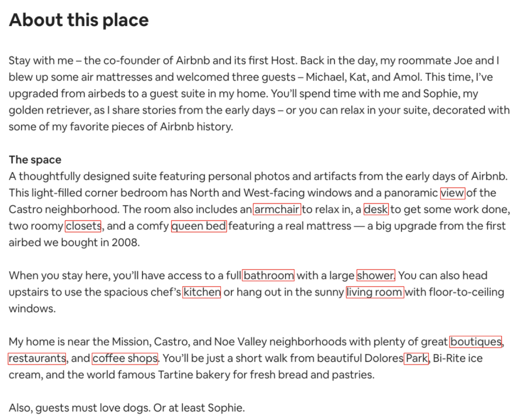
Figure 2: Example of Listing Reviews



3.2 Product Descriptions

Airbnb listing product description page shows important information about a listing such as, neighborhood, ease of transportation, amenities, house rules etc. These are all factors that can contribute to a guest's trip experience/quality. Product descriptions are always available which can help us increase the coverage for quality measurement and can overcome the cold start problem of solely relying on customer feedback for quality measure - when a new listing doesn't have customer review yet. Figure 3 shows an example of listing description page. The advantage of using this data is its broad coverage and amount of information contained. Beyond that, leveraging product description to quantify experience also helps us to cluster similar products that can be used in product recommendation as the future application.

Figure 3: Example of Airbnb Listing Description



About this place

Stay with me – the co-founder of Airbnb and its first Host. Back in the day, my roommate Joe and I blew up some air mattresses and welcomed three guests – Michael, Kat, and Amol. This time, I've upgraded from airbeds to a guest suite in my home. You'll spend time with me and Sophie, my golden retriever, as I share stories from the early days – or you can relax in your suite, decorated with some of my favorite pieces of Airbnb history.

The space
A thoughtfully designed suite featuring personal photos and artifacts from the early days of Airbnb. This light-filled corner bedroom has North and West-facing windows and a panoramic view of the Castro neighborhood. The room also includes an armchair to relax in, a desk to get some work done, two roomy closets, and a comfy queen bed featuring a real mattress — a big upgrade from the first airbed we bought in 2008.

When you stay here, you'll have access to a full bathroom with a large shower. You can also head upstairs to use the spacious chef's kitchen or hang out in the sunny living room with floor-to-ceiling windows.

My home is near the Mission, Castro, and Noe Valley neighborhoods with plenty of great boutiques, restaurants, and coffee shops. You'll be just a short walk from beautiful Dolores Park, Bi-Rite ice cream, and the world famous Tartine bakery for fresh bread and pastries.

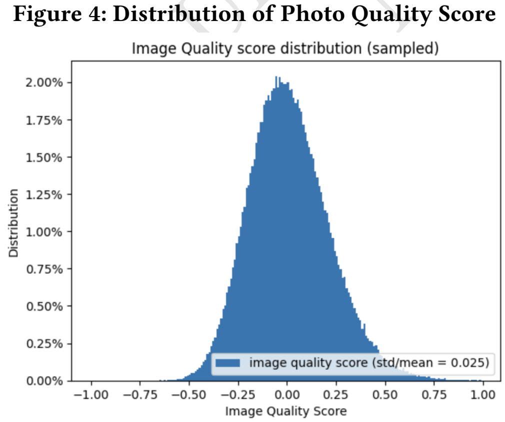
Also, guests must love dogs. Or at least Sophie.

3.3 Listing Photos

Listing photos provide a rich source of information about a listing and are one of the most important listing attributes that are considered by guests when making a decision to book. Among the information provided by listing images, a key signal provided is the level of listing attractiveness and home design. The image scores can then be assigned to provide signals that approximate attractiveness.

For image scoring, we are using existing Airbnb in-house trained model - a multi-class model build on *MobileNetV2*. Figure 4 shows the distribution of listing photo quality scores. [1].

Figure 4: Distribution of Photo Quality Score



[1]We do not discuss the training details of model in this paper. But utilizing listing photo quality signal falls into our theme of leveraging unstructured data for quality understanding.

3.4 Stratified Sampling

As discussed in the previous section, in this research we are trying to find features that are predictive of the customer retention (if they re-book within 1-year period from previous trip) as a quality metric. A churn would suggest bad or subpar service/listing quality from previous trip.

To ensure balanced signal from both churned and retained guests, we stratify-sampled 180k customers stays from 2022 for each churned and retained group separately (total of 360K sampled stays).

4 EMBEDDING AND CLUSTERING OF UNSTRUCTURED DATA

To extract information from unstructred data, we first create text embeddings using product descriptions and review text data. Embeddings are numerical representations of texts which enable us to easily cluster similar listings together. Intuitively, given our balanced stratified sample, if features representing a cluster have no predictive power on retention, we will see no differences in return rate compared to the total population. As a result, clusters that have high retention rate gain are considered having above average quality. The higher the retention rate of a cluster, the higher the quality. We will formally test (using Chi-square test) whether the retention rate of a given cluster is statistically different from the baseline. We describe the details of embedding and clustering below.

4.1 Embedding

There has been extensive researches on the text embedding clustering techniques [11]. In this study, for listing descriptions and published listing reviews we choose a pre-trained model to generate the text embeddings. The choice was based the token limits (given our longer input texts), performance [10], and the integration with the Airbnb's data environment. Testing with different embeddings methods is discussed as next step at the end. For image embedding and scoring we use our in-house trained model mentioned in section 3.3 to generate listing image quality scores.

4.2 Clustering

For clustering, we apply the *K-means* clustering method and tune the numbers of clusters with two goals in mind: 1) making sure there is minimal number of outlier clusters: no cluster is smaller than 1% of the total population; 2) limiting generation of less informative clusters: if increasing k does not add clusters with higher rebooking rates[2] that pass the Chi-square tests, further splitting is unnecessary. We have explored different values of k using a step search. Table 1 [3] shows that when k > 25 the clustering process can no longer generate more clusters that can pass the Chi-square test due to the shrinking cluster size. We conclude under our cluster criteria, k = 25 is the most optimal cluster numbers in this analysis.

Clustering result is presented in Table 2 and Figure 5. Within each cluster, we perform the Chi-square test to see if rebooking rate and churn rate are significantly different. Clusters that passed

[2]Rebooking rate is calculated using 1-yr rebooking likelihood of each cluster.
[3]In practice the difference of rebooking rate is the absolute difference between each cluster and the total sampled population, we demonstrate them in relative percentages in table 1 and table 2 for easy interpretation.

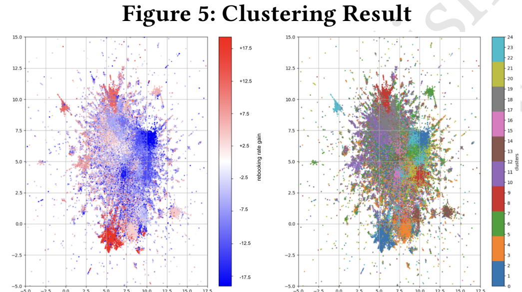
Table 1: Cluster Size

k	clusters pass Chi-Sq test	min cluster size	max diff of rebooking rate
5	1	11.14%	8.38%
10	3	5.36%	15.58%
15	3	3.85%	16.52%
20	4	2.96%	16.52%
25	**6**	**1.32%**	**17.73%**
30	5	0.69%	19.63%
35	5	0.56%	18.83%

Table 2: Clustering Result

cluster	cluster size	cluster size pct	diff of rebooking rate	p-val (Chi-sq)
2	**14,356**	**3.86%**	**17.73%**	**0.00**
8	**11,497**	**3.10%**	**12.19%**	**0.00**
24	**7,766**	**2.09%**	**12.01%**	**0.00**
11	11,054	2.98%	8.62%	0.00
6	10,273	2.77%	7.28%	0.00
3	22,737	6.12%	7.18%	0.00
13	4,923	1.33%	5.61%	0.00
10	21,438	5.77%	5.55%	0.00
4	16,341	4.40%	5.13%	0.00
6	16,954	4.56%	3.87%	0.00
7	20,468	5.51%	2.09%	0.04
1	19,224	5.18%	-0.92%	0.22
12	4,917	1.32%	-1.38%	0.34
21	10,282	2.77%	-1.68%	0.09
19	12,607	3.39%	-2.37%	0.01
20	19,571	5.27%	-2.83%	0.00
17	19,314	5.20%	-2.91%	0.00
14	21,556	5.80%	-3.65%	0.00
5	21,220	5.71%	-5.29%	0.00
22	11,765	3.17%	-5.98%	0.00
18	11,690	3.15%	-4.13%	0.00
9	15,356	4.13%	-8.38%	0.00
23	13,397	3.61%	-8.52%	0.00
15	19,557	5.27%	-11.81%	0.00
0	13,185	3.55%	-13.01%	0.00
ALL	*371,448*	*100.00%*	0	-

Figure 5: Clustering Result



the Chi-square tests would be used to score the listing quality in the next step. [4]

5 RETRIEVING PREDICTIVE FEATURES AND CONSTRUCTING QUALITY METRICS

Traditional ML practice only uses the embeddings and clustering results for model prediction. The embedding results are hard to interpret and generalize. However, for insights generation and product improvement purposes, we would like to extract features from the embeddings from clusters that link to high user retention. This adds an interpretation layer to this approach.

[4]Here to increase the model sensitivity, we only use the top 3 ranked cluster as they have the highest guest rebooking rate (10% of total sampled population).

In this section we will leverage TF-IDF, a common practice in NLP to extract text information from clusters. Then we use these keywords to construct a listing quality measurement at scale.

5.1 Retrieving Keywords and Assigning Weights

After the clustering from section 4, we apply the TF-IDF method to extract keyword features in listing description and listing reviews that are associated with high rebooking clusters. The importance of keyword is calculated by their uniqueness and occurrence within the group of documents in cluster j. This step only involves the top three retention clusters and extracts the important keywords for these high retention clusters, i.e. key words that are predictive of retention.

$$w_{i,j} = tf_{i,j} * log(\frac{N}{df_i})$$

$tf_{i,j}$ = total number of occurrences of keyword i in cluster j. df_i is the number of documents containing keyword i in cluster j and N is the total number of documents in cluster j.

In the previous step, the TF-IDF calculation was based on the high retention clusters. However, high TF-IDF words in those clusters can be common in low retention clusters too. We only want to highlight features (keywords) that are unique in the high retention clusters. Consequently, we remove common keywords and phrases that have high occurrences across all listings. To do that, we apply a feature suppression on keywords that are common across all listings, this is called Discriminative Approach introduced by Luhn 1957 [9]. The goal is to only highlight significant terms when they occur more frequently in target documents compared to other common terms:

$$w_i = max\left\{(\sum_j w_{i,j} - \sum_{common} w_{i,common}), 0\right\}$$

Here, j is still a high retention cluster, but w_i, *common* is the TF-IDF of word i in all documents.

Lastly, in order to make the TF-IDF values comparable, for each keyword, we normalized the TF-IDF values based on the total sum of TD-IDF values among all keywords:

$$\widetilde{w_i} = \frac{w_i}{\sum_i w_i}$$

5.2 Listing Scoring with Weighted Keywords Presence

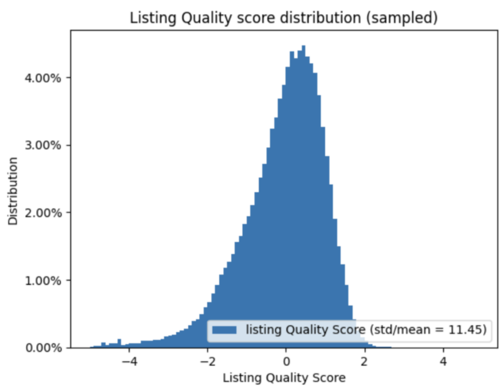
Using the extracted keywords list that are predictive of retention, we can easily evaluate any given listing based on their similarity to the keyword list. Following [12], we sum the weights of the retention predictive keywords in listing l :

$$\sum_{i \in l} \widetilde{w_i}$$

Due to the common feature suppression on keywords in step 2, listings with many common keywords will have sums of weights close to 0. Here we also apply a log transformation to normalize the scores so they're not close to 0. Figure 7 shows the final distribution of the listing score. This is what we call *listing quality score* of listing l:

$$score_l = log(\sum_{i \in l} \widetilde{w}_i)$$

Figure 6: Distribution of Listing Quality Score



For each listing, the higher score it receives, the more similar it is to the listings in the high retention cluster identified in section 4. Hence we believe these listings will generate higher retention stays. We will validate this in the following section.

6 EVALUATION OF THE LISTING QUALITY SCORE

6.1 Correlation with Retention

In the previous section, we constructed the listing quality score metric using the extracted keywords that are predictive of positive retention behaviors (re-booking). In this section, we attempt to validate the linkage between our constructed listing quality score and retention using **out of sample data**. We sampled 500k unique listings and 800k 2022 stays and constructed their listing quality score with the methodology above and assess the predictability of our quality score on retention.

First, let's take a look at the simple correlations between listing quality score and out of sample retention. We bucketize the individual trip level data by quantile from 1 to 100. For each quantile bucket we calculate the average 1-year return rate. We compare our *listing quality score* (Figure 7.c) with two other measurements of the quality: listing review rating score (Figure 7.a) and listing review sentiment score (Figure 7.b). The *listing quality score* we develop has the strongest correlation with 1 year retention (***Pearson corr: 0.903***). This comparison suggests that the proposed *listing quality score* using text and image embedding outperforms the traditional rating metrics in correlation with customer retentions.

We conduct a formal t-test to see whether there are significant difference in the quality metric we develop between the guest group who churned vs guest group who retained. Figure 8 shows the test result. The t-test return stat-sig results (p<0.00) for all 3 quality metrics: (a) listing review rating score; (b) listing review sentiment score and (c) listing quality score. Obviously, we observe a bigger stat-sig difference in the listing quality score.

Figure 7: Score Quantile vs. Rebooking Rate

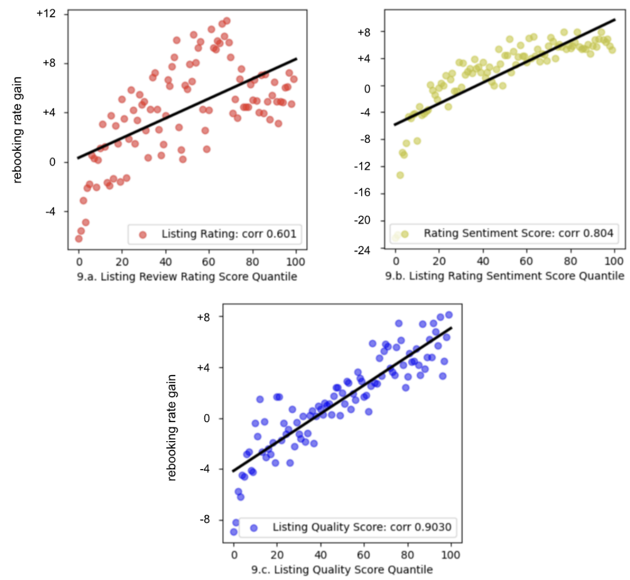
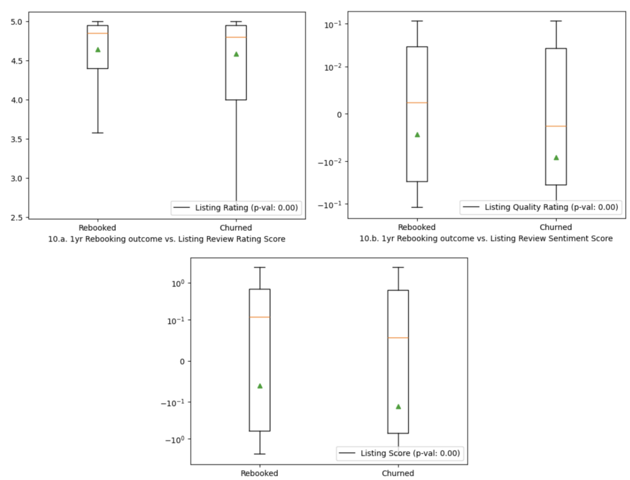


Figure 8: Rebooked User vs. Churned User



6.2 Combining Structured and Unstructured Data

Next, we show how much performance gain we can obtain by combining signals from structured (such as listing review rating score) and and unstructured data (our proposed embedding based listing quality score). Table 3 summarizes the logistic regression results of using different quality metrics to predict out of sample retention. Among all single predictors, the embedding based quality score has the strongest predictive power. Meanwhile, combining signals from structured and unstructured data together largely increases the performance, suggesting great potential in this direction.

Table 3: Logistic Regression Result

	Metrics	Pseudo R-squ.	P-Val
	Listing Review Rating Score	0.055*	0.00
	Listing Review Sentiment Score	0.0566	0.00
	Image Quality Score	0.0564	0.00
	Listing Quality Score	0.0576	0.00
Combined	Listing Review Rating Score	**0.0605**	0.00
	Listing Review Sentiment Score		0.00
	Image Quality Score		0.00
	Listing Quality Score		0.00

** about 30% of listings are missing Listing Ratings, which are removed in the regression*

7 MARKETING APPLICATIONS

As discussed earlier, our listing quality score of a guest's previous trip not only better predicts the retention, but also solves the cold start problem when trying to measure the expected trip quality if a guest books a new listing. This has the potential to benefit the following Airbnb business areas.

Airbnb uses marketing landing pages to showcase listings in a destination. A key challenge has been deciding which listings to feature on the landing page. Review ratings could be leveraged for this selection; however, we know that that can be biased. Instead, we propose using the listing quality score generated in this paper to make more informed quality-price trade-off decisions.

8 CONCLUSION AND FUTURE WORK

We propose a novel approach to assess product quality by leveraging unstructured data. We argue that relying solely on structured data for quality measurement introduces selection bias and results in low data coverage. In this work, we specifically focus on user retention as a metric for product quality. We propose the framework to extract predictive features from unstructured data in relation to user retention. The framework is based on 1) *retentive cluster discovery*: the method which clusters unstructured data and discovers retentive clusters based on k-means clustering and Chi-square tests and 2) *listing scoring*: the method which extracts keywords (i.e. features) based on TF-IDF values and scores the listing quality with features extracted. We perform thorough validation of our proposed framework to demonstrate its effectiveness in assessing product quality using unstructured data.

For future work, we plan to explore and train the in-house multimodal embeddings for better predictive power. We also plan to expand to multilingual corpus for more holistic understanding of product quality across different regions and languages. It would also be interesting to employ state-of-the-art LLMs for keyword extraction.

9 ACKNOWLEDGEMENT

The authors would like to thank Carolina Barcenas (Airbnb.com), Alex Deng (Airbnb.com) for their help and guidance on this project.

REFERENCES

[1] Steven T Berry. 1994. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics* (1994), 242–262.
[2] J Joseph Cronin Jr and Steven A Taylor. 1992. Measuring service quality: a reexamination and extension. *Journal of marketing* 56, 3 (1992), 55–68.
[3] Chrysanthos Dellarocas. 2003. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science* 49, 10 (2003), 1407–1424.
[4] Andrey Fradkin, Elena Grewal, Dave Holtz, and Matthew Pearson. 2015. Bias and Reciprocity in Online Reviews: Evidence From Field Experiments on Airbnb. *EC* 15 (2015), 15–19.
[5] Thorsten Hennig-Thurau and Alexander Klee. 1997. The impact of customer satisfaction and relationship quality on customer retention: A critical reassessment and model development. *Psychology & marketing* 14, 8 (1997), 737–764.
[6] Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. Beyond words: a comparative analysis of LLM embeddings for effective clustering. In *International Symposium on Intelligent Data Analysis*. Springer, 205–216.
[7] Amit Khandelwal. 2010. The long and short (of) quality ladders. *The Review of Economic Studies* 77, 4 (2010), 1450–1476.
[8] Sheetal Kusal, Shruti Patil, Aashna Gupta, Harsh Saple, Devashish Jaiswal, Vaishnavi Deshpande, and Ketan Kotecha. 2023. Sentiment Analysis of Product Reviews Using Deep Learning and Transformer Models: A Comparative Study. In *International Conference on Artificial Intelligence on Textile and Apparel*. Springer, 183–204.
[9] Hans Peter Luhn. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development* 1, 4 (1957), 309–317.
[10] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and Code Embeddings by Contrastive Pre-Training. arXiv:2201.10005 [cs.CL]
[11] Alina Petukhova, Joao P Matos-Carvalho, and Nuno Fachada. 2024. Text clustering with LLM embeddings. *arXiv preprint arXiv:2403.15112* (2024).
[12] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. Citeseer, 29–48.
[13] Puspita Kencana Sari, Andry Alamsyah, and Sulistyo Wibowo. 2018. Measuring e-Commerce service quality from online customer review using sentiment analysis. In *Journal of Physics: Conference Series*, Vol. 971. IOP Publishing, 012053.
[14] Valarie A Zeithaml. 2002. Service excellence in electronic channels. *Managing Service Quality: An International Journal* 12, 3 (2002), 135–139.