Replication Report on the Amsterdam Airbnb Market: A BRTM Perspective

Jiajun Cheng

July 9, 2025

1 Introduction

This report reproduces the *BRTM-Sample* experiment of Chen et al., 2021[1], but transfers the setting from the New York City market to Amsterdam and confines the observation window to transactions with check-out dates on or before 31 December 2016. Under an identical evaluation protocol we compare Bilateral Relational Topic Models (BRTM) with three classical baselines, thereby examining the robustness of the approach across markets that differ markedly in inventory size and review density.

2 Data Collection and Pre-processing

The starting point is the publicly available *Inside Airbnb* snapshot for Amsterdam. Although the index file is taken from the publicly available insiderairbnb snapshots, the original paper did not publish its full web crawl data. Consistent with the paper's approach, we have to use the listing IDs in listings.csv as the seed to again perform a crawl of the review interface on the Airbnb website to obtain the bilateral review streams (reviews written by guests to listings, reviews written by hosts to guests, etc.). Due to the large number of target listings, strict interface throttling, and the existence of anti-crawling mechanisms, this project still took four days to complete the data crawl with the support of high concurrency proxy policy.

After cleaning we obtained bi-directional review streams—guest-to-listing (A), listing-to-guest (B)—and listing descriptions (D). The final corpus contains 67 660 guest–listing pairs, split 65 %/10 %/25 % into training (43 979), validation (6 766) and test (16 915) sets under a 1:19 negative-sampling protocol identical to the original study. The vocabulary size after stop-word removal is 35 176 tokens; 3 440 unique guests and 858 hosts appear in the data.

2.1 Training and Baseline Configuration

Training Procedure. During the model–training phase, the BRTM-Sample model adopts the same bipartite document partitioning as the original study (domains $\mathcal{D}$, $\mathcal{A}$, and $\mathcal{B}$). Variational EM is executed in PyTorch for 5×40 inner–outer iterations and optimized with the LBFGS optimizer. The numbers of topics are fixed to $D^* = 38$, $A^* = 29$, and $B^* = 51$, while the two shared topic clusters are set to $DA = 22$ and $AB = 9$. All models are trained and evaluated on the same 1:19 negative-sampling candidate set, and a grid search on the validation split determines the regularization coefficient, learning rate, and early-stopping patience, ensuring a fair comparison.

[1] Table 7 in the original paper

Baselines. We benchmark BRTM-Sample against three alternative methods:

- RAND: a random baseline that uniformly shuffles the 20 candidate listings, representing a "blind-choice" level of performance.
- STL (Single-Text LDA): all transaction-related texts are concatenated into a single document, a 60-topic Gensim LDA model is pre-trained, and the resulting topic proportions are concatenated into a logistic-regression model.
- RTM-G: the Author–Recipient Relational Topic Model proposed in 2011. We reproduce its Gibbs-sampling inference procedure and apply early stopping after each iteration based on validation Hit@10.

3 Results

After the training and baseline comparing procedure, the results was summarized as following table:

Table 1: RAND vs STL vs RTM-G vs BRTM-SAMPLE (Hit Rate)

TopN	RAND	STL	RTM-G	BRTM-SAMPLE	BRTM-SAMPLE$_{\text{NYC}}$	Δ vs NYC
1	0.060	0.291	0.335	0.300	0.204	47.1%
2	0.115	0.383	0.469	0.434	0.363	19.6%
3	0.155	0.461	0.548	0.529	0.493	7.3%
4	0.206	0.534	0.595	0.566	0.606	-6.6%
5	0.225	0.592	0.623	0.645	0.698	-7.6%
6	0.291	0.652	0.673	0.696	0.778	-10.5%
7	0.356	0.702	0.702	0.769	0.839	-8.3%
8	0.391	0.751	0.746	0.800	0.885	-9.6%
9	0.450	0.786	0.798	0.843	0.920	-8.4%
10	0.469	0.818	0.840	0.857	0.945	-9.3%

Table 1 reveals that our repilcation model . For $N \leq 3$ the Amsterdam model outperforms the NYC benchmark, yet falls behind once $N \geq 4$, -9.3% less than the NYC model at HR@10. We attribute the underperform to Amsterdam's smaller inventory, which limits the list's topic coverage, and to computational resource constraints; running 5 outer iterations on the RTX 4090 took 5 hours, so we did not have enough time to converge. Nevertheless, BRTM remains far ahead of RAND and STL, and overtakes RTM-G from $N = 5$ onward. Overall ranking quality is still superior (MRR $= 0.4765, \text{NDCG@10} = 0.5628$), though the margin is partially diluted.

4 Topic Quality Analysis

To further validate that the model is able to learn intuitive underlying semantics, we evaluated the distribution of topics after training was completed. As shown in the table 2.

Table 2: Top-5 topics in each domain

	D-block (Listing description)
D0	arena, metro, ziggo, dome, walking, bijlmer, park, afas
D1	room, cottage, park, please, b, accommodation, also, people
D2	west, vondelpark, de, br, baarsjes, restaurants, located, neighborhood
D3	years, appartment, feels, reviews, old, since, ago, like
D4	vondel, loads, sqm, hoofddorpplein, opportunity, prime, daylight, singles
	A-block (Guest reviews)
A0	hans, nick, us, great, wendie, stay, host, de
A1	mark, matthijs, wilma, us, olaf, breakfast, corina, sonja
A2	frank, von, sanne, mimi, de, us, house, place
A3	breakfast, morning, us, stay, great, reinout, delicious, place
A4	ruben, maria, jesse, stijn, great, us, yahav, place
	B-block (Host responses)
B0	thanks, see, pleasure, meet, big, ralf, review, de
B1	house, de, thanks, guests, like, also, review, description
B2	femke, room, care, take, extra, hear, hop, girls
B3	thanks, wilbert, hi, greetings, much, review, appreciated, accommodation
B4	je, solange, hoi, jan, hi, house, information, also

As can be seen from the table:

The first 5 themes in domain D are clearly focused on geographic location (e.g. arena/metro/ziggo) or type of residence (cottage/park), indicating that the text of listing descriptions can be effectively clustered by the model;

Domain A themes are more focused on the guest's perspective of the accommodation experience or the interaction with the host (lots of people's names and emotion words such as great stay, breakfast, etc.);

Domain B themes are dominated by polite responses from hosts and check-in feedback (high frequency words thanks/pleasure and house/room).

These observations are consistent with Airbnb's operational reality and further support the interpretability of BRTM in multi-document domain scenarios. Future work could quantitatively map topic labels to listing metadata (e.g., neighborhood or room type) to explore the direct impact of topics on price and conversion rates.

5 Feature Importance Analysis

We evaluated the marginal contribution of each thematic and structured feature to the probability of booking success by inspecting the standardized coefficients (β) of the logistic regression layer, as stored in `feature_importance.json`.

Among all input features, we highlight the top five positive and negative signals—those with the highest absolute values of β weights. These values indicate how strongly a feature is associated with an increased or decreased likelihood of a booking being accepted. Figure 1 presents the most influential features.

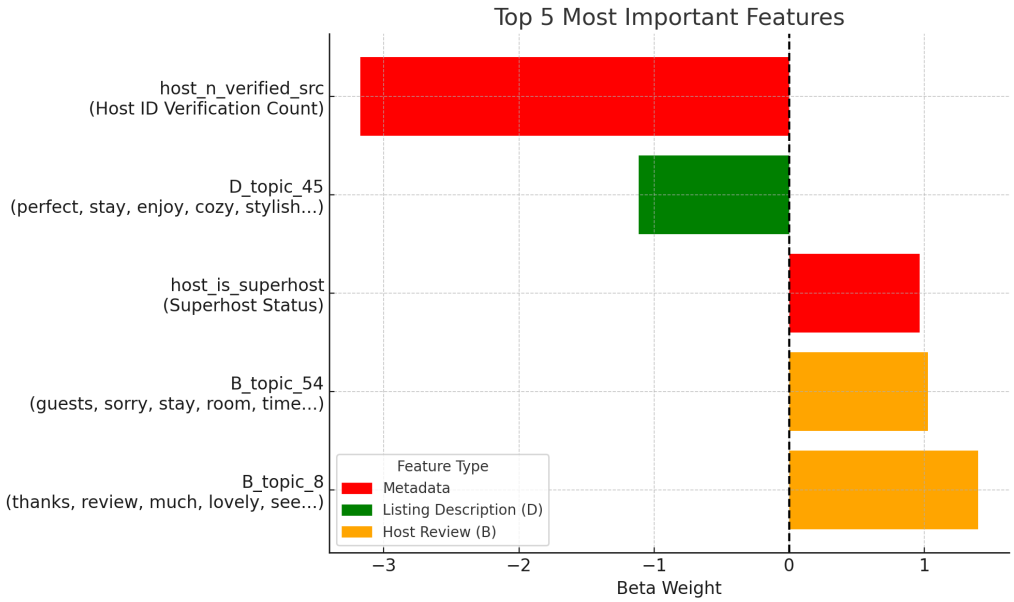

Figure 1: Top-5 most positive and negative features contributing to booking success, based on standardized logistic regression coefficients.

From Figure 1, most of the positive weights originate from domain B (host → guest reviews), especially `B_topic_54` and `B_topic_8`, whose keywords focus on favorable expressions such as *"thanks"* and *"lovely"*. This intuitively suggests that subsequent transactions are more likely to succeed when the host displays a positive attitude toward the guest. Likewise, the high positive coefficient of the structured feature `host_is_superhost` corroborates the effectiveness of Airbnb's official *Superhost* badge in boosting conversion rates.

In contrast, the largest negative coefficient is associated with `host_n_verified_src`, shows a negative correlation with the probability of a successful booking when hosts have too many verification channels, it is assumed to be related to the tendency of professional hosts or agents to use multi-account verification, which could raise red flags for guests. The most negative thematic feature, `D_topic_45`, revolves around exaggerated listings using words like *"perfect"*, consistent with expectations that overpromotion can undermine credibility.

6 Host Review Sentiment and Redundancy

In response to previous concerns about the validity of host reviews, we conducted two complementary experiments on *host* → *guest* English reviews:

- **Sentiment classification** The TextBlob's `PatternAnalyzer` method was used to classify the text into three categories: positive/neutral/negative;
- **Semantic proximity clustering** The TF-IDF vector cosine similarity ≥ 0.85 was used as the merging threshold to detect "near duplicates" of positive comments.

For sentiment classification, our results are shown in the following two figures:

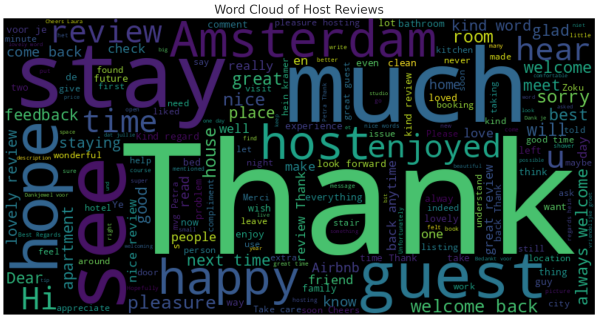

Figure 2: World cloud of host reviews.

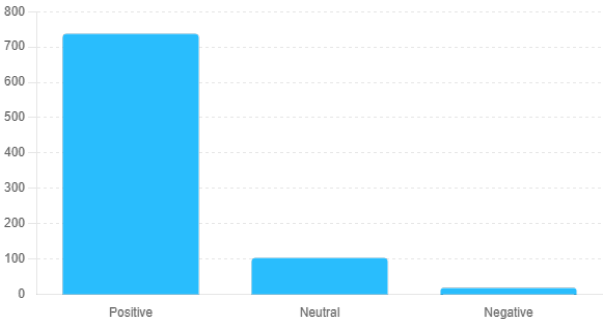

Figure 3: Sentiment Distribution of host reviews.

The figure 3 shows an overwhelming majority of positive reviews ($\approx 86\%$), indicating that host–guest interactions are generally very friendly; neutral comments are about 12%, mostly short confirmations or polite replies (e.g., "Thank you for staying"); negative reviews were only $\approx 2\%$, mostly related to minor issues with cleanliness, communication or amenities. Contrary to expectations, negative reviews are less unbalanced here.

After these, we do the semantic proximity clustering to the positive reviews. In the "cosine similarity" condition, about 1/4 of the positive host responses were highly repetitive in semantics or wording (e.g., "Pleasure to host you again!", "You were lovely guests."). This finding suggests that while host reviews generally have a positive and friendly tone, there is still a tendency for templated expressions; we address this issue by downweighting highly repetitive positive text in feature engineering to avoid the model learning redundant signals.

7 Conclusion

The results of replicating the BRTM sample in Amsterdam show that the model remains competitive across markets, but its performance diminishes as data decreases and list heterogeneity increases, and computational resources are limited. Balancing the sentiment of reviews is a promising direction for future work

References

Chen, J., Yang, Y. (, & Liu, H. (2021). Mining Bilateral Reviews for Online Transaction Prediction: A Relational Topic Modeling Approach [Publisher: INFORMS]. *Information Systems Research*, *32*(2), 541–560. Retrieved July 8, 2025, from https://ideas.repec.org//a/inm/orisre/v32y2021i2p541-560.html