Do Listings Lie? How Airbnb Language Predicts Price

Text Analysis for Business | Group Presentation | 17 April 2025

Team Members:
Aayush Damani | Amer Mulla | Anna Miriam Philip
Srinivas Rangarajan | Varsha Narra

Does listing language explain price variation, beyond property characteristics?

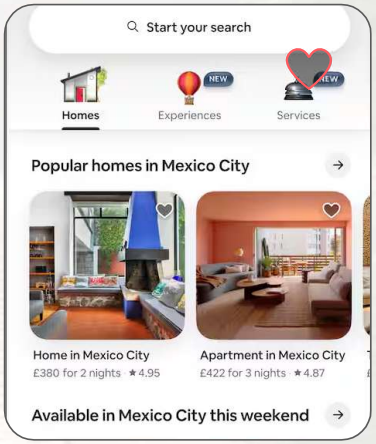
The Problem

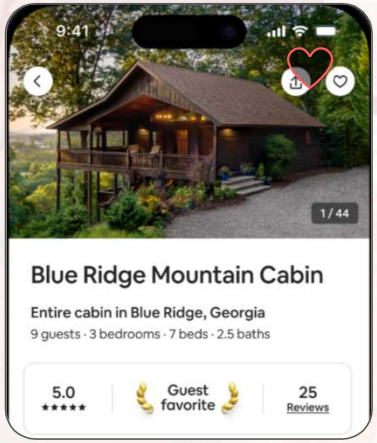
Airbnb hosts compete on price, but listing descriptions vary widely in quality and detail. This raises a key question: does language itself affect pricing, or is price mostly explained by bedrooms, room type, and location?

- *"Nestled in the vibrant heart of the historic Center Square downtown area... charming brownstone... fully furnished living room... spacious bathroom... laundry room."*
- *"Keep it simple at this peaceful and centrally-located place."*

Why It Matters

- ✓ *Hosts:* optimize descriptions to earn more revenue
- ✓ *Platforms:* detect misleading or low-effort listings faster
- ✓ *Researchers:* quantify language as an economic market signal





Cleaning & EDA

Raw Data

- **261,894** original listings
- Missing price, description, city, or room type
- Zero/invalid prices and sparse descriptions
- Mixed city coverage beyond the study scope

Data Cleaning

- Removed listings with missing key fields
- Excluded zero/negative prices
- Filtered descriptions with 20 words or fewer
- Applied IQR trimming on log(price)
- Kept only English-speaking cities

Final Dataset

- **61,521** cleaned listings
- 16 English-speaking cities
- log(price) approximately near-normal
- Ready for regression and NLP analysis

After cleaning, log(price) is far more symmetric, making it suitable for modelling.

Dataset Overview

61,521
clean listings

Final sample after filtering missing fields, short descriptions, invalid prices, outliers, and non-English cities.

16
English-speaking cities

Including London, New York City, Los Angeles, San Francisco, Boston, Chicago, Seattle, Toronto, Vancouver, Sydney, Melbourne, Brisbane, Dublin, Hawaii, San Diego and Austin.

57
variables per listing

Covering price, room type, city, guest capacity, bedrooms, beds, reviews, availability, and listing description text.

Target variable: log(nightly price in USD) | Key text input: listing description (~70 words avg)

Raw train file: 261,894 listings before cleaning

Canonical cleaned dataset used for main analysis: 61,521 listings, 16 cities.

Summary Statistics

61,521 listings

Mean price: $209 | Median price: $160
Large enough for stable cross-city comparisons

~70 words avg. description

Range: 21–207 words
Rich enough for meaningful NLP features

Room type matters

Entire home median: ~$190 | Shared room: ~$44
Clear structural price tiers by listing format

City matters

Hawaii highest: ~$256 | Toronto lowest: ~$120
Strong geographic variation before text enters

Language adds value

Metadata alone: $R^2 = 0.57$ | Text results follow
Substantial variance still remains unexplained

Accommodates r = 0.59

Strongest metadata correlation with log(price)
Capacity is the clearest non-text price driver

Text Preprocessing Steps

1

Clean

Lowercase text in the descriptions, remove stop words and punctuations

“Luxury Loft!!!” → “luxury loft”

2

Lemmatise

Reduce related words to a shared root

“luxurious”, “luxury” → “luxury”

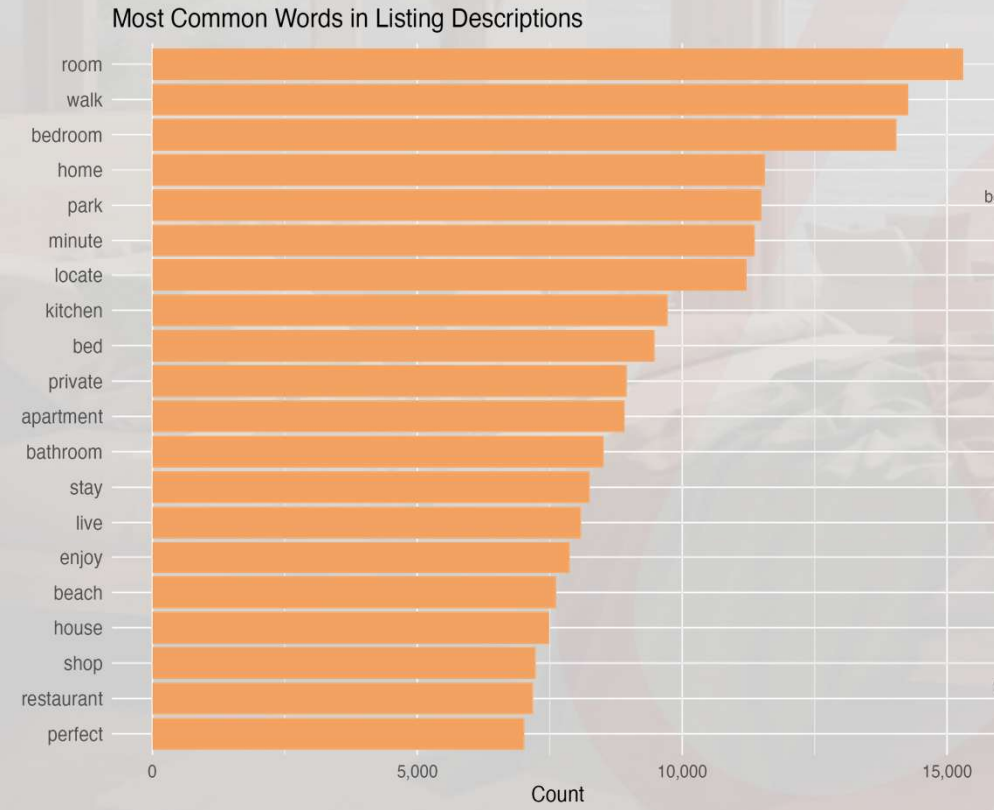
3

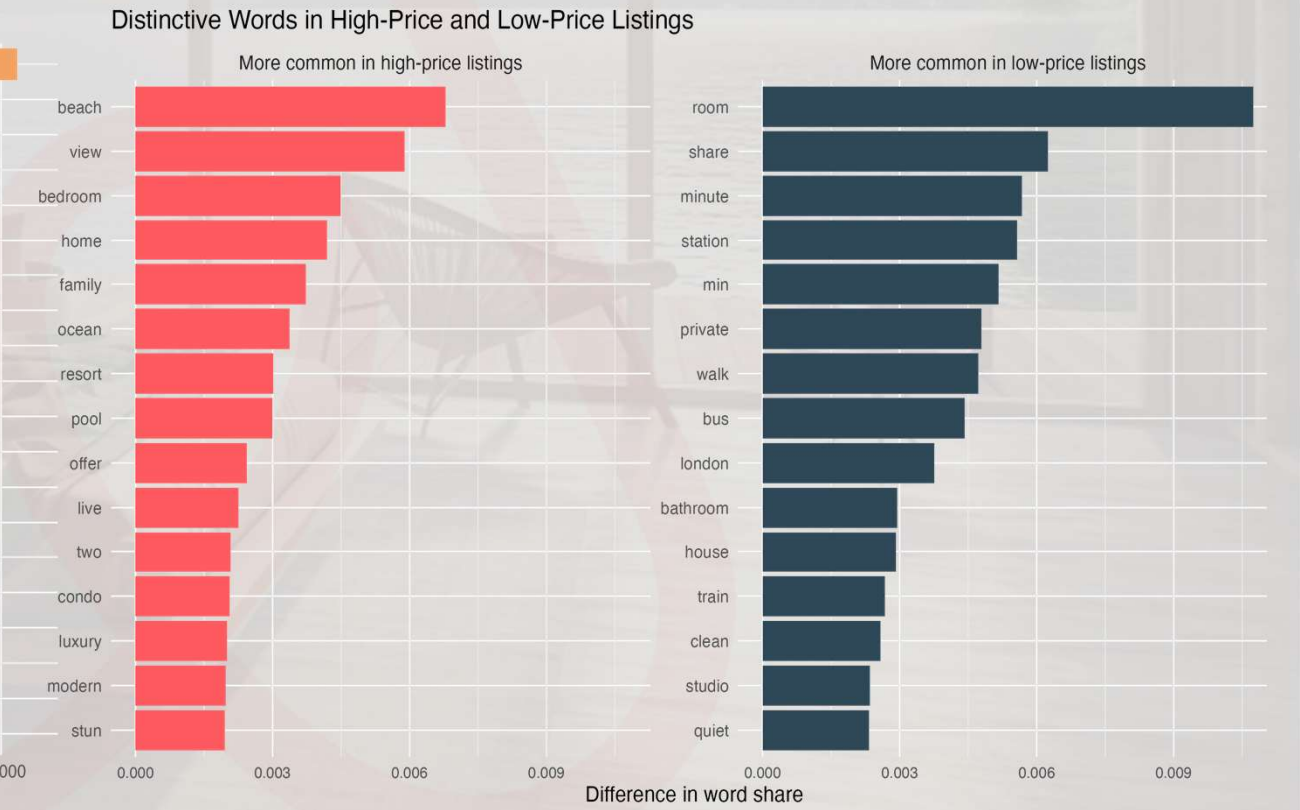
Tokenise

Clean tokens prepared for n-grams, TF-IDF, topics, and sentiment. Used to compare vocabulary patterns in low- vs high-price listings

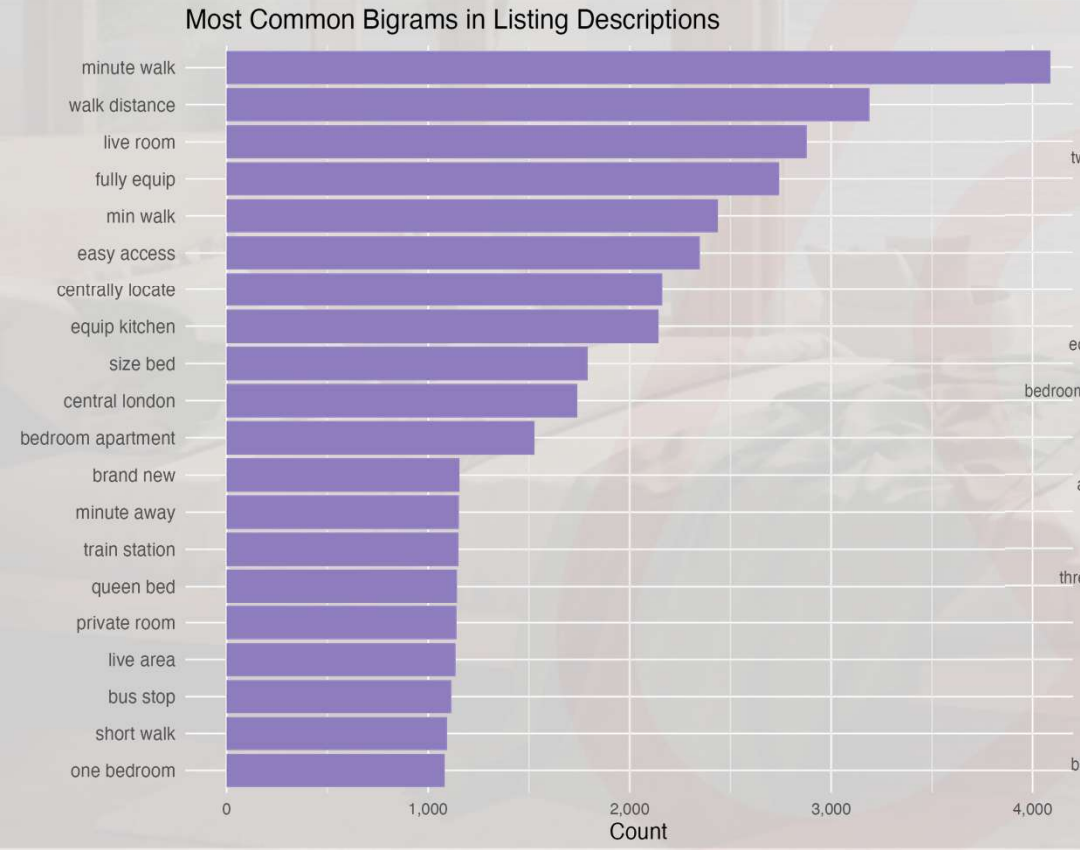
Price groups split by quartiles: bottom quartile = low price (~15k listings), top quartile = high price (~15k listings)

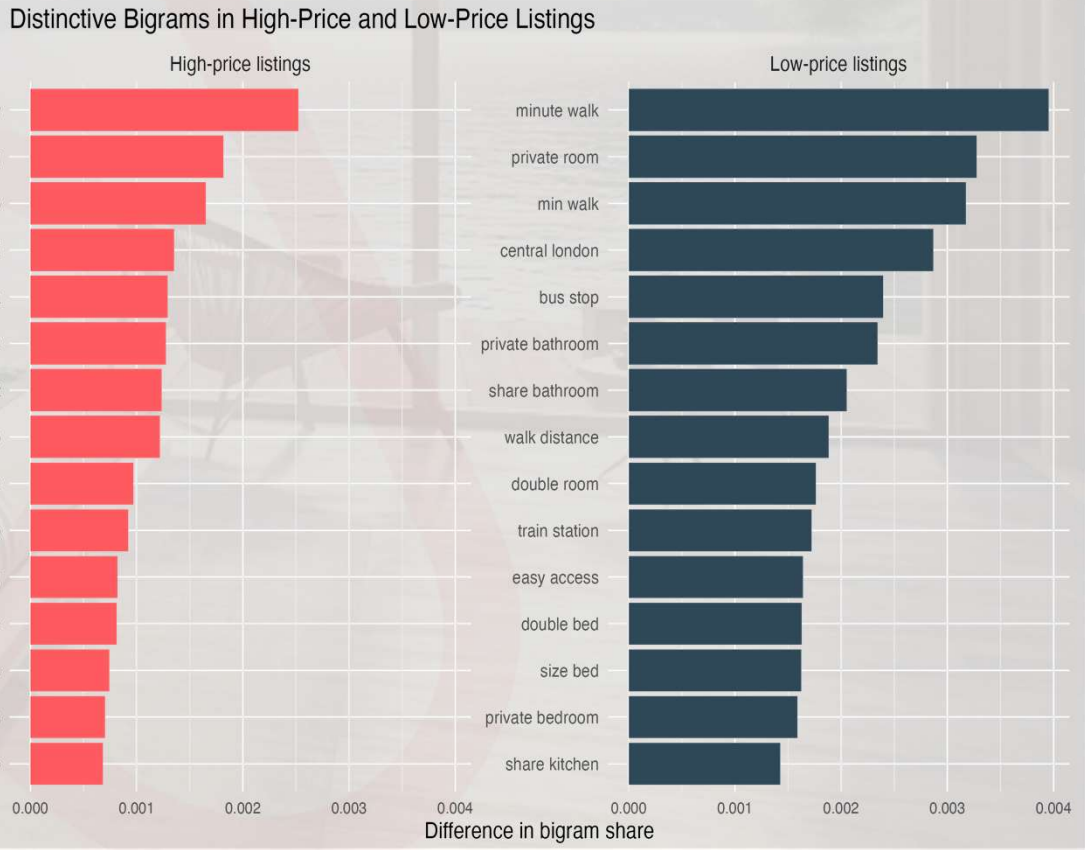
Unigrams: High vs. Low-Price Listings





Bigrams: High vs. Low-Price Listings





High vs. Low Phrases & Benchmark

High-price common phrases

- ocean view
- two bedroom
- fully equip
- king bed
- live area

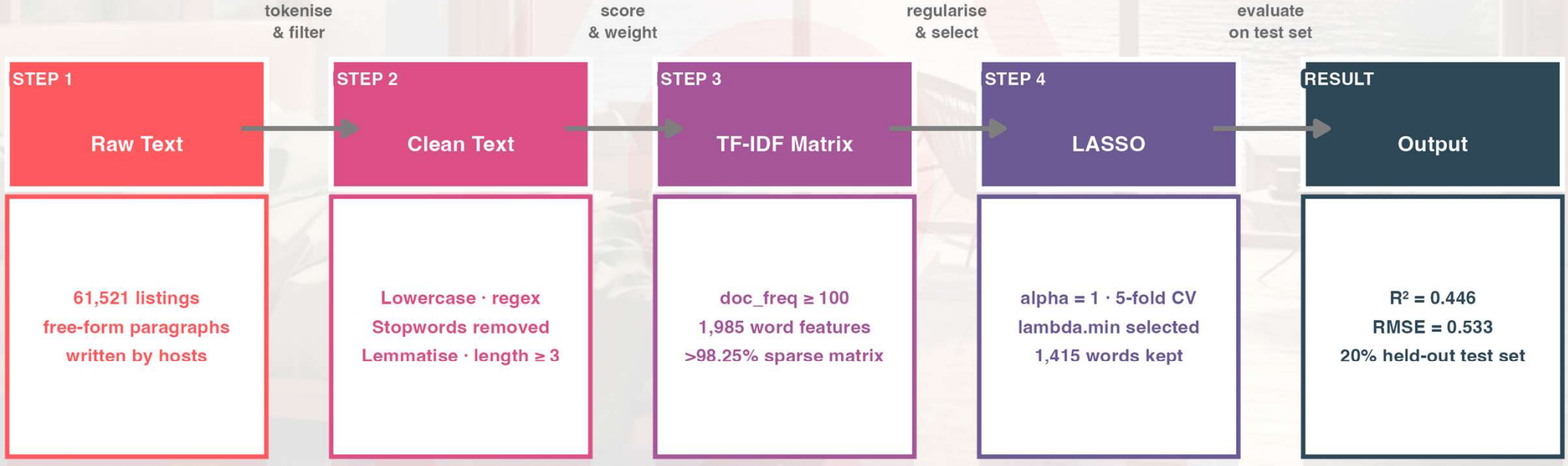
Low-price common phrases

- minute walk
- private room
- central london
- bus stop
- private bathroom

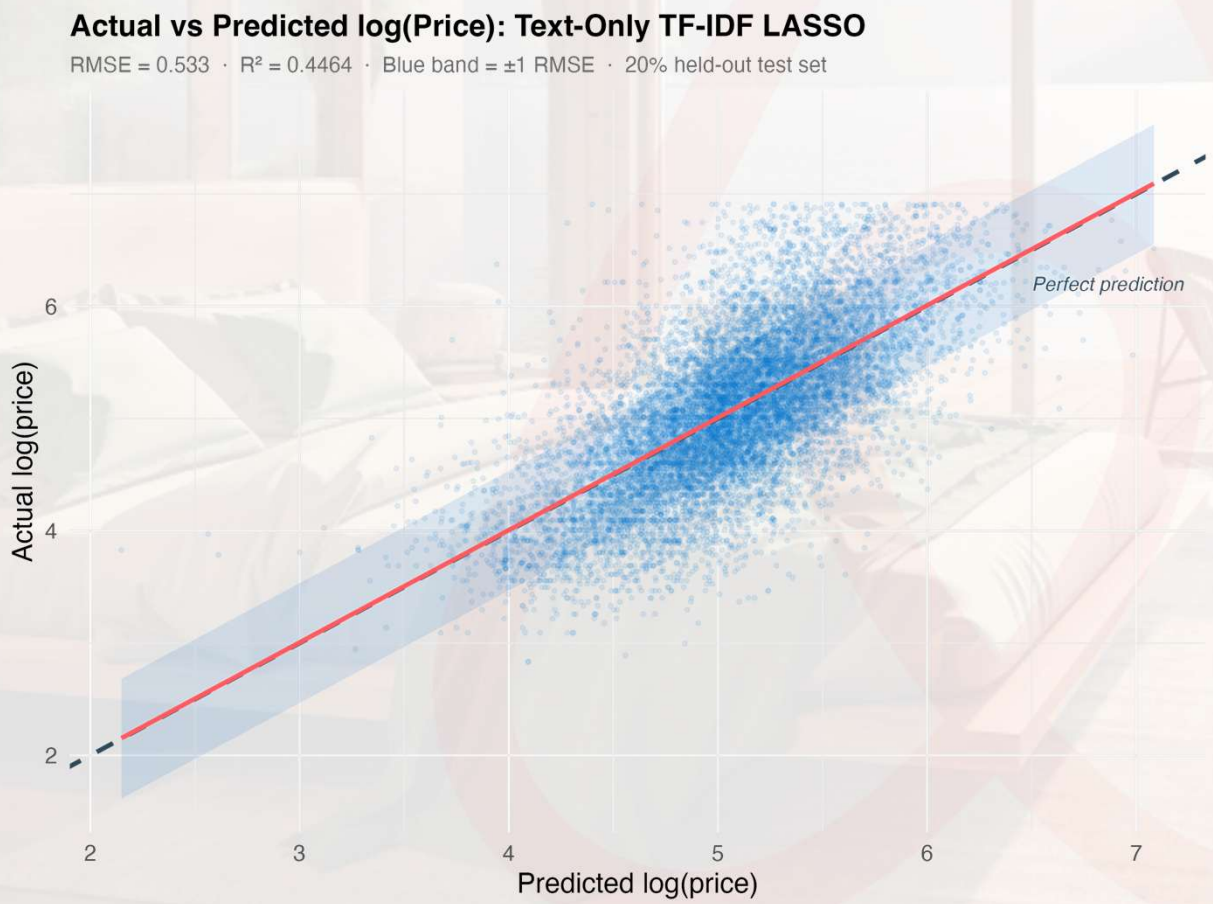
Metadata-only LASSO Benchmark, ***8 variables including:***
room_type, city, accommodates, bedrooms, bathrooms, beds, reviews, superhost

0.5687	0.4649	0.3613
R^2	$RMSE$	MAE

TF-IDF: Turning Descriptions into a Predictive Feature Matrix



Can Words Alone Predict Price? YES!



Every dot is one of the 12,304 listings the model never saw during training

Red line tracks the perfect-prediction line from cheapest listings (log 2) to most expensive (log 7) well-calibrated across the full range

Scatter around the line is the variation that room type, city, and bedroom count recover in the combined model

Can Words Alone Predict Price? Yes!

Metadata-only LASSO

RMSE: 0.4649 | MAE: 0.3613 | **R^2: 0.5687**

8 property variables (Room type, city, size, reviews, etc.)

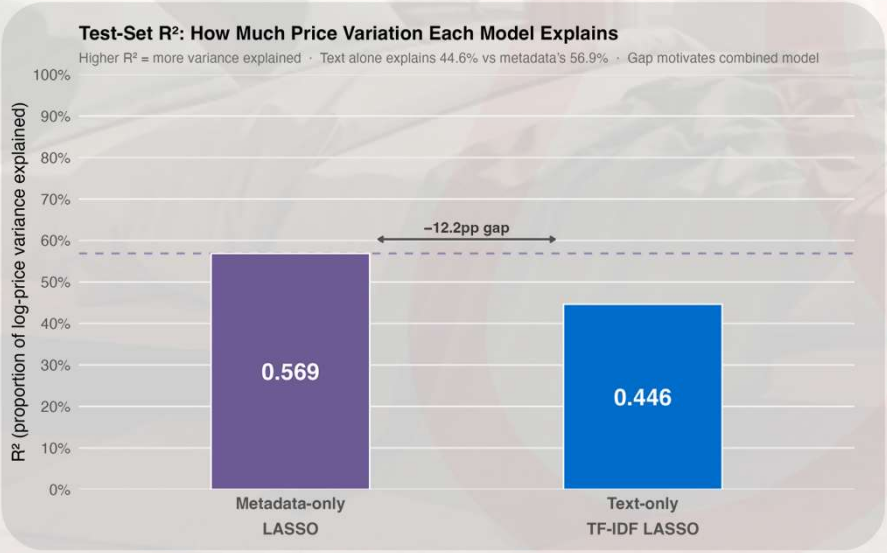
Text-only TF-IDF

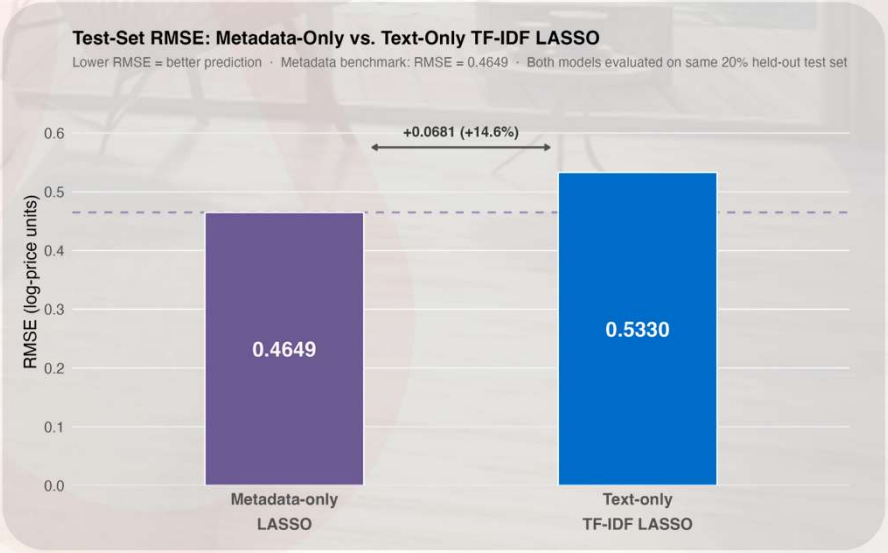
RMSE: 0.5330 | MAE: 0.4159 | **R^2: 0.4464**

98.25% sparsity, text alone – no metadata

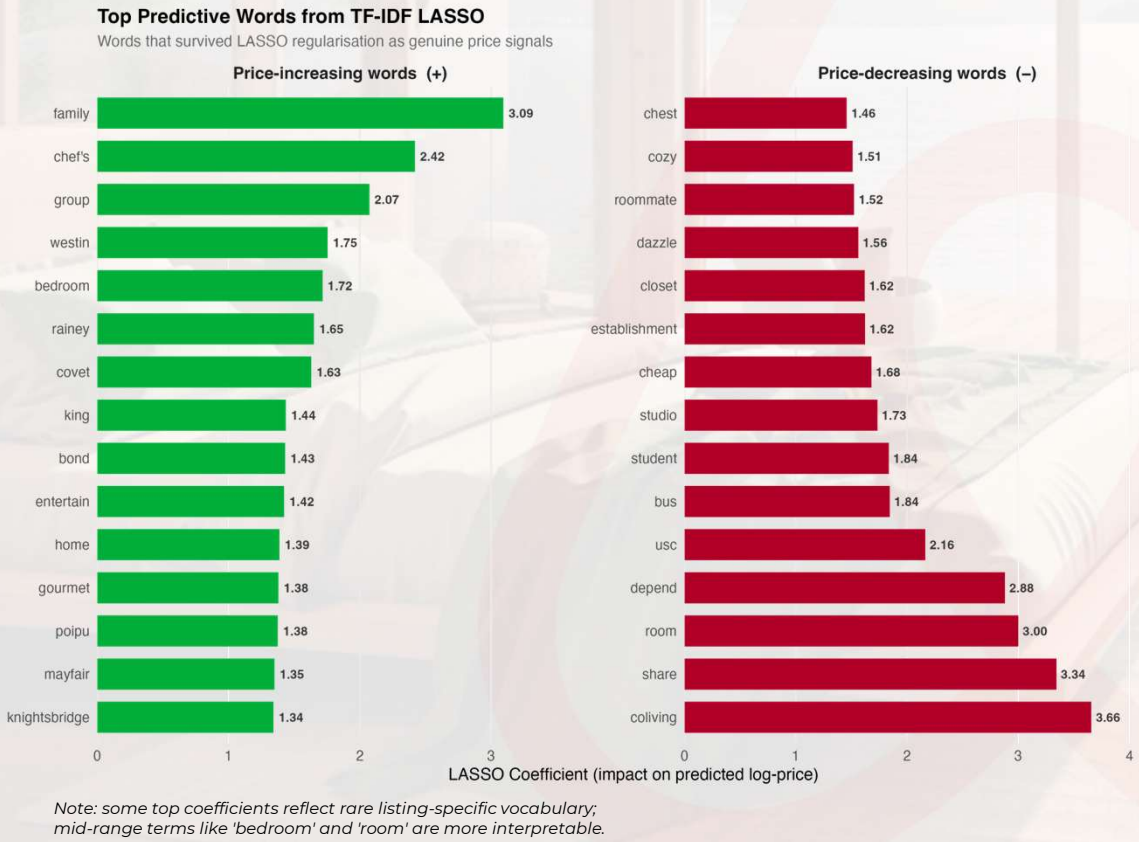
Metadata wins on accuracy but text reaches 78% of metadata's explanatory power using only words

The 12.2pp R^2 gap is not failure, it is the exact gain that the combined model recovers





LASSO Coefficients



Note: some top coefficients reflect rare listing-specific vocabulary; mid-range terms like 'bedroom' and 'room' are more interpretable.

1,415 words selected by LASSO from 61,521 listings and tested on 12,304 unseen data points

Green: premium locations, luxury amenities, whole-property signals

Red: shared arrangements, budget positioning, student/transport signals

Coefficients show association, not causation. The model found these patterns, not invented them

What Themes Emerge from Listing Descriptions?

Highest mean price – Topic 6: Private / Shared space

beach · view · enjoy · pool · beautiful · ocean · condo · unit · outdoor · resort

Mean log(price): 5.582
N = 8,170 listings

Listings that foreground private/shared room arrangements command the highest prices.

Mid-high mean price – Topic 2: Urban / Walkability

home · park · house · locate · family · downtown · new · neighborhood · quiet

Mean log(price): 5.085
N = 7,187 listings

Urban, walkable neighbourhood language aligns with mid-high prices.

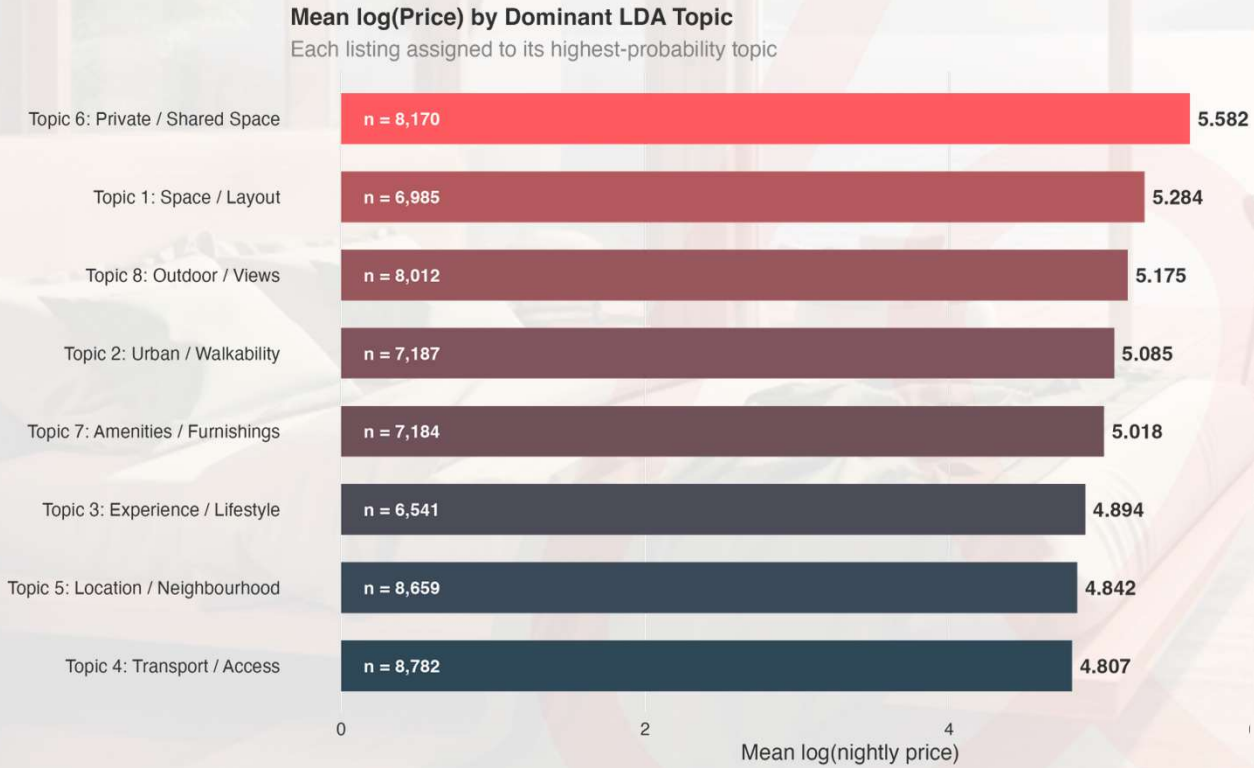
Lowest mean price – Topic 4: Transport / Access

minute · walk · min · station · london · central · flat · bus · street · train

Mean log(price): 4.807
N = 8,782 listings

Transport-focused descriptions cluster at the budget end of the market.

How Does Topic Relate to Price?



~117%

Real-terms price premium: Private/Shared Space vs. Transport / Access

Topic 6 (Private / Shared Space) has the highest mean log-price at 5.582, while Topic 4 (Transport / Access) has the lowest at 4.807.This implies an approximate 117% higher nightly price for listings dominated by private/shared-space language versus transport-access language.

The pattern suggests that description themes capture meaningful market positioning, not just writing style.

Does Language Improve Price Prediction? Yes!

Metadata-only LASSO

RMSE: 0.4649 | MAE: 0.3613 | **R^2: 0.5687**
8 property variables (Room type, city, size, reviews, etc.)

Text-only TF-IDF

RMSE: 0.5330 | MAE: 0.4159 | **R^2: 0.4464**
98.25% sparsity, text alone – no metadata

Combined: Meta + TF-IDF + Topics

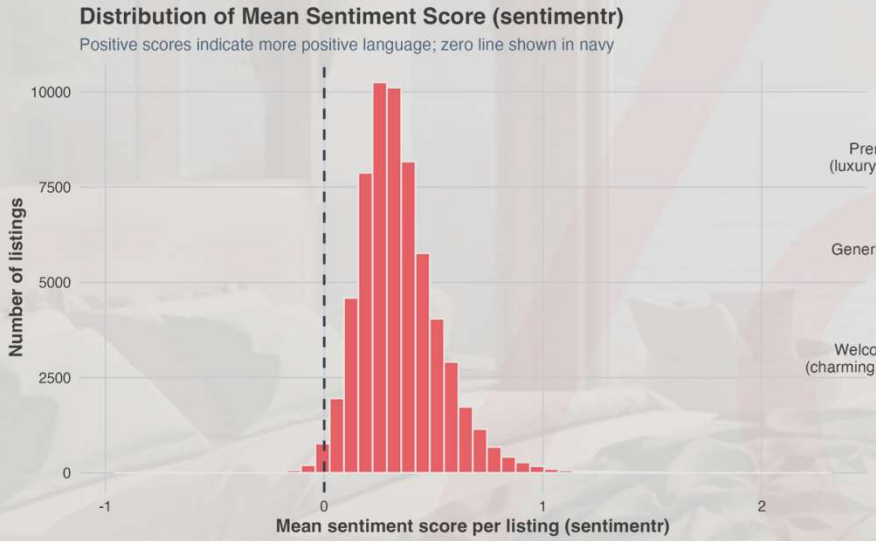
RMSE: 0.4061 | MAE: 0.3150 | **R^2: 0.6604**
Best model: ~16% higher R^2 vs. metadata baseline

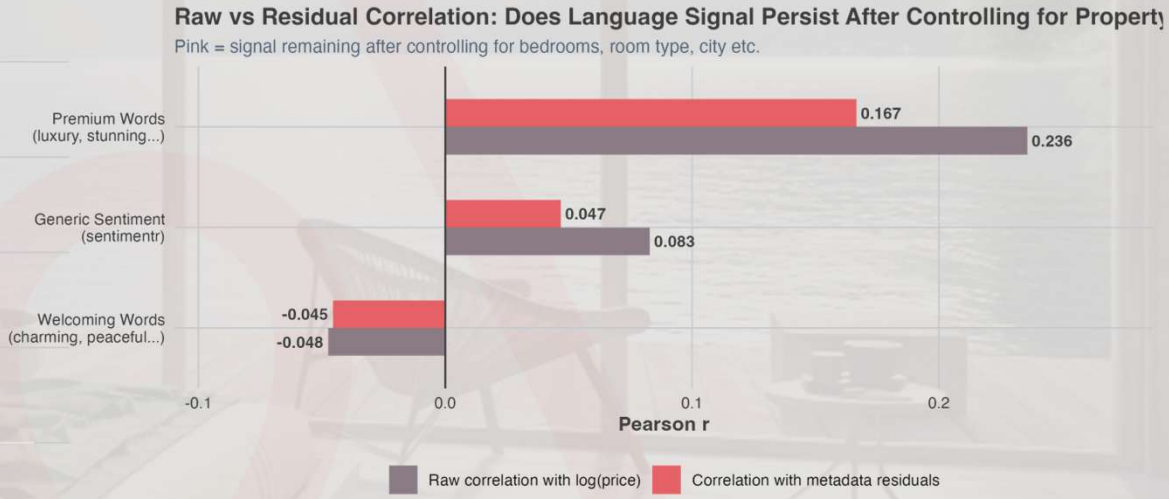
Combined model: 48,032 common listings because it intersects metadata, TF-IDF and topic features.

Research Question Answered:

YES – listing language explains price variation beyond property characteristics.
R^2 rises from 0.57 to 0.66 | RMSE falls by ~13% vs baseline
Language clearly improves price prediction beyond metadata.

Sentiment Analysis

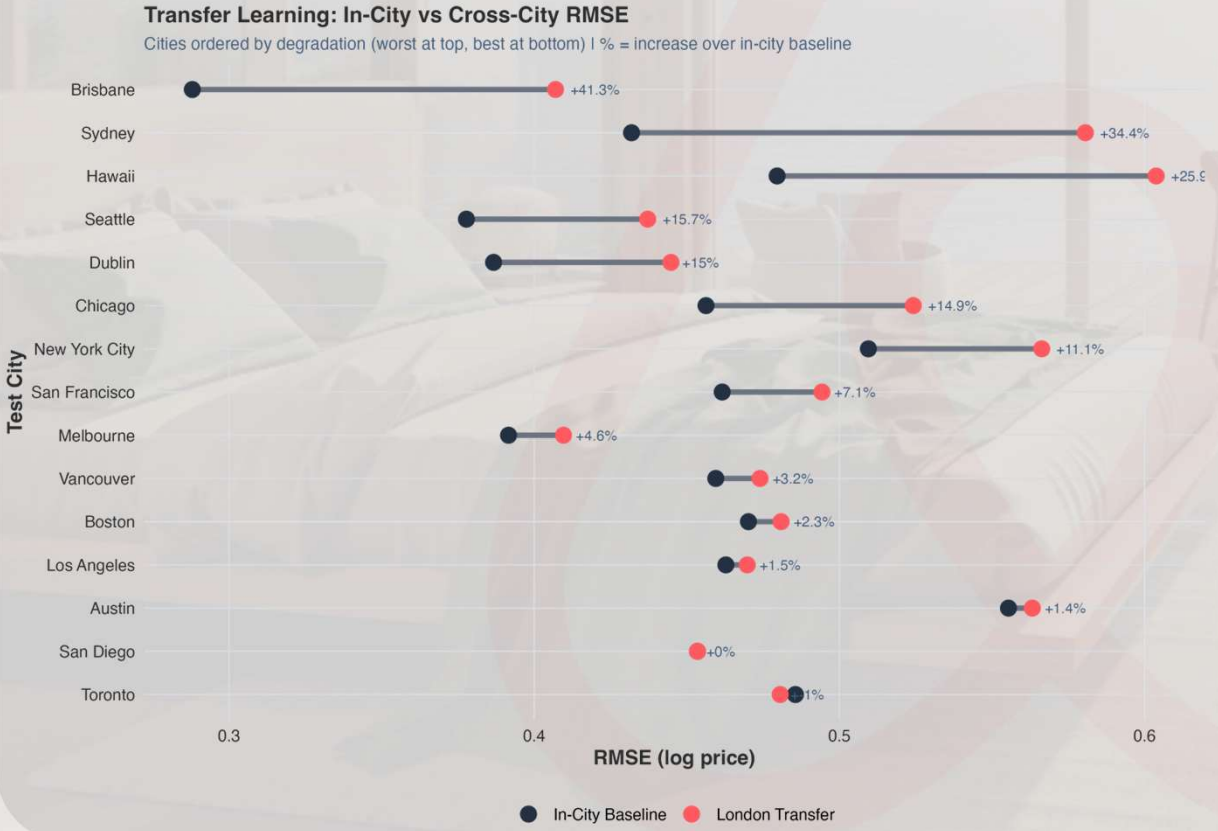




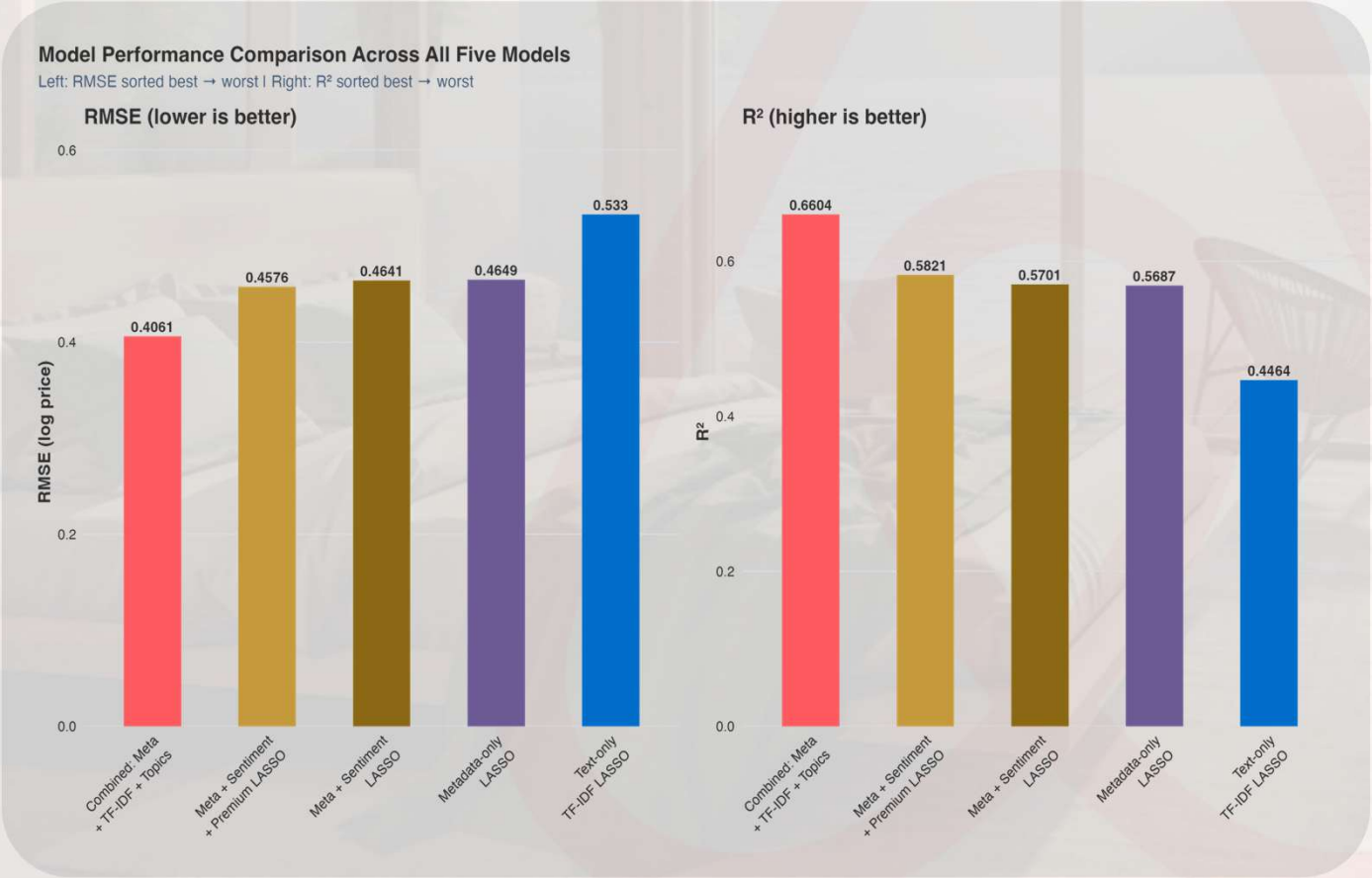
- Using **sentimentr**, we defined **premium** vocabulary (luxury, stunning, breathtaking) and **welcoming** vocabulary (cozy, peaceful, charming)
- Premium words **(r = 0.24)** strongly outperforms generic sentiment **(r = 0.08)** – what hosts say matters more than how positively they say it
- Premium signal survives property controls **(residual r = 0.17)** – luxury language commands prices above what the property alone justifies

Transfer Learning – Do Pricing Patterns Travel?

What we did: *Train on London (11,394 listings), test on 15 held-out cities, compare to each city's in-city RMSE baseline.*

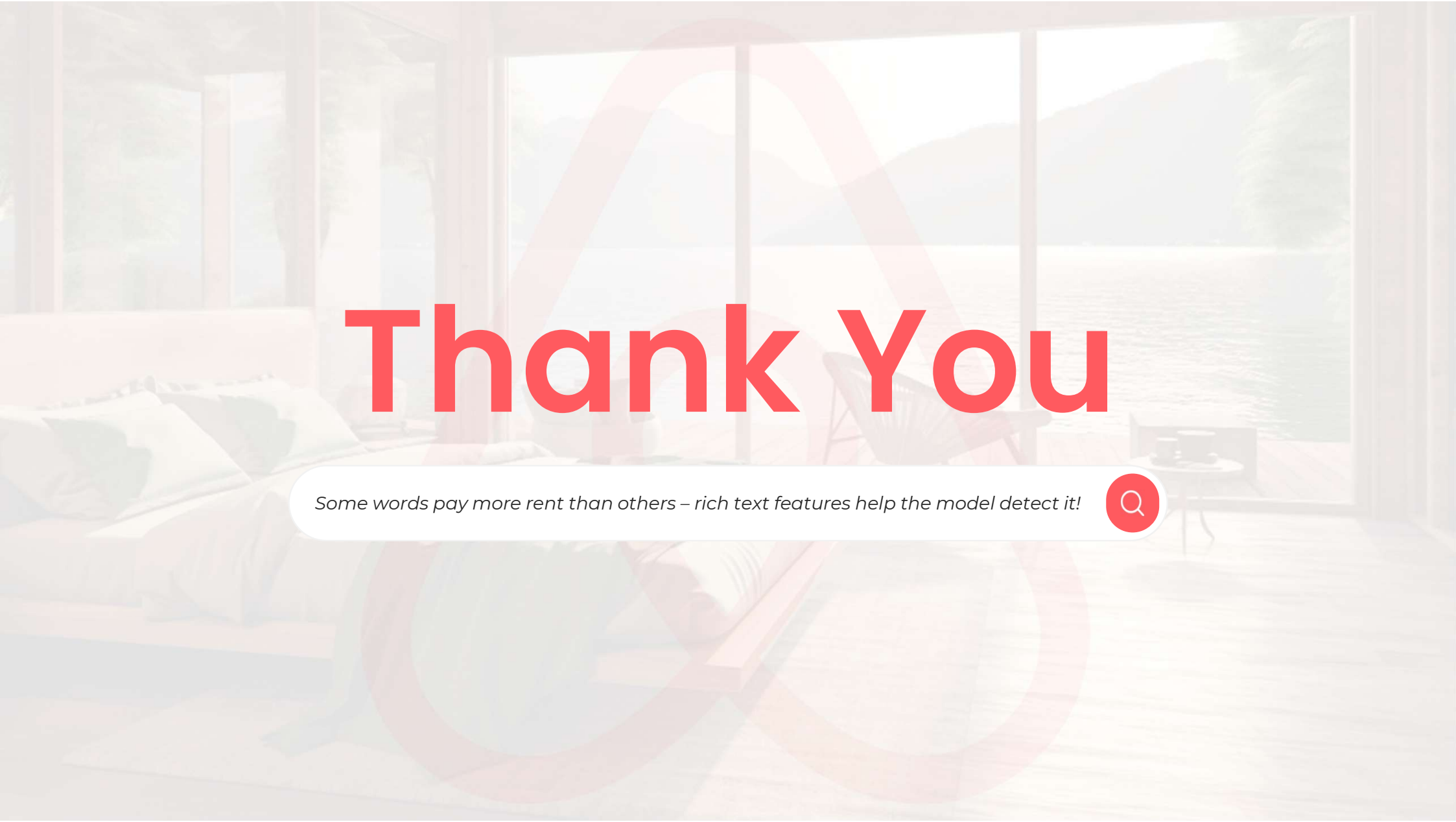


So, do listings lie? Sometimes



Key takeaways:

- **Metadata is strong:** Structured features like room type, city, and capacity already explain most of the variation in log price, giving a solid baseline model.
- **Sentiment is weak:** Once we control for those structural features, a single sentiment score adds little and can even slightly worsen prediction performance.
- **Rich text wins:** Representing language with thousands of word and topic-level features, combined with metadata, delivers the best overall predictive accuracy.


Thank You
Some words pay more rent than others – rich text features help the model detect it!