Quantifying Social Sentiment in Hostels Using A Domain-Specific Transformer Pipeline

Ian W. McMurry
Georgia Institute of Technology
Atlanta, GA, United States
imcmurry3@gatech.edu

Abstract

This paper presents a domain-specific transformer pipeline for quantifying social atmosphere in hostel reviews, an experiential dimension that travelers consistently prioritize but that existing NLP methods and booking platforms fail to capture. We train a cross-encoder on 4,994 manually annotated reviews and use it to pseudo-label 162,840 additional reviews; these labels are then distilled into a sentence-transformer bi-encoder, producing embeddings where proximity reflects social interaction level rather than generic sentiment. On held-out human-labeled data, the domain-adapted embeddings achieve F1 = 0.826, outperforming generic sentence embeddings (0.671) and zero-shot GPT-4o (0.774), with a 40-fold improvement in intra-class versus inter-class similarity. Aggregating predictions to the property level reveals that hostel socialness follows an approximate exponential distribution, confirming that highly social hostels are rare. This work formalizes socialness as a measurable semantic construct and provides a general template for extracting implicit experiential attributes from text at scale.

1 Introduction

For travelers staying in hostels, the social atmosphere, that is, how lively, friendly, and communal a property feels, is often more important than traditional quality indicators such as cleanliness, location, or amenities. Prior research in tourism and backpacking culture has consistently shown that younger and solo travelers select hostels specifically to meet other travelers, form temporary communities, and engage in shared social experiences (Oliveira-Brochado and Gameiro, 2013). Industry data similarly underscores this priority: over half of solo travelers report that "meeting interesting people along the way" is the single most memorable aspect of a trip, ranking above destination or attractions (Hostelworld, 2024, 2025). Despite this clear behavioral pattern, the social dimension of hostels remains poorly captured in both academic hospitality research and commercial booking platforms.

Traveler reviews present a promising yet underutilized source of information about social atmosphere. These reviews often contain rich descriptions of friendliness, group dynamics, shared meals, pub crawls, communal areas, and the ease or difficulty of meeting other guests. However, the signals are embedded in unstructured, subjective language that traditional NLP tools fail to extract meaningfully. General-purpose sentiment analysis models primarily detect positive versus negative valence, overlooking experiential or social nuance. Topic-modeling studies in hospitality typically emphasize concrete attributes such as cleanliness, facilities, location, or price (Li et al., 2015; Krishnan et al., 2024), and existing aspect-based sentiment models for hotels largely exclude social interaction as a defined category. Prior NLP work in the tourism domain repeatedly notes the challenge of modeling experiential, intangible, and socially constructed attributes, which often require domain-specific cues and contextual understanding (Xiang et al., 2015).

Even platforms explicitly designed for socially oriented travelers do not provide a direct measurement of social atmosphere. Hostelworld, for example, includes an "atmosphere" score, but this metric is likely degraded through expectation alignment. According to expectation–disconfirmation theory (EDT) (Oliver, 1980; Pizam and Milman, 1993), satisfaction ratings primarily measure whether a guest's expectations were met, not what the true underlying attribute is. A tranquil hostel can receive a high atmosphere rating if guests expected calmness and it delivered; a party hostel can receive the same high score when it meets expectations of social intensity. Because both extremes are rewarded equally, the rating cannot distinguish high-social from low-social hostels.

The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2026), pages 24–36

Due to the complexity of extracting social cues from unstructured reviews and the expectation-alignment bias embedded in atmosphere ratings under EDT, neither the research community nor the tourism industry currently possesses an operational or data-driven metric for measuring the true social interaction level of a hostel. Consequently, travelers cannot reliably compare hostels by social atmosphere, and operators lack tools to monitor or improve the communal experience that many guests value most.

To address this gap, we develop the first domain-specific transformer model designed to identify social behavior and communal atmosphere in hostel reviews. Our approach begins by training a high-accuracy cross-encoder to detect socialness cues in a curated seed dataset, enabling large-scale pseudo-labeling of more than 160k hostel reviews. Pseudo-labeling has a long history as an effective semi-supervised learning technique for low-density separation and improving classifier confidence (Chapelle et al., 2009; Lee et al., 2013). We use the term *pseudo-labeling* to emphasize that these automatically assigned labels are treated as supervision and used only for downstream representation learning rather than as ground truth annotations.

We then fine-tune a sentence-transformer bi-encoder on the pseudo-labeled corpus to create a vector space where semantic proximity reflects social interaction levels rather than generic sentiment or topic similarity. This mirrors successful strategies in other specialized NLP domains, such as BioBERT for biomedical literature (Lee et al., 2020) and FinBERT for financial communication (Huang et al., 2023), which demonstrate that pre-trained models often need to be adapted to domain-specific corpora to capture specialized meaning.

Together, the cross-encoder supervision and domain-adapted bi-encoder directly address the two core challenges outlined above. First, the cross-encoder enables precise identification of social cues that generic sentiment and topic models fail to capture, allowing the model to learn from examples where social interaction is subtly or implicitly expressed. Second, the fine-tuned sentence-embedding space overcomes the inadequacy of expectation-aligned ratings by providing a representation that reflects the actual semantic content of reviews rather than guests' prior expectations. In combination, this pipeline produces embeddings that encode the underlying social vibrancy of hostels, something neither raw ratings nor out-of-the-box NLP models can isolate.

This work contributes four advances:

1. A new annotated and pseudo-labeled dataset capturing social interaction themes in hostel reviews, the first collection focused explicitly on socialness rather than sentiment, amenities, or destination attributes.

2. A formal introduction of "socialness" as a measurable semantic construct within NLP for tourism, distinct from sentiment polarity, topic categories, or expectation-aligned ratings.

3. A domain-specific transformer embedding that captures hostel social atmosphere more accurately than general-purpose models, enabling fine-grained discrimination of communal versus non-communal environments.

4. Applied empirical evaluation, showing that the resulting embeddings support downstream tasks such as hostel ranking, clustering, retrieval, and social-atmosphere mapping at global scale.

By quantifying an attribute that travelers consistently value but that platforms cannot meaningfully measure, this work fills a methodological and practical gap at the intersection of natural language processing, tourism analytics, and experiential hospitality research. Beyond academic contributions, the resulting domain-specific embeddings enable a wide range of real-world applications: social-based hostel ranking, similarity search (for example, "find hostels with a similar vibe"), clustering of hostels into social typologies, temporal monitoring of a property's social atmosphere, and operator-facing analytics that help hostel managers understand, benchmark, and improve the communal guest experience. These downstream uses demonstrate both the feasibility and the immediate utility of modeling socialness as a quantifiable semantic dimension.

2 Related Work

2.1 Tourism NLP and the Absence of Social Atmosphere Modeling

Tourism and hospitality research has extensively mined online reviews, but this work consistently focuses on amenity-based and operational hotel

attributes rather than experiential or social atmosphere. For example, Li et al. (Li et al., 2015) analyze 118,000 TripAdvisor hotel reviews using emerging pattern mining and find that the dominant extracted features are location, price, room, service, staff, breakfast, and cleanliness, all of which are concrete physical characteristics of the property. Their automatically generated feature list such as room, staff, breakfast, cleanliness, lounge, and pool contains no representations of guest-to-guest interaction or communal social dynamics. Similar patterns appear in alternative accommodations. Krishnan et al. (Krishnan et al., 2024) apply sentiment analysis and topic modeling to homestay reviews and identify topics related to rooms, facilities, bathrooms, cleanliness, location, and host hospitality, but again no indicators of the vibe, communal energy, or interaction among guests that characterize social atmosphere. Across both hotels and homestays, existing tourism NLP reliably extracts tangible service attributes while entirely omitting interpersonal or community-oriented experiences, leaving the social dimension of accommodations computationally unmodeled.

2.2 Subjectivity and Experiential Meaning in NLP

A substantial body of work shows that many subjective or experiential attributes in text are conveyed implicitly rather than through explicit lexical markers. Kim and Klinger (Kim and Klinger, 2018) demonstrate that emotions are often expressed by describing situations rather than affective words, noting that implicit emotion requires reasoning over events and commonsense knowledge rather than surface sentiment cues. This finding highlights a general limitation of polarity-based or keyword-driven approaches when meaning is embedded in context.

Recent work on implicit semantics extends this challenge to modern embedding models. Sun et al. (Sun et al., 2025) show that widely used text embeddings prioritize surface similarity and topical overlap while performing poorly on tasks involving implicature, speaker stance, or social meaning. Their experiments indicate that even state-of-the-art models capture only a small fraction of pragmatic or attitudinal content.

These observations indicate that experiential constructs often require specialized modeling to infer meaning that is not explicitly stated. Social atmosphere in hostel reviews exhibits similar characteristics. Descriptions of social interaction, friendliness, or communal energy are frequently implied through situational cues rather than explicit terminology. This motivates the need for a domain-specific modeling approach capable of capturing these implicit social signals

2.3 Domain-Specific Transformer Models

Large pretrained language models often require domain-specific adaptation when applied to specialized corpora, since general-purpose models do not fully capture the terminology, semantic patterns, or contextual cues present in domain-specific text. BioBERT (Lee et al., 2020) demonstrates this clearly in the biomedical domain. By continuing BERT pretraining on PubMed abstracts and PMC full-text articles, BioBERT achieves substantial performance gains across biomedical named entity recognition, relation extraction, and question answering tasks, as demonstrated by substantial gains across biomedical NER, relation extraction, and question answering. FinBERT (Huang et al., 2023) applies the same principle to financial communication, where nuanced sentiment and domain-specific vocabulary lead general pretrained models to underperform. The authors show that pretrained transformers adapted to financial filings and earnings reports provide significantly more accurate sentiment extraction, outperforming general BERT on financial-specific evaluation tasks. Together, these models show that domain-adapted transformers consistently outperform general-purpose variants when the target domain contains specialized terminology or subtle contextual cues. This motivates our use of a domain-specific sentence-transformer trained on hostel reviews, where social interaction cues are similarly domain dependent and not well captured by general pretrained models.

3 Methodology

Our goal is to learn a representation of hostel reviews in which distances reflect the underlying social interaction level rather than generic sentiment or topic similarity. The pipeline consists of (1) defining socialness and constructing a labeled seed set, (2) training a cross-encoder classifier, (3) pseudo-labeling the full corpus (4) fine-tuning a domain-specific bi-encoder on these labels, (5) calibrating a linear classifier on human-annotated data, and (6) aggregating review-level predictions into hostel-level socialness scores.

3.1 Defining Socialness and Seed Labels

3.1.1 Construct Definition

Socialness refers to the extent to which a review describes guest-to-guest interaction, communal energy, or situational cues indicating that travelers are likely to meet, engage, or form temporary communities. Unlike amenities (e.g., cleanliness, facilities) or evaluative constructs such as sentiment, socialness captures the interpersonal dynamics of the property as experienced by guests. Socialness and overall sentiment are independent dimensions: a review may be strongly positive or negative in tone without being social at all, and a highly social environment may be described unfavorably.

Linguistically, socialness is expressed primarily through implicit situational descriptions rather than explicit labels such as "this is a social hostel." Reviews rarely state socialness directly. Instead, social cues appear through descriptions of activities, shared spaces, and interaction patterns among guests. For example, a review describing "parties every night" or "something happening every day like drinking games and going to the clubs" conveys a highly social environment even when accompanied by complaints about noise or cleanliness. These cases illustrate that high socialness can co-occur with negative sentiment.

Conversely, not all mentions of party-related vocabulary or positive affect indicate genuine social interaction. A reviewer may note that "while it is a party hostel, I found it impersonal and hard to meet people," indicating the presence of a party-oriented setting without meaningful guest-to-guest engagement. Similarly, reviews emphasizing "very friendly staff" or a welcoming host environment provide no evidence of interactions among guests and are therefore considered non-social. These distinctions motivate grounding the construct in observed interpersonal dynamics rather than topical keywords, facilities, or affective polarity.

Formally, we view socialness as a continuous experiential dimension. However, for supervision at the review level, we operationalize it as a binary label: `social = 1` if the review provides clear textual evidence of guest interaction or communal participation, and `social = 0` otherwise. This binary operationalization prioritizes annotation consistency and high-precision supervision. The underlying continuity of socialness emerges downstream through aggregation across reviews and through the geometry of the learned embedding space, rather than from individual review labels.

The full annotation rubric, including decision rules and a labeling flowchart, is provided in Appendix A.

3.1.2 Seed Dataset and Annotation

The cross-encoder (Section 3.3.1) is trained on a manually annotated seed dataset of 4,994 reviews sampled from a relational database of hostel reviews. Because explicitly social reviews are relatively rare in purely random samples, we use keyword-guided enrichment as a sampling strategy only to increase the prevalence of candidate socialness edge cases during annotation. Concretely, we draw a mixed pool consisting of (i) reviews containing social-event related strings (e.g., "pub crawl", "party hostel"), (ii) reviews containing calm or low-sociality strings (e.g., "quiet", "not a party hostel"), and (iii) randomly sampled reviews. Full list of strings that were used to draw out the enrichment can be seen in Appendix B. To avoid collapsing onto narrow lexical triggers, these keyword pools are used solely to select reviews for annotation, and final labels are assigned based on full-text review content using the rubric in Appendix A rather than keyword presence. In practice, enriched candidates constituted a minority of the seed dataset, with the remainder drawn from random sampling to preserve coverage of the natural review distribution.

All reviews in the seed dataset were labeled manually by a single annotator with domain familiarity, prioritizing internal consistency in applying the socialness rubric. Ambiguous cases were resolved conservatively as non-social to maintain high precision in the positive class. This annotated seed set provides the sole human supervision used to train the cross-encoder model.

3.2 Dataset and Preprocessing

The full corpus consists of 162,840 hostel reviews from two major booking platforms, covering 2,230 distinct hostels worldwide. Reviews were programmatically collected from publicly accessible online review pages and stored in a relational database prior to preprocessing. The texts are short but information-dense, with most reviews well within standard transformer context windows, making them suitable for sentence-level modeling.

Because hostel travel is highly international, the corpus is multilingual. We first apply automatic language identification and then translate all non-English reviews into English using the Google

Cloud Translation API. The detected language code is stored alongside the original text. Reviews identified as English are kept verbatim; others are translated in batched requests. If translation fails, the original text is retained. Manual inspection of a random subset confirms that translations remain fluent and preserve the situational cues relevant for socialness.

Preprocessing is intentionally minimal. Aggressive normalization could remove precisely the cues needed to detect guest interaction. We therefore preserve punctuation, capitalization, emojis, and nonstandard spellings. The only substantive cleaning step concerns reviews that are scraped into multiple fields (e.g., separate positive and negative comments); these are concatenated into a single free-text field. Empty or null reviews are discarded. Otherwise, text is kept exactly as written by travelers.

3.3 Modeling Pipeline

3.3.1 Cross-Encoder Socialness Classifier

The first stage of the pipeline is a high-precision cross-encoder that maps each review to the probability that it describes a socially active environment. We fine-tune the publicly available `cross-encoder/ms-marco-MiniLM-L-12-v2` model from Sentence-Transformers, originally trained for passage–query relevance. The model receives as input a pair consisting of (1) the review text and (2) a fixed natural-language query that specifies the target construct, such as:

> *"Does this review describe a lively social environment with friendly guests?"*

The cross-encoder jointly encodes the concatenated review and query and outputs a scalar probability via a sigmoid layer. We train the model with binary cross-entropy loss on the annotated seed set, using standard hyperparameters for learning rate, batch size, and number of epochs. Because social reviews are less frequent than non-social ones, we modestly upsample the positive class during training to expose the model to a wider variety of social cues. To assess generalization across properties, we use both a stratified row-level split and a hostel-level split in which all reviews from held-out hostels appear only in validation or test sets.

3.3.2 Pseudo-Labeling and High-Confidence Subset

After training, we apply the cross-encoder to the entire translated corpus to obtain, for each review x, a probability $p_{\mathrm{CE}}(y = 1 \mid x)$ that the review is social. We then derive a binary pseudo-label $\hat{y}_{\mathrm{CE}} \in \{0, 1\}$ for every review using a fixed decision threshold selected on the validation set.

To reduce noise in downstream representation learning, pseudo-labeling pipelines often apply confidence-based filtering to discard ambiguous cases near the decision boundary. In our setting, however, the cross-encoder produces a sharply bimodal probability distribution, with relatively few reviews assigned intermediate probabilities. We therefore retain the full pseudo-labeled corpus for bi-encoder training, as explicit high-confidence thresholding would remove only a small number of borderline instances and does not materially affect downstream performance. This approach yields a large, automatically labeled dataset while preserving the diversity of social and non-social signals present in the corpus.

3.3.3 Domain-Specific Bi-Encoder Training

Although the cross-encoder yields high-quality predictions, it is computationally costly because each inference requires full joint attention over the review and query. To support large-scale analysis (e.g., unsupervised clustering, embedding-space exploration) and to enable low-latency inference for retrieval or recommendation settings, we distill the cross-encoder into a lightweight sentence-transformer bi-encoder that produces socialness-aware embeddings.

We start from the `sentence-transformers/all-MiniLM-L6-v2` model, which maps each review x independently to a 384-dimensional embedding $f(x)$. Training is performed on the pseudo-labeled reviews. We first create train/validation/test splits with respect to $\hat{y}_{\mathrm{CE}}$, optionally subsampling to maintain a balanced and tractable training set.

Rather than conditioning on a query, the bi-encoder is trained with a pairwise similarity objective over review–review pairs. Positive pairs are constructed from reviews that share the same pseudo-label (both social or both non-social), and negative pairs from reviews with different pseudo-labels. The model is optimized with a cosine-similarity loss that encourages embeddings of same-label reviews to be close and embeddings

of different-label reviews to be far apart. We use standard settings for batch size, learning rate, and number of epochs. The best checkpoint is selected based on validation performance on the pseudo-label classification task.

3.3.4 Linear Classifier Head and Continuous Scores

To align the distilled representation directly with human annotation and to obtain a lightweight classifier for downstream use, we freeze the fine-tuned bi-encoder and train a logistic regression head on top of its embeddings. Each of the 4,994 human-labeled reviews is embedded with the bi-encoder, and the resulting vectors are split into train/validation/test sets using the same proportions as for the cross-encoder. An ℓ_2-regularized logistic regression model with class-balanced weights is then trained to predict the manual socialness label from the embedding.

In addition to binary decisions, both the logistic regression output and the underlying embedding space provide continuous socialness scores. For example, hostel-level socialness can be computed as the fraction of a hostel's reviews predicted as social or by aggregating the continuous probabilities across its reviews. These aggregation procedures are described in the following subsection.

3.3.5 Hostel-Level Aggregation

Finally, we derive property-level socialness scores by aggregating review-level predictions. For each hostel, we collect all associated reviews and compute the fraction of reviews classified as social by the bi-encoder + logistic regression model. This fraction serves as a continuous socialness score in $[0, 1]$, where higher values indicate a higher prevalence of social interaction in the textual record of the property. Alternative aggregation schemes (e.g., averaging probabilities, weighting by recency) are straightforward but beyond the scope of this work.

4 Results

4.1 Dataset Descriptive Statistics

The full corpus consists of 162,840 hostel reviews collected from two major booking platforms: Booking.com (88,656 reviews) and Hostelworld (74,184 reviews). Reviews span 72 cities worldwide and correspond to 2,230 unique hostels, reflecting substantial geographic and property-level diversity.

The dataset is multilingual, reflecting the international nature of hostel travel. Automatic language identification indicates that 68.08% of reviews are written in English, with the remaining reviews distributed across more than twenty languages. Table 1 summarizes the most common languages. Because cues related to social interaction may appear across diverse linguistic contexts, all subsequent analyses operate on a translated English corpus.

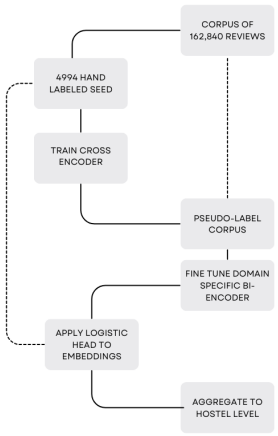

Figure 1: Methodology Workflow

Language	Percent of Corpus
English	68.08%
Spanish	9.25%
French	4.59%
Portuguese	4.06%
German	3.25%
Japanese	1.47%
Italian	1.31%
Russian	1.24%
Korean	0.91%
Dutch	0.89%

Table 1: Most common languages in the 162,840-review corpus prior to translation.

Reviews are generally short to moderate in length. The mean review length is 55.8 tokens, with a median of 43 tokens and a standard deviation of 48.6 tokens. The distribution is right-skewed, with the 90th percentile at 110 tokens, the 95th percentile at 128 tokens, and the 99th percentile at 235 tokens. Extremely long reviews are rare: only 0.05% of reviews exceed 512 tokens, and fewer than one hundred reviews exceed 1,024 tokens. This length profile is well suited to transformer-based encoders without requiring aggressive truncation.

4.2 Human-Labeled Seed Dataset

The cross-encoder is trained on a manually annotated seed set of 4,994 reviews. This dataset provides the only human supervision used in the

pipeline and therefore anchors the definition of socialness. Table 2 summarizes its key properties. Social reviews constitute 25.15% of the dataset, reflecting the relative rarity of explicit guest–guest interaction cues in hostel reviews. Review lengths are short (median 50 tokens), and the dataset spans both major platforms, ensuring coverage of diverse writing styles and traveler perspectives.

Statistic	Value
Total reviews	4,994
Positive class (social = 1)	25.15%
Negative class (social = 0)	74.85%
Median length	50 tokens
95th percentile length	146 tokens
Platforms represented	Hostelworld & Booking.com

Table 2: Summary statistics of the human-labeled seed dataset used to train the cross-encoder.

4.3 Evaluation Metrics

We report standard metrics for binary classification, chosen to reflect both thresholded classification performance and ranking quality:

- **F1 score** is the harmonic mean of precision and recall and provides a single summary measure of classification performance at a fixed decision threshold. Because we care equally about avoiding false positives and false negatives, F1 serves as the primary metric for comparing models in this setting.
- **AUC** (area under the ROC curve) measures ranking quality independent of any particular threshold.
- **Average Precision (AP)** summarizes the precision–recall curve by weighting precision at different recall levels.

4.4 Cross-Encoder Performance

We first evaluate the cross-encoder on the 4,994 human-labeled reviews. Table 3 reports performance under two evaluation regimes: (a) a standard row-level split and (b) a stricter hostel-level split that tests generalization to unseen properties. The model achieves high discrimination (AUC > 0.97) and ranking quality (AP $= 0.954$) in both settings. The slightly higher F1 on the hostel-level split indicates that the classifier captures generalizable cues of social interaction beyond property-specific phrasing.

Split	AP	AUC	F1	Thresh.
Row-level	0.9540	0.9819	0.8829	0.0733
Hostel-level	0.9540	0.9797	0.8916	0.0972

Table 3: Cross-encoder performance on human-labeled data.

4.4.1 Pseudo-Labeling the Full Corpus

Applying the cross-encoder to all 162,840 reviews yields continuous socialness probabilities and pseudo-labels. The model predicts that 20.34% of reviews describe socially active environments, representing a downward shift from the hand-labeled seed set. This difference is expected because the seed dataset was constructed using a sampling enrichment procedure that increased the prevalence of candidate socialness edge cases (in addition to random reviews) during annotation, whereas the full corpus reflects the natural distribution of review content. Figure 2 shows the resulting probability distribution, which is sharply bimodal: most reviews lie near 0 or near 1.

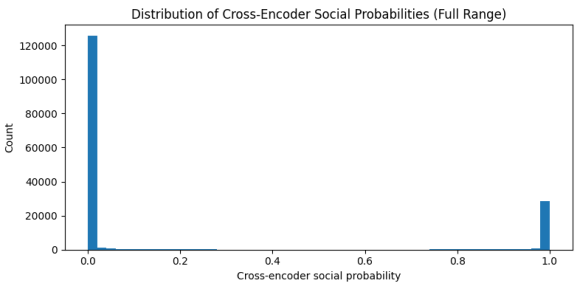

Figure 2: Cross-encoder probability distribution for all reviews.

The cross-encoder probability distribution is sharply bimodal, with most reviews assigned probabilities near 0 or 1. As a result, we do not apply explicit high-confidence filtering during pseudo-labeling, since removing borderline cases would discard only a small fraction of the data and does not materially affect downstream performance.

Platform differences. Hostelworld reviews exhibit a substantially higher proportion of social-labeled content (27.87%) compared to Booking.com reviews (14.04%). This pattern aligns with differences in platform focus: Hostelworld targets backpackers and hostel-goers, for whom social interaction is often a primary motivation for accommodation choice, while Booking.com serves a broader traveler population. The result provides correlational evidence that social atmosphere is a

meaningful and behaviorally relevant dimension in hostel travel.

4.5 Bi-Encoder Distillation and Classification

4.5.1 Classification Performance

We embed each review with the fine-tuned bi-encoder and train a logistic regression classifier on the human-labeled dataset. Table 4 compares performance against two baselines: (1) generic sentence-transformer embeddings produced by `sentence-transformers/all-mpnet-base-v2` without domain adaptation, and (2) a zero-shot classification prompt to GPT-4o-mini. Domain-adapted embeddings yield the strongest overall performance, improving F1 by 15.5 points and AUC by 8 points over the generic embedding baseline.

Model	F1	Prec.	Rec.	AUC
Generic embeddings + LogReg	0.671	0.594	0.772	0.872
GPT-4o-mini (zero-shot)	0.774	0.983	0.638	0.817
Fine tuned Bi-encoder + LogReg	**0.826**	**0.790**	**0.865**	**0.952**

Table 4: Classification performance on the human-labeled test set.

4.5.2 Embedding Structure

To assess whether socialness forms a coherent dimension in embedding space, we project embeddings of the 4,994 labeled reviews using UMAP. As shown in Figure 3, generic embeddings exhibit substantial overlap between classes, whereas the domain-specific bi-encoder produces a clear separation between social and non-social reviews.

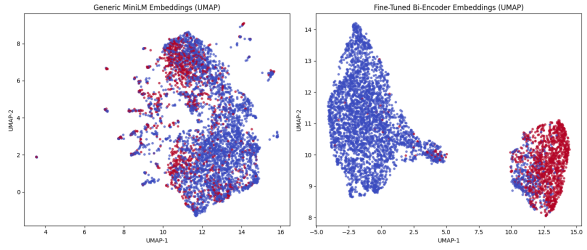

Figure 3: UMAP projection of generic embeddings (left) and domain-specific embeddings (right).

Pairwise cosine similarities confirm this structure: the intra–inter class similarity gap increases from 0.018 (generic) to 0.704 (fine-tuned), a 40-fold improvement. Unsupervised k-means clustering (with $k = 2$) achieves purity of 0.909 on the bi-encoder embeddings, compared to 0.748 for the generic baseline, indicating strong alignment between the learned representation and the socialness construct.

4.6 Hostel-Level Socialness Distribution

For each hostel, we aggregate review-level predictions to compute the fraction of its reviews classified as social, yielding a continuous property-level socialness score in $[0, 1]$. Prior to distribution fitting, we remove hostels with `socialness` $= 0$ (i.e., no reviews predicted as social) and drop any invalid or missing values, resulting in $n = 2{,}072$ hostels retained out of 2,230.

Figure 4 shows the empirical distribution of these nonzero hostel-level scores. Fitting a one-parameter exponential model via maximum likelihood yields $\hat{\lambda} = 5.39$ (mean $= 0.186$). The exponential provides a strong descriptive approximation in the upper tail: the empirical survival curve is close to linear on a log scale ($R^2 = 0.9748$). However, formal goodness-of-fit tests reject the exponential assumption (KS: $p = 3.8 \times 10^{-10}$; χ^2: $p = 4.2 \times 10^{-4}$), indicating statistically detectable deviations from an exact exponential model.

Despite this rejection, the near-linear log-survival behavior suggests a clear rarity structure: highly social hostels are uncommon, while most properties exhibit limited or intermittent social cues. We therefore interpret the exponential fit as a compact summary of the decay pattern rather than a strictly correct generative model.

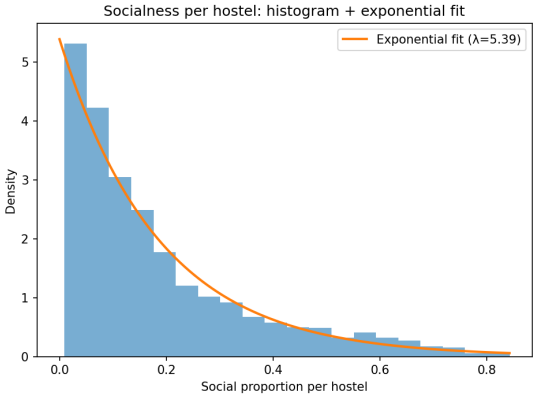

Figure 4: Distribution of hostel-level socialness scores with exponential fit.

5 Conclusion

This work introduces a data-driven approach for quantifying the social atmosphere of hostels from unstructured traveler reviews. By formalizing *socialness* as a semantic construct distinct from sentiment, amenities, and expectation-aligned ratings, we address a long-standing gap in both tourism

research and commercial travel platforms: the absence of an operational metric for guest-to-guest interaction and communal experience.

Empirically, we show that socialness can be reliably inferred from text using a domain-specific modeling pipeline. A cross-encoder trained on a small, carefully annotated seed set generalizes well to unseen hostels and produces sharply bimodal probability distributions when applied at scale. Distilling these judgments into a sentence-transformer bi-encoder yields an embedding space in which socialness emerges as a coherent latent dimension, enabling both accurate classification and unsupervised structure discovery. The resulting representations substantially outperform generic embeddings and zero-shot prompted language models, demonstrating the value of domain adaptation for capturing implicit experiential meaning.

At the property level, aggregating review-level predictions reveals a striking and well-behaved distribution: hostel socialness follows an exponential decay, with highly social properties being rare and most hostels exhibiting limited or intermittent social cues. This structure provides independent validation that the learned signal is not arbitrary, and it offers a compact statistical characterization of social atmosphere at global scale.

Beyond academic contributions, the learned embeddings enable practical applications that are not supported by existing rating systems. These include social-based hostel ranking, similarity search (e.g., discovering hostels with comparable social profiles), unsupervised clustering into social typologies, and temporal monitoring of changes in communal atmosphere. By shifting the measurement of social experience from expectation-aligned ratings to text-derived semantic structure, this work demonstrates how implicit social behavior can be operationalized at scale.

More broadly, the approach illustrates a general strategy for modeling experiential constructs that are difficult to define explicitly but are consistently described in natural language. The combination of targeted supervision, pseudo-labeling, and domain-specific representation learning provides a template for extracting other forms of social and experiential meaning from reviews, opening new directions for NLP in tourism analytics and beyond.

Limitations

While the proposed pipeline provides a coherent and scalable representation of hostel social atmosphere, several limitations remain.

Construct ambiguity and absence of ground truth. Socialness is an inherently fuzzy and context-dependent construct, with no external ground-truth benchmark against which model predictions can be definitively validated. Many reviews lie near conceptual boundaries. Examples include, party-oriented hostels with limited mingling, quiet hostels with occasional communal encounters, or reviews emphasizing staff friendliness rather than guest interaction. These ambiguities introduce unavoidable subjectivity into the labeling process and make fine-grained distinctions difficult for both humans and models. As a result, evaluation at the hostel level necessarily relies on internal consistency and distributional structure (e.g., rarity patterns) rather than comparison to a trusted external reference. This limitation reflects the broader challenge of operationalizing socially constructed experiences from text rather than a deficiency of the proposed model.

Single-annotator supervision. The seed dataset used to train the cross-encoder was annotated by a single domain-familiar annotator, prioritizing internal consistency in the initial definition of socialness. While this choice reduces inter-annotator disagreement during early development, it limits the ability to quantify annotation reliability or capture alternative interpretations of borderline cases. Future work could substantially strengthen the empirical foundation of this task by incorporating multiple annotators, measuring inter-annotator agreement, and introducing verification or adjudication layers for ambiguous reviews.

Data coverage and translation effects. The corpus is drawn exclusively from Hostelworld and Booking.com, which constrains representativeness across the broader hostel ecosystem and reflects the preferences of travelers who choose to leave reviews. In addition, approximately one-third of the corpus relies on machine translation. Although manual inspection suggests that translations preserve the situational cues most relevant to socialness, subtle shifts in tone, pragmatics, or culturally specific expressions of social interaction may affect classification. These factors introduce systematic biases that future work could address through expanded platform coverage and multilingual model-

ing.

References

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542.

Hostelworld. 2024. State of solo travel report 2024. Accessed: 2025-12-13.

Hostelworld. 2025. State of solo travel report 2025. Accessed: 2025-12-13.

Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jay Krishnan, Biplab Bhattacharjee, Maheshwar Pratap, Janardan Krishna Yadav, and Moinak Maiti. 2024. Survival strategies for family-run homestays: Analyzing user reviews through text mining. *Data Science and Management*, 7(3):228–237.

Dong-Hyun Lee and 1 others. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Gang Li, Rob Law, Huy Quan Vu, Jia Rong, and Xinyuan Zhao. 2015. Identifying emerging hotel preferences using emerging pattern mining technique. *Tourism Management*, 46:311–321.

Ana Oliveira-Brochado and Carla Gameiro. 2013. Toward a better understanding of backpackers' motivations. *Tékhne*, 11(2):92–99.

Richard L. Oliver. 1980. A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*, 17(4):460–469.

Abraham Pizam and Ady Milman. 1993. Predicting satisfaction among first-time visitors to a destination by using the expectancy disconfirmation theory. *International Journal of Hospitality Management*, 12(2):197–209.

Yiqun Sun, Qiang Huang, Anthony K. H. Tung, and Jun Yu. 2025. Text embeddings should capture implicit semantics, not just surface meaning. *Preprint*, arXiv:2506.08354.

Zheng Xiang, Zvi Schwartz, John H. Gerdes, and Muzaffer Uysal. 2015. What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management*, 44:120–130.

A Annotation Rubric for Socialness

This appendix documents the review-level annotation rubric used to construct the human-labeled seed dataset. The goal of the rubric is to operationalize *socialness* as evidence of guest-to-guest interaction and communal participation, while explicitly distinguishing it from sentiment polarity, staff friendliness, facilities, and topic keywords.

A.1 Task Framing and Unit of Annotation

Unit of annotation. Each *review* is annotated independently. Labels are assigned based only on the textual content of the review, without using metadata (e.g., rating, reviewer nationality, hostel name).

Label set. Socialness is annotated as a binary label:

- `social = 1`: the review provides clear evidence of guest-to-guest interaction or communal participation.
- `social = 0`: the review does not provide such evidence, or explicitly indicates low sociality.

Core principle. A review is labeled `social = 1` only when it contains *direct or strongly implied* evidence that guests interact with other guests or participate in shared social activities. Mentions of facilities (e.g., bar, common room), staff friendliness, or general positivity are not sufficient on their own.

A.2 Decision Rules

A.2.1 Conditions for `social = 1`

A review receives `social = 1` if it satisfies at least one of the following:

- **Explicit sociality.** The review directly states that the hostel is social or that it is easy to meet other travelers (e.g., "very social," "easy to meet people," "met lots of people").
- **Guest interaction described.** The review describes guests meeting, talking, hanging out, cooking/eating together, playing games, or spending time together in shared spaces.
- **Hostel-organized activities with participation.** The review mentions naturally social group events (e.g., pub crawls, group dinners, drinking games, dance classes, or parties) (e.g., "pub crawl was super fun").
- **Lively atmosphere with interpersonal content.** The review describes a vibrant communal environment where social interaction is clearly occurring, even if overall sentiment is negative (e.g., complaints about noise while describing nightly social activity).

A.2.2 Conditions for `social = 0`

A review receives `social = 0` if any of the following apply:

- **Explicit statements of low sociality.** The review states that it is hard to meet people, that the hostel is not social, boring, impersonal, or lacks atmosphere.
- **Facilities mentioned without interaction.** The review mentions a bar, rooftop, common area, or events *without* describing guests interacting (e.g., "nice bar" with no interaction evidence).
- **Staff-only friendliness.** The review focuses on friendly staff/hosts or a welcoming reception but provides no evidence of guest-to-guest interaction.
- **Keyword-only or vague cues.** The review contains isolated or vague atmosphere descriptions (e.g., "good vibe") without evidence of interaction. These cases are labeled `social = 0` unless interpersonal participation is described.

A.3 Representative Examples

The examples below illustrate how socialness can be expressed implicitly and independently from sentiment polarity, staff friendliness, or topical keywords. Excerpts are lightly edited for brevity.

Example Type	Review Excerpt
High social, implicit (`social=1`)	"There are parties every night, and the music can be heard in the rooms."
High social, negative sentiment (`social=1`)	"Don't go if you want a clean place, it's only good for partying. There is something happening every day like drinking games and going to the clubs."
Party keywords but low interaction (`social=0`)	"While it is a party hostel I found it impersonal and hard to meet people."
Low social, positive sentiment (`social=0`)	"Amazing hostel! Very friendly staff and they gave great advice on things to do."

A.4 Labeling Decision Flow

Figure 5 summarizes the decision process used during annotation.

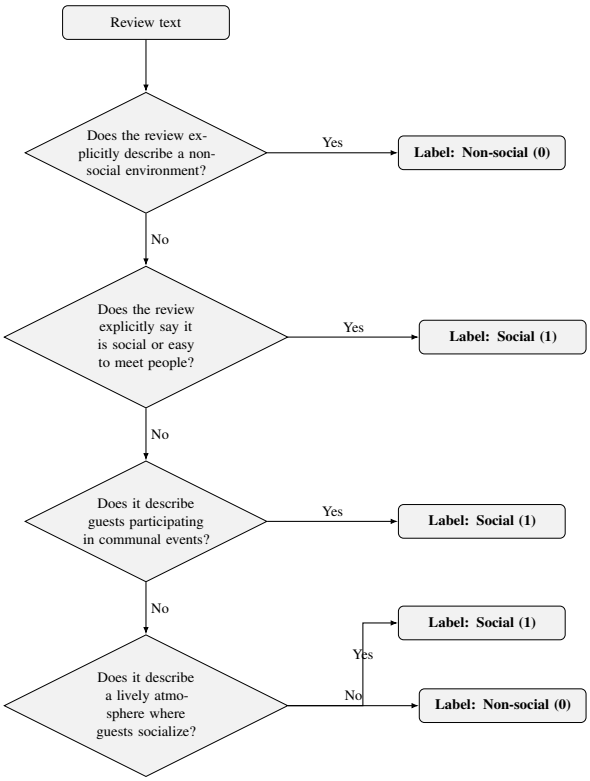

Figure 5: Decision flow for assigning binary socialness labels at the review level.

B Keyword-Guided Sampling Enrichment

This appendix documents the keyword lists used for sampling enrichment when selecting candidate reviews for annotation. Keywords were used solely to increase the probability of sampling reviews likely to contain high- or low-socialness cues and were not used as labeling rules, model features, or inference heuristics.

B.1 Party and Calm Token Lists

Party-related strings.

> party hostel; pub crawl; bar crawl; club crawl; beer pong; drinking games; DJ; nightlife; shots; happy hour; karaoke; music until late; wild party; parties every night;

Calm or low-sociality strings.

> quiet; peaceful; calm; good sleep; sleep well; not a party hostel; relaxing; chill; no noise; quiet hours; silent