Tourism Mining: Aligning Sentiment & Topic Representations of Sentence Transformer Embeddings for Tourism Opinion Mining

Anonymous ACL submission

Abstract

We introduce **TourCSE**, a Sentence Transformer tailored for tourism opinion mining, leveraging training data from topic modeling with negative filtering using GISTEmbed (Solatorio, 2024). TourCSE enables identification of key aspects in large-scale exploratory analyses.

Beyond improved results, we investigate the sources of these improvements through ablation studies, including Low-Rank Adaptation, dropout augmentation, sentiment sampling, and negative filtering. Our experiments demonstrate the necessity of negative filtering, in combination with either topic modeling or sentiment sampling, to substantially improve topic–sentiment alignment. We also compare TourCSE with other available models in terms of both visualization and retrieval performance, confirming the value of domain-specific models. Finally, we advocate reconsidering the NEUTRAL sentiment in current benchmarks.

1 Introduction

When social tourism researchers seek to analyze customer opinions, a fundamental question arises: *Which dimensions of the hospitality experience are most relevant to extract?* The challenge is that in most cases social researchers do not have predefined annotation schemes, as their ground truth depends on their individual expertise and research objectives (Schofield et al., 2025). This limitation reduces the applicability of supervised learning approaches (Grimmer et al., 2021). In contrast, unsupervised approaches, although less accurate, offer a compelling alternative by allowing the labeling scheme to emerge directly from the data. Yet, despite their potential, systematic reviews of tourism opinion mining from both computer science and tourism fields underline the lack of unsupervised methods for large-scale analyses involving millions of customer reviews (Ameur et al., 2023; Mehraliyev et al., 2022).

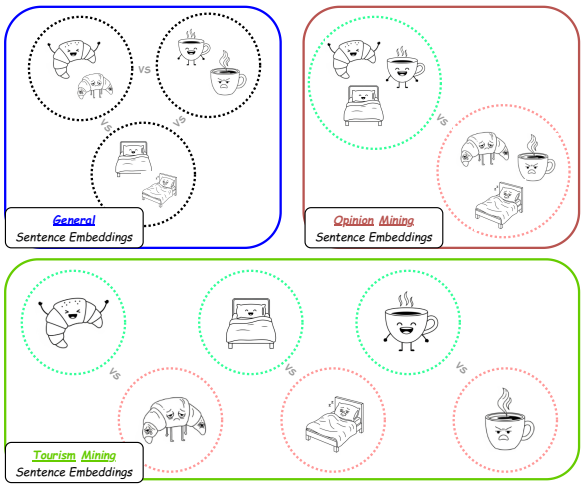


Figure 1: We categorize sentence embeddings into three groups: GENERAL, OPINION MINING, and TOURISM MINING. GENERAL embeddings are typically biased toward the sentence subject (Nikolaev and Padó, 2023; Ghafouri et al., 2024), whereas OPINION MINING embeddings tend to discard it (Kim et al., 2024). To overcome these limitations, we introduce TourCSE, which reorganizes the embedding space to capture both topic and sentiment specific to tourism reviews.

In our study, we focus on Sentence Transformers, motivated by their growing potential for opinion mining (Ghafouri et al., 2024). A Sentence Transformers can be used in two main ways: (1) leveraging their embeddings to perform pairwise comparisons between N sentences and T topics with linear complexity $O(N + T)$, offering significant scalability advantages for large datasets; or (2) navigating the embedding space organized through thematic grouping (see Figure 6), improving the comprehension of patterns in the data. Sentence Transformers should be regarded as preliminary and complementary approaches to more computationally intensive methods (see Section 6.2). They are particularly valuable for social scientists who aim to run models on personal computers, in contrast to large

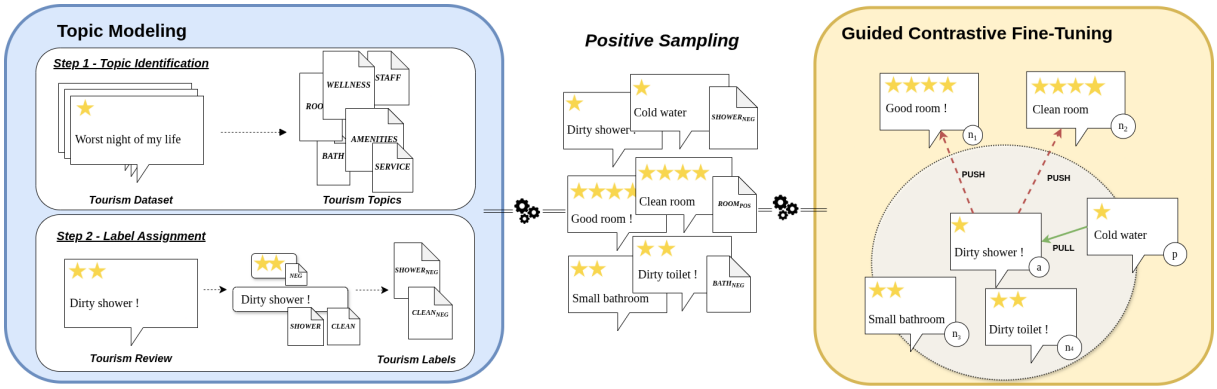


Figure 2: **TourCSE.** We mine positive sampling using Topic Modeling (**SWETT**) by selecting sentences that share the same sentiment and topic (a, p) through our annotation process. Then, we treat other sentences in the batch (n_1, n_2, n_3, n_4) as negatives. In this case, n_3 and n_4 share the same topic (BATHROOM) and sentiment (NEGATIVE) as anchor a but differ in their specific aspects (BATHROOM & TOILET). To address negatives highly similar to a like n_3 and n_4, we introduce **Guided Contrastive Fine-Tuning** (Solatorio, 2024). This method leverages a guidance model to ignore negative pairs (a, n_i) with higher similarity to the current positive pair (a, p) in the batch training.

language models (LLMs) or more complex neural architectures (Hoyle et al., 2022, 2021; Li et al., 2025). Nonetheless, as shown in Figure 5, embeddings from general-purpose or opinion-focused models remain insufficiently specialized to fully capture both **tourism-specific topics** and corresponding **sentiments**. Therefore, we demonstrate that using simple data-driven approaches, it is possible to reorganize the embeddings space, thus improving alignment between topics and sentiment (Section 3). Our contributions are summarized as follows:[1]

- An overall guideline for sentiment learning based on our ablation study (Section 6.1): using topic-model pseudo-labeling with GISTEmbed negative filtering (Solatorio, 2024) effectively aligns domain-specific topics and sentiments, while topic-model can also be replaced by simple sentiment sampling.
- Demonstrating the relevance of **TourCSE** (Section 5.2 and Figure 5), we pave the way for further development on domain-specific opinion-mining Sentence Transformers.
- We position our work in light of the Venkit et al. (2023) critical sentiment analysis survey, adding a discussion in sentiment labeling (Section 3.2) and clearly defining the targeted application scope to avoid unintended biases. Also, we underline some annotation inconsistencies with the NEUTRAL sentiment supported by linguistic research (Kamoen et al., 2015) and Liu (2020) survey.

[1]The code is available at `https://anonymous.4open.science/r/TourCSE-8372/README.md`

2 Related Work

Opinion Mining Sentence Embeddings As illustrated in Figure 5, general-purpose sentence embeddings fail to encode sentiment information effectively due to their bias toward nominal subject (Nikolaev and Padó, 2023). To address this limitation, `SentiCSE` (Kim et al., 2024) fine-tunes a RoBERTa model using contrastive learning with positive and negative sentence pairs whose sentiment is defined via SentiWordNet (Baccianella et al., 2010). `StanceSBERT` (Ghafouri et al., 2024), in turn, fine-tunes `all-mpnet-base-v2` by combining contrastive and triplet loss, applying LoRA to mitigate catastrophic forgetting, mining debate-forum data to construct positive pairs, and designing a custom pipeline for negative-sample filtering. **However**, in line with our recommendations (Section 6.1), negative filtering can be automated using GISTEmbed (Solatorio, 2024). Moreover, applying LoRA does not mitigate catastrophic forgetting. Additionally, we explicitly define the domain scope of our model, rather than claiming broad, domain-general opinion-mining performance.

Social Media Analysis Beyond model development, several studies demonstrate the utility of Sentence Transformers for opinion mining. Representative examples using pairwise comparison include Introne (2023), who fine-tune `all-mpnet-base-v2` to quantify opinions on climate change, and Sarkar et al. (2025), who

combine LLMs to decompose sentences with `all-mpnet-base-v2` embeddings to score political attitudes in online communities. **In the tourism domain**, Sánchez-Franco and Rey-Moreno (2022) explore `all-MiniLM-L6-v2` embeddings through BERTopic to uncover travelers' motivational patterns underlying tourism destinations. We evaluate the performance of pairwise comparisons in Section 5 and illustrate the advantages of TourCSE embeddings for visualization in Figure 5, with a practical example shown in Figure 6.

3 Proposed Approach : TourCSE

TourCSE (**Tour**ism **C**ontrastive **S**entence **E**mbeddings) is our Sentence Transformer designed for tourism opinion mining. Like other Sentence Transformers, it requires training data pairs: some that share the same topic and sentiment (positive samples) and others that differ (negative samples). For positive sampling, we introduce a simple topic modeling approach, SWETT (**S**imple **W**ord **E**mbeddings-based **T**ourism **T**opic Model). For negative samples, we demonstrate that simple negative filtering using `GISTEmbed` (Solatorio, 2024) significantly improves the alignment between domain topics and sentiment representations. The general pipeline is illustrated in Figure 2 and is performed at the sentence level.

3.1 Topic Modeling : SWETT

Definition Although a topic is a nebulous concept, it is often represented as a group of keywords that share the same "*meaning*" (Doogan and Buntine, 2021). In opinion mining, this corresponds to explicit aspects which directly refer to the same concept (e.g., bed, mattress, pillow) (Liu, 2020). A topic may also represent an implicit aspect (e.g., "*Walking to the museums takes less than 15 minutes*" refers to LOCATION, though not explicitly mentioned) (Wang et al., 2023a). We rely on explicit notions for identification, and on both implicit and explicit for assignment and extraction.

Topic Identification First, we employ part-of-speech tagging to extract the most frequent nouns, as these correspond to aspects evaluated by customers (Hu and Liu, 2004). Then, to capture semantic relationships between aspects and extract thematic information, we train a `Word2Vec` model on tourism reviews (Mikolov et al., 2013). The visualization of aspect embeddings in Figure 3 reveals a natural topic clustering tendency.

Figure 3: UMAP visualization of Word2Vec representations of sample aspects from HotelRec (Antognini and Faltings, 2020). Aspects related to BEDDING in red.

Topic Assignment To extract both implicit and explicit aspects, we implement CAt (Tulkens and van Cranenburgh, 2020). An simple method that relies exclusively on tourism domain `Word2Vec` embeddings and an extracted set of aspects, which have already been introduced in the previous Section 3.1. As target, we use cluster centroids derived from identified topics (Figure 3).

3.2 Sentiment Assignment

The Sentiment Problem Venkit et al. (2023) and Rebora (2023) highlight a lack of interdisciplinary reflection in sentiment analysis research. For example, Kamoen et al. (2015) demonstrate in their human studies that indirect positive expressions (e.g, *not bad*) are perceived as less positive than direct ones. More pragmatically, a sentiment is inherently context-dependent, yet general models exhibit a bias toward NEGATIVE or NEUTRAL. For instance, a review stating *"Hotel near to Eiffel Tower"* may be POSITIVE in tourism contexts, but models trained on social media data could classify it as NEUTRAL.[2] Although LLMs perform well on explicit aspect-based sentiment extraction, they struggle with subjective sentiment or implicit aspects (Zhang et al., 2024). Moreover, Voutsa et al. (2025), indicates a bias toward NEUTRAL in LLMs that deserves further investigation. We provide additional errors examples in Appendix C.1.

Rethinking Sentiment Because, our objective is to remain as faithful as possible to the sentiment expressed by customers while minimizing biases introduced by subjective interpretation. We adopt star ratings as a coarse-grained proxy for sentiment

[2] cardiffnlp/twitter-roberta-base-sentiment-latest

polarity. We also follow the ethical-sheet guidelines of Venkit et al. (2023) (see Appendix C).

3.3 Guided Contrastive Fine-Tuning

Training Objective As only positive pairs are available at this stage, we follow state-of-the-art practice (Wang et al., 2024) and employ the InfoNCE loss (van den Oord et al., 2019), with negatives randomly sampled from the batch:

$$\mathcal{L} = \frac{e^{cosim(q_i,p_i)/\tau}}{e^{cosim(q_i,p_i)/\tau} + \sum_{j\in B} e^{cosim(q_i,n_j)/\tau}} \quad (1)$$

The objective is to minimize the distance between the embeddings of positive pairs (q_i, p_i), and maximize the distance between negative pairs (q_i, n_j), selected from the batch sample B, using cosine similarity ($cosim$). The parameter τ controls the temperature, allowing adjustment of the scale of cosine similarity.

Negative Filtering The main issue with randomly sampling negatives from the batch is the introduction of false negatives. For instance, a review describing a *"small bathroom"* might be incorrectly paired as negative with a review containing a *"dirty shower"*. Although these reviews differ semantically, both express negative sentiment about the same topic BATHROOM. To address this, we adopt `GISTEmbed` (Solatorio, 2024).[3] This method refines the fine-tuning process by discarding negative samples that exhibit higher similarity to the anchor and its corresponding positive pair. All discarding occurs during training by setting the similarity between the anchor and discarded negative to $-\infty$, effectively reducing their contribution to the cross-entropy loss to zero. The only requirement is a *guided* pre-trained Sentence Transformer to compute these similarities.

4 Experimental Settings

4.1 Implementation Details

Topic Modeling We tokenize text, remove stopwords, and extract the 1000 most frequent common nouns as explicit aspects using NLTK (Bird and Loper, 2004). Word2Vec models are trained with Skip-Gram via Gensim (Řehůřek and Sojka, 2010). Topics are extracted via hierarchical clustering on aspect-vector cosine similarities, using the desired number of clusters as the threshold. For CAt, we made some modification to the implementation of Tulkens and van Cranenburgh (2020) by using the centroid vectors of aspect clusters as targets, and developing a slightly more stable attention.[4]

Sentiment Assignment Sentiment scores are assigned based on the review rating of each extracted sentence (1–5). Although simplistic, this method performs surprisingly well in practice.

Positive Sampling We generate 300,000 positive pairs by coupling each sentence with a random sentence sharing at least one topic and sentiment.

Negative Sampling and Hyperparameters All models are trained with a batch size of 128 (1 positive + 127 negative pairs per step) for 3 epochs using AdamW, a learning rate of 5×10^{-5}, no warm-up, and a temperature $\tau = 20$.

4.2 Model Selection

We adopt the same fine-tuning setup (base model and LoRA configuration) as Ghafouri et al. (2024), which introduces StanceSBERT, a opinion-mining sentence transformers. Using GISTEmbed loss and LoRA on a single H100 GPU, training completes in 1–2 hours.

Model We use `all-mpnet-base-v2` as the base model for domain adaptation. It provides a strong balance between model size and performance on MTEB (Muennighoff et al., 2023) and had been trained on a large-scale corpus of approximately 1B data pairs. For negative filtering, we employ a copy of the selected base model, `all-mpnet-base-v2` like Ghafouri et al. (2024).

Low-Rank Adaptation Ghafouri et al. (2024) suggests that integrating adapter modules mitigate catastrophic forgetting. We apply LoRA (Hu et al., 2022) with a rank of 32 to assess its effectiveness.

4.3 Datasets

We focus on the hotel and restaurant domains due to their importance in the tourism sector. For the training process in **TourCSE**, we select large publicly available datasets avoiding data contamination caused by possible overlap between the training and test sets. For evaluation, we use English-annotated datasets that follow established ABSA terminology (Zhang et al., 2023).

[3] See sbert.net for documentation.

[4] An introduction to the original motivation, evaluation protocol, and example code for our reimplementation of CAt is provided in Appendix B.

Training Dataset We select two publicly available TripAdvisor datasets, each corresponding to a specific domain:

- **SixTripAdvisorDatasets** for the restaurant domain (Botana et al., 2022)[5]
- **HotelRec** for the hotel domain (Antognini and Faltings, 2020)[6]

Both datasets include sentiment metadata derived from star ratings ranging from 1 to 5. To ensure a balanced distribution, we sampled 100,000 English-language reviews per dataset, with equal representation of 20,000 reviews for each rating.

Evaluation Dataset We select three annotated datasets: Rest14 (Pontiki et al., 2014), Rest16 (Pontiki et al., 2016), and HotelOATS (Chebolu et al., 2024). These datasets provide sentence-level topics with their associated sentiments. We retain the full train and test splits from each dataset.

Issue with Neutral Annotation We observe inconsistencies in the annotation of NEUTRAL sentiment, providing some examples from Rest16. As noted in linguistic studies, indirect negative expressions such as *"Food was okay"* or *"Service was decent"*, currently labeled as NEUTRAL, may in fact convey a negative opinion expressed politely (Kamoen et al., 2015). Similarly, a sentence like *"Plan on waiting 30–70 minutes"* was also annotated as NEUTRAL. However, this example should be categorized as NOT OPINIONATED, since it does not express any sentiment (Liu, 2020). They also misannotated unambiguous NEUTRAL sentences, such as "*Food was mediocre*". To better align the dataset with our research goals, we removed all sentences with NEUTRAL aspect from the evaluation set in our main experiments.[7]

4.4 Detail of Downstream Tasks

Retrieval Setup We evaluated our embeddings by retrieving topic-sentiment pairs using a single-query approach, simulating real-world scenarios (Wang et al., 2023a). The retrieved topics are as follows:

- **Restaurant**: FOOD, SERVICE, DRINKS, LOCATION, AMBIANCE, PRICE
- **Hotel:** ROOMS, AMENITIES, FACILITIES, SERVICE, LOCATION, FOOD & DRINKS

To incorporate sentiment, we use the sentiment words *excellent* for positive polarity and *terrible* for negative polarity. Queries are constructed in the format "`{sentiment} {topic}`". For example, "*excellent food*" retrieves positive opinions about food, while "*terrible food*" retrieves negative ones. Retrieval is performed based on cosine similarity between the query and sentence embeddings.

Evaluation Metric Following Booking.com (Wang et al., 2023a), we adopt the macro-averaged precision score from Scikit-Learn as the evaluation metric (Pedregosa et al., 2011).

5 Results

5.1 Ablation Study

Number of Topics Results are presented in Figure 4. As positive sampling generation is based on Topic Modeling, we aim to evaluate the impact of topic number, denoted as K. Since there is no universally optimal method to determine the ideal number of topics, we test six different values: 10, 50, 100, 150, 200, and 250. The fine-tuning is performed with Guided Contrastive Fine-Tuning and LoRA. Our analysis shows that **TourCSE consistently outperforms the base model across all values of** K. However, increasing the number of topics does not necessarily lead to better performance. Instead, for HotelOATS and Rest16, the optimal K is 10, beyond which performance declines and then stabilizes.

Data Sampling and LoRA The ablation study presented in Table 1. We aim to evaluate the impact of both positive, negative filtering and LoRA. Additionally, we compare the effect of our positive sampling, refereed as Topic Modeling, with two others possible strategies:

- **Dropout**: Following `SimCSE` (Gao et al., 2021), this strategy generates positive pairs by using the same sentence twice, applying dropout within layers as a form of data augmentation.
- **Sentiment**: Following `SentiCSE` (Kim et al., 2024), this approach selects positive pairs based solely on sentiment, ignoring specific aspect mentioned in the sentence.

[5]https://zenodo.org/records/6583422
[6]https://github.com/Diego999/HotelRec
[7]Discard NEUTRAL is a common practice, as shown in previous work (Nguyen et al., 2023) but it has not been explicitly justified. Additional results are provided in Appendix A.

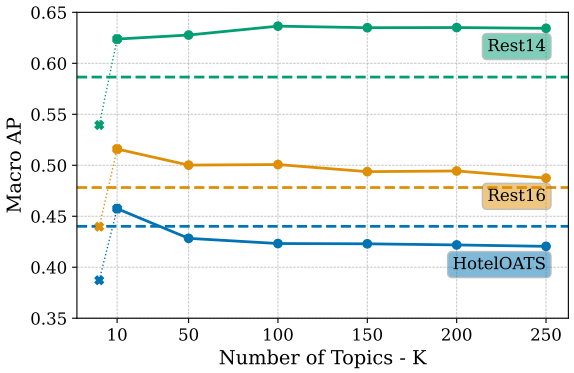


Figure 4: Macro Average Precision scores for different values of K during positive pair generation. Fine-tuning is performed using Guided Contrastive Fine-Tuning and LoRA. The cross markers represent the score of the base model `all-mpnet-base-v2`. The horizontal dashed line indicates the score obtained with randomly sampled positive pairs sharing the same sentiment, as discussed in Section 5.1. Overall, topic mining achieves better performance than both the base model and, with the right choice of K, the sentiment-only strategy.

Topic modeling with LoRA, GCFT, or both significantly outperforms all other configurations. The inclusion of GCFT has a more noticeable positive impact on Sentiment sampling, **positioning it as a strong alternative when only sentiment information is available**. In contrast, Dropout sampling consistently underperforms. The small performance gap between models without LoRA in Topic Modeling suggests that negative filtering plays a greater role.

	HotelOATS	Rest16	Rest14	Avg.
all-mpnet-base-v2	38.72	43.98	53.94	45.55
Dropout	20.25	24.47	31.78	25.50
Δ w/o LoRA	-2.34	-4.08	-3.58	-3.33
Sentiment	44.01	47.82	58.65	50.16
Δ w/o GCFT	-21.87	-23.98	-26.17	-24.01
Δ w/o LoRA	-0.51	-0.70	-1.33	-0.85
Δ w/o LoRA & GCFT	-21.99	-22.55	-23.23	-22.59
Topic Modeling $k = 10$	**45.75**	51.59	62.38	53.24
Δ w/o GCFT	-4.56	-5.30	**+0.86**	-3.00
Δ w/o LoRA	-0.13	**+0.78**	-0.16	**+0.16**
Δ w/o LoRA & GCFT	-19.11	-9.91	-10.93	-13.32

Table 1: Ablation study of different positive and negative sampling strategies and the impact of LoRA. Overall, topic modeling sampling yields the best results. However, sentiment sampling with GCFT remains a strong alternative.

5.2 Model Comparison

We compare **TourCSE** with other Sentence Transformers, including a variety of architectures, training data, and objective loss:

	HotelOATS	Rest16	Rest14	Avg.
General Sentence Embeddings				
all-MiniLM-L6-v2	30.73	34.46	45.38	36.86
all-MiniLM-L12-v2	33.61	36.20	47.71	39.17
all-mpnet-base-v2	38.72	43.98	53.94	45.55
E5-small-v2	34.26	37.34	54.14	41.91
E5-base-v2	35.43	40.96	56.57	44.32
E5-large-v2	34.29	38.32	53.52	42.04
E5-Mistral-7B-Instruct	31.87	36.88	47.48	38.74
GTE-base-v1.5	36.75	40.13	56.50	44.46
GTE-large-v1.5	39.78	44.01	57.69	47.16
GTE-ModernBERT	42.18	49.48	67.32	52.99
ST5-XXL	43.70	47.20	67.20	52.70
GTR-T5-XXL	31.30	40.16	50.13	40.53
Opinion Mining Sentence Embeddings				
SentiCSE	31.27	34.29	45.37	36.98
StanceSBERT	37.00	37.53	49.65	41.39
Tourism Mining Sentence Embeddings				
$\text{Word2Vec}_{\text{CBOW}}$	31.58	31.73	40.73	34.68
$\text{Word2Vec}_{\text{SG}}$	29.78	32.86	41.26	34.63
$\text{FastText}_{\text{CBOW}}$	30.23	28.93	38.27	32.48
$\text{FastText}_{\text{SG}}$	28.19	32.10	39.63	33.31
TourBERT	18.68	20.72	30.10	19.26
TourMPNet	11.78	17.47	23.20	17.48
TourCSE $K = 10$	45.75	51.59	62.38	53.24
GTE-TourCSE $K = 10$	**46.80**	**55.34**	**73.80**	**58.65**

Table 2: Performance comparison of sentence embedding models across datasets using Macro Average Precision. Best overall results are shown in **bold**, and the best within each category are underlined. Given the strong performance of `GTE-ModernBERT`, we also fine-tune it (`GTE-TourCSE`). Overall, **TourCSE** outperforms both general-purpose and opinion-mining models.

- **General Sentence Embeddings** trained on large-scale, multi-domain data, including `all-MiniLM-L6-v2`, `all-MiniLM-L12-v2`, `all-mpnet-base-v2`, the `E5` series (Microsoft) (Wang et al., 2024, 2023b), `GTE` (Alibaba) (Li et al., 2023), and `T5-XXL` (Google) (Ni et al., 2021b,a).
- **Opinion Mining Sentence Embeddings**, including `SentiCSE` (Kim et al., 2024) and `StanceSBERT` (Ghafouri et al., 2024).
- **Tourism Mining Sentence Embeddings**, including `Word2Vec` (Mikolov et al., 2013), `FastText` (Bojanowski et al., 2017), and `TourBERT` (Arefeva and Egger, 2022). We also fine-tuned `all-mpnet-base-v2` using MLM (`TourMPNet`).[8]

[8] `TourBERT` is a BERT model fine-tuned on tourism data using MLM objective. `Word2Vec`, `FastText` and `TourMPNet` were trained on the same dataset as TourCSE.

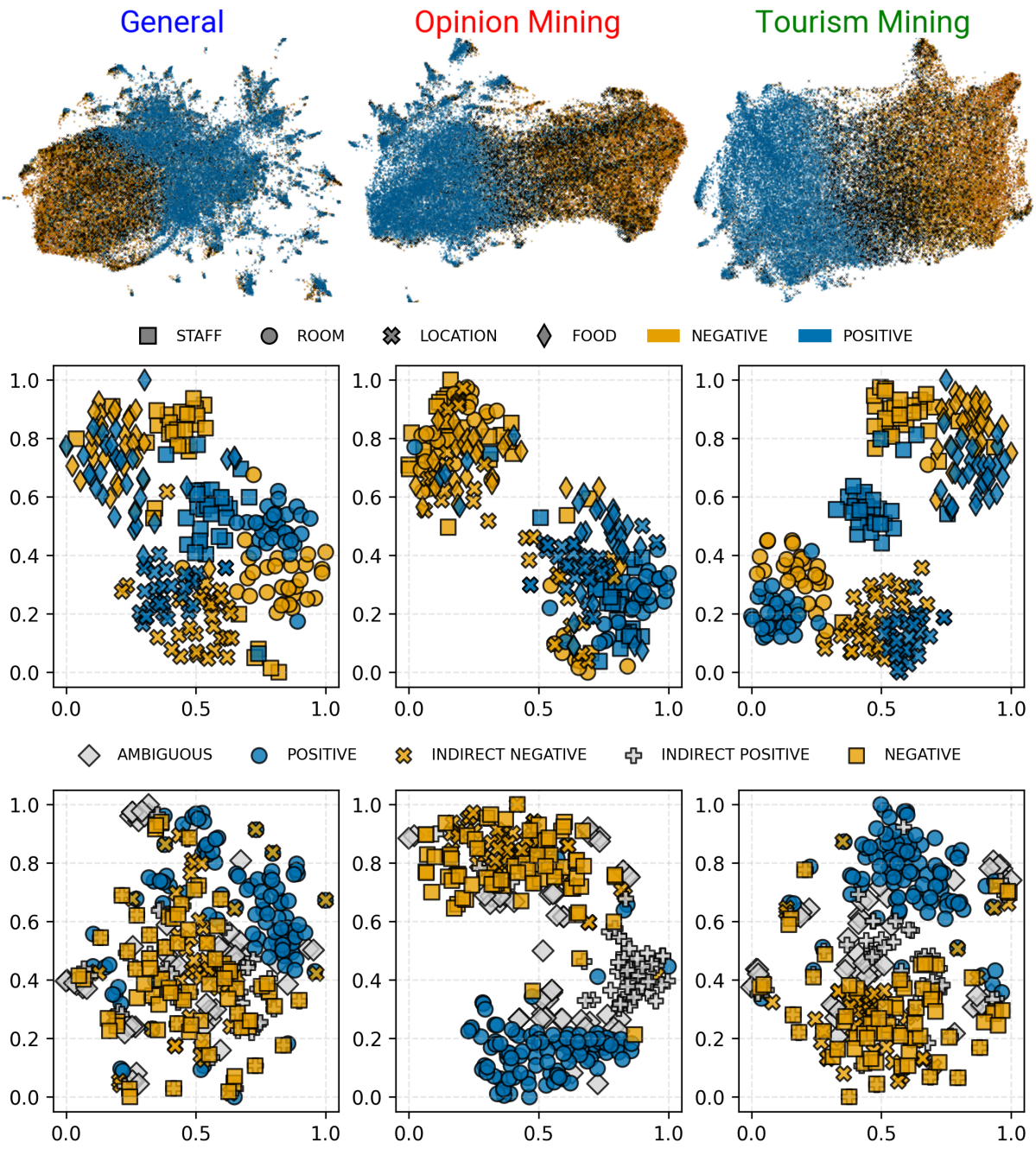


Figure 5: TSNE visualization of different `all-mpnet-base-v2` embeddings. GENERAL: `all-mpnet-base-v2`; OPINION MINING: **StanceBERT**, a LoRA adapter fine-tuned on debate forums (Ghafouri et al., 2024); TOURISM MINING: **TourCSE**, **our** LoRA adapter fine-tuned on HotelRec ($K = 250$). We report clustering metrics Table 3.
TOP: Reviews distribution from HotelRec: POSITIVE (★★★★★, ★★★★☆), NEGATIVE (★★☆☆☆, ★☆☆☆☆), and NEUTRAL (★★★☆☆).
MIDDLE: Visualization of 4 majors distinct topics and their associated sentiment: LOCATION, STAFF, ROOM and FOOD sampled from HotelOATS, Rest16 and Rest14.
BOTTOM: We build pseudo-sentences: "room is ...", where "..." is an opinion. We examine how indirect positives (e.g.,"*not bad*"), indirect negatives (e.g.,"*not great*"), and ambiguous sentiments (different sentiments separated by "*but also*", e.g.,"*good but also noisy*") are positioned relative to explicit positive and negative opinions. As opinion, we use the 100 most frequent adjectives in HotelRec. Sentiment orientation is defined according to the Opinion Lexicon (Liu, 2004). Indirect positives are colored like ambiguous because they are less positive that direct statement (Kamoen et al., 2015).
Takeway: GENERAL already have a reasonable representation of sentiment. Nevertheless, the **top** and **bottom** visualizations indicate that TOURISM MINING achieves better separation of both positive and negative sentiments than GENERAL. In the **middle**, TOURISM MINING topic and sentiment representation is slightly better than GENERAL, whereas OPINION MINING appears to overfit to sentiment.

	Sil. ↑		CH ↑		DB ↓	
	S.	S.T	S.	S.T	S	S.T
G.	0.08	0.05	19.64	9.40	3.30	3.70
O.M	**0.24**	0.01	**79.13**	13.88	**1.63**	3.69
T.M	0.19	**0.09**	56.55	**17.05**	1.92	**2.76**

Table 3: Clustering metrics (Figure 5). "S" refers to bottom visualization (clusters indicated by colors; ambiguous/indirect positive discarded) "S.T" refers to middle visualization (clusters indicated by shape and color). Sil. = Silhouette, CH = Calinski-Harabasz, DB = Davies-Bouldin. ↑ = higher is better, ↓ = lower is better. While Opinion Mining models provide a clear distinction in their representation of sentiment, Tourism Mining achieve a better balanced representation between sentiment and topic.

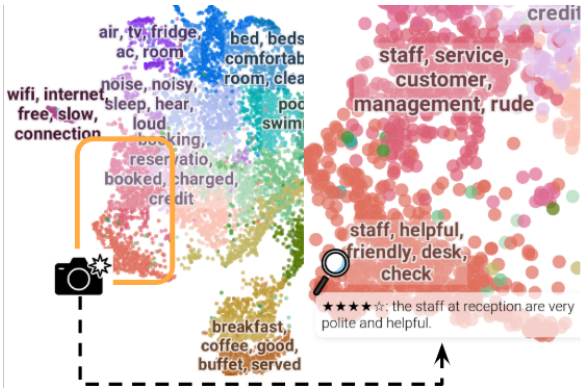


Figure 6: Snapshot of TourCSE embedding visualization using BERTopic (Grootendorst, 2022) (See Appendix Figure 15c). Sentences related to STAFF cluster together in the embedding space **and** are separated according to sentiment.

The results in Table 2 show that TourCSE consistently outperforms models with comparable backbones, such as `E5` (BERT), `GTE` (Transformers++), `SentiCSE` (RoBERTa), and `StanceSBERT` (MPNet). However, `GTE-ModernBERT` demonstrates comparable performance to TourCSE and even surpasses it on Rest14. Therefore, we fine-tune it using the same procedure (LoRA, $k = 10$, and a copy of `GTE-ModernBERT` as guided model) which we refer to as `GTE-TourCSE`. As it achieves the best performance across all datasets, **we confirm the effectiveness of our pipeline**. Conversely, `TourBERT` and `TourMPNet` performs worse than `Word2Vec`, highlighting the limitations of domain adaptation through MLM for learning sentence representations. Despite that Arefeva and Egger (2022) reported improved sentence representations for tourism reviews. These findings underscore the **importance of incorporating domain-specific topics and sentiment into contrastive objective training**.

6 Discussion

6.1 Implementation Guideline

For positive sampling, practitioners can use their own topic models (e.g., TopicGPT (Pham et al., 2024)), but they should test multiple values of K or manually refine the topics. Alternatively, it is possible to rely solely on sentiment-based sampling. Importantly, false negatives should be discarded using negative filtering from GISTEmbed (Solatorio, 2024) to improve sentiment-related representations without degrading topic alignment. As shown in our ablation study (Section 5.1), LoRA does not mitigate catastrophic forgetting, as evidenced by the negligible performance gains observed.

6.2 Further Research

As Sentence Transformers remain a component within a broader system, we suggest the following additional research directions: (1) mining training data with TourCSE to develop more robust ABSA models (de Souza P. Moreira et al., 2025; Nguyen et al., 2023); and (2) leveraging TourCSE as a retrieval component within retrieval-augmented generation pipelines for opinion summarization (Nayeem and Rafiei, 2025). Finally, we advocate for re-evaluating the NEUTRAL category to better benchmark supervised SLMs and LLMs (Kenyon-Dean et al., 2018; Rebora, 2023; Venkit et al., 2023).

7 Conclusion

We introduce TourCSE, a Sentence Transformer model specifically designed for opinion mining in the tourism sector, enabling scalable analysis of millions of reviews. We empirically demonstrate the effectiveness of using topic models as a pseudo-labeling approach, combined with GISTEmbed, to align domain-specific topics and sentiment in embeddings. Our results highlight the limitations of general-purpose and existing opinion-mining sentence embedding models in retrieving and clustering domain-specific topics and sentiment. These findings underscore the need for explicitly domain-targeted approaches. Direct applications include assisting experts in corpus exploration by integrating TourCSE with BERTopic.

Limitations

As a limitation, we note that TourCSE was developed exclusively for tourism opinion mining, and its evaluation focuses on a limited set of high-level aspects of hospitality reviews. A more fine-grained evaluation would also be possible, for instance by distinguishing room-related aspects from bathroom-related ones, or separating views from location. Moreover, the evaluation of topic models, and more broadly unsupervised methods, remains an active area of research (Li et al., 2024, 2025; Hoyle et al., 2025).

The Domain Scope of SWETT Recent work has highlighted the limitations of general LLM-based topic models (Pham et al., 2024; Lam et al., 2024), particularly their difficulty in extracting informative topics when the document collection originates from a single, specific domain (Li et al., 2025). To address this limitation, we developed **SWETT** specifically for tourism reviews. Practitioners who wish to align sentiment embeddings in their own domain may rely on existing topic models, such as LDA (Blei et al., 2003) or TopicGPT (Pham et al., 2024), although better results can be expected when the topic model is adapted to the target domain. See Laureate et al. (2023); Hoyle et al. (2025) for a discussion of the research agenda on topic model development.

The K Parameters Issues with traditional automated evaluation metrics helping select K are now well documented (Hoyle et al., 2021; Chang et al., 2009; Doogan and Buntine, 2021; Shadrova, 2021). Although recent methods such as TopicGPT (Pham et al., 2024) use LLMs to automatically set K, the resulting topics still require expert refinement (Choi et al., 2024; Schofield et al., 2025), and they are not necessarily more informative than those extracted by traditional methods such as LDA (Li et al., 2025). To assist in selecting K, Stammbach et al. (2023) explored the use of LLMs, reporting promising results when the LLM was prompted with a relevant hypothesis (e.g., "*Which dimensions of the hospitality experience are most relevant to extract?*"). They also acknowledge that "*The optimal number of topics is a vague concept, dependent on a practitioner's goals and the data under study*"

The Ground Truth Subjectivity As noted by Grimmer et al. (2022), the primary goal of unsupervised methods is not to reproduce an existing ground truth, but to enable new interpretations through exploratory analysis. In our study, we evaluated only a limited set of predefined topics. By contrast, Booking.com uses a much more fine-grained taxonomy with 239 distinct topics to train a Sentence Transformer in a supervised manner (Wang et al., 2023a). Nevertheless, unsupervised Sentence Transformers should not be used to extract preconceived topics envisioned by experts. Rather, their goal is to explore embeddings through visualization (as shown in Figure 6) to assist experts in defining topic sets, which can then guide the development of supervised methods or LLM-based analyses.

Ethical Considerations

Considering the work of Venkit et al. (2023), we respond to the ethics sheet for sentiment analysis systems, provided in Appendix C. We acknowledge potential biases in training data, which may influence the outputs of our system, and we strive to mitigate them through **careful dataset selection** and a **well-defined application scope**. The datasets used in our study may contain identifiable information; however, we are not responsible for such content, as the data is publicly available and originates from previous research. Generative AI was used solely for rephrasing, improving grammar, and providing coding assistance.

Acknowledgments

References

Asma Ameur, Sana Hamdi, and Sadok Ben Yahia. 2023. Sentiment analysis for hotel reviews: A systematic literature review. *ACM Comput. Surv.*, 56(2).

Diego Antognini and Boi Faltings. 2020. HotelRec: a novel very large-scale hotel recommendation dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4917–4923, Marseille, France. European Language Resources Association.

Erik Arakelyan, Zhaoqi Liu, and Isabelle Augenstein. 2024. Semantic sensitivities and inconsistent predictions: Measuring the fragility of NLI models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–444, St. Julian's, Malta. Association for Computational Linguistics.

Veronika Arefeva and Roman Egger. 2022. When bert started traveling: Tourbert—a natural language processing model for the travel industry. *Digital*, 2(4):546–559.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Iñigo López-Riobóo Botana, Verónica Bolón-Canedo, Bertha Guijarro-Berdiñas, and Amparo Alonso-Betanzos. 2022. Explain and conquer: Personalised text-based reviews to achieve transparency. *Preprint*, arXiv:2205.01759.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2024. OATS: A challenge dataset for opinion aspect target sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12336–12347, Torino, Italia. ELRA and ICCL.

Alexander Choi, Syeda Sabrina Akter, J.p. Singh, and Antonios Anastasopoulos. 2024. The LLM effect: Are humans truly using LLMs, or are they being influenced by them instead? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22032–22054, Miami, Florida, USA. Association for Computational Linguistics.

Gabriel de Souza P. Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2025. Nv-retriever: Improving text embedding models with effective hard-negative mining. *Preprint*, arXiv:2407.15831.

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? re-evaluating semantic interpretability measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848, Online.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *WebDB*, volume 9, pages 1–6.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aitor García-Pablos, Montse Cuadros, and German Rigau. 2018. W2vlda: almost unsupervised system for aspect based sentiment analysis. *Expert Systems with Applications*, 91:127–137.

Vahid Ghafouri, Jose Such, Guillermo Suarez-Tangil, and 1 others. 2024. I love pineapple on pizza!= i hate pineapple on pizza: Stance-aware sentence transformers for opinion mining. In *Empirical Methods in Natural Language Processing*.

Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2021. Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24(1):395–419.

Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2022. *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *Preprint*, arXiv:2203.05794.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada. Association for Computational Linguistics.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc.

Alexander Miserlis Hoyle, Lorena Calvo-Bartolomé, Jordan Lee Boyd-Graber, and Philip Resnik. 2025. ProxAnn: Use-oriented evaluations of topic models and document clustering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15872–15897, Vienna, Austria. Association for Computational Linguistics.

Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2022. Are neural topic models

broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344, Abu Dhabi, United Arab Emirates.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

C. J. Hutto and Eric E. Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, Ann Arbor, MI.

Joshua Introne. 2023. Measuring belief dynamics on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 387–398.

Naomi Kamoen, Maria B.J. Mos, and Willem F.S. Dekker (Robbin). 2015. A hotel that is not bad isn't good. the effects of valence framing and expectation in online reviews on text, reviewer and product appreciation. *Journal of Pragmatics*, 75:28–43.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, and 1 others. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895.

Jaemin Kim, Yohan Na, Kangmin Kim, Sang-Rak Lee, and Dong-Kyu Chae. 2024. SentiCSE: A sentiment-aware contrastive sentence embedding framework with sentiment-guided textual similarity. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14693–14704, Torino, Italia. ELRA and ICCL.

Michelle S Lam, Janice Teoh, James A Landay, Jeffrey Heer, and Michael S Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–28.

Caitlin Doogan Poet Laureate, Wray Buntine, and Henry Linger. 2023. A systematic review of the use of topic models for short text social media analysis. *Artificial Intelligence Review*, 56(12):14223–14255.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Zongxia Li, Lorena Calvo-Bartolomé, Alexander Hoyle, Paiheng Xu, Alden Dima, Juan Francisco Fung, and Jordan Boyd-Graber. 2025. Large language models struggle to describe the haystack without human help: Human-in-the-loop evaluation of llms. *Preprint*, arXiv:2502.14748.

Zongxia Li, Andrew Mao, Daniel Stephens, Pranav Goel, Emily Walpole, Alden Dima, Juan Fung, and Jordan Boyd-Graber. 2024. Improving the TENOR of labeling: Re-evaluating topic models for content analysis. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 840–859, St. Julian's, Malta. Association for Computational Linguistics.

Weng Marc Lim. 2024. What is qualitative research? an overview and guidelines. *Australasian Marketing Journal*, page 14413582241264619.

Bing Liu. 2004. Opinion lexicon. Accessed: 2025-06-24.

Bing Liu. 2020. *Sentiment analysis*, 2 edition. Studies in Natural Language Processing. Cambridge University Press, Cambridge, England.

Ling Luo, Xiang Ao, Yan Song, Jinyao Li, Xiaopeng Yang, Qing He, and Dong Yu. 2019. Unsupervised neural aspect extraction with sememes. In *IJCAI*, pages 5123–5129.

Tingting Ma, Jin-Ge Yao, Chin-Yew Lin, and Tiejun Zhao. 2021. Issues with entailment-based zero-shot text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 786–796, Online. Association for Computational Linguistics.

Fuad Mehraliyev, Irene Cheng Chu Chan, and Andrei Petrovich Kirilenko. 2022. Sentiment analysis in hospitality and tourism: a thematic and methodological review. *International Journal of Contemporary Hospitality Management*, 34(1):46–77.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Christopher Mitcheltree, Skyler Wharton, and Avneesh Saluja. 2018. Using aspect extraction approaches to generate review summaries and user profiles. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 68–75, New Orleans - Louisiana. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia.

Mir Tafseer Nayeem and Davood Rafiei. 2025. OpinioRAG: Towards generating user-centric opinion highlights from large-scale online reviews. In *Second Conference on Language Modeling*.

Thi-Nhung Nguyen, Hoang Ngo, Kiem-Hieu Nguyen, and Tuan-Dung Cao. 2023. A self-enhancement multitask framework for unsupervised aspect category detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8043–8054, Singapore. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021a. Large dual encoders are generalizable retrievers. *Preprint*, arXiv:2112.07899.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2021b. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *Preprint*, arXiv:2108.08877.

Dmitry Nikolaev and Sebastian Padó. 2023. Representation biases in sentence transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3701–3716, Dubrovnik, Croatia.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland.

Simone Rebora. 2023. Sentiment analysis in literary studies. a critical survey. *DHQ: Digital Humanities Quarterly*, 17(3).

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Anna Rogers. 2021. Changing the world by changing the data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2182–2194, Online.

Manuel J Sánchez-Franco and Manuel Rey-Moreno. 2022. Do travelers' reviews depend on the destination? an analysis in coastal and urban peer-to-peer lodgings. *Psychology & marketing*, 39(2):441–459.

Rupak Sarkar, Patrick Y Wu, Kristina Miler, Alexander Miserlis Hoyle, and Philip Resnik. 2025. Pairscale: Analyzing attitude change with pairwise comparisons. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1722–1738.

Alexandra Schofield, Siqi Wu, Theo Bayard de Volo, Tatsuki Kuze, Alfredo Gomez, and Sharifa Sultana. 2025. "my very subjective human interpretation": Domain expert perspectives on navigating the text analysis loop for topic models. *Proceedings of the ACM on Human-Computer Interaction*, 9(1):1–30.

Anna Shadrova. 2021. Topic models do not model topics: epistemological remarks and steps towards best practices. *Journal of Data Mining & Digital Humanities*, 2021:4.

Tian Shi, Liuqing Li, Ping Wang, and Chandan K Reddy. 2021. A simple and effective self-supervised contrastive learning framework for aspect detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13815–13824.

Aivin V. Solatorio. 2024. Gistembed: Guided in-sample selection of training negatives for text embedding fine-tuning. *Preprint*, arXiv:2402.16829.

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357, Singapore.

Stéphan Tulkens and Andreas van Cranenburgh. 2020. Embarrassingly simple unsupervised aspect extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3182–3187.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Pranav Venkit, Mukund Srinath, Sanjana Gautam, Saranya Venkatraman, Vipul Gupta, Rebecca Passonneau, and Shomir Wilson. 2023. The sentiment problem: A critical survey towards deconstructing sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13743–13763, Singapore.

Maria C Voutsa, Nicolas Tsapatsoulis, and Constantinos Djouvas. 2025. Biased by design? evaluating bias and behavioral diversity in llm annotation of real-world and synthetic hotel reviews. *AI*, 6(8):178.

Fengjun Wang, Moran Beladev, Ofri Kleinfeld, Elina Frayerman, Tal Shachar, Eran Fainman, Karen Lastmann Assaraf, Sarai Mizrachi, and Benjamin Wang. 2023a. Text2Topic: Multi-label text classification system for efficient topic detection in user generated content with zero-shot capabilities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 93–103, Singapore.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023b. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Trans. on Knowl. and Data Eng.*, 35(11):11019–11038.

A Additional Experiments

Aspect Category Detection (LEFT) We evaluate aspect extraction using queries constructed solely on the `{aspect}` without sentiment words. An ablation study on different values of K is provided in Figure 7 and a model comparison is shown in Table 4.

Aspect Category Sentiment Analysis (Neutral) (RIGHT) We added the NEUTRAL sentiment and extracted it using queries of the form "`okay {aspect}`". An ablation study on different values of K is provided in Figure 7 and a model comparison is shown in Table 4.

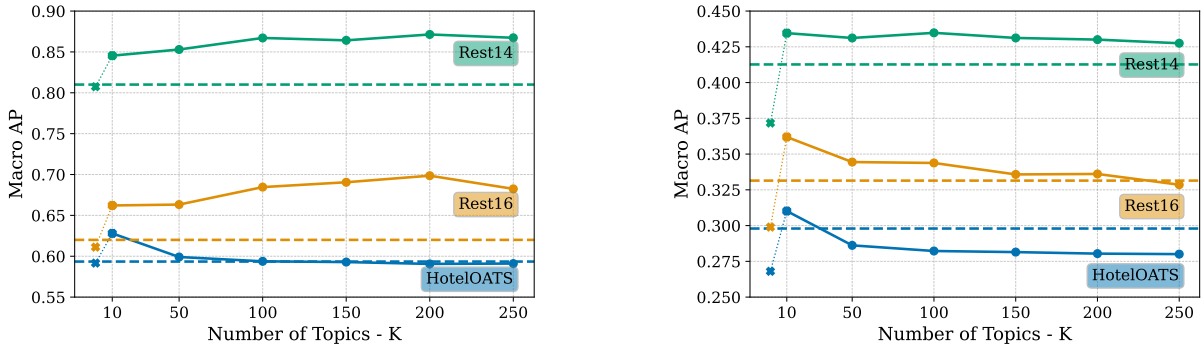


Figure 7: (LEFT) Randomly sampling by sentiment does not show any improvement in topic retrieval, while SWETT improves or at least does not degrade performance. (RIGHT) We observe results consistent with those obtained without the neutral sentiment, as described in Section 5.1.

.

	HotelOATS	Rest16	Rest14	Avg.
General Sentence Embeddings				
all-MiniLM-L6-v2	58.28	62.94	80.86	67.36
all-MiniLM-L12-v2	59.87	63.15	81.16	68.06
all-mpnet-base-v2	59.16	61.11	80.74	67.00
E5-small-v2	52.53	50.60	72.96	58.70
E5-base-v2	51.66	56.61	76.35	61.54
E5-large-v2	49.91	51.55	74.17	58.54
E5-Mistral-7B-Instruct	59.40	68.31	84.12	70.61
GTE-base-v1.5	55.82	59.17	78.11	64.37
GTE-large-v1.5	57.27	58.53	80.03	65.28
GTE-ModernBERT	54.03	64.60	82.49	67.04
ST5-XXL	56.05	58.19	76.14	63.46
GTR-T5-XXL	52.57	60.10	78.27	63.65
Opinion Mining Sentence Embeddings				
SentiCSE	39.47	35.66	49.23	41.45
StanceSBERT	44.96	40.26	57.54	47.59
Tourism Mining Sentence Embeddings				
$\text{Word2Vec}_{\text{CBOW}}$	62.73	63.18	72.83	66.25
$\text{Word2Vec}_{\text{SG}}$	60.27	65.23	76.46	67.32
$\text{FastText}_{\text{CBOW}}$	59.73	55.15	71.26	62.05
$\text{FastText}_{\text{SG}}$	59.93	61.40	74.89	65.41
TourBERT	29.70	32.13	44.86	35.56
TourMPNet	21.56	24.51	34.14	26.74
TourCSE $K = 10$	**62.80**	66.21	84.54	71.18
GTE-TourCSE $K = 10$	58.01	**71.34**	**85.46**	**71.60**

	HotelOATS	Rest16	Rest14	Avg.
General Sentence Embeddings				
all-MiniLM-L6-v2	21.05	23.39	31.27	25.24
all-MiniLM-L12-v2	23.42	24.95	32.86	27.08
all-mpnet-base-v2	26.80	29.90	37.17	31.29
E5-small-v2	23.81	26.28	37.86	29.32
E5-base-v2	24.70	30.05	40.10	31.62
E5-large-v2	24.17	29.18	39.21	30.85
E5-Mistral-7B-Instruct	23.00	27.53	33.84	28.12
GTE-base-v1.5	26.05	31.22	42.62	33.30
GTE-large-v1.5	27.61	33.74	42.78	34.71
GTE-ModernBERT	29.67	36.99	49.06	38.57
ST5-XXL	31.04	36.56	48.29	38.63
GTR-T5-XXL	22.26	29.21	35.53	29.00
Opinion Mining Sentence Embeddings				
SentiCSE	21.07	22.87	30.89	24.94
StanceSBERT	24.59	25.03	33.96	27.86
Tourism Mining Sentence Embeddings				
$\text{Word2Vec}_{\text{CBOW}}$	22.45	23.64	28.59	24.89
$\text{Word2Vec}_{\text{SG}}$	21.46	24.44	28.96	24.95
$\text{FastText}_{\text{CBOW}}$	21.54	21.92	26.74	23.40
$\text{FastText}_{\text{SG}}$	20.56	24.08	27.82	24.15
TourBERT	12.73	14.67	21.13	16.18
TourMPNet	0.08	0.09	12.30	4.16
TourCSE $K = 10$	31.02	36.19	43.45	36.89
GTE-TourCSE $K = 10$	**32.74**	**42.26**	**54.40**	**43.13**

Table 4: (LEFT) **TourCSE** outperforms both general and opinion mining models. (RIGHT) There is no difference between **TourCSE** and both general and opinion mining models.

B CAt

B.1 Background

Between 2018 and 2020, several studies aimed to improve implicit aspect extraction using neural network architectures based on Word2Vec features (He et al., 2017; García-Pablos et al., 2018; Luo et al., 2019). Tulkens and van Cranenburgh (2020) argued that such architectures are unnecessary, showing that relying solely on Word2Vec embeddings achieve competitive results. Their method is based on two main concepts:

- Incorporating an attention mechanism to assign higher weights to tokens that are more similar to the most frequent aspect embeddings.
- Computing cosine similarity between the mean weighted token embeddings of a sentence and a target topic embedding.

Is CAt Outdated? Although improvements in unsupervised implicit aspect extraction have been reported for pipelines using pseudo-labeling to train transformer-based language models (Nguyen et al., 2023; Shi et al., 2021), direct comparisons with CAt are not entirely fair. CAt can generalize beyond annotated aspects in evaluation because Word2Vec embeddings naturally group semantically related words into thematic clusters (see Figure 3). In contrast, pseudo-labeling pipelines from related work are over-optimized to extract only the aspects present in the evaluation set. Furthermore, since CAt was not evaluated using pseudo-labeling, it remains unclear whether these newer pipelines (somewhat complex and less reproducible) truly outperform a basic Word2Vec clustering approach.

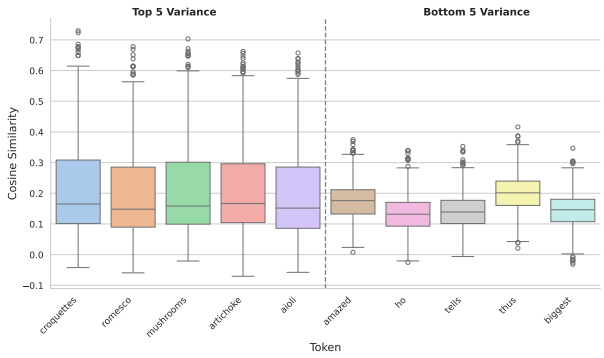


Figure 8: Boxplot of token similarities. The top 5 tokens with the highest variance are related to the topic FOOD, while the bottom 5 tokens exhibit low variance and are not informative for any specific aspect.

Method	FOOD	STAFF	AMBIANCE
ABAE	82.8	**75.7**	74.0
Mean*	86.2 / 85.6	66.9 / 67.5	71.3 / 57.4
Softmax*	86.5 / 88.1	71.1 / 69.3	68.6 / 66.4
RBF*	86.2 / 92.1	67.3 / 78.8	70.9 / 76.6
CosineVar	88.4	72.2	**76.8**
$\text{DeBERTa}_{\text{NLI}}$	**89.1**	67.8	68.9

Table 5: F1-macro scores on the **CitySearch** dataset. For Mean, Softmax, and RBF, scores are reported as "ours / Tulkens and van Cranenburgh (2020)". Differences may stem from Word2Vec hyperparameter settings and the candidate aspect sets.

Re-implementation We compare our reimplementation of CAt with the results reported by Tulkens and van Cranenburgh (2020) and ABAE (He et al., 2017). Additionally, we evaluate a state-of-the-art Natural Language Inference (NLI) model [9] using the template "`this review discusses [ASPECT]`".

[9] https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli

We follow the original experimental setup, tuning hyperparameters on the SemEval datasets (Pontiki et al., 2014, 2016) and evaluating performance on CitySearch (Ganu et al., 2009). Furthermore, we introduce a novel attention mechanism, **CosineVar**, detailed in Appendix B.2. As illustrated in Figure 8, the intuition behind CosineVar is that aspect-relevant tokens, such as *mushrooms*, tend to have high similarity with their corresponding candidate aspect (e.g., *FOOD*) and low similarity with others (e.g., *STAFF*). In contrast, non-aspect tokens generally exhibit relatively uniform similarity across all candidate aspect terms. Results are reported in Table 5. Overall, we find that NLI models do not outperform `Word2Vec` embeddings on the topics STAFF and AMBIANCE[10]. Furthermore, no single method consistently achieves the best performance across all aspects. Nevertheless, our proposed **CosineVAR** attention mechanism outperforms our implementations of SoftMax, Mean, and RBF attention. Importantly, CosineVAR requires only the candidate aspect set as a hyperparameter, unlike RBF, which also depends on tuning a gamma value.

B.2 CosineVar : Algorithm

Given a matrix $S \in R^{n \times d}$ of sentence token input vectors and a matrix $C \in R^{m \times d}$ of candidate aspect vectors, we first compute the cosine similarity matrix $Z \in R^{n \times m}$, where each entry is:

$$Z_{ij} = \cos(\vec{s}_i, \vec{c}_j)$$

For each input vector $\vec{s}_i$, we compute the variance of its cosine similarities with all candidate vectors:

$$\text{var}_i = \text{Var}(Z_{i1}, Z_{i2}, \ldots, Z_{im})$$

The attention weights are then obtained by normalizing these variances:

$$\alpha_i = \frac{\text{var}_i}{\sum_{k=1}^{n} \text{var}_k}$$

The final attention vector $\boldsymbol{\alpha} \in R^{1 \times n}$ reflects the relative variability of each input vector's similarity to the candidates.

B.3 Code

```python
from cat_aspect_extraction import CAt, RBFAttention  # for using the model
from reach import Reach  # for loading word embeddings

# Load in-domain word embeddings and create a CAt instance
r = Reach.load("path/to/embeddings", unk_word="UNK")
cat = CAt(r)

# Initialize candidate aspects
candidates = [
    "food",
    "service",
    "ambiance",
    "price",
    "location",
    "experience"
]

for aspect in candidates:
    cat.add_candidate(aspect)

# Add topics
cat.add_topic("food", ["taste", "flavor", "quality", "portion", "menu", "dish", "cuisine", "ingredient"])
cat.add_topic("service", ["staff", "waiter", "waitress", "service", "server", "host", "manager", "bartender"])
cat.add_topic("ambiance", ["atmosphere", "decor", "interior", "design", "lighting", "music", "noise", "vibe"])

# Compute topic score
sentence = "The food was great !".split()  # tokenize your sentence
cat.get_scores(sentence, attention=RBFAttention())
# Output: [('food', 1), ('service', 0.5), ('ambiance', 0.0)]
```

[10] See (Ma et al., 2021; Arakelyan et al., 2024), which empirically show that NLI models are sensitive to slight changes.

C Definition of Sentiment

C.1 Misclassifications of General Sentiment Models

Review	V	R-T	DB-N	Mist.L	Stars
The room was fine.	**pos.**	**pos.**	**pos.**	neu.	4
nothing special, but it was clean, big enough for two and decent private bathroom.	**pos.**	**pos.**	**pos.**	neu.	4
I called the front desk to ask for the manager on duty.	neu.	neu.	pos.	**neg.**	1
Talk about "fresh" food.	neu.	neu.	**pos.**	neg.	2
The location within the emirates mall is very convenient and the movie theater is very close to the hotel.	neu.	**pos.**	**pos.**	neu.	4
The service is fine.	**pos.**	**pos.**	**pos.**	neu.	4
service is friendly... but smell.... oh, boy!	neu.	pos.	pos.	**neg.**	2
The hotel is located perfectly.	**pos.**	**pos.**	**pos.**	neu.	4
The restaurant / cafe staff were great and the food was fine.	**pos.**	**pos.**	**pos.**	neu.	4
At check-in, the staff was polite.	**neu.**	pos.	neg.	**neu.**	2
Breakfast by the the beach every morning.	neu.	neu.	**pos.**	**pos.**	4

Table 6: V. refers to VADER (Hutto and Gilbert, 2014), R-T to a RoBERTa model trained on Twitter data (`cardiffnlp/twitter-roberta-base-sentiment-latest`), DB-N to a DistilBERT model trained on NLI data (`lxyuan/distilbert-base-multilingual-cased-sentiments-student`), and Mistral.L to Mistral Large (using the prompt from (Zhang et al., 2024)).

C.2 Ethics Sheet

Q1: What is the framework and definition of sentiment utilized? We adopt the definition proposed by Liu (2020):

> *"A subjective statement, view, attitude, emotion, or appraisal about an entity or an aspect of an entity from an opinion holder."*

Q2: What framework is employed for sentiment analysis in the measurement of sentiment? Sentiment is modeled using a regression scale (1–5), aligned with user star ratings, and evaluated as a binary classification: `positive` or `negative`.

Q3: Will this study be made available for public use in measuring sentiment in NLP? Yes.

- Q3.1: **Is the training dataset publicly available without access restrictions?** Yes. All datasets used for training, visualization, and evaluation are publicly released.
- Q3.2: **Is the model algorithm publicly available without access restrictions?** Yes. The algorithm is openly accessible.

Q4: Is this system primarily designed for users outside the field of NLP? Yes. This research is intended to support transdisciplinary investigation.

Q5: What are the specific use cases for this system? The primary use case is content analysis of tourism reviews (Lim, 2024). Notably trough embeddings visualization with BERTopic (Grootendorst, 2022). Applications beyond this domain scope may yield misleading interpretations.

Q6: Who are the intended users of this system? The system targets social scientists aiming to analyze and interpret the meanings conveyed in tourism-related reviews.

Q7: Were tests conducted to identify explicit and implicit biases in sentiment analysis models, particularly regarding sociodemographic factors? No, such bias evaluations were not performed.

Q8: Were experts from interdisciplinary fields involved in discussing the use and evaluation of sentiment analysis models in social applications? Yes, the system is actually used by domain experts. It is also easier to navigate in embedding visualizations compared to general and opinion mining embeddings.

Q9: Did the study consider potential cultural or contextual variations in sentiment interpretation? No. Sentiment was treated as directly inferred from the reviewer's star rating, without accounting for cultural variation in expression.

Q10: Were any measures implemented to mitigate potential biases in the model? Yes. We used star ratings as sentiment proxies and trained models on domain-specific data, whose biases are more interpretable than those of deep learning models (Rogers, 2021).

D Tourism Opinion Mining

In tourism review opinion mining, methods typically rely on lightweight supervised or unsupervised approaches that prioritize scalability.

D.1 Academia

Similar to Laureate et al. (2023), Mehraliyev et al. (2022) note that scholars predominantly use LDA (Blei et al., 2003) to analyze tourism reviews due to the lack of available training data. Interestingly, Sánchez-Franco and Rey-Moreno (2022) extract aspects using BERTopic (Grootendorst, 2022) with `all-MiniLM-L6-v2` to study travelers' destination choices between coastal and urban locations. Their approach is motivated by the computational efficiency and interpretability provided by Sentence Transformer embeddings.

D.2 Industry

Airbnb (Mitcheltree et al., 2018) applies ABAE, a neural topic model based on Word2Vec embeddings (He et al., 2017), to analyze customer reviews. However, they find that simple `k-means` clustering with ABAE does not significantly improve the identification of common aspects, consistent with Hoyle et al. (2022), who report that neural topic models are less stable and do not consistently outperform traditional algorithmic methods. Booking.com (Wang et al., 2023a) uses Sentence Transformers trained in a supervised manner with an annotation scheme based on user search behavior. This approach is motivated by the desire to avoid reliance on prompt engineering while ensuring scalability.

D.3 Additional Information on Data (Selection, Evaluation, Extraction)

Model	Objective Loss	Positive Sampling	Negative Sampling
SentiCSE	Contrastive	Sentiment	Sentiment
StanceSBERT	Contrastive & Triplet	Sentiment & Topic	Sentiment & Topic & Filtering
TourCSE	**InfoNCE**	Sentiment & Topic	**Random & In-Batch Filtering**

Table 7: Comparison of contrastive learning methods for sentiment-aware sentence embeddings. We report the loss function and the strategies used for constructing positive and negative pairs. `Filtering` refers to the use of a Sentence Transformer to discard sentences based on a similarity threshold. Overall, **the creation of the TourCSE training data is simpler, as it requires only positive samples.**

Topic	Pos.	Neg.	Neu.
LOCATION	1119	48	18
FACILITIES	280	105	18
ROOMS	835	328	42
ROOM AMENITIES	219	110	8
SERVICE	1044	181	16
FOOD & DRINKS	350	82	17

Topic	Pos.	Neg.	Neu.
AMBIENCE	217	32	17
FOOD	730	213	58
LOCATION	32	1	8
DRINKS	95	17	2
SERVICE	261	269	19

Topic	Pos.	Neg.	Neu.
AMBIENCE	314	112	86
FOOD	1113	260	193
SERVICE	410	252	61
PRICE	212	123	28

Table 8: Distribution of sentiment polarity (positive, negative, neutral) by topic across HotelOATS (LEFT), Rest16 (MIDDLE), and Rest14 (RIGHT) datasets.

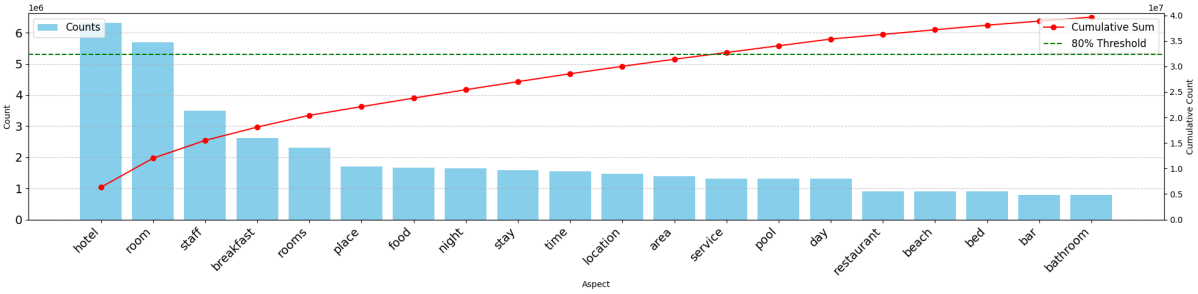


Figure 9: Top 20 most frequent aspects extracted from HotelRec (Hotel Domain)

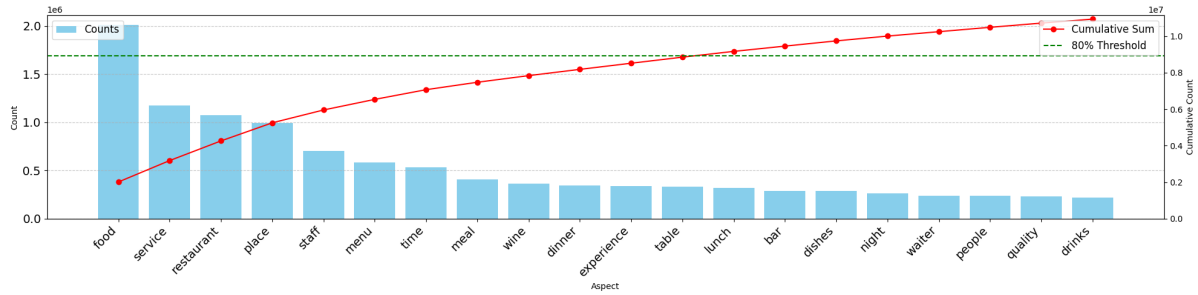


Figure 10: Top 20 most frequent aspects extracted from SixTripAdvisor (Restaurant Domain)

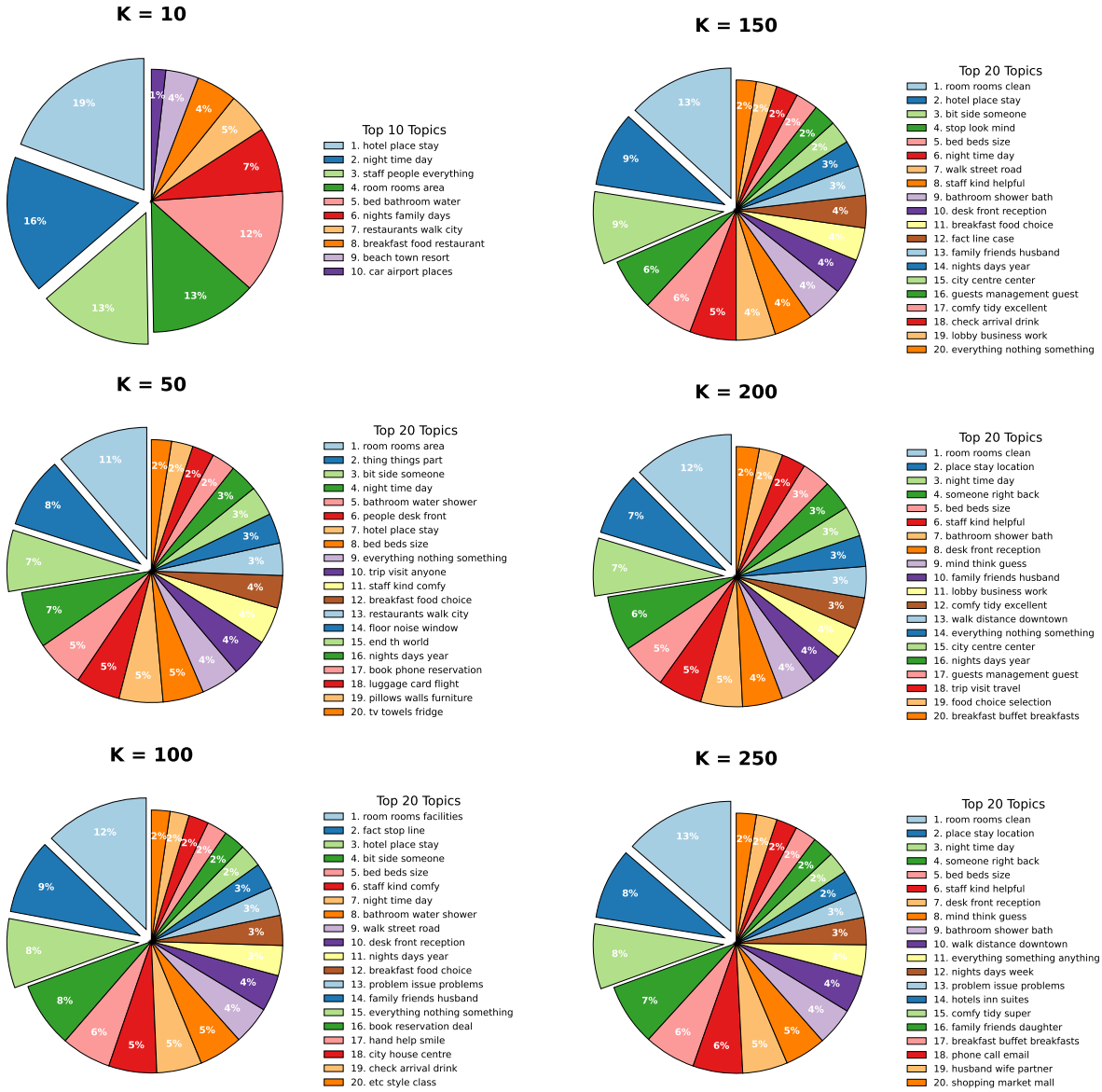


Figure 11: Distribution of extracted topics from HotelRec (Hotel Domain) for different values of K.

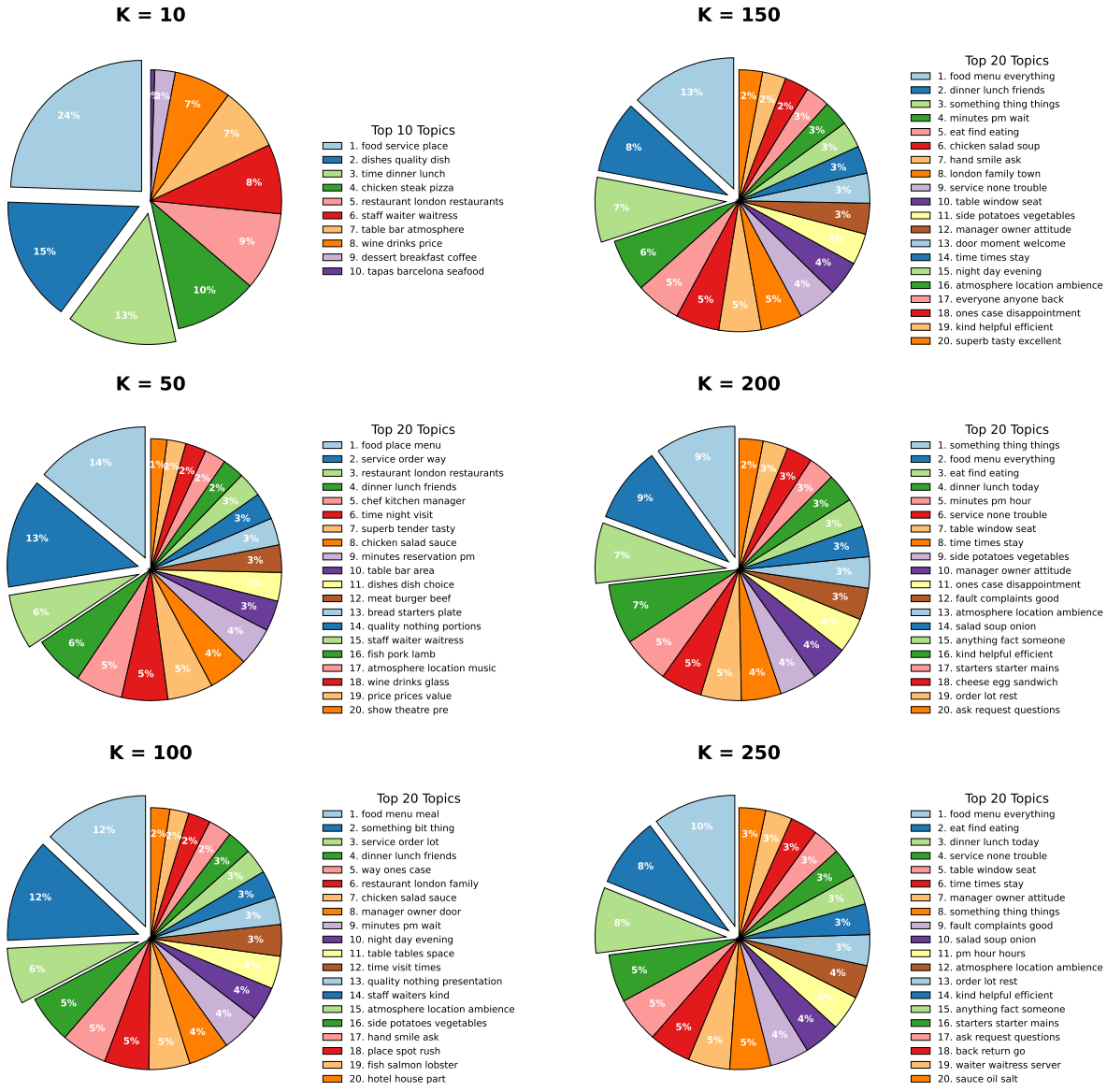


Figure 12: Distribution of extracted topics from SixTripAdvisor (Restaurant Domain) for different values of K.

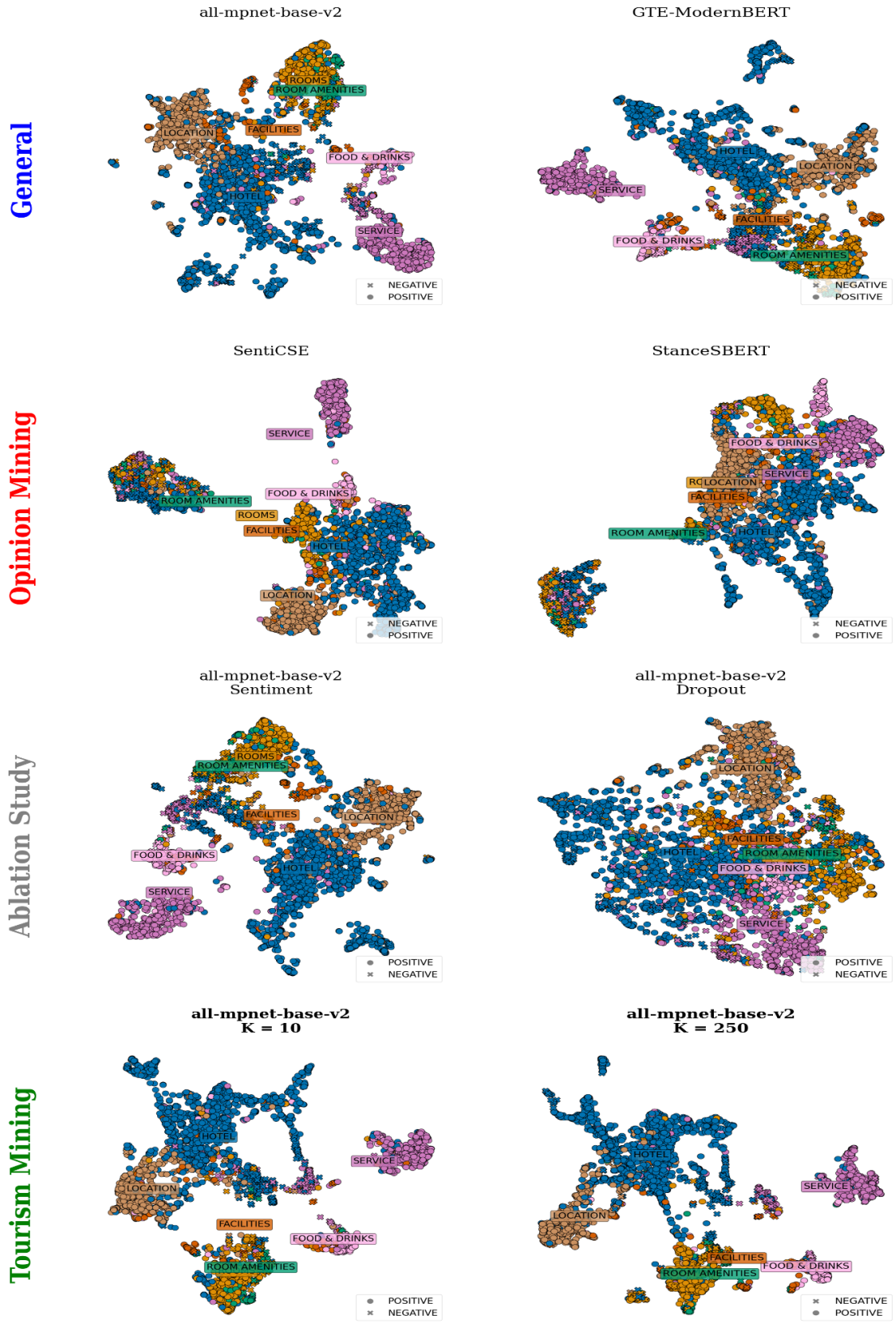


Figure 13: UMAP projections of sentence embeddings from the HotelOATS datasets (Hotel Domain)

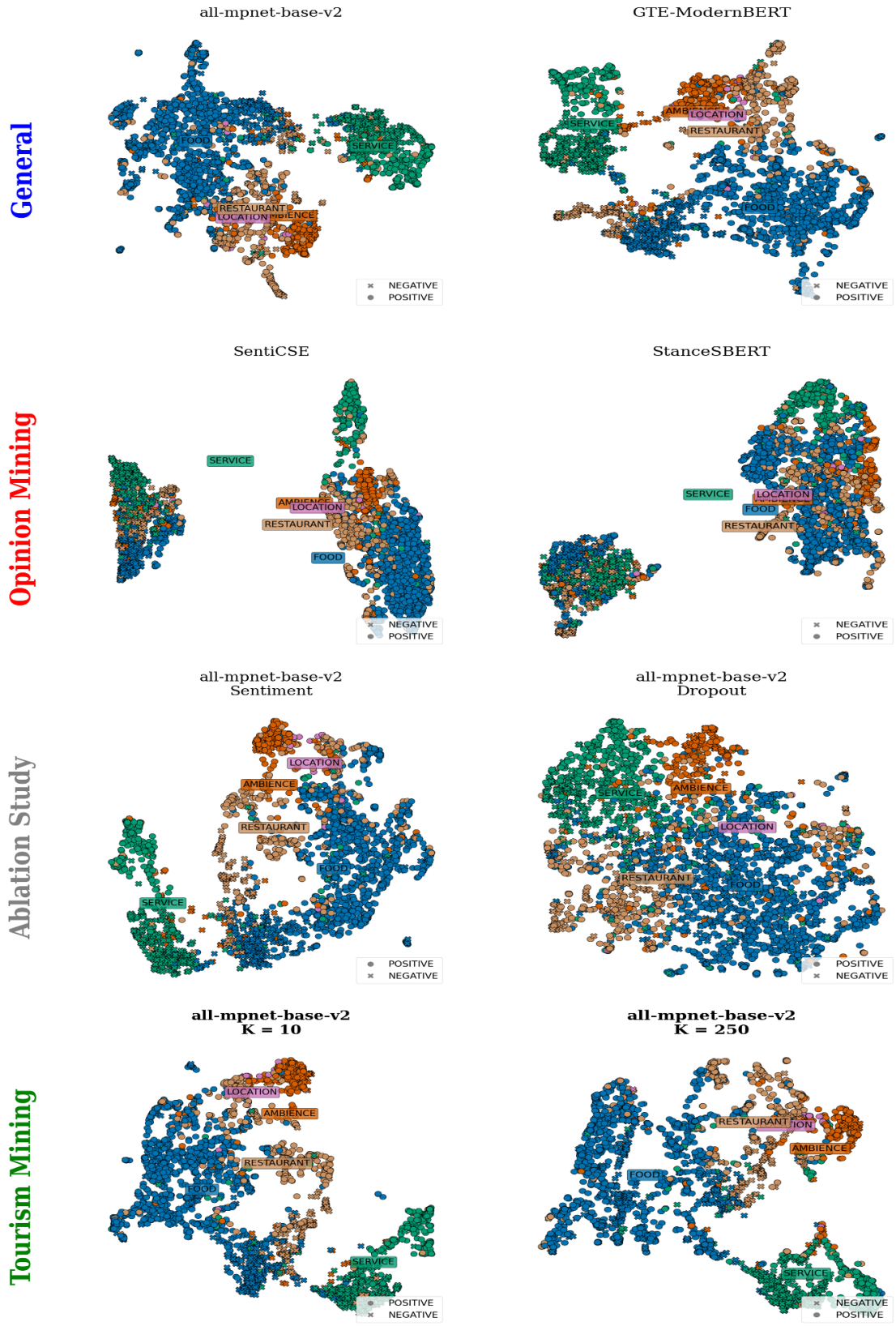
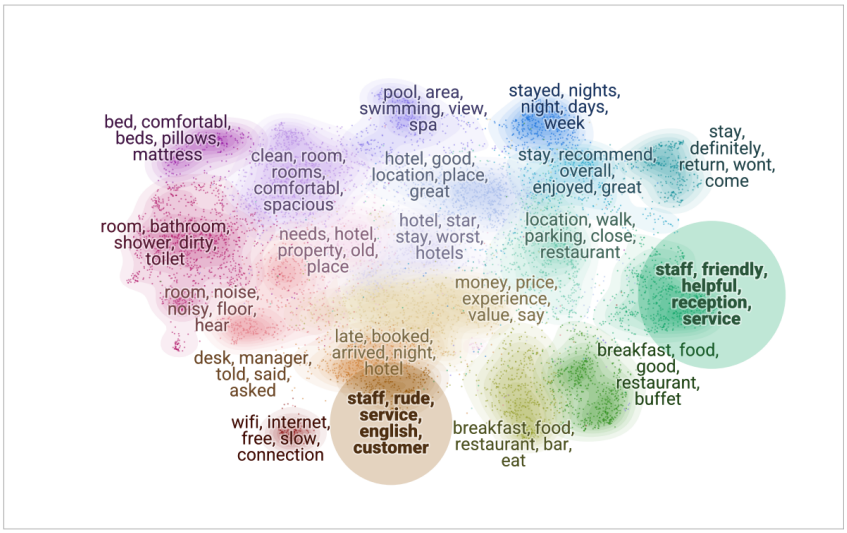
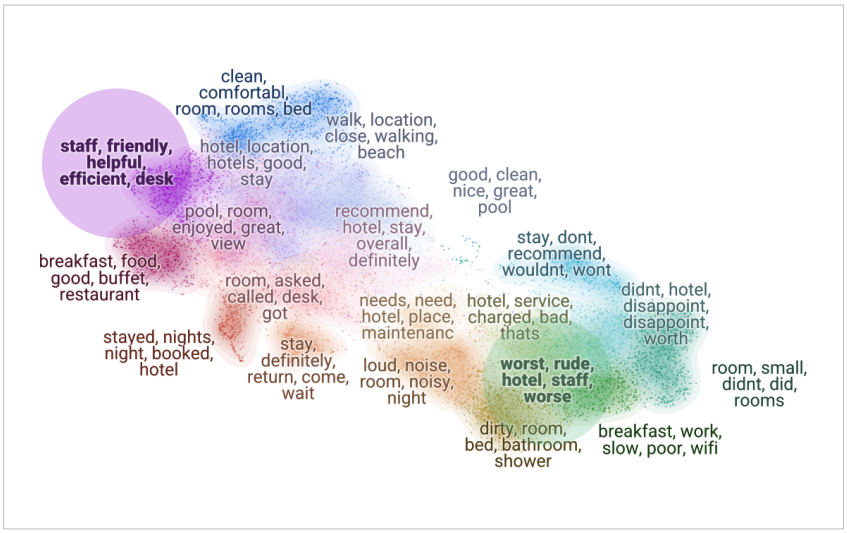


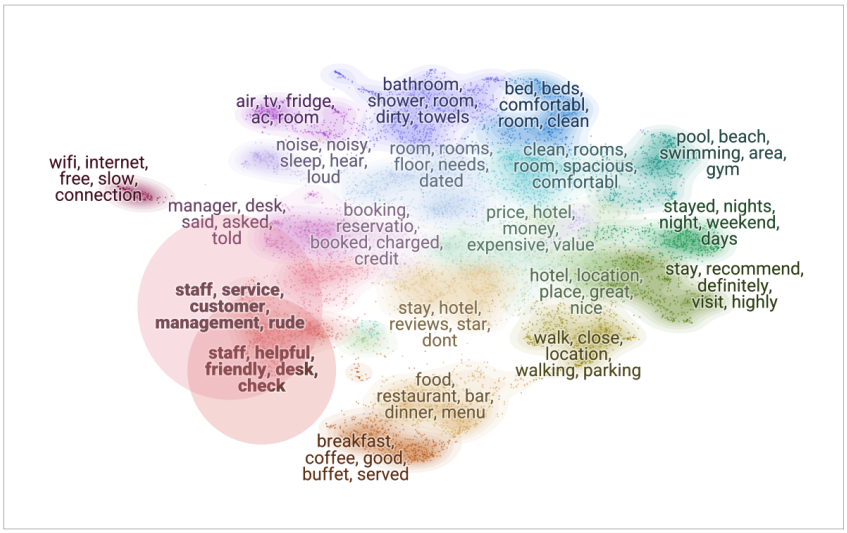
Figure 14: UMAP projections of sentence embeddings from the Rest16 and Rest14 datasets (Restaurant Domain)



(a) General : all-mpnet-base-v2



(b) Opinion Mining : StanceSBERT



(c) Tourism Mining : TourCSE (LoRA, K=250)

Figure 15: Full visualization of BERTopic outputs (KMeans, $K = 20$). Clusters related to STAFF are highlighteds.