Are (LLM) Sentence Transformers really useful for unsupervised hospitality opinion mining?

Anonymous ACL submission

Abstract

Despite growing interest in opinion mining for the hospitality industry, the lack of benchmarks aligned with real-world use cases limits the development of robust classifiers. Additionally, recent advancements in dense retrieval methods using Sentence Transformers, which enable zero-shot text classification, have not been thoroughly explored. This study evaluates embedding models for classifying hospitality reviews using publicly available human-annotated datasets to assess their limitations and applicability for opinion mining. Our findings indicate that dense retrieval models based on large language models either underperform or show only marginal improvements over a simple continuous bag of words model trained on in-domain data. While fine-tuning pre-trained sentence transformers perform strongly in extracting both sentiment and topic information, the lack of sufficient training data limits the development of effective solutions. Finally, we offer recommendations, based on key surveys in the literature, to bridge the gap between domain-specific needs and recent NLP advancements, thereby enhancing opinion mining in the hospitality sector.

1 Introduction

Recently, (Ameur et al., 2023) conducted a systematic review of the literature on opinion mining–also referred to as sentiment analysis (Liu, 2020)–in the hospitality domain, analyzing over 700 articles from the past 20 years. An important subtask covered is *Aspect-Based Sentiment Analysis* (Zhang et al., 2023), which classifies sentiment polarity in relation to specific aspects. In the hospitality domain, these aspects range from core services, such as *room quality*, to more subjective attributes, such as *ambience*. This survey revealed the lack of expert-annotated training data, which significantly limits the development of effective opinion mining systems[1] (Rogers, 2021). In particular, most publicly annotated datasets are based on a limited set of topics that fail to capture the full diversity of the corpus[2], while alternative unsupervised approaches, such as topic modeling, "give often inaccurate classification results" (Ameur et al., 2023, p. 19). In practice, the approaches discussed in this literature review fail to provide management scholars with a publicly reliable framework for testing their hypotheses about customer feedback.

However, this survey does not cover recent advances of topic modeling using Sentence Transformers (Reimers and Gurevych, 2019), which have enabled modern approaches through sentence clustering based on embedding representations (Grootendorst, 2022). These methods also offer a low-computation solution for zero-shot multi-label text classification (Eden et al., 2023) by computing cosine similarity between sentence embeddings and target labels, their representative queries, or the centroid of previously extracted clusters. As a result, only n calls to the model are needed, where n is the number of labels and sentences.

Still, these models do not provide a universal definition of sentence representation. First, their embeddings are biased toward undefined topics often represented by common nouns present in the sentence (Nikolaev and Padó, 2023). This might overlook the specificities of the hospitality domain, such as the distinction between concepts like *service* and *staff* or *facilities* and *amenities*. Additionally, the lack of documentation on the full training dataset of some models[3] hinders the interpretabil-

[1] See (Ha et al., 2021) for a qualitative study showing how annotators' domain knowledge affects the performance of BERT-based classification models.

[2] SemEval annotation guidelines use 34 topics for hotel reviews, 5 of which are miscellaneous (Pontiki et al., 2016). Booking.com uses 239 topics (Wang et al., 2023a).

[3] Notably, E5 (Wang et al., 2024) and GTR embeddings (Zhang et al., 2024b) are pre-trained on unspecified "web data" without details on its composition.

ity of their representations (Rogers, 2021). As a direct consequence, these models tend to be ineffective in retrieving sentences based on both specific sentiment and topic relevance to a given query[4].

In this paper, we explore a specific research question: "*Do Sentence Transformers offer useful features for developing unsupervised methods for hospitality opinion mining?*"

2 Experimental Studies

We evaluated the ability of various models to retrieve all relevant documents for a given query, a common use case in opinion mining systems (Wang et al., 2023a; Eden et al., 2023; Introne, 2023).

2.1 Datasets

We selected four human-annotated english datasets from the hotel and restaurant domains: Rest14 (Pontiki et al., 2014), Hotel15 (Pontiki et al., 2015), Rest16 (Pontiki et al., 2016), and HotelOATS (Chebolu et al., 2024). These datasets provide sentence-level topic labels and their associated sentiments (*positive*, *negative* or *neutral*), with annotation guidelines available to clarify labels[5]. We made several modifications to these datasets[6] by removing sentences with *neutral* or *conflicting* sentiment, as well as those labeled with the MISCELLANEOUS aspect, since these categories are often ambiguous or subjective. Additionally, we excluded entities such as RESTAURANT and HOTEL, as these broad classifications are too general and could be better categorized into more specific sub-entities. Finally, we focused exclusively on retrieving topics at the entity level, following previous research (Huang et al., 2020), as this provides a more accurate representation of model performance in an unsupervised setting. All of these modifications[7] are driven by the "fact that sentiment analysis is a very subjective task" (Chebolu et al., 2022, p. 7).

2.2 Models

We selected several models available on the Hugging Face Hub and classified them based on their backbone size (static embeddings, sentence transformers, and LLM-based sentence transformers). We also built two static embeddings using CBOW and Skip-Gram (Mikolov et al., 2013) with Gensim (Řehůřek and Sojka, 2010), trained on a large-scale in-domain dataset[8]. Sentence embeddings are computed as the mean embedding of all tokens.

Following the training procedure in (Zhao et al., 2023), we introduced a supervised category, OpinionCSE, by fine-tuning *all-mpnet-base-v2* with InfoNCE loss[9]. Although this training objective only requires a list of positive sentence pairs, we considered all possible sentence combinations sharing the same labels from the cleaned training sets of HotelOATS (for the hotel domain) and Rest16 (for the restaurant domain), sampling negative examples randomly within the batch.

2.3 Details of Downstream Tasks

For each dataset, we generated two versions of the test set following ABSA terminology: (i) Aspect Category Detection (ACD) focuses on identifying aspects alone, (ii) Aspect Category Sentiment Analysis (ACSA), extends this by retrieving aspects with their associated sentiment. A single query is used for each aspect, derived from the annotation guidelines. For ACD, the query is directly derived from these guidelines, while for ACSA, it is adjusted to include a sentiment representation of the targeted aspect within the description[10].

Following (Wang et al., 2023a), we use the macro-averaged precision score from Scikit-Learn (Pedregosa et al., 2011) to compare the models.

3 Results

Results are provided in Table 1. **OpinionCSE** outperforms in both ACD and ACSA tasks, except for ACD on HotelOATS, highlighting the strength of its training architecture. However, the evaluation process, based on low quality data and lacking

[4]This problem was highlighted in (Introne, 2023, p. 391) and investigated by (Ghafouri et al., 2024) who propose a solution for adding stance in pre-trained Sentence Transformers.

[5]HotelOATS and Hotel15 use the same labeling scheme, while Rest14 lacks annotation guidelines but uses the same aspects as Rest16. These datasets have been used in a fully supervised setting with BERT pre-trained on in-domain datasets, and a pre-training performed via MLM and contrastive learning tasks, achieving over 90% F1 scores (Chebolu et al., 2024; Sun et al., 2019; Li et al., 2021; Liang et al., 2021). LLM(s) have also been used on Rest16 (Zhang et al., 2024a), yielding lower performance compared to previous studies using BERT on the full training set, despite the discussion on potential emergent properties of LLMs (Rogers and Luccioni, 2024).

[6]The modified versions of the datasets are available at `https://anonymous.4open.science/r/dataset-acl-2025-A180/README.md`.

[7]More justifications are given in appendix A.1.

[8]HotelReC (Antognini and Faltings, 2020) for Hotel domain and SixTripAdvsisorReview (López-Riobóo Botana et al., 2022) for the Restaurant domain.

[9]Implemented as *MultipleNegativeRankingLoss* in the Sentence Transformers library. See Table 2 for the hyperparameters used in fine-tuning training.

[10]See Tables 4 and 5 for the queries used.

real-world scenarios, limits the insights on its performance for hospitality opinion mining. Figure 1 gives a visualization of the embedding from OpinionCSE[11].

On the other hand, static embeddings such as CBOW and SG achieve competitive results for aspect retrieval while offering greater interpretability compared to sentence transformers. Among pre-trained sentence embeddings, GTE-modernbert provides the best trade-off between inference time and comprehension of both aspects and sentiments. It outperforms SentiCSE, which is designed exclusively for sentiment analysis, and competes with LLM-based sentence transformers in most scenarios. However, it remains unclear whether this advantage stems from its training architecture or from the data on which it was trained, as the full training set is not publicly available. We conducted an alternative evaluation, presented in Appendix A.4, which yields similar conclusions.

From these results we can state that **for the hospitality use case, domain-specific CBOW appears to be a better choice than large LLM-based sentence transformers**, despite their strong performance on MTEB (Muennighoff et al., 2023). This underlines the subjectivity of retrieval performance, which varies by use case, and emphasizes the potential value of fine-tuning sentence embeddings. This brings us to the next question: *"How can we fine-tune these models and evaluate their performance effectively?"*

Figure 1: Embeddings representation (using t-SNE) of hotel domain with OpinionCSE. We plotted the most representative embedding points for each label, using the label's embeddings.

[11]Embeddings with others models are presented in Appendix A.3.

4 Related Work

Debiasing Sentence Transformers by fine-tuning is an emerging topic (Ramesh Kashyap et al., 2024, p. 1749). It is typically done by pseudo-labeling or using metadata present in the dataset (Schopf et al., 2023; Ghafouri et al., 2024). The most common approach requires domain-specialized cross-encoders (Wang et al., 2022a; Dai et al., 2022), **but this does not resolve the fundamental issue of the quality of training data**. Other pseudo-labeling approaches explore natural language inference (Vacareanu et al., 2024) or synthetic data generation via LLMs (Ma et al., 2021a), **but it seems difficult to assess whether subjective topics, such as ambiance, are perceived in the same way by both humans and the model** (Ma et al., 2021b; Bhargava et al., 2021; Tan et al., 2024; Rogers, 2021; Rogers and Luccioni, 2024; Yu et al., 2024; Bender et al., 2021). Alternatively, they exist easy-to-use and explainable methods for pseudo-labeling, including dynamic rule-based approaches for opinion mining (Qiu et al., 2011) and in-domain word embeddings (Tulkens and van Cranenburgh, 2020). Also, specialized topic models have been developed for hotel reviews to extract both aspects and their associated sentiments (Mukherjee and Liu, 2012; Lu et al., 2011)[12] despite their relative efficiency for content analysis (Laureate et al., 2023). All these approaches have their advantages and drawbacks but researchers need to focus more on the **choice of training data used** and the **evaluation of expected output** rather than the training architecture. For example, (Zhao et al., 2023) proposed a sentence representation model for hospitality opinion mining. However, it remains unclear why the sentiment is extracted using the domain-agnostic rule-based VADER[13] method (Hutto and Gilbert, 2014) instead of star ratings, as suggested by (Xu et al., 2020)? This may be symptomatic of a broader issue where NLP researchers do not explicitely define their understanding of sentiment (Venkit et al., 2023). Also, we do not have any information about the topics extracted using *all-MiniLM-L6-v2* and the definitions that are provided to human evaluators. This raises the challenge of comparing and evaluating unsupervised models, as they all seem to produce effective results despite their different designs.

[12]More examples are described in (Ameur et al., 2023).

[13]For a discussion on potential issues with VADER, see (Rebora, 2023).

Model	HotelOATS		Hotel15		Rest16		Rest14	
	ACD	ACSA	ACD	ACSA	ACD	ACSA	ACD	ACSA
Pre-Trained								
Static Embeddings								
CBOW	**72.09**	47.08	**62.93**	**50.47**	57.97	39.13	71.84	52.19
SG	64.94	38.63	57.73	45.69	52.56	38.29	68.14	50.71
Potion-base-8M (Tulkens and van Dongen, 2024)	56.29	36.54	52.03	39.99	54.59	33.46	66.22	44.35
Sentence Transformers								
SentiCSE (Kim et al., 2024)	32.64	33.17	26.61	35.65	31.4	32.38	39.61	35.77
all-MiniLM-L6-v2	59.09	38.81	51.54	38.67	51.78	42.05	71.42	47.23
all-MiniLM-L12-v2	59.64	44.11	54.64	39.25	50.71	41.02	68.34	45.85
all-mpnet-base-v2	51.60	42.12	46.09	38.74	47.56	39.26	60.98	43.32
e5-small-v2 (Wang et al., 2022b)	52.44	41.44	47.24	42.04	40.99	40.51	59.34	51.32
e5-base-v2 (Wang et al., 2022b)	56.39	44.42	49.30	41.32	43.5	41.86	60.64	51.9
e5-large-v2 (Wang et al., 2022b)	49.18	45.4	46.00	45.85	41.23	40.38	60.39	50.79
GTE-base-en-v1.5 (Li et al., 2023)	62.71	42.11	57.81	40.91	58.93	45.60	**77.23**	57.19
GTE-large-en-v1.5 (Li et al., 2023)	67.74	47.00	56.78	45.25	59.17	46.40	76.13	62.13
GTE-modernbert-base (Li et al., 2023)	62.88	45.01	59.78	50.11	**60.02**	46.86	75.12	58.18
LLM Sentence Transformers								
Sentence-T5-xxl (Ni et al., 2021b)	66.90	**50.96**	54.74	48.44	59.93	**49.74**	75.66	**65.89**
GTR-T5-xxl (Ni et al., 2021a)	57.55	40.23	50.23	34.96	47.66	38.86	71.15	47.16
e5-mistral-7b-instruct (Wang et al., 2023b)	60.33	40.88	54.33	42.37	56.88	43.83	76.89	55.36
Supervised								
OpinionCSE	66.31	62.24	64.39	70.99	79.41	64.73	94.01	74.68

Table 1: Performance comparison of models across datasets for ACD and ACSA tasks using annotation guideline as query. Underlined values represent the top models in each category, bolded values indicate the best unsupervised or pre-trained models, colored cells give the best model overall.

5 Conclusion and Discussion

How to evaluate the validity of topic modeling, especially when research suggests that mathematical metrics are not aligned with human judgment (Chang et al., 2009; Hoyle et al., 2021)? This raises concerns about objectivity, as the definition of the "*best topic model*" depends on the chosen metric and the dataset used (Doogan and Buntine, 2021)[14]. Typically, (Ameur et al., 2023) does not define the topics discussed by customers, aside from a brief mention: "*For instance, in hospitality, we are interested in 'rooms,' 'Food_Drinks,' 'service,' etc.*". If the expected output is not discussed, it is difficult to assess any ground truth and compare approaches. More generally we can wonder whether the NLP research community truly knows which topics should be extracted[15]? Booking.com (Wang et al., 2023a) delegates this task to experts with explanatory tools to assist annotation, which can reduce belief bias but does not eliminate it (González et al., 2021). It is possible that most models and datasets for hospitality opinion mining **remain agnostic to any collaborative and interdisciplinary research questions**, as noted by (Laureate et al., 2023), who also offers guidance on constructing "*good*" topic models[16]. This includes integrating social science hypotheses, as "the focus should be on developing and validating alternative performance measures that reflect the needs of researchers applying topic models to SMD" (Laureate et al., 2023), and considering the specifics of in-domain datasets (Hu and Liu, 2004; Liu, 2020). This calls for a **more rigorous qualitative analysis of the expected outputs**. Addressing such epistemological challenges requires a clear and well-defined understanding of the extracted topics and sentiments. This would reduce confirmation bias and ensure that identified patterns are functional and meaningful, particularly by aligning more precisely with the multifaceted notion of sentiment (Venkit et al., 2023).

To address the question in the title, we have shown that simple models–for instance CBOW with sentiment extraction by star ratings– may suffice, but a truly meaningful answer will remain out of reach until the previous issues are addressed.

[14]The reliability of automatic evaluation has been questioned from a digital humanities perspective (Shadrova, 2021).

[15]This might reflect how NLP is taught, giving the impression that "data is not part of the job" (Rogers, 2021).

[16]See (Hoyle et al., 2022) for an explanation of why some of our complex topic models are "broken" and do not perform better that a simple LDA as regard to content analysis.

6 Limitations

One limitation of this study is the **scope of evaluation and optimization of models**. We did not test all LLMs listed on the MTEB leaderboard, which might have produced different results, nor did we explore all potential methods for optimizing LLM performance, such as leveraging synergies through query modification by another LLM (Feng et al., 2024).

There is also a **lack of practical assessment**. We did not evaluate if the expected output from the tested models was truly informative or useful, as this type of evaluation has not been conducted in prior work. As suggested in (Gu et al., 2025), we could have validated our experiments using LLMs-as-a-judge, however we cannot critique which category assignments should be validated, as we ourselves do not know.

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.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in NLI: Ways (not) to go beyond simple heuristics. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 125–135, Online and Punta Cana, Dominican Republic.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: how humans interpret topic models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, NIPS'09, page 288–296, Red Hook, NY, USA. Curran Associates Inc.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2022. A review of datasets for aspect-based sentiment analysis. In *International Joint Conference on Natural Language Processing*.

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.

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *Preprint*, arXiv:2209.11755.

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.

Lilach Eden, Yoav Kantor, Matan Orbach, Yoav Katz, Noam Slonim, and Roy Bar-Haim. 2023. Welcome to the real world: Efficient, incremental and scalable key point analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 483–491, Singapore.

Jiazhan Feng, Chongyang Tao, Xiubo Geng, Tao Shen, Can Xu, Guodong Long, Dongyan Zhao, and Daxin Jiang. 2024. Synergistic interplay between search and large language models for information retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9571–9583, Bangkok, Thailand. Association for Computational Linguistics.

Vahid Ghafouri, Jose Such, Guillermo Suarez-Tangil, et al. 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*.

Ana Valeria González, Anna Rogers, and Anders Søgaard. 2021. On the interaction of belief bias and explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2930–2942, Online.

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

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Sooji Ha, Daniel J. Marchetto, Sameer Dharur, and Omar I. Asensio. 2021. Topic classification of electric vehicle consumer experiences with transformer-based deep learning. *Patterns*, 2(2):100195.

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, 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.

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.

Jiaxin Huang, Yu Meng, Fang Guo, Heng Ji, and Jiawei Han. 2020. Weakly-supervised aspect-based sentiment analysis via joint aspect-sentiment topic embedding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6989–6999, Online.

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. *Proceedings of the International AAAI Conference on Web and Social Media*, 17(1):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.

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.

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.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. 2021. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 246–256, Online and Punta Cana, Dominican Republic.

Bin Liang, Wangda Luo, Xiang Li, Lin Gui, Min Yang, Xiaoqi Yu, and Ruifeng Xu. 2021. Enhancing aspect-based sentiment analysis with supervised contrastive learning. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 3242–3247, New York, NY, USA. Association for Computing Machinery.

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

Bin Lu, Myle Ott, Claire Cardie, and Benjamin K. Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 81–88.

I. L. López-Riobóo Botana, A. Alonso-Betanzos, V. Bolón-Canedo, and B. Guijarro-Berdiñas. 2022. A tripadvisor dataset for dyadic context analysis (1.0).

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021a. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1075–1088, Online.

Tingting Ma, Jin-Ge Yao, Chin-Yew Lin, and Tiejun Zhao. 2021b. 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.

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*.

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.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 339–348, Jeju Island, Korea.

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.

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, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado.

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.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1):9–27.

Abhinav Ramesh Kashyap, Thanh-Tung Nguyen, Viktor Schlegel, Stefan Winkler, See-Kiong Ng, and Soujanya Poria. 2024. A comprehensive survey of sentence representations: From the BERT epoch to the CHATGPT era and beyond. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1738–1751, St. Julian's, Malta.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China.

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.

Anna Rogers and Alexandra Sasha Luccioni. 2024. Position: Key claims in llm research have a long tail of footnotes. *Preprint*, arXiv:2308.07120.

Tim Schopf, Emanuel Gerber, Malte Ostendorff, and Florian Matthes. 2023. AspectCSE: Sentence embeddings for aspect-based semantic textual similarity using contrastive learning and structured knowledge. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1054–1065, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA.

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.

Stephan Tulkens and Thomas van Dongen. 2024. Model2vec: Fast state-of-the-art static embeddings.

Robert Vacareanu, Siddharth Varia, Kishaloy Halder, Shuai Wang, Giovanni Paolini, Neha Anna John, Miguel Ballesteros, and Smaranda Muresan. 2024. A weak supervision approach for few-shot aspect based sentiment analysis. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2734–2752, St. Julian's, Malta.

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.

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.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022a. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022b. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

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*.

Hu Xu, Lei Shu, Philip Yu, and Bing Liu. 2020. Understanding pre-trained BERT for aspect-based sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 244–250, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. *ACM Comput. Surv.*, 56(12).

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024a. 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.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024b. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *Preprint*, arXiv:2407.19669.

Runcong Zhao, Lin Gui, and Yulan He. 2023. Cone: Unsupervised contrastive opinion extraction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1066–1075, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Further Justification For Datasets Modifications

In the Rest16 test set, sentences like "*Yum!*" and "*Salads are a delicious way to begin the meal*" are annotated as FOOD#QUALITY, but the definition of quality is overly broad: "*opinions focusing on the taste, freshness, texture, consistency, temperature, preparation, authenticity, cooking, or general quality of the food and drinks served in the restaurant.*" This encompasses multiple distinct concepts that should ideally be defined separately. At the same time, for hotel domain, should we distinguish the general notion of cleanliness from its specific application to hotels and rooms, or should they be treated as three separate entities?

Similarly, sentences such as "*I can't wait to go back*" and "*Will absolutely visit again*", labeled as RESTAURANT#GENERAL in Rest16, and "*I recommend this hotel*", labeled as HOTEL#GENERAL in HotelOATS, seem to reflect intent (an other research area (Liu, 2020)) rather than evaluating a specific aspect. Additionally, the value of making a distinction between GENERAL and MISCELLANEOUS in these cases remains highly unclear.

Furthermore, "*Service was decent*" and "*Food was okay, nothing great*" are labeled as *neutral*, but management studies suggest that a more appropriate classification would be *negative* (Kamoen et al., 2015). There also seems to be confusion between the sentiment explicitly stated by the author and how it is perceived by readers. For example, "*Waited 35 minutes for a table for 8, which was ok for such a big crowd*" may seem neutral but could be interpreted as negative by readers.

A.2 OpinionCSE Fine-Tuning Configuration

Table 2 gives the hyperparameters used for the fine-tuning training of OpinionCSE.

Parameter	Value
Training epochs	3
Batch size	128 (hotel)
	32 (restaurant)
Learning rate	5×10^{-5}
Temperature for softmax	0.05
Floating precision	bf16

Table 2: OpinionCSE Fine-Tuning Configuration

A.3 Embeddings Visualization

Figure 2 provides some embeddings visualizations similar to Figure 1. These visualizations demonstrate the efficiency of our fine-tuning of *all-mpnet-base-v2* (OpinionCSE) and the effectiveness of *CBOW* in modeling our corpus.

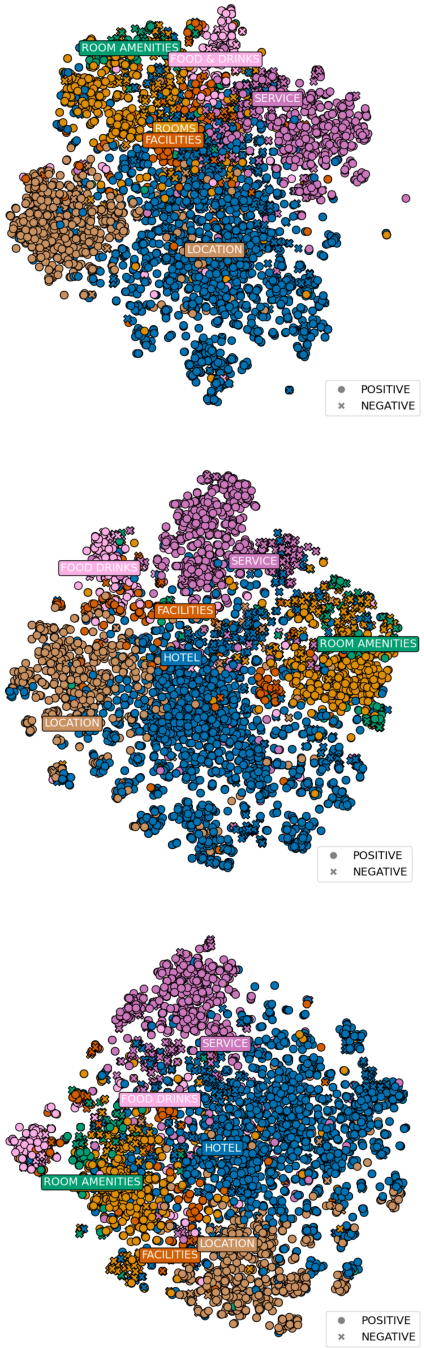

Figure 2: Embeddings representation of hotel domain using t-SNE. From top to bottom: CBOW, all-mpnet-base-v2, e5-mistral-7b-instruct. We plotted the most representative embedding points for each label, using annotation guideline descriptions for CBOW and the label's own embeddings for all other models.

A.4 Additional Experimentation

An alternative evaluation is presented in Table 3. This evaluation also demonstrates that some models are more effective, particularly SentiCSE. For Aspect Category Detection (ACD), we used the entity itself as a query, categorizing aspects as follows:

- For restaurants: food, service, drinks, location, ambiance, and price.
- For hotels: rooms, room amenities, facilities, service, location, and food & drinks.

For Aspect Category Sentiment Analysis (ACSA), we prefixed a sentiment word ("excellent" or "horrible") to the entity. For instance, the query "*excellent food*" retrieves sentences with positive mentions of food, while "*horrible food*" retrieves negative ones.

Model	HotelOATS		Hotel15		Rest16		Rest14	
	ACD	ACSA	ACD	ACSA	ACD	ACSA	ACD	ACSA
Pre-Trained								
Static Embeddings								
CBOW	63.18	39.95	**57.77**	39.78	64.28	34.93	72.83	40.84
SG	61.77	37.2	56.02	36.70	63.15	34.27	76.46	40.9
Potion-base-8M	51.44	37.33	48.47	40.61	58.31	33.91	75.29	42.22
Sentence Transformers								
SentiCSE	42.24	38.46	37.81	44.64	38.5	37.53	49.28	45.36
all-MiniLM-L6-v2	62.16	37.92	51.58	37.34	61.98	35.44	80.86	46.13
all-MiniLM-L12-v2	65.04	42.24	55.15	41.25	62.67	38.01	81.16	48.52
all-mpnet-base-v2	**63.64**	48.39	52.08	41.57	59.9	41.57	80.74	53.34
e5-small-v2	55.13	44.09	47.07	43.49	48.6	38.73	72.96	54.06
e5-base-v2	54.82	43.11	46.90	44.04	53.32	40.83	76.35	56.48
e5-large-v2	52.78	43.93	44.36	43.23	50.1	39.66	74.17	53.48
GTE-base-en-v1.5	57.36	45.86	49.39	43.23	56.83	40.86	78.11	58.54
GTE-large-en-v1.5	59.83	47.14	50.42	47.34	56.23	43.1	80.04	57.99
GTE-modernbert-base	51.35	52.21	54.43	**55.10**	67.11	**49.63**	82.5	**67.44**
LLM Sentence Transformers								
Sentence-T5-xxl	59.76	**53.10**	48.04	48.67	55.29	47.74	76.14	67.41
GTR-T5-xxl	55.77	40.13	51.18	37.31	60.35	44.93	78.27	50.25
e5-mistral-7b-instruct	61.99	38.14	54.82	39.81	**69.32**	42.46	**84.12**	48.86
Supervised								
OpinionCSE	76.94	73.75	67.98	71.94	75.34	58.54	93.31	75.23

Table 3: Performance comparison of models across datasets for ACD and ACSA tasks using entity label as query. Underlined values represent the top models in each category, bolded values indicate the best unsupervised or pre-trained models, colored cells give the best model overall.

Topic	Sentiment	Description
ROOMS	Positive	excellent opinions praising the rooms in terms of size, general condition, view, furniture, bathroom, sleep quality, or availability of extra features / amenities
	Negative	horrible opinions criticizing the rooms for being small, poorly maintained, lacking amenities, or uncomfortable
	Description	opinions evaluating the rooms in terms of their size, general condition, view, furniture, bathroom, sleep quality and the lack or presence of extra features / amenities
ROOM AMENITIES	Positive	excellent opinions praising the amenities in terms of functionality, quality, or availability (e.g. air condition, refrigerator, microwave, mini bar, hair dryer, tv, toiletries, safe, balcony, coffee maker, linen)
	Negative	horrible opinions criticizing the amenities for being non-functional, of poor quality, or missing (e.g. air condition, refrigerator, microwave, mini bar, hair dryer, tv, toiletries, safe, balcony, coffee maker, linen)
	Description	opinions evaluating the rooms in terms of the amenities they include (e.g. air condition, refrigerator, microwave, mini bar, hair dryer, tv, toiletries, safe, balcony, coffee maker, linen)
FACILITIES	Positive	excellent opinions praising the hotel facilities (e.g. swimming pool, spa&sauna, beauty salon, restaurants, café, night club, casino, business center, gymnasium, access facility for the differentlyabled, parking) or guest services (e.g. shuttle, laundry, baby sitting or wake up services, sports activities, 24-hour concierge &front desk, information desk, in-room dining, internet access, availability of touristic material) for being well-maintained, diverse, or convenient
	Negative	horrible opinions criticizing the hotel facilities (e.g. swimming pool, spa&sauna, beauty salon, restaurants, café, night club, casino, business center, gymnasium, access facility for the differentlyabled, parking) or guest services (e.g. shuttle, laundry, baby sitting or wake up services, sports activities, 24-hour concierge &front desk, information desk, in-room dining, internet access, availability of touristic material) for being inadequate, unavailable, or poorly maintained
	Description	opinions focusing on the hotel facilities in terms of specific installations / areas (e.g. swimming pool, spa&sauna, beauty salon, restaurants, café, night club, casino, business center, gymnasium, access facility for the differentlyabled, parking) or guest services offered by a hotel (e.g. shuttle, laundry, baby sitting or wake up services, sports activities, 24-hour concierge &front desk, information desk, in-room dining, internet access, availability of touristic material)
SERVICE	Positive	excellent opinions praising the staff's attitude, promptness, problem-solving ability, or quality of service
	Negative	horrible opinions criticizing the staff's attitude, lack of promptness, inability to solve problems, or poor service quality
	Description	opinions focusing on the staff's attitude and promptness, easiness to problem solving, execution of service in time, or the rooms/ check-in / check-out / reception service
LOCATION	Positive	excellent opinions praising the hotel's location for its convenience, surroundings, or views
	Negative	horrible opinions criticizing the hotel's location for being inconvenient, unattractive, or poorly situated
	Description	opinions focusing on the location of the reviewed hotel in terms of its position, the surroundings, the view
FOOD & DRINKS	Positive	excellent opinions praising the food and drinks for their quality, variety, or presentation
	Negative	horrible opinions criticizing the food and drinks for poor quality, lack of variety, or unappealing presentation
	Description	opinions focusing on the breakfast, the food and the drinks in general or in terms of specific dishes and drinks, dining / drinking options

Table 4: Descriptions provided for each category as query in the HOTEL domain

Topic	Sentiment	Description
FOOD	Positive	excellent opinions praising the food for its exceptional taste, freshness, creative presentation, or diverse menu options
	Negative	horrible opinions criticizing the food for being bland, stale, poorly prepared, or lacking variety
	Description	opinions focusing on the food in general or in terms of specific dishes, dining options
DRINKS	Positive	excellent opinions highlighting the quality, freshness, variety, or creative presentation of drinks
	Negative	horrible opinions complaining about the drinks being poorly prepared, lacking options, or served at an inappropriate temperature
	Description	opinions focusing on the drinks in general or in terms of specific drinks, drinking options
SERVICE	Positive	excellent opinions appreciating the promptness, friendliness, professionalism, or attentiveness of the restaurant staff
	Negative	horrible opinions criticizing the staff for being rude, slow, unprofessional, or inattentive to customer needs
	Description	opinions focusing on the (customer / kitchen / counter) service, on the promptness and quality of the restaurant's service in general, the food preparation, the staff's attitude and professionalism, the wait time, the options offered (e.g. takeout)
AMBIENCE	Positive	excellent opinions praising the atmosphere for being cozy, elegant, lively, or well-decorated with pleasant entertainment options
	Negative	horrible opinions criticizing the ambiance for being noisy, poorly lit, uncomfortable, or unattractive
	Description	opinions focusing on the atmosphere or the environment of the restaurant's interior or exterior space (e.g. terrace, yard, garden), the decor, entertainment options
LOCATION	Positive	excellent opinions highlighting the convenient location, beautiful surroundings, or stunning views of the restaurant
	Negative	horrible opinions criticizing the location for being hard to reach, poorly situated, or lacking appealing surroundings
	Description	opinions focusing on the location of the reviewed restaurant in terms of its position, the surroundings, the view
PRICE	Positive	excellent opinions appreciating the restaurant for offering good value, reasonable pricing, or affordability
	Negative	horrible opinions complaining about the restaurant being overpriced, charging excessively, or not delivering value for money
	Description	opinions that refer to the prices

Table 5: Descriptions provided for each category as query in the RESTAURANT domain