

Shoes-ACOSI: A Dataset for Aspect-Based Sentiment Analysis with Implicit Opinion Extraction

Joseph J. Peper¹ Wenzhao Qiu¹ Ryan Bruggeman²
Yi Han² Estefania Ciliotta Chehade² Lu Wang¹

¹Computer Science and Engineering, University of Michigan, Ann Arbor, MI

²College of Arts, Media and Design, Northeastern University, Boston, MA

{jpeper, qwzhao, wangluxy}@umich.edu

{bruggeman.r, han.yi1, e.ciliottachehade}@northeastern.edu

Abstract

We explore *implicit opinion extraction* as a new component of aspect-based sentiment analysis (ABSA) systems. Prior work in ABSA has investigated opinion extraction as an important subtask, however, these works only label concise, *explicitly*-stated opinion spans. In this work, we present **Shoes-ACOSI**, a new and challenging ABSA dataset in the ecommerce domain with implicit opinion span annotations, the first of its kind. Shoes-ACOSI builds upon the existing Aspect-Category-Opinion-Sentiment (ACOS) quadruple extraction task, extending the task to quintuple extraction—now localizing and differentiating both implicit and explicit opinion. In addition to the new annotation schema, our dataset contains paragraph-length inputs which, importantly, present complex challenges through increased input length, increased number of sentiment expressions, and more mixed-sentiment-polarity examples when compared with existing benchmarks. We quantify the difficulty of our new dataset by evaluating with state-of-the-art fully-supervised and prompted-LLM baselines. We find our dataset presents significant challenges for both supervised models and LLMs, particularly from the new implicit opinion extraction component of the ACOSI task, highlighting the need for continued research into implicit opinion understanding.

1 Introduction

Aspect-based sentiment analysis (ABSA) is an important type of fine-grained sentiment analysis, with critical applications in areas such as healthcare, product review analysis, and argument analysis (Zhang et al., 2022; Han et al., 2023). Popular ABSA subproblems include extracting *aspect terms* and their corresponding *aspect categories*, finding supporting *opinion spans*, and classifying the *sentiment polarity*. In particular, the Aspect-Category-Opinion-Sentiment (ACOS) ABSA task

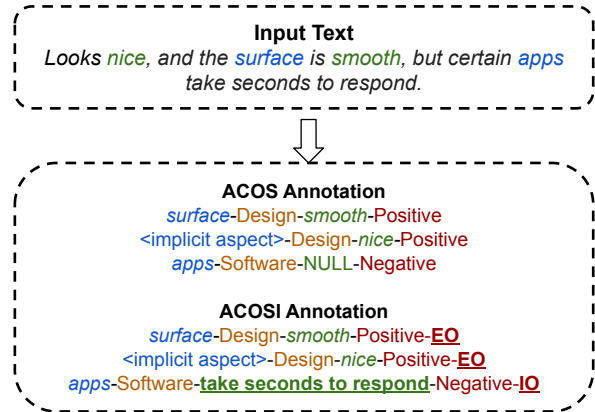


Figure 1: Comparison of ACOSI (*aspect, category, opinion, sentiment, implicit v. explicit opinion flag*) and ACOS (Cai et al., 2021) tasks. ACOSI extends ACOS to additionally extract implicit opinion spans that are localizable. In addition, we add an Implicit/Explicit opinion flag to differentiate between the two. Figure is adapted from Cai et al. (2021).

(Cai et al., 2021; Zhang et al., 2021) has proven challenging, requiring precise extraction of sentiment quadruples from a provided text. Cai et al. (2021) emphasize the need for *implicit* aspect and opinion consideration with their Restaurant-ACOS and Laptop-ACOS datasets, finding a significant proportion of sentiment expressions contain implicit sentiment. Interestingly, these existing datasets consider the existence of *implicit opinion*, however, they simply designate them as ‘null’ text spans, failing to record any meaningful opinion information. We observe that *implicit opinions can be localized within text inputs* (see examples in Fig. 1 and Appendix E). While less an issue for short, sentence-level inputs (one could simply scan the entire input to glean the implicit opinion), opinion span localization is particularly beneficial in settings with long and dense passages where such manual parsing is infeasible.

In this work, we introduce **Shoes-ACOSI**¹, a new ABSA dataset with emphasis on **Implicit** opinion identification and localization. As part of this process, we introduce the **ACOSI** (*aspect, category, opinion, sentiment, implicitness*) quintuple extraction task to the language community, which builds off the existing ACOS quadruple extraction schema. Concretely, our format supports labeling all supporting opinion spans, adding a binary flag to differentiate between the **Implicit** and **Explicit** cases. This schema is backwards-compatible with existing ACOS datasets, and supports richer information extraction.

In addition to providing novel implicit-opinion span annotations, Shoes-ACOSI provides further challenges through its *longer inputs, increased sentiment tuple quantity* (an avg. of 4.3 tuple annotations per example versus ~ 1.6 for the annotations in existing datasets), and many *more mixed-polarity examples* (an avg. of 1.73 unique sentiment polarities/example, versus 1.08 as in previous work). We evaluate our dataset in both the fully supervised setting—using two strong comparison ABSA models—as well as in the few-shot setting with prompted LLMs.

2 Related Work

Many efforts have been made to grow the ABSA task, with significant effort placed into developing datasets (Cai et al., 2021, 2023; Chebolu et al., 2023; Zhang et al., 2021; Zhou et al., 2023) as well as improved methods (Gao et al., 2022; Gou et al., 2023; Hu et al., 2022, 2023; Peper and Wang, 2022; Varia et al., 2023; Xu et al., 2023b) for the burgeoning ACOS task. These works have been instrumental in pushing the boundaries of the ABSA field. Implicit opinion has been an area of research within ABSA (Ouyang et al., 2024; Li et al., 2021). Lazhar and Yamina (2016) and Han et al. (2023) note the need for implicit opinion extraction, with real-world applications including implicit opinion mining and identifying latent needs expressed in customer reviews with the goal of improving product designs. However, ABSA datasets which identify implicit opinion ignore the task of *extracting* localized spans (Cai et al., 2021), introducing a gap between existing evaluation benchmarks and real-world ABSA applications. Our work addresses this gap by expanding the scope of ABSA to distin-

guish and extract both explicit and implicit opinions within text.

3 Shoes-ACOSI Dataset

Shoes-ACOSI is the first ABSA dataset to include implicit opinion span annotations. Below we outline the dataset, curation process, and notable characteristics. For the purposes of this task, we define an implicit sentiment opinion as a span containing nuanced and potentially ambiguous sentiment language whose meaning can be resolved only via thorough contextualization from the domain and full input.

3.1 ACOSI Task

We introduce the ACOSI sentiment task (Figure 1), which extends ACOS in two key ways:

- Implicit opinions are (when possible) localized to a sub-span of the input text. As such, we support both explicit and implicit opinion span extraction.
- To differentiate between explicit and implicit opinions, an implicit opinion indicator (i.e. the ‘I’ term) is added as the fifth tuple element.

3.2 Data Collection + Annotation Process

Shoes-ACOSI is compiled from several thousand product reviews sourced from online footwear retailer websites². Reviews were filtered via simple length heuristics (keeping those with 2-5 sentences) and sampled uniformly by review star rating. We hired native-English-speaking product design students to perform the annotation task, assigning two annotators per example and using a project lead to resolve any conflicts. Using Cohen’s Kappa (Cohen, 1960) as our inter-annotator agreement metric, we observed moderate 52.8% chance-adjusted agreement on the categorical components, and 40.8% on the span extraction tasks (aspect and opinion) between annotators. Each sample is annotated in the ACOSI format, producing a set of (*aspect term, aspect category, opinion, sentiment polarity, implicit/explicit opinion flag*) ground-truth quintuples. Annotators were instructed to annotate implicit opinion spans by labeling the minimal supporting span which covered the implicit sentiment expression for the current tuple. If no localization

¹Dataset is available at https://github.com/jpeper/shoes_acosi.

²<https://www.asics.com/>
<https://www.newbalance.com/>
<https://www.finishline.com/>

was possible (i.e., the full input is needed) then the span was marked as ‘null’. See Appendix B for more details on the annotation process and Appendix E for examples of opinion spans from our dataset.

3.3 Dataset Characteristics

Table 1 displays key dataset characteristics and compares with the popular Restaurant-ACOS and Laptop-ACOS datasets (Cai et al., 2021). We observe the following characteristics that stand out in comparison to Restaurant and Laptop-ACOS:

- **Longer inputs and more tuples per sample:** The Shoes-ACOSI average input length is 2.3x longer than Restaurant and Laptop, and also have significantly more tuples per example.
- **Higher concentration of implicit aspect+opinion expressions:** Shoes-ACOSI contains higher proportions of tuples with implicit sentiment phenomena. This is likely attributable to the domain (shoe reviews are inherently very subjective and contain diverse opinions+perspectives), and the review sampling process (using only longer multi-sentence reviews which are more nuanced than short reviews).
- **Increased tuple diversity:** We sample reviews uniformly by star rating (1-5), surfacing a wide range of review sentiment polarities. Notably, Shoes-ACOSI contains an average 1.73 unique sentiment polarities (+/-/=) per example, whereas the Restaurant and Laptop examples are largely homogeneous in polarity (~ 1.1 unique sentiments per example).

4 Evaluation Setup

We benchmark our dataset, comparing two fully supervised ABSA models that perform strongly on the ACOS task, as well as few-shot prompted LLMs using an ABSA-specific prompting method.

Baseline Models

- **MvP** (Gou et al., 2023) is an ABSA-specific supervised model which addresses limitations with autoregressive decoding by aggregating outputs over several unique inference ‘views’, with each view prompting the model to produce tuples with a different ordering (e.g.

$\langle a,c,o,s,i \rangle$, or $\langle c,a,o,i,s \rangle$). We analyze both **MvP-unified**, which is jointly trained on several ABSA tasks (including ours), and **MvP-main**, which is trained solely on our desired task.

- **GEN-SCL-NAT** (Peper and Wang, 2022) An ACOS-specific supervised model, combines a novel structured generation format with a supervised contrastive learning objective designed to improve implicit aspect and opinion handling.
- **OpenAI GPT-3.5³, GPT-4.0⁴**. We explore LLM performance in the few-shot prompting scenario, providing a task description and k labeled exemplars as in-context learning demonstrations. We follow Xu et al. (2023a), who find a lightweight tf-idf KNN-based similarity heuristic is effective in sampling appropriate ABSA demonstrations from the training set. We use $k = 10$ in our experiments.

For the supervised models (MvP, GEN-SCL-NAT), we modify their output target formats slightly to account for the new Implicit/Explicit flag that forms the fifth ACOSI tuple component. Appendix C and D.1.1 respectively contain more details for the fully supervised and LLM setups.

Evaluation Metrics We evaluate on the task of exact match tuple extraction (ACOSI extraction for our new dataset and ACOS extraction when comparing with existing datasets). A predicted tuple is deemed correct only if it exactly matches with a ground-truth tuple. We report precision, recall and F1 metrics, comparing the predicted set of tuples with the ground truth set.

5 Results & Analysis

Table 2 reports model performance on Shoes-ACOSI. To directly compare model performance with the existing ACOS datasets, we also create a simplified *Shoes-ACOS* dataset variant, removing the implicit opinion span annotations and simply marking them as ‘null’ spans. We report the ACOS comparison in Table 3.

Shoes Results The Shoes domain proves challenging for all models; for Shoes-ACOSI (Table 2), *MVP-unified performs the best* (17.24 quintuple

³<https://platform.openai.com/docs/models/gpt-3-5>

⁴<https://platform.openai.com/docs/models/gpt-4>

	Restaurant	Laptop	Shoes
Dataset Format	ACOS	ACOS	ACOSI
Num. Examples	2,284	4,076	1,147
Tokens / Example	15.1 (9.8)	15.7 (9.9)	37.4 (19.4)
Tuples / Example	1.6 (1.1)	1.4 (0.8)	4.3 (1.8)
Total Tuples	3,661	5,773	4,877
IA/IO Tuples	9.56%	5.92%	16.67%
IA/EO Tuples	14.48%	15.80%	33.71%
EA/IO Tuples	9.56%	21.50%	13.55%
EA/EO Tuples	66.40%	56.78%	36.07%
Pos. Tuples	68.37%	61.98%	37.52%
Neu. Tuples	4.12%	5.47%	11.11%
Neg. Tuples	27.51%	32.55%	51.36%
Avg. Polarities / Example	1.08 (0.27)	1.05 (0.22)	1.73 (0.68)

Table 1: ACOS* Dataset Statistics. On average, Shoes-ACOSI examples have more sentiment tuples, more implicit sentiment expressions, and are more likely to be mixed-polarity than existing implicit-labeled ACOS datasets.

Dataset		Shoes-ACOSI		
Score		f1	precision	recall
MvP	main	16.92	18.60	15.52
	unified	17.24	18.94	15.83
GEN-SCL-NAT		15.21	16.03	14.48
GPT	3.5	10.98	11.16	10.81
	4	14.14	14.61	13.71

Table 2: Model performance on ACOSI extraction task. We report the metrics obtained from exact quintuple match. **Bold** and underline respectively refer to the **best** and **second-best** performers.

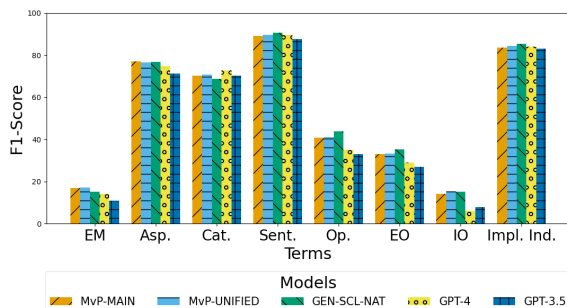


Figure 2: Component-level performance breakdown for Shoes-ACOSI dataset. We aggregate all predicted and ground-truth values for each individual ACOSI component, performing set-wise comparison between the predicted and ground-truth sets. For further opinion analysis we consider three different breakdowns: **Opinion** considers general opinion span extraction performance, *ignoring the implicit/explicit label correctness*. In contrast, **IO** and **EO** *do* consider the spans’ implicit/explicit opinion labels, and predictions are correct only if the both the predicted span and its implicit/explicit flag match a ground-truth tuple. **Impl.Ind.** refers to the implicit/explicit opinion indicator flag. **EM** refers to overall exact match performance for reference.

F1), perhaps due to its joint-task training that encourages faster task adaption. The *LLMs struggle noticeably with this task*, with GPT 3.5 obtaining only 10.98 F1. As expected, all models perform better on the simpler Shoes-ACOS (Table 3) than Shoes-ACOSI, GEN-SCL-NAT having the largest delta of 3.9 points. Interestingly, MvP-main outperforms MVP-unified on the Shoes-ACOS formulation, perhaps due to the relative ease of the task requiring less general ABSA prowess. The LLM methods struggle similarly on both tasks, with GPT 3.5 and 4.0 respectively performing 0.15 and 0.79 points worse on ACOSI than ACOS.

ACOSI Component-level Analysis Figure 2 outlines the Shoes-ACOSI exact match performance broken down per ACOSI component. We see *opinion extraction is especially challenging, particularly so for implicit opinions which are longer in nature*. Notably, *LLMs struggle heavily with the exact match implicit opinion extraction task (IO)*, likely due to seeing only 10 in-context demonstrations, while the supervised models (though still struggling) perform better. Interestingly, we see the Category component is handled well by prompted LLMs, where providing the category names in-context yields strong performance (refer to Appendix D.1.1 for the LLM prompts used).

Dataset Comparison (ACOS) We compare the ACOS datasets in Table 3. We see all models struggle heavily on Shoes-ACOS compared with the Restaurant and Laptop datasets. As outlined in Section 3.2, our domain and input setup present a number of unique challenges due to relative complexity of the input, with the shoes domain having 1) longer inputs 2) more tuples per example, and 3) much more diverse sentiment polarities within

Datasets		Restaurant-ACOS			Laptop-ACOS			Shoes-ACOS		
Score		f1	precision	recall	f1	precision	recall	f1	precision	recall
MvP	main	59.86	60.72	59.02	44.09	44.57	43.62	19.66	21.36	18.21
	unified	59.74	60.05	59.43	44.01	44.25	43.77	19.22	20.56	18.06
GEN-SCL-NAT		62.26	63.32	61.24	46.17	46.68	45.67	19.09	19.55	18.64
GPT	3.5	43.54	41.63	45.63	28.26	26.69	30.02	11.13	11.41	10.87
	4	51.13	49.14	53.28	32.75	31.81	33.74	14.93	15.77	14.17

Table 3: Model performance on ACOS extraction task. We report the F1, precision, and recall scores obtained from exact tuple match. **Bold** and underline respectively refer to the **best** and second-best performers.

examples.

Impact of Implicit Sentiment on Overall Sentiment Polarity

Given its significance to this task, we explore the influence of implicit sentiment tuples on the document-level sentiment distribution observed in our dataset. First, we calculate the average sentiment at the review level, using a simple mean of the constituent sentiment tuples (with -1 representing negative, 0 for neutral, and +1 for positive sentiment). For the full Shoes-ACOSI dataset, the average review-level sentiment is -0.126. When implicit opinion tuples were excluded from this calculation (i.e., considering only explicit sentiments), the average sentiment shifted to -0.003, indicating *explicit tuples alone convey a relative balanced sentiment*. In contrast, excluding explicit opinion tuples and considering only implicit sentiments resulted in a more negative average sentiment of -0.214, suggesting that *implicit sentiment tuples capture a stronger degree of negative sentiment than explicit ones* in our dataset. This noticeable distribution shift highlights the significance of capturing subtle implicit language phenomena.

6 Conclusion

Shoes-ACOSI and the ACOSI task introduce a new and challenging problem for the aspect-based sentiment analysis community. Notably, Shoes-ACOSI introduces the challenge of implicit opinion extraction, which poses problems for both fully supervised models and, in particular, LLM approaches which are confined to few-shot in-context demonstrations. Our dataset presents a novel challenge and motivates a new and important direction of research for aspect-based sentiment analysis.

Limitations

Though our emphasis was on the ACOSI and ACOSI-adjacent (ACOS) tasks, it is worth exploring other ABSA task formulations such as Aspect-Sentiment-Triplet Extraction (ASTE) (Peng et al., 2020) or Aspect-Opinion-Pair-Extraction (AOPE) (Yu et al., 2019) using sub-components of our Shoes dataset. Our evaluation was also focused on two conventional evaluation settings—supervised model fine-tuning, and prompting-based LLM evaluation. LLM fine-tuning is certainly a direction worth exploring, however, we deprioritized this due to compute constraints.

Acknowledgements

This work is supported in part by the National Science Foundation through grants CMMI-2050130 and IIS-2046016. We thank the ARR reviewers for their valuable comments.

References

- Hongjie Cai, Nan Song, Zengzhi Wang, Qiming Xie, Qiankun Zhao, Ke Li, Siwei Wu, Shijie Liu, Jianfei Yu, and Rui Xia. 2023. [Memd-absa: A multi-element multi-domain dataset for aspect-based sentiment analysis](#).
- Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. [Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions](#). 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 340–350, Online. Association for Computational Linguistics.
- Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Tamar Solorio. 2023. [Oats: Opinion aspect target sentiment quadruple extraction dataset for aspect-based sentiment analysis](#).
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20:37 – 46.

- Tianhao Gao, Jun Fang, Hanyu Liu, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Yongjun Bao, and Weipeng Yan. 2022. [LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7002–7012, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. [Mvp: Multi-view prompting improves aspect sentiment tuple prediction](#).
- Yi Han, Ryan Bruggeman, Joseph Peper, Estefania Ciliotta Chehade, Tucker Marion, Paolo Ciuccarelli, and Mohsen Moghaddam. 2023. [Extracting latent needs from online reviews through deep learning based language model](#). *Proceedings of the Design Society*, 3:1855–1864.
- Chengwei Hu, Deqing Yang, Haoliang Jin, Zhen Chen, and Yanghua Xiao. 2022. [Improving continual relation extraction through prototypical contrastive learning](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1885–1895, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Mengting Hu, Yin hao Bai, Yike Wu, Zhen Zhang, Liqi Zhang, Hang Gao, Shiwan Zhao, and Minlie Huang. 2023. [Uncertainty-aware unlikelihood learning improves generative aspect sentiment quad prediction](#).
- Farek Lazhar and Tlili Yamina. 2016. [Mining explicit and implicit opinions from reviews](#). *International Journal of Data Mining, Modelling and Management*, 8:75.
- 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. Association for Computational Linguistics.
- Jihong Ouyang, Zhiyao Yang, Silong Liang, Bing Wang, Yimeng Wang, and Ximing Li. 2024. [Aspect-based sentiment analysis with explicit sentiment augmentations](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18842–18850.
- Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. [Knowing what, how and why: A near complete solution for aspect-based sentiment analysis](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8600–8607.
- Joseph J. Peper and Lu Wang. 2022. [Generative aspect-based sentiment analysis with contrastive learning and expressive structure](#).
- Siddharth Varia, Shuai Wang, Kishaloy Halder, Robert Vacareanu, Miguel Ballesteros, Yassine Benajiba, Neha Anna John, Rishita Anubhai, Smaranda Muresan, and Dan Roth. 2023. [Instruction tuning for few-shot aspect-based sentiment analysis](#).
- Xiancai Xu, Jia-Dong Zhang, Rongchang Xiao, and Lei Xiong. 2023a. [The limits of chatgpt in extracting aspect-category-opinion-sentiment quadruples: A comparative analysis](#).
- Xiancai Xu, Jia-Dong Zhang, Lei Xiong, and Zhishang Liu. 2023b. [iacos: Advancing implicit sentiment extraction with informative and adaptive negative examples](#).
- Jianfei Yu, Jing Jiang, and Rui Xia. 2019. [Global inference for aspect and opinion terms co-extraction based on multi-task neural networks](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):168–177.
- Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. [Aspect sentiment quad prediction as paraphrase generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. [A survey on aspect-based sentiment analysis: Tasks, methods, and challenges](#).
- Junxian Zhou, Haiqin Yang, Ye Junpeng, Yuxuan He, and Hao Mou. 2023. [An empirical study of benchmarking chinese aspect sentiment quad prediction](#).

A Shoes-ACOSI Category Ontology

The category ontology of the Shoes dataset (as provided to annotators) is shown in Figure 3. This is provided as reference during the annotation process.

B Annotation Procedure

In addition to the category ontology, we provide annotators with an annotation guidebook containing instructions on labeling for the ACOSI task (Figure 4 and 5).

C Supervised Model Details

We use GEN-SCL-NAT and MvP as baseline supervised models for our evaluation.

- **MvP** (Gou et al., 2023), a 220M parameter (T5-base base model), addresses the limitations of autoregressive structured prediction through **Multi-view Prompting**, prompting

the model to generate structured ABSA tuples in several component "views" (e.g. possible tuple element orderings). The $k = 5$ diverse views are then aggregated via majority vote, including any generated tuples which appear in the majority of sampled views. MvP achieves the strongest performance on the ACOS tasks amongst comparable fully-supervised models, albeit with higher compute cost as it requires sampling and aggregating five different outputs. As the MvP-unified model is co-trained on several ABSA tasks, we simply incorporated our ACOSI dataset within these during training.

- **GEN-SCL-NAT** (Peper and Wang, 2022), a 770M parameter (T5-large base model) model which performs structured ACOS generation, combining a novel structured generation format with a supervised contrastive learning objective during training. It utilizes two key techniques for the structured generation of ACOS quadruples: GEN-SCL, a supervised contrastive learning objective, and GEN-NAT, a novel structured generation format.

C.1 Modifications for ACOSI-Extract

We slightly modify the output targets for each model in order to adapt them to the ACOSI-Extract task (Fig. 6 and 7).

D LLM Evaluation Details

We utilize the popular OpenAI GPT-3.5 and GPT-4 models in our LLM evaluation. For both, we use the '0613' model checkpoints. We briefly explored using LLaMA-2-13B⁵, a popular open-source large language model developed by Meta; however, we found it performed extremely poorly on the ACOS and ACOSI tasks, with numerous formatting and category hallucination issues causing many invalid outputs.

D.1 Few-shot Prompting Method(s)

As described in Section 4, we adapt the few-shot prompting method developed by Xu et al. (2023a). There are two variations of this method; in the KNN variant, we sample and demonstrate k similar reviews from the training set based on tf-idf similarity. As seen in Xu et al. (2023a), we found this method noticeably outperforms a simple random

example selection. To ensure a fair comparison across datasets with varying sizes, we standardize the magnitude of the pool of available training examples across all datasets. Namely, for both the Restaurant-ACOS and Laptop-ACOS datasets, we randomly select 906 training as the sampling pool on which we run the similarity-driven sampling, matching the size of the Shoes-ACOSI train set.

D.1.1 Example Prompts

The prompt used for the ACOS task is displayed in Figure 8. We adapt this for ACOSI, slightly modifying the generation targets for each model (Figure 9).

E Opinion Comparison

We analyze opinion spans across different datasets in Fig. 10. We see the (explicit) opinion spans from the Laptop and Restaurant-ACOS datasets are generally brief, typically just one token. In contrast, the Shoes dataset features implicit opinion spans which are longer and more opaque. Additionally, we see explicit opinion spans within the Shoes dataset also tend to be slightly longer on average than those in Laptop and Restaurant.

⁵<https://llama.meta.com/llama2>

Category Glossary:	
Appearance: The aesthetic appearance of the shoe; the way that someone or something looks.	Material: The quality or state of being material; an product quality relating to, derived from, or consisting of matter.
Color: A quality such as red, blue, green, yellow, etc., that you see when you look at something; visual perception that enables one to differentiate otherwise identical objects.	Miscellaneous (Misc.): A specific category, not yet defined, that doesn't fall into the other categories.
Construction: The process, art, or manner of constructing a product; to make, or form, by combining or arranging parts or elements.	Place: Where the product is being used; the user's environment.
Context of Use: To do something with (an object, machine, person, method, etc.) in order to accomplish a task, do an activity, etc.; the situation in which something happens; the group of conditions that exist where and when something happens.	Purchase Context: Other factors that led the user to purchasing the shoe.
Comfort: Producing or affording physical comfort, support, or ease.	Review Temporality: A statement on the product and the user's time of use with it.
Cost/Value: The price paid to acquire, produce, accomplish, or maintain anything; relative worth, merit, or importance.	Shoe Component: A constituent part; one of the parts that form something.
Durability: Able to exist for a long time without significant deterioration in quality or value.	Stability: The quality, state, or degree of being stable; steady in purpose; firmly established, not changing or fluctuating.
Fit: The quality of a distinct object or body of specific form or figure to be sound physically.	Style Choice: The amount of style options that a product has. Is there a sufficient amount of styles to choose from or is there only one option?
Form: A particular form or shape of an object, the spatial form or contour of the object.	Usage Frequency: Any manner of doing or handling something; treatment; the state or fact of being frequent; frequent occurrence.
Functional Applicability: To perform a specified action or activity; work; operate.	Use Case: For what is the product used for.
General: A category or subcategory that captures the essence of something, but does not have a definable category name that can express it. The category or subcategory is implied.	Versatility: The state or quality of being useful for or easily adapted to various tasks, styles, fields of endeavor.

Category Format Options:	Context_of_Use#																
<table border="0"> <tr> <td>Performance#</td> <td>Appearance#</td> </tr> <tr> <td>Performance#Fit</td> <td>Appearance#General</td> </tr> <tr> <td>Performance#General</td> <td>Appearance#Form</td> </tr> <tr> <td>Performance#Comfort</td> <td>Appearance#Color</td> </tr> <tr> <td>Performance#Durability</td> <td>Appearance#Material</td> </tr> <tr> <td>Performance#Stability</td> <td>Appearance#Shoe_Component</td> </tr> <tr> <td>Performance#Functional_Applicability</td> <td>Appearance#Misc</td> </tr> <tr> <td>Performance#Misc</td> <td></td> </tr> </table>	Performance#	Appearance#	Performance#Fit	Appearance#General	Performance#General	Appearance#Form	Performance#Comfort	Appearance#Color	Performance#Durability	Appearance#Material	Performance#Stability	Appearance#Shoe_Component	Performance#Functional_Applicability	Appearance#Misc	Performance#Misc		<ul style="list-style-type: none"> • Usage_Frequency <ul style="list-style-type: none"> ◦ "I use them a couple times a week." <ul style="list-style-type: none"> ■ (Context_of_Use#Usage_Frequency, NEU, week, I use them a couple times) • Use_Case <ul style="list-style-type: none"> ◦ "For running." <ul style="list-style-type: none"> ■ (Context_of_Use#Use_Case, NEU, running, For running) • Place <ul style="list-style-type: none"> ◦ "I use them on the trails outdoors." <ul style="list-style-type: none"> ■ (Context_of_Use#Place, NEU, trails outdoors, I use them on the trails outdoors.) • Review_Temporality <ul style="list-style-type: none"> ◦ "Still working them in." <ul style="list-style-type: none"> ■ (Context_of_Use#Review_Temporality, NEU, Still working them, Still working them in.) • Purchase_Context <ul style="list-style-type: none"> ◦ "Dad bought me these for xmas." <ul style="list-style-type: none"> ■ (Context_of_Use#Purchase_Context, POS, [Dad][xmas], bought me these)
Performance#	Appearance#																
Performance#Fit	Appearance#General																
Performance#General	Appearance#Form																
Performance#Comfort	Appearance#Color																
Performance#Durability	Appearance#Material																
Performance#Stability	Appearance#Shoe_Component																
Performance#Functional_Applicability	Appearance#Misc																
Performance#Misc																	
<table border="0"> <tr> <td>Context_of_Use#</td> <td>General#</td> </tr> <tr> <td>Context_of_Use#Usage_frequency</td> <td></td> </tr> <tr> <td>Context_of_Use#Use_Case</td> <td>Cost_Value#</td> </tr> <tr> <td>Context_of_Use#Place</td> <td></td> </tr> <tr> <td>Context_of_Use#Review_Temporality</td> <td>Misc#</td> </tr> <tr> <td>Context_of_Use#Purchase_Context</td> <td></td> </tr> </table> <p>Annotation Structure: (Category#Category, Sentiment, Aspect Term, Opinion Term) <category>'s sentiment is <sentiment> because <aspect term> is <opinion>.</p>	Context_of_Use#	General#	Context_of_Use#Usage_frequency		Context_of_Use#Use_Case	Cost_Value#	Context_of_Use#Place		Context_of_Use#Review_Temporality	Misc#	Context_of_Use#Purchase_Context		<p>Appearance#</p> <ul style="list-style-type: none"> • General <ul style="list-style-type: none"> ◦ "Looks tacky." <ul style="list-style-type: none"> ■ (Appearance#General, NEG, NULL, tacky) • Form <ul style="list-style-type: none"> ◦ "Shape is like another shoe." <ul style="list-style-type: none"> ■ (Appearance#Form, NEU, Shape, like another shoe) • Color <ul style="list-style-type: none"> ◦ "Needs to be red!" <ul style="list-style-type: none"> ■ (Appearance#Color, NEU, red!, Needs to be) • Material <ul style="list-style-type: none"> ◦ "Fabric is super smooth on my feet." <ul style="list-style-type: none"> ■ (Appearance#Material, POS, Fabric, super smooth) • Shoe_Component <ul style="list-style-type: none"> ◦ "The sole is so cool." <ul style="list-style-type: none"> ■ (Appearance#Shoe_Component, POS, sole, so cool) • Misc <ul style="list-style-type: none"> ◦ "Walking down the street my feet look fly!" <ul style="list-style-type: none"> ■ (Appearance#Misc, POS, my feet, look fly) <p>Cost_Value#</p> <ul style="list-style-type: none"> • No attribute <ul style="list-style-type: none"> ◦ "Was it worth the price?" <ul style="list-style-type: none"> ■ (Cost_Value#, NEU, price, Was it worth the price?) <p>Misc#</p> <ul style="list-style-type: none"> • No attribute <ul style="list-style-type: none"> ◦ "The way I see it, I should be able to catch my balance." <ul style="list-style-type: none"> ■ (Misc#, NEG, [I][my], [The way I see it, I should be able to catch my balance.]) 				
Context_of_Use#	General#																
Context_of_Use#Usage_frequency																	
Context_of_Use#Use_Case	Cost_Value#																
Context_of_Use#Place																	
Context_of_Use#Review_Temporality	Misc#																
Context_of_Use#Purchase_Context																	

Performance#
<ul style="list-style-type: none"> • Fit <ul style="list-style-type: none"> ◦ "Nice and snug when I run." <ul style="list-style-type: none"> ■ (Performance#Fit, POS, snug, Nice and snug) • General <ul style="list-style-type: none"> ◦ "Sports feel better!" <ul style="list-style-type: none"> ■ (Performance#General, POS, Sports, better!) • Comfort <ul style="list-style-type: none"> ◦ "Don't hurt when I play basketball." <ul style="list-style-type: none"> ■ (Performance#Comfort, POS, play basketball, Don't hurt) • Durability <ul style="list-style-type: none"> ◦ "I have run 24 marathons and no rips." <ul style="list-style-type: none"> ■ (Performance#Durability, POS, 24 marathons, no rips.) • Stability <ul style="list-style-type: none"> ◦ "My foot was wobbling everywhere during my match!" <ul style="list-style-type: none"> ■ (Performance#Stability, NEG, foot, wobbling everywhere) • Functional_Applicability <ul style="list-style-type: none"> ◦ "Nice and snug when I run." <ul style="list-style-type: none"> ■ (Performance#Functional_Applicability, POS, run, Nice and snug) • Misc <ul style="list-style-type: none"> ◦ "Walk along with others." <ul style="list-style-type: none"> ■ (Performance#Misc, POS, Walk, with others) <p>General#</p> <ul style="list-style-type: none"> • No attribute <ul style="list-style-type: none"> ◦ "These are crazy awesome!" <ul style="list-style-type: none"> ■ (General#, POS, These, crazy awesome!)

Figure 3: Shoes-ACOSI Category Ontology

Introduction:

Our goal in this work is to take online user product reviews, in this case discussing shoes, and annotate the reviews in such a way that we can identify the users sentiment and that sentiment's relationship to an aspect of that shoe. The following begins with a description of what the annotation procedure we are employing is and the four components that it is made of (hereby referred to as *quadruple*): aspect term, opinion term, category, and sentiment. Included in these descriptions will be the conventions that each of these components follow. The remainder of the document will consist of a step by step process by which the annotating of the reviews will be done and the quadruple comes together to identify aspect based user sentiment. Additionally, there is a glossary of category definitions included and a list of formatting options that each category can have.

Descriptions:

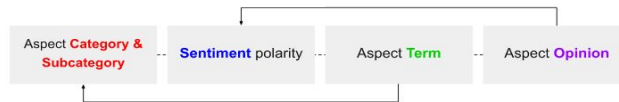
AOCS Annotation: The annotation schema that we are employing is a quadruple, meaning four component, procedure. The annotation components that make up AOCS are aspect term, opinion term, category, and sentiment. The goal of AOCS is to break down the language of a review in such a way that we are able to identify an aspect of a product, in this case shoes, and the sentimental relationship that the user has with it. AOCS begins by identifying an *aspect term* of a product mentioned in the review, as well as an *opinion term* that is found in that review. The annotator then identifies which category the aspect term is being used in this review and whether it has a positive, negative, or neutral sentiment which is identified by the opinion term.

Aspect Term: Denotes an entity and its aspect and is the opinion term target. This will be a term with an explicit reference to an aspect category. The aspect term will consist of a *noun word* (ex. person i.e. Dad, place i.e. outside, or thing i.e. color), a *noun phrase* (a string of noun words and adjectives i.e. blue running shoe), or a *verb* (i.e. running). If an aspect term cannot be identified in a review sentence it will be identified as IMPLICIT.

Opinion Term: Refers to the subjective statement on an aspect term, which is normally a *subjective word* (ex. good) or *phrase* (ex. so narrow I couldn't get my foot in them). Additionally, the opinion term can be either direct or indirect. A direct opinion term is that which has an explicit term that can be identified (in this case highlighted in yellow), the aspect term is contained in square brackets, "[the shoe] is really good." An indirect opinion term is that which is implied. "[This shoe] reduced my pain significantly." For the indirect case there was no identifiable word or phrase explicitly stated that signaled the opinion of the shoe, but as a reader we know that the opinion is positive as the shoe reduced the user pain, which we think is a good thing, and is thus to the annotator an implied, and indirect, opinion.

Category: Represents a unique predefined category for the aspect term to be represented by. A category is part of the ontology of the product that the annotator must decide the aspect term is associated with. For example, "The sole of the shoe has great traction for running." We begin by identifying the aspect terms [sole of the shoe], [traction]; next we can identify the category that these aspect terms are representing, and as annotator the choice is yours, so in this case we can say it is a part of the *performance* category, and the subcategory *functional applicability* (*performance#functional_applicability*). Many times a review can't be represented by just one category and must be annotated more than once. For example if we now had, "The sole of the shoe has great traction for running.", we still have the same aspect terms and can include our original category (*performance#functional_applicability*), but instead of stopping there we now have the "for running" portion of the review and thus must include an additional category (*context_of_use#use_case*).

Sentiment: Is the predefined semantic orientation (i.e., Positive, Negative, or Neutral) toward the aspect term. The annotator identifies the sentiment through the use of the opinion term, for example, "the sole of the shoe has terrible traction." In this case the opinion term being the direct [terrible] lets us know that the sentiment is *negative*. We can easily imagine what a positive sentiment may look like, but on the other hand, for the case of neutral sentiment, we cannot identify a direct or indirect opinion term, or if there is an opinion term it doesn't say anything about the users sentiment towards the aspect term, "the shoe absorbs rocks when I use them to run." This is important for annotation because the sentence gives context to the shoe and the use of it, but the sentiment associated with the sentence does not itself tell us whether this is a good or bad thing.



Annotation Process:

Note: Please see Inception annotation interface overview at bottom of this document for guidance.

Annotation Server URL:

- You will be assigned a series of shoe reviews, each broken down to the sentence level.
- For each sentence you annotate, you will identify user sentiment expressions and their attachments to aspect terms by generating the AOCS quadruple for each expression you see. This means that you may have to annotate a sentence more than once.

1 When I received the shoes were so narrow I could not even get my foot into them.

- Begin by identifying the aspect term.

1 When I received the shoes were so narrow I could not even get my foot into them.

- Now, identify the opinion term.

1 When I received the shoes were so narrow I could not even get my foot into them.

In our case we have an *indirect opinion phrase*. The string of words from "so...them" has been identified as containing the opinion term.

- You will now click and hold the aspect term and drag an arrow to the opinion term to create a sentiment connection.

1 When I received the shoes were so narrow I could not even get my foot into them.

- In the right hand column identify the category that the sentence is related to.

Category
Performance#Sizing/Fit x

1 When I received the shoes were so narrow I could not even get my foot into them.

- In the right hand column identify the sentiment of the sentence.

Polarity
Negative
Neutral
Positive

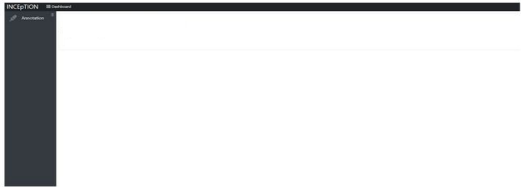
1 When I received the shoes were so narrow I could not even get my foot into them.

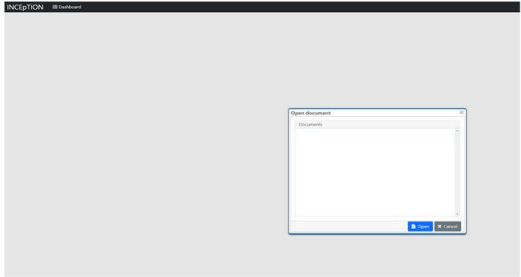
- Once the sentiment component is complete you are done and can move on to the next sentence, beginning the procedure with step (3).

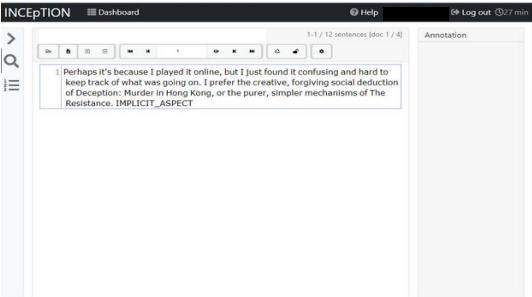
Figure 4: Shoes-ACOSI Annotation Guidebook, Pt. 1

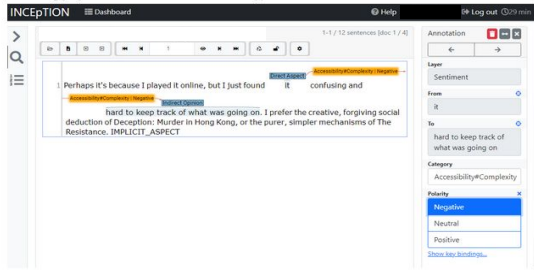
Annotation Interface Walkthrough

1. Main page after login. Click 'Annotation' to view files-to-be-annotated

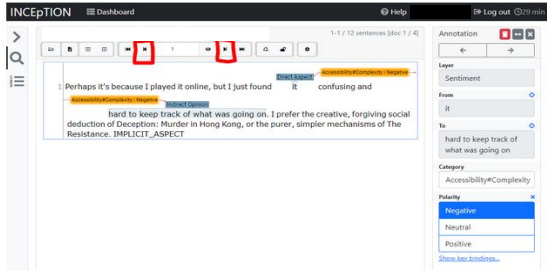

2. Select an assign file and press 'Open'


3. Example review


4. Drag and highlight a span of text to identify an aspect or opinion term, and then label the text span accordingly by selection one of the 'Mention Types' in the sidebar


5. To label a quadruple, you must:

 - Identify the aspect category (e.g. "Accessibility#Complexity")
 - The sentiment associated with that aspect (e.g. "Negative")
 - Any keyword referring to the product or category (e.g. "it")
 - The associated opinion term that justifies the user's sentiment (e.g. "hard to keep track of what was going on" implies the sentiment is negative)
6. The process of doing this in the annotation tool is as follows:

 - Highlight and label an aspect term and an opinion term
 - Draw a connection between the two terms by holding and dragging
 - Then, assign a Polarity and a Category to this link (as seen in the sidebar of this figure)
7. Navigate between reviews using the left and right arrows (circled in red above)

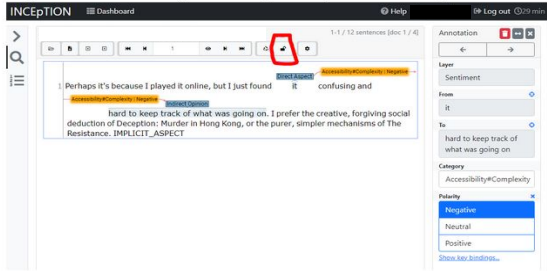

8. After completing all of the annotations in the assigned document, press the lock button to submit your annotations for review.

Figure 5: Shoes-ACOSI Annotation Guidebook, Pt. 2

Input: new balance did great with the design of these shoes . they look great in all the color combos they released . i love that they ' re low key and not crazy neon colors that aren ' t practical for street wear .

ACOS Output: [A] shoes [C] appearance general [S] positive [O] great with the design [SSEP]
[A] NULL [C] appearance color [S] positive [O] look great in all the color combos [SSEP]
[A] NULL [C] contextofuse use case [S] positive [O] NULL [SSEP]
[A] neon colors [C] appearance color [S] positive [O] NULL

ACOSI Output: [A] shoes [C] appearance general [S] positive [O] great with the design [I] explicit [SSEP]
[A] NULL [C] appearance color [S] positive [O] look great in all the color combos [I] explicit [SSEP]
[A] NULL [C] contextofuse use case [S] positive [O] i love that they ' re low key and not crazy [I] implicit [SSEP]
[A] neon colors [C] appearance color [S] positive [O] i love that they ' re low key and not crazy [I] implicit

Figure 6: MvP Output Target Format Example. Note: These are the resultant tuples after the MvP multi-view aggregation is performed.

Input: new balance did great with the design of these shoes . they look great in all the color combos they released . i love that they ' re low key and not crazy neon colors that aren ' t practical for street wear .

ACOS Output: the appearance overall | the shoes is great with the design | pos [SEP]
the color | the it is look great in all the color combos | pos [SEP]
the use case | the it is i love that they ' re low key and not crazy | pos [SEP]
the color | the neon colors is i love that they ' re low key and not crazy | pos

ACOSI Output: the appearance overall | the shoes is great with the design EXPLICIT | pos [SEP]
the color | the it is look great in all the color combos EXPLICIT | pos [SEP]
the use case | the it is i love that they ' re low key and not crazy IMPLICIT | pos [SEP]
the color | the neon colors is i love that they ' re low key and not crazy IMPLICIT | pos

Figure 7: GEN-SCL-NAT Output Target Format Example.

```

Instruction: Extract aspect-category-sentiment-opinion quadruples from
input data.
Context: An aspect or opinion must be a term existing in input data or
null if non-existing; the category is one in the predefined list
['appearance color', 'appearance form', 'appearance general',
'appearance material', 'appearance misc', 'appearance shoe component',
'contextofuse place', 'contextofuse purchase_context', 'contextofuse
review_temporality', 'contextofuse usage frequency', 'contextofuse use
case', 'cost/value', 'general', 'misc', 'performance comfort',
'performance durability', 'performance general', 'performance misc',
'performance sizing/fit', 'performance support/stability',
'performance use case applicability', 'versatility']; the sentiment is
positive, negative or neutral; do not ask me for more information, I
am unable to provide it, and just try your best to finish the task.
You can learn from the following examples.
Output format: (aspect, category, sentiment, opinion)
Input: it ' s really comfortable and fitting for my feet . thanks for
the width choices , i took 2e in 9 . 5 . none of other shoes can feel
like this , extremely perfect !
Output: [('NULL', 'performance sizing/fit', 'positive', 'fitting for
my feet'), ('NULL', 'performance comfort', 'positive', 'really
comfortable'), ('NULL', 'contextofuse purchase_context', 'positive',
'thanks for the width choices'), ('NULL', 'appearance form',
'positive', 'thanks for the width choices'), ('NULL', 'performance
general', 'positive', 'none of other shoes can feel like this ,
extremely perfect')]
// other examples
Input: really great shoe , live the color as well . only thing the run
tight so i got the wide size .
Output: [('shoe', 'general', 'positive', 'really great'), ('shoe',
'appearance color', 'positive', 'live the color'), ('NULL',
'performance sizing/fit', 'negative', 'run tight so i got the wide
size')]
Input:the design is great poor color choices too bland . color choices
from previous shoes was much better .
Output:

```

Figure 8: ACOS Few-shot Prompting Example. We provide intructions, category information, and 10 examples sampled from the training set.

```

Instruction: Extract
aspect-category-sentiment-opinion-implicitIndicator quintuples from
input data.
Context: An aspect must be a term existing in input data or null if
non-existing; an opinion must be a term existing in input data; the
category is one in the predefined list ['appearance color',
'appearance form', 'appearance general', 'appearance material',
'appearance misc', 'appearance shoe component', 'contextofuse place',
'contextofuse purchase_context', 'contextofuse review_temporality',
'contextofuse usage frequency', 'contextofuse use case', 'cost/value',
'general', 'misc', 'performance comfort', 'performance durability',
'performance general', 'performance misc', 'performance sizing/fit',
'performance support/stability', 'performance use case applicability',
'versatility']; the sentiment is positive, negative or neutral, the
implicitIndicator is direct or indirect; do not ask me for more
information, I am unable to provide it, and just try your best to
finish the task. You can learn from the following examples.
Output format: (aspect, category, sentiment, opinion,
implicitIndicator)
Input: it ' s really comfortable and fitting for my feet . thanks for
the width choices , i took 2e in 9 . 5 . none of other shoes can feel
like this , extremely perfect !
Output: [( 'NULL', 'performance sizing/fit', 'positive', 'fitting for
my feet', 'direct'), ('NULL', 'performance comfort', 'positive',
'really comfortable', 'direct'), ('NULL', 'contextofuse
purchase_context', 'positive', 'thanks for the width choices',
'direct'), ('NULL', 'appearance form', 'positive', 'thanks for the
width choices', 'direct'), ('NULL', 'performance general', 'positive',
'none of other shoes can feel like this , extremely perfect',
'direct')]
// other examples
Input: really great shoe , live the color as well . only thing the run
tight so i got the wide size .
Output: [('shoe', 'general', 'positive', 'really great', 'direct'),
('shoe', 'appearance color', 'positive', 'live the color', 'direct'),
('NULL', 'performance sizing/fit', 'negative', 'run tight so i got the
wide size', 'direct')]
Input:the design is great poor color choices too bland . color choices
from previous shoes was much better .
Output:

```

Figure 9: Few-shot Prompting Example. We provide intructions, category information, and 10 examples sampled from the training set.

<p>Shoes: Explicit Opinions true to size unwearable for me hassle is a classic felt cheap mediocre fit me irritated my ankle sole coming off they run small</p>	<p>Shoes: Implicit Opinions feel like they are swimming without water i ' d rather have the way to go wanted to love this is the 3rd time i ' m buying didn ' t look maroon to me looked like clown shoes my foot dr suggested these bought for my daughter i am usually very pleased</p>
<p>Laptop: Explicit Opinions unproductive functional defect flimsy upgradable secure lightweight smoothly sharp glitchy</p>	<p>Restaurant: Explicit Opinions sticky sweet not even apologetic diverse fresh tender worth staying for courteous die for down - to - earth decadent</p>

Figure 10: Examples of opinion spans from each dataset.