It is Simple Sometimes: A Study On Improving Aspect-Based Sentiment Analysis Performance

Laura Cabello[*]
Department of Computer Science,
University of Copenhagen
lcp@di.ku.dk

Uchenna Akujuobi
Sony AI, Japan
uchenna.akujuobi@sony.com

Abstract

Aspect-Based Sentiment Analysis (ABSA) involves extracting opinions from textual data about specific entities and their corresponding aspects through various complementary subtasks. Several prior research has focused on developing *ad hoc* designs of varying complexities for these subtasks. In this paper, we present a generative framework extensible to any ABSA subtask. We build upon the instruction tuned model proposed by Scaria et al. (2023), who present an instruction-based model with task descriptions followed by in-context examples on ABSA subtasks. We propose PFInstruct, an extension to this instruction learning paradigm by appending an NLP-related task prefix to the task description. This simple approach leads to improved performance across all tested SemEval subtasks, surpassing previous state-of-the-art (SOTA) on the ATE subtask (Rest14) by +3.28 F1-score, and on the AOOE subtask by an average of +5.43 F1-score across SemEval datasets. Furthermore, we explore the impact of the prefix-enhanced prompt quality on the ABSA subtasks and find that even a noisy prefix enhances model performance compared to the baseline. Our method also achieves competitive results on a biomedical domain dataset (ERSA).

1 Introduction

User-generated reviews on e-commerce and social media platforms benefit both consumers and stakeholders. With the exponential growth of data, developing reliable tools for understanding the sentiment of online review texts is essential to moderate online content, enable effective decision-making and customer satisfaction. Liu (2012) proposed Aspect-Based Sentiment Analysis (ABSA) as a step towards fine-grained sentiment analysis of specific aspects. ABSA involves the detection of opinions

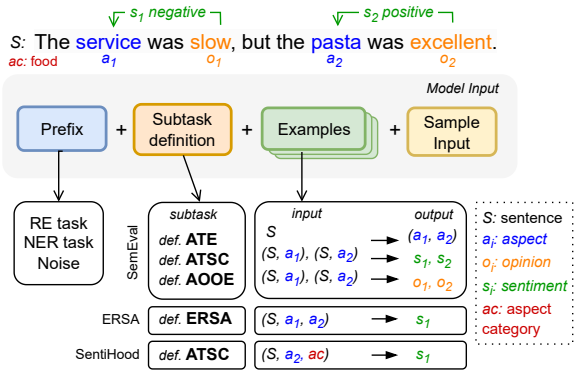


Figure 1: Illustration of model input and ABSA subtasks examined in this paper. The **prefix** can vary between NLP-related tasks (instruction) or textual noise (random words), followed by the **subtask definition**, few **examples** and the corresponding **sample input** for each subtask. The model is expected to follow the instructions and generate a prediction. Subtasks belong to three distinct data sources: SemEval, ERSA and SentiHood from different domains.

(o) and sentiment (s) associated with particular aspects (a) in a text (S). Figure 1 summarizes the five ABSA subtasks considered in this paper.

Instruction-based learning has emerged as a promising paradigm to successfully tune large language models (LLM) on a variety of tasks (Wei et al., 2022; Wang et al., 2022; Singhal et al., 2022; Gupta et al., 2023). The attraction of instruction-based learning is the ability to steer the base model behaviour to follow instructions (Ouyang et al., 2022; Bowman, 2023). In the context of ABSA, Scaria et al. (2023) proposed InstructABSA, an instruction-based model based on a 200M-parameter Tk-Instruct model (Wang et al., 2022). Their best performing setting, dubbed as InstructABSA2, frames the task instruction as a task definition followed by two positive, negative, and neutral examples. We build our experiments upon the same setting.

In this paper, we propose PFInstruct, an exten-

[*] Work done during internship at Sony AI.

Findings of the Association for Computational Linguistics: ACL 2024, pages 6597–6610

sion to the InstructABSA framework with the introduction of prefix prompt. Specifically, we append a prefix to the task definition, extend the evaluation to domains like biomedicine and urban neighbourhoods, and formulate *all* the subtasks as a generative task. The prefix aims at instructing the model on a related NLP task, namely Relation Extraction (RE) or Named Entity Recognition (NER), so the target text S is seen on a different task context. We postulate this approach helps to collect richer semantic information about the main entities in S, which allows the model to make a more informed prediction about the ABSA subtasks. We also consider to use a randomly generated (noise) prefix. We observe that not only it boosts –to a lesser degree– average performance, but it also makes the model more robust to out-of-domain data.

Contributions We introduce a simple approach to solve ABSA subtasks, that is, a prefix prompt followed by instructions. Our approach outperforms previous SOTA on several tasks despite being based on a 200M model. We conduct extensive analysis on five subtasks from different domains: customer reviews from SemEval 2014, 15 and 16 (Pontiki et al., 2014, 2015, 2016), on a biomedical domain dataset, ERSA (Young and Akujuobi, 2023), and on user comments from SentiHood (Saeidi et al., 2016). We assess the effect of the prefix prompt in terms of prompt quality (RE, NER or noise) and domain generalization (out-of-domain performance).

We make our code publicly available to ensure reproducibility and foster future research.[1]

2 Related Work

Current state-of-the-art (SOTA) solutions on SemEval datasets include LSA (Yang and Li, 2023) on ATSC, which leverage the use of gradient descent to design a differential-weighted approach, BARTABSA (Yan et al., 2021) on AOOE, based on an end-to-end generative framework, and Sun et al. (2019) on SentiHood (ATSC). On a similar line of research to Scaria et al. (2023) and our work, Varia et al. (2023) propose IT-MTL, multi-task prompting with instructional prompts to tackle ABSA subtasks as a question-answering problem. For a review of popular ABSA datasets, we refer to Chebolu et al. (2023).

To compare to related work (Yan et al., 2021; Li et al., 2021; Mao et al., 2021; Varia et al., 2023; Yang and Li, 2023), we report macro-averaged F1-score.

3 Background and Methodology

3.1 Background: ABSA

ABSA subtasks can be classified into single output and compound output subtasks. In single output subtasks such as ATE (Aspect-Term Extraction), ATSC (Aspect-Term Sentiment Classification), and AOOE (Aspect-Opinion Extraction), the output is limited to either the aspect a, opinion o or sentiment s from a given text S. Compound output subtasks ask for a combination of the $\{a, o, s\}$ entity types. Many researchers (Xue and Li, 2018; Wu et al., 2020; Pouran Ben Veyseh et al., 2020; Yang and Li, 2023) focus only on the former. For convenience, we do the same.[2]

In addition, we evaluate our method on ERSA (Young and Akujuobi, 2023) and SentiHood (ATSC) (Saeidi et al., 2016) tasks. ERSA is an extension of the ABSA task, where, given a text S and two aspects (or entities) a_1 and a_2, where $a_1 \neq a_2$, the goal is to determine the sentiment polarity s of the relationship between the aspects given S. ERSA targets biomedical texts but it does not require to have factual knowledge about the entities. SentiHood, defined as an extension of ATSC, requires to classify the sentiment s towards each aspect a_i of one or more aspect categories ac. Appendix A describes the datasets in more detail.

3.2 Methodology

We propose an extension of the InstructABSA framework. Given a sample S, we construct a prompt that consists of 4 components which we detail below.

Prefix. An initial instruction to explicitly ask the model to solve an NLP task on the sample S. The purpose of this prefix is to involve the main entities in S in an preliminary NLP task, which can inform the subsequent ABSA subtask on the main entities in S to determine the correct output. This NLP task can be Named Entity Recognition (NER) or Relation Extraction (RE). RE is applied if the aspects a_i are part of the task input and the sample contains at least two entities (or aspects). We also analyse the effect of having a noisy prefix prompt composed of random words.

[1] https://github.com/lautel/PFInstruct

[2] The application of our method to compound-output subtasks is straightforward.

Task definition. A succinct overview of the ABSA subtask. In sentiment classification tasks, we also include the set of pre-defined classes.

Examples. A set of two positive, negative, and neutral in-domain examples. Scaria et al. (2023) carry out an extensive analysis on the effect of different task definitions and example manipulations. As our method extends their approach, we fix the task definition and set of examples to match their best performing set-up, namely InstructABSA2.

Sample input. Similar to the in-context examples, we provide the model the input *S* and expect the model will follow the instructions and generate the corresponding output.

At inference time, we repeat the same structure with sample inputs from the test split. Appendix B shows specific examples of the final prompts.

4 Results

We present the main results of our experimental set-up. report macro-averaged F1-score averaged across five random initialization seeds and the standard deviation. Details about fine-tuning settings are provided in Appendix A.

4.1 Analysis of SemEval subtasks

Tables 1–3 show results of ATE, ATSC[3] and AOOE subtasks respectively. Our method achieves superior performance (F1-scores) when compared with previous SOTA methods across all subtasks. Specifically, we observe that setting an NLP-related task prefix outperforms previous models in 6/12 cases. Interestingly, adding a noise prefix surpasses previous approaches in 4 of the remaining 6 cases. These results validate our initial hypothesis: instructing the model to solve a related NLP task for the target text *S* seems to complement the model's understanding of the main entities in *S*, which leads to more accurate predictions.[4]

In general, providing a random prefix improves model performance compared to not including a prefix at all (see results from InstructABSA2 rows) in the three subtasks. This is more pronounced in ATE. Contrary to ATSC and AOOE, ATE requires the model to make a prediction based solely in the input text *S*, *i.e.*, it does not include a target aspect *a* as input. This setting causes the effect of focusing on entity recognition or having random prefixes to be similar on average across datasets: F1 = 90.35 (PFInstruct-NER) compared to F1 = 90.53 (PFInstruct-Noise). However, the disparate variance in PFInstruct-Noise makes PFInstruct-NER an overall better model choice.

Model	Lapt14	Rest14	Rest15	Rest16	Avg.
BARTABSA†	83.52	87.07	75.48	-	-
InstructABSA2†	92.30	92.10	76.64	80.32	85.34
PFInstruct-NER	92.65±0.70	**95.38**±0.10	82.86±1.15	90.51±0.72	90.35±0.67
PFInstruct-Noise	**92.90**±3.95	94.92±0.46	**83.58**±0.61	**90.73**±2.41	90.53±1.86

Table 1: F1-scores for ATE subtask. Avg stands for average across datasets. †Results from original papers.

Model	Lapt14	Rest14	Rest15	Rest16	Avg.
LSA_T-X†	**83.93**	86.26	-	-	-
Dual-MRC†	75.97	82.04	73.59	-	-
BARTABSA†	76.76	75.56	73.91	-	-
InstructABSA2‡	81.66±0.80	86.70±0.63	85.06±1.05	**93.01**±0.71	86.61±0.80
PFInstruct-RE	82.57±1.05	86.68±0.71	**86.16**±1.17	92.51±0.20	86.98±0.78
PFInstruct-NER	81.63±0.28	86.66±0.92	85.61±1.90	86.60±0.35	85.13±0.86
PFInstruct-Noise	80.88±1.02	**86.88**±1.50	84.32±0.59	91.54±0.51	85.91±0.91

Table 2: F1-scores for ATSC subtask. Avg stands for average across datasets. †Results from original papers. ‡Results are reproduced by us, since Scaria et al. (2023) report accuracy.

Model	Lapt14	Rest14	Rest15	Rest16	Avg.
Dual-MRC†	79.90	83.73	74.50	83.33	80.37
BARTABSA†	80.55	85.38	80.52	87.92	83.59
InstructABSA2†	77.16	81.08	81.34	83.27	80.71
PFInstruct-RE	**84.04**±0.31	90.10±0.36	**89.56**±0.55	88.51±0.38	88.05±0.40
PFInstruct-NER	83.43±1.61	**91.47**±0.33	89.11±0.28	**92.08**±0.81	89.02±0.65
PFInstruct-Noise	81.06±1.17	91.00±0.55	87.56±0.72	90.70±0.31	87.58±0.68

Table 3: F1-scores for AOOE subtask. Avg stands for average across datasets. †Results from original papers.

We postulate that the additional prefix (random or NLP related) enhances the model's ability to selectively filter out irrelevant information to the final task, thereby bolstering its resilience to textual inaccuracies such as misspellings or grammatical errors. However, including a random prefix has the negative side effect of making the model more sensitive to its initial random weights, as shown by the higher variance of PFInstruct-Noise across settings.

Error Analysis To gain a better insight on the benefits of introducing a random prefix (PFInstruct-Noise) compared to not introducing a prefix at all, we perform a case study on the incorrectly classified examples by PFInstruct without prefix, i.e., we reproduce results with the same settings from

[3]To maintain consistency with existing methods, we also remove instances labelled as 'conflict' (Chen et al., 2017; Li et al., 2021; Scaria et al., 2023).

[4]We look into the subset of predictions for input samples with two or more aspects in ATSC and AOOE tasks and observe a similar trend to what is reported here.

	Sample	Prefix	Output
i	i highly recommend this place to all that want to try indain food for the first time.	No prefix Noise	place indain food
ii	Screen - although some people might complain about low res which I think is ridiculous.	No prefix Noise	screen screen, res
iii	Really Lovely dining experience in the midst of buzzing midtown area.	No prefix Noise	dining dining experience
iv	They have homemade pastas of all kinds – I recommend the gnocchi – yum!	No prefix Noise	pastas, gnocchi homemade pastas, gnocchi

Table 4: Case study on the ATE subtask on examples where the model fails in the absence of a prefix, but PFInstruct-Noise outputs the correct target aspect(s). We observe that PFInstruct-Noise is more robust to misspellings errors and chatspeak (i, ii) and extracts more detailed answers (iii, iv).

InstructABSA2. From these errors, 23% are correct in ATE with PFInstruct-Noise and 28% in AOOE. In sentiment classification (ATSC), both setups missclassify the same samples. Focusing on ATE, Table 4 showcases examples where the introduction of a noise prefix seems beneficial for a better understanding of misspelling errors and especial jerga. We also find cases where aspect term extraction with PFInstruct-Noise is more comprehensive, including descriptive adjectives (rows i and iii).

Other work in NLP also find beneficial the addition of noise. Amongst others, Jain et al. (2024) show that noisy embeddings improve instruction fine-tuning, and Cuconasu et al. (2024) prove that including irrelevant documents can enhance performance of retrieval-augmented generation (RAG) systems.

4.2 Analysis of ERSA and SentiHood

Table 5 shows results on both datasets[5], where we can see that ERSA is the clear exception to the trend observed in Section 4.1: the choice of a prefix is important as it can negatively affect model performance.

The inherent nature of ERSA presents a more significant challenge than other ABSA subtasks, since the sentiment expressed in a text S may not necessarily reflect the sentiment of the relationship between the target entities, a_1 and a_2 (Young and Akujuobi, 2023). In this case, in-context noise

[5]To obtain comparable results to existing methods (Young and Akujuobi, 2023; Saeidi et al., 2016), we utilize only the four most frequent aspects in SentiHood.

Model	ERSA	SentiHood (ATSC)
CERM[†]	71.0	88.50
BERT-pair-QA-M[†]	-	93.60
InstructABSA2	**70.76**±0.41	94.90±0.07
PFInstruct-RE	70.31±0.14	-
PFInstruct-NER	70.00±0.50	93.83±0.03
PFInstruct-Noise	64.72±0.59	**95.11**±0.02

Table 5: F1-scores for ERSA and SentiHood tasks. Results from InstructABSA2 are reproduced by us.

hurts model performance the most. The model needs to adapt to a specialised domain and learn the nuances of the task. In terms of NLP-task prefix, leveraging the knowledge of a_1 and a_2 to reason about their semantic relationship (PFInstruct-RE) improves model performance over general entity recognition (PFInstruct-NER). However, it does not surpass the model performance achieved without prompt prefixes (InstructABSA2 setup).

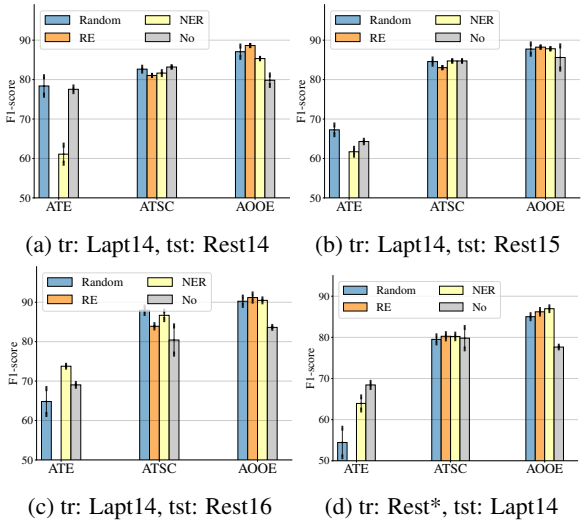
Error Analysis We examine the misclassified examples by PFInstruct-RE to better understand why, contrary to what we observed in Section 4.1, the absence of prefixes appears to be beneficial (see InstructABSA2 in Table 5). We observe that in $\sim 50\%$ of these cases, annotators have labelled the sentiment based on the *meaning of the full sentence* instead of focusing on the relationship between the given entities *in* the context. For instance, *'Treatment with ERY resulted in fewer inflammatory cells and cytokines in the BALF, and fewer emphysema-associated changes...'* is labelled as *negative* despite the relationship between the target entities being *neutral/none*.

4.3 Domain generalization

Results from Tables 1–5 demonstrate the viability of our method with in-domain data in SemEval and SentiHood, while remaining competitive in ERSA. In this section, we explore the robustness of our models when evaluated on out-of-domain data in SemEval. Figure 2 shows results when training the models on laptops domain (Lapt14) and evaluating on restaurants domain (Rest14, Rest15, Rest16), and vice versa.

As expected, we observe a general drop in performance compared to training in-domain, which is especially large in ATE. However, when training in restaurant domain and evaluating in Lapt14 for AOOE –see Figure 2 (d)–, all model variants sur-

[†]Results from the original papers (acc).



(a) tr: Lapt14, tst: Rest14 (b) tr: Lapt14, tst: Rest15

(c) tr: Lapt14, tst: Rest16 (d) tr: Rest*, tst: Lapt14

Figure 2: Out-of-domain evaluation. F1-scores are averaged across five random initialization seeds; error bars show the standard deviation. Models are trained ('tr') on one domain and evaluated ('tst') on a distinct domain. Legends indicate the prefix prompt used. 'No' stands for no use of prefix. RE is not evaluated for ATE (see Section 4).

pass their respective in-domain results with a major improvement of $+3.53$ F1 in PFInstruct-NER. We conclude that the addition of more training data is beneficial for this task.

The addition of any kind of prefix helps to make the model more robust to out-of-domain data for AOOE, while it does not significantly hurt performance for ATSC. Interestingly in this task, the large variance shown by the model without prefix is reduced by the addition of a prefix, especially with RE and NER prefixes –see Figure 2 (c), (d)–.

While the strategy of adding a noisy prefix seem beneficial to out-of-domain data performance too, looking closer we make two observations: *i)* PFInstruct-Noise models show large variance regardless of the domain, and *ii)* the drop in performance when evaluated on out-of-domain data (compared to in-domain) is larger for PFInstruct-Noise models. Therefore, these results suggest that an NLP task prefix makes the model more robust to domain shifts.

5 Discussion

As shown in Section 4, the addition of an NLP task prefix can boost model performance, especially in ATE and AOOE. While the addition of a noise prefix also seems beneficial in most of the cases tested, it also comes with high fluctuations in performance depending on the random initialisation of the model's weights. For this reason, we would caution against this choice for a final application. However, the a priori effectiveness of PFInstruct-Noise echoes the question posed by Kung and Peng (2023), do models really learn to follow instructions? Our analysis provides evidence that performance gains in ABSA subtasks can come from seeing the target entities (or aspects) in a preliminary NLP task instruction, suggesting the utility of this instruction. Nevertheless, our results with PFInstruct-Noise also highlights the need for more in depth analysis of instruction based learning and evaluation.

6 Conclusion

In this paper, we present PFInstruct, a simple yet effective prefix prompting strategy to instruction fine-tune a language model on ABSA subtasks. We analyse the impact of the prefix prompt's quality on in-domain and out-of-domain data and observe that even a random prefix improves average model performance compared to the InstructABSA baseline. We evaluate our method on domains such as customer reviews, biomedical text and user comments, and show that it outperforms previous SOTA approaches on most of the tasks tested and achieves competitive performance (F1-score) in the rest.

Limitations

Our study builds upon an instruction tuned language model, Tk-Intruct, and therefore it inherits its limitations. However, we have done an extensive analysis on a variety of domains and task settings, namely SemEval (customer reviews of laptops and restaurants) ERSA (healthcare) and SentiHood (user comments about urban neighbourhoods), proving the generalizability of our method. Despite our method can be easily applied to other language models, the effectiveness of PFInstruct have not been tested with other model architectures nor model sizes.

Our approach reduces the effective input sequence length of the model, since we need to allocate input tokens for the prefix prompt. While this side-effect is worth noting, it has not supposed an issue for the current experiments (maximum sequence length for Tk-Instruct: 512 tokens, average prompt length excluding input sentences: 348 tokens with Relation Extraction or random (noise) prompt, 304 tokens with Named Entity Recognition prompt). In addition, our work is limited to an

English language model and English texts. Future studies should prove the validity of our approach in languages other than English.

Ethics Statement

The models and datasets used in this study are publicly available, and we strictly follow their terms of use. We meet the ethical implications of previous research related to the data sources. It is important to acknowledge the presence of inherent biases to the data and models used in this study, but we do not anticipate other ethical risks derived from our work.

References

Samuel R. Bowman. 2023. Eight things to know about large language models.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2023. Survey of aspect-based sentiment analysis datasets.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 452–461, Copenhagen, Denmark. Association for Computational Linguistics.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems.

Himanshu Gupta, Saurabh Arjun Sawant, Swaroop Mishra, Mutsumi Nakamura, Arindam Mitra, Santosh Mashetty, and Chitta Baral. 2023. Instruction tuned models are quick learners.

Neel Jain, Ping yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R. Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. NEFTune: Noisy embeddings improve instruction finetuning. In *The Twelfth International Conference on Learning Representations*.

Po-Nien Kung and Nanyun Peng. 2023. Do models really learn to follow instructions? an empirical study of instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1317–1328, Toronto, Canada. Association for Computational Linguistics.

Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. 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 6319–6329, Online. Association for Computational Linguistics.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 19–30.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Nasim Nouri, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2020. Introducing syntactic structures into target opinion word extraction with deep learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8947–8956, Online. Association for Computational Linguistics.

Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1546–1556, Osaka, Japan. The COLING 2016 Organizing Committee.

Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Arjun Sawant, Swaroop Mishra, and Chitta Baral. 2023. Instructabsa: Instruction learning for aspect based sentiment analysis.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge.

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. Association for Computational Linguistics.

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. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 19–27, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Zhen Wu, Fei Zhao, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2020. Latent opinions transfer network for target-oriented opinion words extraction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9298–9305.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523, Melbourne, Australia. Association for Computational Linguistics.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. 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 2416–2429, Online. Association for Computational Linguistics.

Heng Yang and Ke Li. 2023. Improving implicit sentiment learning via local sentiment aggregation.

Julio Christian Young and Uchenna Akujuobi. 2023. *CERM: Context-Aware Literature-Based Discovery via Sentiment Analysis*. IOS Press.

A Experimental details

A.1 Data

Distribution of datasets used:

Lapt14: 3045 Train, 800 Test

Rest14: 3041 Train, 800 Test

Rest15: 1315 Train, 685 Test

Rest16: 2000 Train, 676 Test

Hotel15: 266 Test

ERSA: 8183 Train, 909 Validation, 2274 Test

SentiHood: 5215 Train (2460 from top-4 aspect categories), 610 Validation, 1216 Test

A.2 Experiments

We instruction fine-tune the model checkpoint `Tk-Instruct-base-def-pos` with the following hyprparameters:

- N. epochs: 4
- Batch Size: 16 for ATE, ATSC, ERSA; 8 for AOOE and SentiHood. Batch sizes explored: {8, 16}
- Learning rate: 1e-4 for ATE, ATSC, AOOE and SentiHood; 5e-5 for ERSA. Learning rates explored: {1e-5, 5e-5, 1e-4, 5e-4}
- Warmup ratio: 0.1
- Regularization: weight decay, 0.01

Experiments were performed in 2 A10G GPUs. Hyperparameter tuning was performed based on validation performance in each dataset. If a validation split was not originally provided, we held out 10% of the train split.

B Prompt Examples

Table 6 and Table 7 provide examples of prefixes for two given input texts *S*. Each table illustrate the three prefix types defined in the paper. We set the Noise-prefix length to 50 words to match the average length of the RE-prefix.

In Tables 8–12, we provide details of complete instruction prompts for all five subtasks. Task definition and in-context examples in ATE, ATSC and AOOE subtasks are from (Scaria et al., 2023).

Input *S*	I am pleased with the fast log on, speedy WiFi connection and the long battery life (>6hrs).
RE-prefix	"Definition: Solve a relation extraction (RE) task. Given the context, output the most precise semantic relation between the entities 'log on' and 'WiFi connection'. In cases where there is no relationship the output should be NONE. Reason the answer step-by-step. Context: I am pleased with the fast log on, speedy WiFi connection and the long battery life (>6hrs)."
NER-prefix	"Definition: Given the following context, output the relevant entities in it. Reason the answer step-by-step. Context: I am pleased with the fast log on, speedy WiFi connection and the long battery life (>6hrs)."
Noise-prefix	"Definition: elegantly messier nordin fulke wantonness defile sills newland sbu lena hoff nubia cobblestones caddis disliking gaster domicil martialed sylvestre chagall enquires delphic haring niobe intrusive mnes scolex counterpoise detoxification tanglewood sedgwick vintner anker northfield thrilled transvestite echeverria radula lengths abdullah kiri unhinged minefields cloaked restrictive humored refractometer troy cargoes cordate"

Table 6: Illustration of three prefix types for an input sentence with two aspects (*log on* and *WiFi connection*).

Input	Food is always fresh and hot- ready to eat!
RE-prefix	-
NER-prefix	"Definition: Given the following context, output the relevant entities in it. Reason the answer step-by-step. Context: Food is always fresh and hot-ready to eat!"
Noise-prefix	"Definition: longmans propulsive kirchen cofactor encoders granitic description carlist yorick accosted outgoings flathead metallization ings surrounds cunliffe relevant quagmire hacked castellana extenders railwaymen windbreak stichting sepia stg jewess bashfulness engrossing fiberboard passionless deb vicente hilbert firft independently inconvenient bloodhound complexed eglantine ricardo casts kebir exoneration undernourishment kerygma extenuate englishmen porridge legitimize"

Table 7: Illustration of three prefix types for an input sentence with one aspect (*food*).

"**Definition:** Given the following context, output the relevant entities in it. Reason the answer step-by-step.
Context: I recommend this place to everyone.
Afterwards solve the following task
Definition: The output will be the aspects (both implicit and explicit) which have an associated opinion that are extracted from the input text. In cases where there are no aspects the output should be noaspectterm.
Positive example 1-
input: With the great variety on the menu , I eat here often and never get bored.
output: menu
Positive example 2-
input: Great food, good size menu, great service and an unpretensious setting.
output: food, menu, service, setting
Negative example 1-
input: They did not have mayonnaise, forgot our toast, left out ingredients (ie cheese in an omelet), below hot temperatures and the bacon was so over cooked it crumbled on the plate when you touched it.
output: toast, mayonnaise, bacon, ingredients, plate
Negative example 2-
input: The seats are uncomfortable if you are sitting against the wall on wooden benches.
output: seats
Neutral example 1-
input: I asked for seltzer with lime, no ice.
output: seltzer with lime
Neutral example 2-
input: They wouldnt even let me finish my glass of wine before offering another.
output: glass of wine
Now complete the following example-
input: I recommend this place to everyone.
output: "

Table 8: Illustration of an input prompt with NER-prefix for ATE subtask. Words in **boldface** to ease visualization.

"**Definition:** Given the following context, output the relevant entities in it. Reason the answer step-by-step.
Context: Boot time is super fast, around anywhere from 35 seconds to 1 minute.
Afterwards solve the following task
Definition: The output will be 'positive' if the aspect identified in the sentence contains a positive sentiment. If the sentiment of the identified aspect in the input is negative the answer will be 'negative'.
Otherwise, the output should be 'neutral'. For aspects which are classified as noaspect-term, the sentiment is none.
Positive example 1-
input: I charge it at night and skip taking the cord with me because of the good battery life. The aspect is battery life.
output: positive
Positive example 2-
input: Easy to start up and does not overheat as much as other laptops. The aspect is start up.
output: positive
Negative example 1-
input: Also kinda loud when the fan was running. The aspect is fan.
output: negative
Negative example 2-
input: but now i have realized its a problem with this brand. The aspect is brand.
output: negative
Neutral example 1-
input: I took it back for an Asus and same thing, it required me to remove the battery to reset. The aspect is battery.
output: neutral
Neutral example 2-
input: I can always buy and install a camera. The aspect is camera.
output: neutral
Now complete the following example-
input: Boot time is super fast, around anywhere from 35 seconds to 1 minute. The aspect is Boot time. output: "

Table 9: Illustration of an input prompt with NER-prefix for ATSC subtask. Words in **boldface** to ease visualization.

"**Definition:** Solve a relation extraction (RE) task. Given the context, output the most precise semantic relation between the entities 'spicy tuna roll' and 'asian salad'. In cases where there is no relationship the output should be NONE. Reason the answer step-by-step.
Context: BEST spicy tuna roll , great asian salad .
Afterwards solve the following task
Definition: The output will be the opinion/describing word of the aspect terms in the sentence. In cases where there are no aspects the output should be none.
Positive example 1-
input: I charge it at night and skip taking the cord with me because of the good battery life . The aspect is battery life.
output: good
Positive example 2-
input: it is of high quality , has a killer GUI , is extremely stable , is highly expandable , is bundled with lots of very good applications , is easy to use , and is absolutely gorgeous. The aspect is GUI.
output: killer
Negative example 1-
input: One night I turned the freaking thing off after using it , the next day I turn it on , no GUI , screen all dark , power light steady , hard drive light steady and not flashing as it usually does . The aspect is GUI.
output: no
Negative example 2-
input: I can barely use any usb devices because they will not stay connected properly . The aspect is usb devices.
output: not stay connected properly
Neutral example 1-
input: However , the multi-touch gestures and large tracking area make having an external mouse unnecessary (unless you 're gaming) . The aspect is external mouse.
output: unnecessary
Neutral example 2-
input: I wanted to purchase the extended warranty and they refused , because they knew it was trouble . The aspect is extended warranty.
output: refused
Now complete the following example-
input: BEST spicy tuna roll , great asian salad . The aspect is spicy tuna roll.
output: "

Table 10: Illustration of an input prompt with RE-prefix for AOOE subtask. Words in **boldface** to ease visualization.

"**Definition:** Solve a relation extraction (RE) task. Given the context, output the most precise semantic relation between the entities 'brain disease' and 'neurotrophic factor'. In cases where there is no relationship the output should be NONE. Reason the answer step-by-step.
Context: The loss of neurotrophic factors such BDNF and CNTF may be associated with the pathogenesis of brain diseases (Chauhan, Siegel, & Lee, 2001; Jeon et al., 2015; Jeong et al., 2015; Phillips et al., 1991; Sopova, Gatsiou, Stellos, & Laske, 2014)
Afterwards solve the following task
Definition: The output will be 'positive' if the aspects identified in the sentence express a positive sentiment for human's health.
If the sentiment of the identified aspects in the input is negative for human's health the answer will be 'negative'.
Otherwise, the output should be 'neutral'.
Positive example 1-
input: Clonidine and tizanidine have been used in the treatment of chronic pain disorders. The aspects are tizanidine and chronic pain.
output: positive
Positive example 2-
input: One of the recommended standard treatments for diabetic neuropathy is pregabalin. The aspects are pregabalin and diabetic neuropathy.
output: positive
Negative example 1-
input: NFAT promotes carcinoma invasive migration through glypican-6. The aspects are carcinoma and glypican.
output: negative
Negative example 2-
input: frutescens on streptozotocin induced diabetes in neonatal (n-STZ)-type II Wistar rats. The aspects are streptozotocin and diabetes.
output: negative
Neutral example 1-
input: He had not noted any insect bites and denied ingesting any raw meat, freshwater fish, or liver. The aspects are liver and insect bite.
output: neutral
Neutral example 2-
input: Non-heme iron concentrations in cooked sausages are shown in Figure 1. The aspects are sausages and heme.
output: neutral
Now complete the following example-
input: The loss of neurotrophic factors such BDNF and CNTF may be associated with the pathogenesis of brain diseases (Chauhan, Siegel, & Lee, 2001; Jeon et al., 2015; Jeong et al., 2015; Phillips et al., 1991; Sopova, Gatsiou, Stellos, & Laske, 2014) The aspects are brain disease and neurotrophic factor.
output: "

Table 11: Illustration of an input prompt with RE-prefix for ERSA subtask. Words in **boldface** to ease visualization.

"**Definition:** teases faunas aap pills esophagus overplus bhp inches diplococcus thatcher substituting sama lancey cortex directness pastoralism illicitly urbain inserting educated belonging unde fragmented pepperell scoria chee altogether purifications purchase prostatic neuralgia sternum germinate maneuvering holier gabble conjuration rac dolley hartshorn illam postgate macdonnell tripoli pitot massa radford porcine iiib laufer
Afterwards solve the following task
Definition: Solve the following task. The output will be 'positive' if the identified aspect of a given entity in the input sentence contains a positive sentiment. If the sentiment of the identified aspect in the input is negative the answer will be 'negative'. Otherwise, the output should be 'neutral'.
Positive example 1-
input: Of course LOCATION1 is also very central. The entity is LOCATION1, the aspect is transit-location.
output: positive
Positive example 2-
input: If I were you I would look nearby LOCATION1. The entity is LOCATION1, the aspect is general.
output: positive
Positive example 3-
input: LOCATION1 is an ugly cold place but it isn't dangerous. The entity is LOCATION1, the aspect is safety.
output: positive
Negative example 1-
input: I'd stay away from LOCATION1. The entity is LOCATION1, the aspect is general.
output: negative
Negative example 2-
input: LOCATION1 is a nice area, but apartments are very pricey. The entity is LOCATION1, the aspect is price.
output: negative
Negative example 3-
input: LOCATION1 is all junkies. The entity is LOCATION1, the aspect is safety.
output: negative
Now complete the following example-
input: LOCATION1 is in Greater London and is a very safe place. The entity is LOCATION1, the aspect is safety.
output: "

Table 12: Illustration of an input prompt with Noise-prefix for SentiHood (ATSC) subtask. Words in **boldface** to ease visualization.