Deep Content Understanding Toward Entity and Aspect Target Sentiment Analysis on Foundation Models

Vorakit Vorakitphan [1] **Milos Basic** [1] **Guilhaume Leroy Meline** [1]

Abstract

Introducing Entity-Aspect Sentiment Triplet Extraction (EASTE), a novel Aspect-Based Sentiment Analysis (ABSA) task which extends Target-Aspect-Sentiment Detection (TASD) by separating aspect categories (e.g., food#quality) into pre-defined entities (e.g., meal, drink) and aspects (e.g., taste, freshness) which add a fine-gainer level of complexity, yet help exposing true sentiment of chained aspect to its entity. We explore the task of EASTE solving capabilities of language models based on transformers architecture from our proposed unified-loss approach via token classification task using BERT architecture to text generative models such as Flan-T5, Flan-Ul2 to Llama2, Llama3 and Mixtral employing different alignment techniques such as zero/few-shot learning, Parameter Efficient Fine Tuning (PEFT) such as Low-Rank Adaptation (LoRA). The model performances are evaluated on the SamEval-2016 benchmark dataset representing the fair comparison to existing works. Our research not only aims to achieve high performance on the EASTE task but also investigates the impact of model size, type, and adaptation techniques on task performance. Ultimately, we provide detailed insights and achieving state-of-the-art results in complex sentiment analysis.

1. Introduction

Sentiment Analysis (SA) is the field of Natural Language Processing (NLP) that aims to extract and analyze sentiments, opinions, attitudes and emotions expressed towards certain entities from a textual data. Traditionally, SA focused on detecting the overall polarity e.g., positive, negative or neutral. As such, it brings limited amount of information, often insufficient for most of the real world applications (Liu, 2012). Hence, in the past decade, the research recognized fine-grained SA frequently named as Aspect Based Sentiment Analysis (ABSA) (Pontiki et al., 2014; 2015; 2016) whose task fundamentally consists of detecting two components which are targets and its corresponding sentiments. According to Zhang et al. (2022) the target can be described as either *aspect category* or *aspect term*, sentiment can be described as *sentiment polarity* that often attaches by *opinion term*. Depending on the motives, not all authors aim to extract these elements simultaneously, but they would rather focus on identifying specific subgroups and the relations among its elements which are consequently leading to the creation of various so called ABSA subtasks.

However, the transformer architectures (Vaswani et al., 2017) with pre-trained knowledge (aka Foundation Models (FMs)) influence most of the NLP tasks to not require a specific architecture to train from scratch on a large corpus of data, but rather the best performing solutions rely on adaptation of FMs (Bommasani et al., 2022) for downstream tasks. Generally, Large Language Models (LLMs) can achieve high performance on a downstream task relying only on their emergent abilities (Zhao et al., 2023), without their weights being updated. Such adaption for a downstream task is called zero- or few-shot learning (Brown et al., 2020) toward the text generation approaches (aka Generative AI or GenAI). In recent years, GenAI including models like Llama2 (Touvron et al., 2023), Llama3 [1], Mistral (Jiang et al., 2023), Mixtral (Jiang et al., 2024). These models have gained the popularity in research areas and industries that revolutionized NLP applications by enabling high performance on various tasks through optimized prompting techniques rather than extensive fine-tuning. These models are pre-trained on vast and diverse datasets and can efficiently handle both simple and complex NLP tasks by simply refining prompts. This approach simplifies deployment and enhances adaptability, making instruct models particularly effective for tasks such as classification, text generation, summarization, and translation tasks.

In this work, we propose a fine-grainer detection of *aspect category* into what we called *entity* (i.e., meal) and

[1] IBM Consulting France. Correspondence to: Vorakit Vorakitphan <vorakit.vorakitphan1@ibm.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1] https://github.com/meta-llama/llama3

aspect (i.e., freshness) and focused on sentiment polarity revealed based on detected entity and aspect as a triplet output. This introduces a new ABSA subtask called "Entity Aspect Sentiment Triplet Extraction (EASTE)" task. We benchmark our experimentation settings on the ABSA subtask on the dataset published in a shared task; SemEval16 (Pontiki et al., 2016) which is a dataset based on restaurant reviews in English. Our work also tackle these challenges following EASTE settings in different appraoches: 1) a token classification task like Named-entity recognition (NER) by proposing a unified-loss solution based on BERT model (Devlin et al., 2018) with full fine-tuning, 2) zero/few shot learning and fine-tuning text generation models and 3) model alignments via parameter efficient fine-tuning such as LoRA (Hu et al., 2021) and Prefix-tuning (Li & Liang, 2021b) to observe the impacts and performances on complex tasks such as EASTE on instruct-tuned encoder-decoder LLMs.

We position these contributions on multiple fields.

- Creating an EASTE, a novel ABSA task that profoundly targets *entity*, *aspect* elements for comprehensive analysis of *sentiment*.
- Proposing a novel unified-loss approach to solve EASTE task via token classification that is suitable for multi-class classification.
- Deeply exploring various techniques to adapt Language Models for a complex downstream task and achieving a State-of-the-Art (SoTA) results.

2. Related Work

In this section, we highlight relevant research that shares the SoTA target sentiment analysis work. Regarding the ABSA subtasks which have been introduced as a task for SemEval14 (Pontiki et al., 2014) and continued to be present on both SemEval15 (Pontiki et al., 2015) and SemEval16 (Pontiki et al., 2016) competitions, most of complex SA tasks were introduced as compound ABSA subtasks (Zhang et al., 2022) which aims to extract simultaneously multiple elements – pairs, triplets or quadruplets from a given sentence or a text. Among for our research relevant tasks such as Aspect Category Sentiment Analysis (ACSA), Aspect Sentiment Triplet Extraction (ASTE), and Aspect Sentiment Quad Prediction (ASQP), we highlight Target-Aspect-Sentiment Detection (TASD).

Target-Aspect-Sentiment Detection (TASD). TASD aims to extract target, aspect, sentiment triplets. Target Aspect Sentiment Detection was introduced by Wan et al. (2020) and for a review sentence S aims to extract all (t, a, s) triplets:

$$TASD(S) = \{(t_1, a_1, s_1), ..., (t_n, a_n, s_n)\}$$

where t_i stands for target and is a subsequence of S if explicit and $NULL$ if implicit, $a_i \in \{a_1, ..., a_n\}$ stands for aspect and $s_i \in \{positive, negative, neutral\}$ for sentiment polarity. For example, in the restaurant review sentence *'The food arrived 20 minutes after I called, cold and soggy.'*, output of TASD should be {*(NULL, SERVICE#GENERAL, negative), (food, FOOD#QUALITY, negative)*}

To our knowledge, Brun & Nikoulina (2018) and Wan et al. (2020) have addressed this problem using available parsers and domain-specific semantic lexicons and pre-trained BERT-based architecture respectively. In additional, the work of Zhang et al. (2021) defines the TASD as sequence-to-sequence learning problem and solves it in a generative manner using encoder-decoder T5 architecture (Raffel et al., 2019).

3. Task Definition

Pontiki et al. (2015) define an entity *e*, for instance in the domain of restaurants, as either the reviewed entity itself (restaurant), or another relevant entities directly related to it (e.g. food, service, ambience), while the aspect *a* is a particular attribute (e.g., quality, price, general) of the entity in question. The sentiment polarity *s* is defined as a polarity of the opinion(s) expressed towards the target. All three pieces of information are extracted from a predefined set of entities $e \in \{e_1, ..., e_n\}$ and aspects $a \in \{a_1, ..., a_n\}$ for every specific domain of interest, while sentiment polarity $s \in \{positive, negative, neutral\}$ remains the same per domain. Wan et al. (2020) combine what Pontiki et al. (2015) defined as entity and aspect into one element as "aspect", taking into account only *e* and *a* combinations that appear in the analyzed datasets, which might not cover all true possible combinations.

Therefore, we follow the definitions by Pontiki et al. (2015) and define new ABSA subtask which considers entity and aspect as two different elements. Hence, for a given sentence S extracts all (e, a, s) triplets but taking into account that each of the triplets relates to the corresponding target:

$$EASTE(S) = \{(t_1, e_1, a_1, s_1), ..., (t_n, e_n, a_n, s_n)\}$$

where t_i stands for target and is a subsequence of S if explicit and $NULL$ if implicit, $e_i \in \{e_1, ..., e_n\}$ stands for entity, $a_i \in \{a_1, ..., a_n\}$ stands for aspect and $s_i \in \{positive, negative, netural\}$ for sentiment polarity. If we take the same restaurant sentence *'The food arrived 20 minutes after I called, cold and soggy.'*, output of EASTE should be {*(NULL, SERVICE, GENERAL, negative), (food, FOOD, QUALITY, negative)*}. Hence, we break the aspect into two elements as "entity" and "aspect" where each of them comes from different, independent dataset and is combined after its individual prediction in what Wan et al. (2020) define as aspect. This scenario of extraction creates one more level of ABSA task complexity.

4. Dataset

Our work relies on the latest benchmark in this category which is SemEval16 (Pontiki et al., 2016). We adopt the review sentences for restaurants domain. For each given sentence (e, a, s) triplets are annotated where $e \in \{e_1, e_2, \ldots e_n\}$ stands for entity, $a \in \{a_1, a_2, \ldots a_n\}$ for aspect category, and $s \in \{positive, negative, neutral\}$ for sentiment. e, a and s are chosen from different predefined inventories. To each of the triplets, a *target term* is attached towards which the opinion is expressed. If it is expressed explicitly, the *target term* is a subset of the words of corresponding review sentence, and NULL if implicitly. The dataset contains 2000 review sentences for train set, and 676 review sentences for test set.

5. Methodology and Experimentation

In this section, we discuss our experimental settings to explore EASTE task while the token classification tasks were run locally using Apple CPU and MPS devices. For text generative tasks via LLMs like Llama2, Llama3, and Mixtral models, we use API calls for model inference hosted via IBM watsonx.ai[2].

5.1. Classification Approach

We solve EASTE task using a proposed unified-loss approach toward token classification task which is applied on a sentence S using BERT architecture. Figure 1 demonstrates our implementation of the unified-loss approach based on token classification where a triple classifier is introduced after a loss function per gate.[3] Each token in S represents $\{t_1, t_2, ..., t_n\}$, every t provides single or multiple $(entity, aspect, sentiment)$ triplets where $t = \{(e_1, a_1, s_1), ..., (e_n, a_n, s_n)\}$. We modify the last linear layer of BERT-based-uncased architecture and consequently adapt model's loss function to obtain three classification results where the losses l are computed as an average loss $l(joint)$. The final loss is calculated as a mean of losses via the additional of loss logits then divided by number of output gate of each $entity, aspect, sentiment$ as following:

$$l(joint) = \frac{l(entity) + l(aspect) + l(sentiment)}{Number\ of\ output\ gates}$$

where $l(entity), l(aspect), l(sentiment)$ are individual cross entropy losses for entity, aspect and sentiment obtained per token.

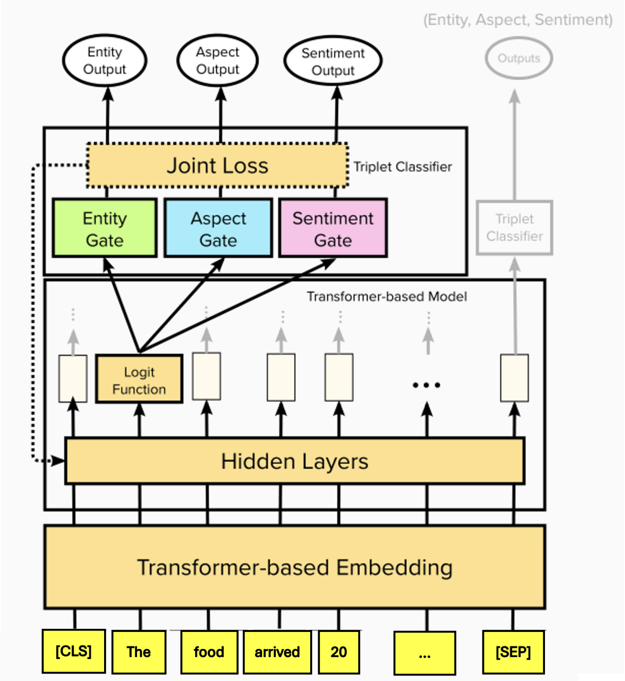

Figure 1: Unified-loss architecture based on BERT token classification where each loss function of $entity, aspect, sentiment$ are weighted thoroughly.

Setting 1: Entity-Aspect Sentiment Triplet Extraction For the unified-loss approach explained aforementioend, we adopted BERT base uncased (110M) for token classification task and fine-tuned it for 50 epochs. The batch size we used for this approach was 1, combined with stochastic gradient descent (SGD) optimization algorithm with learning rate of 1e-3. Batch size, optimization algorithm used are not very common in practice nowadays, but in this combination they proved to obtain the best results. The measure of quality are precision, recall and F1 score.. The best results on token level which are presented in the Table 3 are obtained in the epoch 48.

5.2. Text Generation Approach

In the Text Generation Task framework, models generate an output in dictionary format from an input prompt[4] which consist of input sentence S and instruction $Inst$. GenAI models for EASTE task aim to capture information on specific triplets containing entities, aspects, and sentiment polarity. For each input sentence S, the model generates $\{t_1, t_2, ..., t_n\}$ tokens where n varies for different input t as output, then we encode them into text format. Using zero and few-shot learning approaches and fine-tuning for a downstream task, the model is trained

[2] https://www.ibm.com/products/watsonx-ai

[3] Unified-loss code snippet is shared via GitHub link: https://github.com/vvorakit/Entity-Aspect-Sentiment-Triplet-Extraction

[4] All prompts from GenAI models used in EASTE task can be found in Appendix.

so that the combined outputs $\{t_1, t_2, ..., t_n\}$ contain information $\{(e_1, a_1, s_1), ..., (e_n, a_n, s_n)\}$ triplets only. Unlike Token Classification Tasks, where tokens are assigned to classes, text generative models ensure that the extracted triplets correspond accurately to target terms within review sentences. Evaluating zero and few-shot learning approaches, we assess the capability of LLMs to solve complex tasks like EASTE solely based on pretraining. Zero-shot learning involves providing the model with instructions and review sentences, expecting output to contain relevant triplets referencing target terms $LM(S, Inst) = \{(t_1, e_1, a_1, s_1), ..., (t_n, e_n, a_n, s_n)\}$. Few-shot learning extends this by including input-output examples, maintaining a constant set of examples while varying input sentences $LM(S, (Inst, \{e_1, ..., e_n\})) = \{(t_1, e_1, a_1, s_1), ..., (t_n, e_n, a_n, s_n)\}$.

Setting 2: Text Generation with Zero/Few-shot Technique In the zero/few-shot setting, we conducted numerous experiments searching for the instruction that provides the best results. We experimented with the instructions that provide from 0 (zero-shot) gradually up to 40 (few-shot) review sentence examples from a benchmark dataset to observe learning capabilities of the model.

Setting 3: Text Generation with Fine-tuning Technique

In the fine-tuning setting, we conducted numerous experiments searching for the overall best practice in terms of the model size, memory and time consumption, type of the instruction, as well as the performance on the task. The approaches that we use for fine-tunning are full model fine-tuning which updates all model's trainable parameters via back-propagation, Prefix-tuning (Li & Liang, 2021a) and LoRA (Hu et al., 2021) that update only the small percentage of them (form 0.1% to 0.3% of number of trainable parameters).

For full model fine tuning Flan-T5-XL (3B) trained for 6 epochs with batch size of 4 and AdamW optimizer with learning rate of 2e-4 performed the best. It is important to note that we did not apply the full model fine-tuning on models larger than 3B parameters due to the memory constraints. Furthermore, in the Prefix-tuning setting, our best performing model was Flan-UL2 (20B) which was trained for 50 epochs, using as well the batch size of 4 combined with AdamW optimizer with 2e-4 learning rate, and setting the number of virtual tokens to 30. Finally, the best performing LoRA tuned model was Flan-T5-XXL (11B) which was trained for 10 epochs as well with the batch size of 4 and AdamW optimizer with learning rate of 2e-4. In LoRA setting, we approximate only query and value update matrices with rank 8, and set alpha to 16. All other non-discussed training parameters are set to default values. The best performing results are represented in Table 1

EASTE Task	P	R	F1
Unified-loss-BERT-uncased	57.41	54.25	55.78
Zero-shot prompting			
Llama2-13b-chat	45.72	99.13	43.50
Llama3-8b-instruct	45.79	97.36	62.28
Mixtral-8x7B-Instruct-v0.1	46.32	99.47	63.21
One-shot prompting			
Llama2-13b-chat	37.50	99.35	54.45
Llama3-8b-instruct	57.17	99.55	70.16
Mixtral-8x7B-Instruct-v0.1	54.03	**100**	70.16
Few-shot prompting			
Llama2-13b-chat	45.72	**100**	62.75
Llama3-8b-instruct	54.28	**100**	70.36
Mixtral-8x7B-Instruct-v0.1	55.01	**100**	70.98
Full fine-tuning			
Flan-T5-XL	**76.87**	75.90	**76.38**
Prefix fine-tuning			
Prefix tuning Flan-UL2	68.04	62.74	65.28
LoRa fine-tuning			
LoRA-Flan-T5-XXL	69.87	63.47	66.52

Table 1: Reporting only best results for EASTE task per approach across multiple techniques on SemEval2016 restaurant test set.

TASD Task	P	R	F1
Entity#Aspect	69.32	58.82	63.64
Sentiment	66.96	68.80	87.87

Table 2: Results on token classification task based on TASD settings and chained evaluation of 2 outputs as Entity#Aspect and sentiment via unified-loss approach using BERT architecture on the SemEval2016 restaurants test set.

6. Results Analysis

The main indicator of model performance in our experimentation is F1 score. For each of methodology settings, we defined a way to calculate F1 score in order to get comparable results. In general, the (e, a, s) triplets are considered predicted correctly if and only if the prediction is attached to the correct corresponding *target term*.

Setting 1: As the *target term* is composed of either one or multiple words if the opinion expressed explicitly towards entity, we consider prediction correct only if 50% or more tokens of the *target term* are predicted correctly. On the other hand, if the opinion is expressed implicitly we expect model to attach (e, a, s) triplet to *cls* token.

Settings 2 and 3: We consider prediction correct only if (e, a, s) triplet is generated attached to correct corresponding *target term* if opinion expressed explicitly towards an entity, otherwise we expect a *target term* to be predicted as *'NULL'*.

ABSA Task	P	R	F1
Entity	75.90	69.19	72.39
Aspect	53.10	52.74	52.92
Sentiment	62.51	55.33	58.70

Table 3: Results on token classification task based on ABSA settings for individual evaluation of entity, aspect, and sentiment via unified-loss approach using BERT architecture on the SemEval2016 restaurants test set.

Overall, our best performing model is fully fine-tuned Flan-T5-XL achieving F1 score of 76.38 based on proposed EASTE experimental settings. However, a few-shot prompted on recently released LLMs such as Llama2, Llama3, and Mixtral obtain outstanding results on recall, but considerably lower on precision which impact the final F1-score. Additionally, Mixtral performs as second ranked, up close by Llama3 with 70.98 and 70.36 F1 scores respectively. On the other hand, prefix-tuning yielded competitive results where Flan-UL2 emerged as a top-performing model, achieving a F1 score of 65.28. Our exploration of LoRA tuning did not provide the big gap over prefix-tuning, with the best scoring model being Flan-T5-XXL with the F1 score of 69.87. However, the full-tuned encoder models for token classification using unified-loss approach tends to provide reasonably good results due to joint layers giving the high complexity of the prediction. On the other hand, in Table 2 when we cast the unified-loss approach to evaluate in TASD settings where entity and aspect are merged as a single element, it shows that there is a significant improvement up to almost 10% in the classification report. Ultimately, if we split an evaluation category into individual inspection of Entity, Aspect, Sentiment like in Table 3, we can observe that the unified-loss approach using BERT is well performed on pre-defined entity detection with F1 score of 72.39 in token classification settings where aspect and sentiment remains a challenging detection process.

7. Conclusion

This research introduces EASTE, a novel and complex task for detecting ABSA settings through various NLP techniques. Our results underscore the critical importance of selecting appropriate fine-tuning techniques and prompting strategies tailored to the size and type of LLMs. By employing diverse approaches such as token classification, text generation, and fine-tuning alignments, we assessed model performance across different architectures and sizes. Our findings demonstrate the efficacy of the proposed unified-loss approach in token classification, particularly for less complex tasks like TASD. Additionally, the potential of text generative models combined with advanced prompting strategies is evident. We also highlight the necessity of choosing suitable fine-tuning techniques and model architectures based on task complexity and available resources. This research offers significant contributions to the field of sentiment analysis, providing deeper insights into sophisticated NLP techniques and models for complex sentiment analysis tasks. Our work advances the understanding of deep content analysis for sentiment detection and sets a foundation for future explorations in the realm of natural language processing.

Acknowledgements

We would like to thank IBM Client Engineering EMEA for providing the GPU infrastructure that made our research possible. Their support enabled us to carry out our experiments efficiently and effectively. The computational resources provided by IBM were instrumental in advancing our work, especially in model alignments and the diverse fine-tuning approaches utilized in this study.

References

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the opportunities and risks of foundation models, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Brun, C. and Nikoulina, V. Aspect based sentiment analysis into the wild. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 116–122, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6217. URL https://aclanthology.org/W18-6217.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts, 2024.

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. *CoRR*, abs/2101.00190, 2021a. URL https://arxiv.org/abs/2101.00190.

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation, 2021b.

Liu, B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, May 2012. doi: 10.2200/s00416ed1v01y201204hlt016. URL http://dx.doi.org/10.2200/S00416ED1V01Y201204HLT016.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 27–35, Dublin, Ireland, August 2014. Association for Computational Linguistics. doi: 10.3115/v1/S14-2004. URL https://aclanthology.org/S14-2004.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 486–495, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2082. URL https://aclanthology.org/S15-2082.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., and Eryiğit, G. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 19–30, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1002. URL https://aclanthology.org/S16-1002.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019. URL http://arxiv.org/abs/1910.10683.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Wan, H., Yang, Y., Du, J., Liu, Y., Qi, K., and Pan, J. Z. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *AAAI Conference on Artificial Intelligence*, 2020.

Zhang, W., Li, X., Deng, Y., Bing, L., and Lam, W. Towards generative 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 2: Short Papers)*, pp. 504–510, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.64. URL https://aclanthology.org/2021.acl-short.64.

Zhang, W., Li, X., Deng, Y., Bing, L., and Lam, W. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges, 2022.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. A survey of large language models, 2023.

A. Appendix.

Prompts used in Entity Aspect Sentiment Triplet Extraction (EASTE) task are listed below.

1. Flan-T5, Tk-Instruct and Flan-UL2 models:

```
Definition: In this task you are given a review sentence and your task is to
   extract the triplet of information ’entity’:’aspect’:’sentiment’ for each ’
   term’ (implicit or explicit) the opinion is expressed towards in the given
    review sentence. The final output should be in shape ’term’:’entity’:’
   aspect’:’sentiment’. Every implicit ’term’ should be classified as ’NULL’.

Example 1-
Input: great food, great wine list, great service in a great neighborhood...
Output: food:food:quality:positive, wine list:drinks:style_options:positive,
   service:service:general:positive, neighborhood:location:general:positive

Example 2-
Input: Rather than preparing vegetarian dish, the chef presented me with a
   plate of steamed vegetables (minus sauce, seasoning, or any form or
   aesthetic presentation).
Output: vegetarian dish:food:quality:negative, vegetarian dish:food:
   style_options:negative, chef:service:general:negative

Example 3-
Input: The chicken lollipop is my favorite, most of the dishes (I have to
   agree with a previous reviewer) are quite oily and very spicy, especially
   the Chilli Chicken.
Output: chicken lollipop:food:quality:positive, dishes:food:quality:negative,
    Chilli Chicken:food:quality:negative

Example 4-
Input: Also, they do not take credit card so come with cash!
Output: NULL:restaurant:miscellaneous:neutral

Example 5-
Input: The appetizers we ordered were served quickly - an order of fried
   oysters and clams were delicious but a tiny portion (maybe 3 of each).
Output: fried oysters and clams:food:quality:positive, fried oysters and
   clams:food:style_options:negative, NULL:service:general:positive

Example 6-
Input: The service was spectacular as the waiter knew everything about the
   menu and his recommendations were amazing!
Output: service:service:general:positive, waiter:service:general:positive

Example 7-
Input: I book a gorgeous white organza tent which included a four course prix
    fix menu which we enjoyed a lot.
Output: white organza tent:ambience:general:positive, four course prix fix
   menu:food:quality:positive

Example 8-
Input: The place is beautiful!
```

```
Output: place:ambience:general:positive

Example 9-
Input: MY husbands birthday and my sons was not as it was intended... and we
   drove two hours to spend too much money to be treated terribly!
Output: NULL:restaurant:general:negative, NULL:restaurant:prices:negative,
   NULL:service:general:negative

Now complete the following example-
Input: {sentence}
Output:
```

2. Llama2-13b-chat model:

```
<s>[INST] <<SYS>> You are a cautious assistant.You follow strictly the prompt.
    You carefully follow instructions. You are helpful and harmless and you
   follow ethical guidelines and promote positive behavior. If you don’t know
    the answer to a question, please don’t share false information. <</SYS>>

A triplet is a set of three elements: an entity (E), an attribute (A), and a
   sentiment (S). Your task is to generate only one (entity, attribute,
   sentiment) found in the given sentence. In each sentence , you must find
   exactly one triplet.

The entity must be chosen from the list [’FOOD’, ’RESTAURANT’, ’SERVICE’, ’
   AMBIENCE’, ’DRINKS’, ’LOCATION’].

The attribute must be chosen from the list [’QUALITY’, ’STYLE\_OPTIONS’, ’
   GENERAL’, ’PRICES’, ’MISCELLANEOUS’].

The sentiment must be chosen from [’positive’, ’negative’, ’neutral’].

Don’t generate any text other than the JSON dictionnary.
JSON Format for triplet prediction:
{
  "triplet":
   {
     "entity": "ENTITY_TYPE",
     "attribute": "ATTRIBUTE_TYPE",
     "sentiment": "SENTIMENT_TYPE"
   }
}
Replace ENTITY_TYPE with exactly one of the predefined entity types ( [’FOOD’,
    ’RESTAURANT’, ’SERVICE’, ’AMBIENCE’, ’DRINKS’, ’LOCATION’]),
   ATTRIBUTE_TYPE with exactly one of the attribute types([’QUALITY’, ’
   STYLE_OPTIONS’, ’GENERAL’, ’PRICES’, ’MISCELLANEOUS’]), and SENTIMENT_TYPE
    with exactly one of the sentiment types (["positive", "negative", "
   neutral"]).
Respect the given format.
Sentence:
```

3. Llama3-8b-instruct:

```
<|begin_of_text|> <|start_header_id|> system <|end_header_id|> You are a
   cautious assistant. You follow strictly the prompt. You carefully follow
   instructions. You are helpful and harmless and you follow ethical
   guidelines and promote positive behavior. If you don’t know the answer to
   a question, please don’t share false information. <|eot_id|>

<|begin_of_text|> <|start_header_id|> user <|end_header_id|>
A triplet is a set of three elements: an entity (E), an attribute (A), and a
   sentiment (S). Your task is to generate only one (entity, attribute,
   sentiment) found in the given sentence. In each sentence , you must find
   exactly one triplet.

The entity must be chosen from the list [’FOOD’, ’RESTAURANT’, ’SERVICE’, ’
   AMBIENCE’, ’DRINKS’, ’LOCATION’].

The attribute must be chosen from the list [’QUALITY’, ’STYLE_OPTIONS’, ’
   GENERAL’, ’PRICES’, ’MISCELLANEOUS’].

The sentiment must be chosen from [’positive’, ’negative’, ’neutral’].

Don’t generate any text other than the JSON dictionary.
JSON Format for triplet prediction:
{
 "triplet":
   {
     "entity": "ENTITY_TYPE",
     "attribute": "ATTRIBUTE_TYPE",
     "sentiment": "SENTIMENT_TYPE"
   }
}
Replace ENTITY_TYPE with one of the predefined entity types, ASPECT_TYPE with
    one of the attribute types, and SENTIMENT_TYPE with one of the sentiment
   types.
Respect the given format.
Sentence: <|eot_id|>
```

4. Mixtral-8x7B-Instruct-v0.1 model:

```
<|system|> You are a cautious assistant. You carefully follow instructions.
   You are helpful and harmless and you follow ethical guidelines and promote
    positive behavior. If a question does not make any sense, or is not
   factually coherent, explain why instead of answering something not correct.
    If you don’t know the answer to a question, please don’t share false
   information.

<|user|> A triplet is a set of three elements: an entity (E), an attribute (A)
   , and a sentiment (S). Your task is to generate only one triplet (entity,
   attribute, sentiment) from the given sentence.

The entity must be chosen from the predefined entity types [’FOOD’, ’
   RESTAURANT’, ’SERVICE’, ’AMBIENCE’, ’DRINKS’, ’LOCATION’].
The attribute must be chosen from the list [’QUALITY’, ’STYLE_OPTIONS’, ’
   GENERAL’, ’PRICES’, ’MISCELLANEOUS’].
```

```
The sentiment must be chosen from ['positive', 'negative', 'neutral'].

Your response must be in JSON format, correctly written and complete. Don't
    forget the braces. Don't add any comments at all. Only the triplet is
    required.
Format for triplet prediction:
{
  "triplet":
    {
      "entity": "ENTITY_TYPE",
      "attribute": "ATTRIBUTE_TYPE",
      "sentiment": "SENTIMENT_TYPE"
    }

}
Replace ENTITY_TYPE with one of the predefined entity types, ATTRIBUTE_TYPE
    with one of the attribute types, and SENTIMENT_TYPE with either "positive",
     "negative", or "neutral".
Sentence:
```