Gridlocked Opinions: A Tagging Scheme Unraveling Targets, Holders, Expressions, and Polarities

Javier Rondon
javier-rondon@berkeley.edu

Dominic Lim
limdo@berkeley.edu

Abstract

Our team revisited the results from the SemEval-2022 Task 10: Structured Sentiment Analysis. We leveraged a Grid Tagging Scheme (GTS) which extracted *target*, *holder*, *expression*, and *polarity*. We adapted a pre-existing pipeline solution, which consisted of several steps to extract information and transformed into a single-step model that extracts all aspects of the sentiment. The proposed model demonstrated compelling performance when compared against the more complex and resource-intensive model it was initially based on.

1 Introduction

Sentiment analysis automatically extracts and quantifies a text's emotional tone or attitude, such as social media posts, online reviews, news articles, and customer feedback and is used for applications in business and policy planning. Structured Sentiment Analysis (SSA) is a more advanced form of sentiment analysis that considers the structure and context of text data and the relationships between entities, aspects, and sentiments. Unlike traditional sentiment analysis which focuses on assigning a single sentiment score to a text, SSA aims to identify and extract the sentiment associated with different aspects of a given entity or topic.

For example, in a product review, SSA can identify the sentiment associated with specific aspects of the product, such as its price, performance, or design. This enables businesses to gain more detailed insights into customer sentiment and preferences and make more informed decisions about product development and marketing.

The SemEval-2022 Task 10 was organized in an effort to consolidate previous partial and overlapping investigations into SSA (Barnes et al., 2022). Whereas previous work into Aspect-Based Sentiment Analysis focused on target, expression and polarity classification or just the polarity, Barnes et al. (2021) proposed unifying these extraction and classification tasks into a single sentiment graph.

After reviewing the different submissions by the SemEval-2022 Task 10 participants, our team developed a one-step model that utilizes a Grid Tagging Scheme to extract all aspects of the sentiment graph. While there are many possible approaches to the SSA problem, we believe that our approach takes on a simpler and elegant graph representation of the sentiment parsing problem.

2 Background

The SemEval-2022 Task 10 participants predicted all sentiment graphs in a text, where a single sentiment graph is composed of a sentiment holder, target, expression, and polarity. The task included two sub-tracks (monolingual and cross-lingual) with seven datasets available in five languages, and submissions were evaluated on Sentiment Graph F_1 score. Structured Sentiment Analysis is an information extraction problem in which we attempt to find all of the opinion tuples $O = O_i \text{ , . . . } O_n$ in a text. Each opinion O_i is a "quad-tuple" (h, t, e, p) where h is a holder who expresses a polarity p towards a target t through a sentiment expression e, implicitly defining pairwise relationships between elements of the same tuple.

One such approach proposed by Barnes et al. (2021) leverages dependency graph parsing, where the nodes are spans of sentiment holders, targets and expressions, and the arcs are the relations between them. Figure 1 shows an illustration of a sample review being parsed as a sentiment graph.

One of the SemEval-2022 submissions that was of particular interest was the ISCAS team submission. The ISCAS team (Lu et al., 2022) submitted a three-part, extraction and validation pipeline architecture. First, a Grid Tagging Scheme (GTS) model extracted *target*, *expression*, and *polarity*. The pipeline then extracted the *holder* by formulating a Question Answering model given the previ-

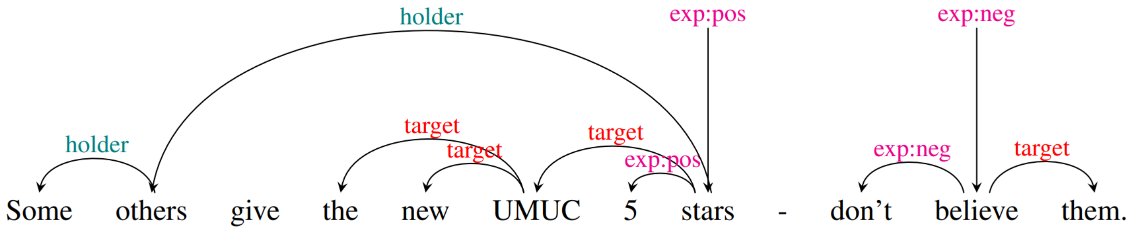


Figure 1: Illustration of the Baseline Graph Dependency Model for a sample review (Barnes et al., 2021)

ously extracted aspects. Finally, the classification step funneled only "valid" quad-tuples that took the shape of (holder, target, expression, polarity). Appendix 6 provides an illustration of the ISCAS pipeline architecture.

The ISCAS team produced sentiment F_1 scores that were compelling against their monolingual participant peers. However, the ISCAS pipeline architecture is a considerable departure from the original GTS tagging paper in which Wu et al. (2020) utilized GTS model to extract all aspects of *target*, *expression*, and *polarity*. Wu et al. (2020) argue that a primary advantage of a single end-to-end solution is avoiding error propagation that might plague multi-model pipeline architectures.

3 Methods

3.1 Datasets

The SemEval Task provided the following datasets in JSON format in five languages: Basque (MultiBookedEU), Catalan (MultiBookedCA), English, with data from OpenER-EN , Multi-Perspective Question Answer corpus (MPQA) and Darmstadt Universities corpus (DSUnis), Norwegian (NoReCFine) and Spanish (OpenER-ES) (Barnes et al., 2022). The SemEval Task was composed of two subtasks – monolingual and cross-lingual structured sentiment. Participants that pursued the monolingual task trained and tested on the same language whereas participants that pursued the cross-lingual task trained on other languages while testing on Spanish, Catalan, and Basque test-set.

For the scope of our project, we decided to pursue the monolingual sub-task and limit our training and testing on the English (OpenER-EN) and Spanish (OpenER-ES) datasets (Agerri et al., 2013). The SemEval-2022 Task 10 organizers prepared train/dev/test splits of the datasets. Our team decided not to augment the dataset in any way.

An initial exploration of the annotated training dataset provides two major insights. The *holders* for the majority of reviews are missing or the opinion can be considered as having an *implict holder* while missing *targets* can also be considered as having an *implicit target*. Roughly 6 % of opinions also had an implicit holder and an implicit target. See Table 1 for an accounting of train and development splits for the OpenER-EN and OpenER-ES datasets.

The visualizations provided in Appendix 7 indicate that the distribution of token counts per review exhibits a pronounced positive skew. This means that most reviews tend to have a smaller number of tokens, while a few reviews in the long tail possess a significantly higher token count. Furthermore, the OpenER-ES dataset displays a more pronounced tail compared to the OpenER-EN dataset, suggesting a greater prevalence of reviews with higher token counts in the Spanish version.

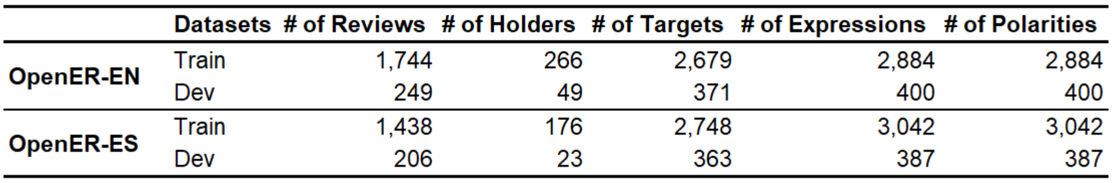
3.2 Evaluation

In accordance with the SemEval Task 10 specifications, participants were evaluated on a Sentiment Graph F_1 Score (Barnes et al., 2022). A true positive is defined as an exact match at graph-level, weighting the overlap in predicted and true spans for each element, averaged across all three spans (holder, target, expression). Precision was calculated as the number of correctly predicted tokens divided by the total number of predicted tokens whereas the denominator was the number of label tokens for recall.

We used the baseline models provided by the organizers of SemEval-2022 Task 10 for comparison. We also used the results obtained by the ISCAS team to assess the difference in performance between our single-step model and a more complex architecture.

3.3 Proposed Model Architecture

We propose to modify the ISCAS team pipeline architecture, removing the QA model and the validation model and instead, extracting all elements of the opinion with a single model.



	Datasets	# of Reviews	# of Holders	# of Targets	# of Expressions	# of Polarities
OpenER-EN	Train	1,744	266	2,679	2,884	2,884
	Dev	249	49	371	400	400
OpenER-ES	Train	1,438	176	2,748	3,042	3,042
	Dev	206	23	363	387	387

Table 1: Train/Development Split statistics for the OpenER-EN and OpenER-ES datasets.

Our proposed model architecture leverages the GTS originally proposed by Wu et al. (2020) and further developed by Lu et al. (2022). The GTS tagging consists of an upper triangular grid, whose length and width is the tokenized sequence length l. Specifically, for i,j $\in$ $[0, l]$, cell (i, j) contains the tag for token-pair (t_i, t_j). The tag represents the relationship between the members of the token-pair. This is illustrated in Figure 2 which shows an example of the application of the GTS scheme for a sentence.

Our implementation of GTS brings in additional tags for *holder*, *implicit holder*, and *implicit holder/implicit target*. Each pair of tokens can be labelled with the following tags labels: $Y = \{H, A, O, IA, IO, IH, IHIA, Pos, Neu, Neg, N\}$. Table 2 lists the tags utilized in our extended GTS.

Tags	Meaning
A	words in pair (w_i,w_j) belong to the same target term
O	words in pair (w_i,w_j) belong to the same opinion term
H	words in pair (w_i,w_j) belong to the same holder term
IA	indicates and implicit target term
IO	indicates and implicit expression term
IH	indicates and implicit holder term
IHIA	indicates implicit holder and target term
Pos Neg Neu	the words belong to a target, expression and holder term and form a positive/neutral/negative relation
N	No relation between the word pair

Table 2: Meaning of tags in GTS

Figure 3 shows an illustration of our all-in-one GTS model. Given a sentence $s = \{w1, w2, ..., w_n\}$ we used the BERT encoder to generate a representation of $\boldsymbol{r}_{ij}$ of the word-pair (w_i, w_j). This contextual representation of the sentence, $s = \{\boldsymbol{H_1}, \boldsymbol{H_2}, ..., \boldsymbol{H_n}\}$ is concatenated as $\boldsymbol{r}_{ij} = [\boldsymbol{H_i} : \boldsymbol{H_j}]$ to represent the word-pair (w_i, w_j). An inference block with a specific number of layers extracts the possible tags for the (w_i, w_j). The number of layers in this inference block is the number of hops. A larger number indicates a deeper network.

This proposed architecture extracts all terms of the opinion tuple in a single model.

3.4 Language Models

We trained the GTS model with a variety of BERT style models which are performative for classification tasks. In the ISCAS submission, Lu et al. (2022) found the best performance with the BERT and XLM-RoBERTa models for their OpenER-EN and OpenER-ES datasets, respectively.

For the English-written OPENER-EN dataset, we also explored using the post-trained model on e-commerce reviews, ReviewBERT[1] (Xu et al., 2019) as well as all-RoBERTa-large[2] which was trained on very large sentence datasets.

For the Spanish-written OPENER-ES dataset, we also explored using the RoBERTa-BNE Model[3], a RoBERTa model pretrained on the largest Spanish language corpus (570 GB) (Fandiño et al., 2022). We also experimented using a multi-lingual distilbert[4] model to explore performance of a model with fewer parameters.

Dataset		Language Model	F_1 Score	Precision	Recall
OpenER-EN	Dev	BERT large uncased	0.66	0.67	0.65
	Test		0.62	0.65	0.59
OpenER-EN	Dev	BERT_review	0.65	0.69	0.62
	Test		0.63	0.66	0.6
OpenER-EN	Dev	all-RoBERTa-large-v1	0.66	0.7	0.62
	Test		**0.66**	**0.68**	**0.64**
OpenER-ES	Dev	XLM_RoBERTa_large	0.67	0.74	0.62
	Test		**0.61**	**0.71**	**0.54**
OpenER-ES	Dev	RoBERTa-large-bne	0.61	0.64	0.58
	Test		0.58	0.62	0.54
OpenER-ES	Dev	distilbert-base	0.48	0.63	0.48
	Test	-multilingual-cased	0.37	0.58	0.28

Table 3: Results of Development and Test files according to Sentiment Graph F_1

[1] https://huggingface.co/activebus/BERT_Review

[2] https://huggingface.co/sentence-transformers/all-RoBERTa-large-v1

[3] https://huggingface.co/PlanTL-GOB-ES/RoBERTa-large-bne

[4] https://huggingface.co/distilbert-base-multilingual-cased

[CLS]	Fantastic	food	and	breathtaking	view	
Implicit Holder	Positive	0	0	Positive	0	[CLS]
	Opinion	Positive	0	0	0	Fantastic
		Target	0	0	0	food
			0	0	0	and
				Opinion	Positive	breathtaking
					Target	view

Figure 2: Example of Grid Tagging Scheme with sample hotel review

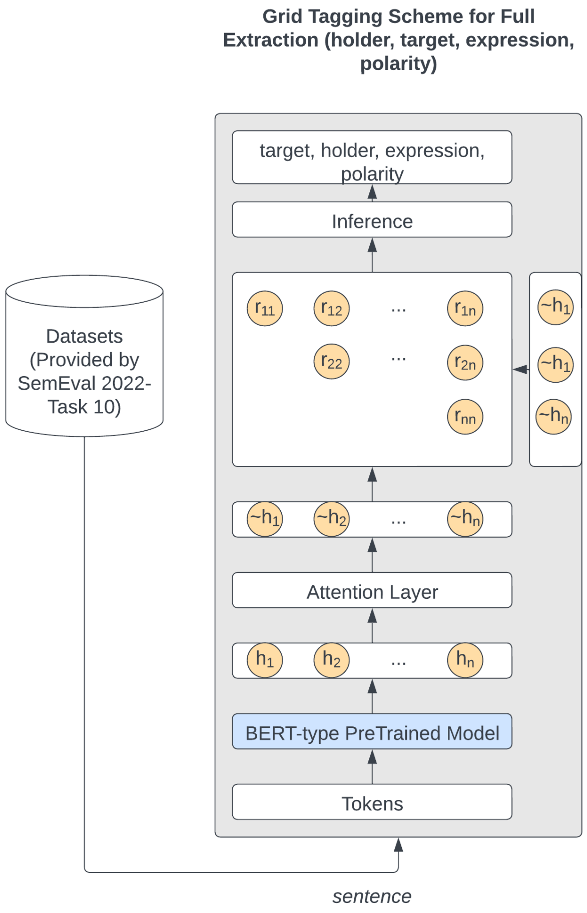

Grid Tagging Scheme for Full Extraction (holder, target, expression, polarity)
target, holder, expression, polarity
Inference
r_{11}
r_{12}
...
r_{1n}
r_{22}
...
r_{2n}
r_{nn}
$\sim h_1$
$\sim h_1$
$\sim h_n$
$\sim h_1$
$\sim h_2$
...
$\sim h_n$
Attention Layer
h_1
h_2
...
h_n
BERT-type PreTrained Model
Tokens
Datasets (Provided by SemEval 2022-Task 10)
sentence

Figure 3: Proposed GTS Extraction Model

Model	OpenER-EN	OpenER-ES
Graph Baseline	0.521	0.495
Seq Baseline	0.329	0.24
ISCAS GTS	0.71	0.669
Our results	**0.66**	**0.61**

Table 4: Comparison of results from this study and the baseline models in terms of F_1 score

3.5 Hardware and Computing Resources

We trained the selection of models on a NVIDIA A10 GPU (24 GB PCIe).

4 Results and discussion

4.1 Train and Test Results

In our extended GTS study, we fine-tuned models based on Sentiment F_1 scores from the development split files provided by the SemEval task 10 organizers. We explored the performance of our system using a variety of high-performing pretrained language models.

During the tuning process, we examined the available hyperparameter ranges. We set the maximum sequence length equal to the longest sequence found in the training and development sets. Due to GPU memory constraints, we limited the maximum number of hops to four. Although investigating a larger number of hops might have improved the model's performance, hardware limitations restricted this exploration. We observed that the model was highly sensitive to the learning rate and identified parameter ranges that could cause the model to fail, potentially due to vanishing gradients.

We applied the trained GTS models to the test datasets to evaluate their performance. Table 3 presents a summary of the extended GTS model's results using different pretrained language models for each dataset. The models were assessed using the Sentiment Graph F_1 score, as described in Section 3.2, with results for both development and test files displayed.

In general, we found that the model performs better on the OpenER-EN dataset compared to the Opener-ES dataset. The model may not have a good capacity to extract information from long sentences, and as shown in Appendix 1, the Spanish language dataset contains a much larger number of these sentences. It was interesting to note that the performance using the XLM-RoBERTa-Large in the Spanish dataset was better than using the language model pretrained exclusively on the Spanish language. This may be a result the difference in size of training corpus between these pretrained language models.

For the OPENER-EN dataset, the performances using different language models were similar, with the All-RoBERTa-Large-v1 model exhibiting the best results. This pretrained model was designed for tasks like clustering or semantic search. Using BERT-Review did not improve the scores, despite it being post-trained with e-commerce reviews.

We compared our findings with the baseline models provided by the organizers of SemEval-2022 Task 10 and the results obtained by the ISCAS team. Table 4 shows the F_1 scores of our study and the baseline. Our study's outcomes not only surpass the baseline but also demonstrate compelling performance against the ISCAS models, all while using a fraction of the computational resources and time.

4.2 Analysis of Errors

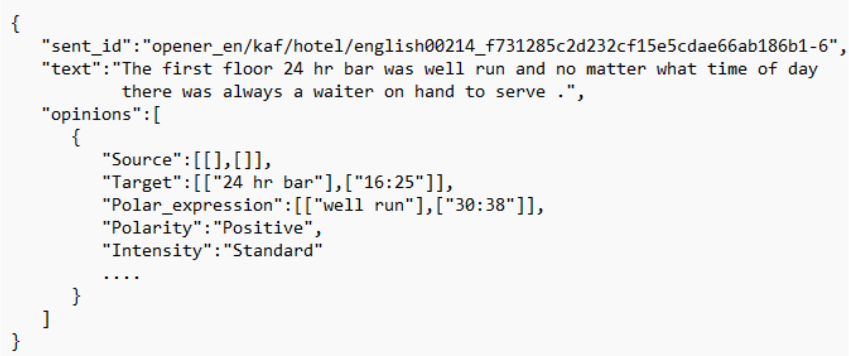
Our all-in-one Grid Tagging Scheme model performed well compared to the SemEval-2022 Task 10 baseline graph and baseline seq2seq models. We performed a review of the most common errors and discrepancies between our model predictions and labels.

One main source of error is in the target span start and end. Figure 4 illustrates an example of a review with 24 tokens and the annotated target `24 hr bar`. As seen in Figure 5, the trained GTS model produced an target inference of `bar`. While the holder is not an exact match between the predicted target and labelled target, the predicted target is still correct and could still be useful in sentiment analysis. This source of error is more frequent for longer and more complex reviews. Longer reviews have multiple targets, holders and opinions, and the model specification did not allow to learn how to extract all sentiment graphs correctly.

Another potential source of errors is the quality of the annotated datasets themselves. As an example one of the provided datasets contains the following sentence:

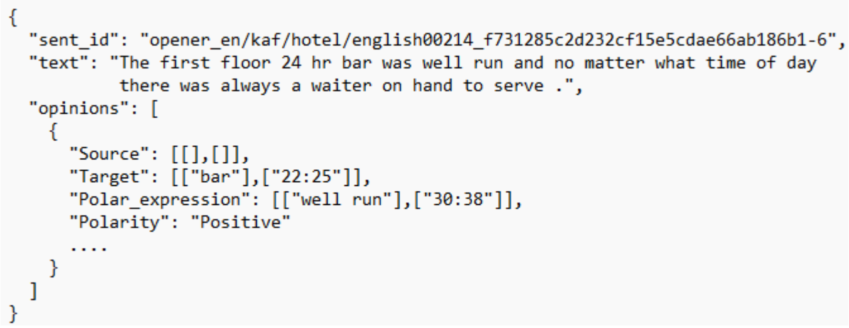
```
Minibus goes from the airport to a hotel
```

The above review in the training data was annotated as having a positive sentiment with the target `minibus` and the expression `goes from the airport to the hotel`. We would argue that the review does not constitute an



```
{
   "sent_id":"opener_en/kaf/hotel/english00214_f731285c2d232cf15e5cdae66ab186b1-6",
   "text":"The first floor 24 hr bar was well run and no matter what time of day
           there was always a waiter on hand to serve .",
   "opinions":[
      {
         "Source":[[],[]],
         "Target":[["24 hr bar"],["16:25"]],
         "Polar_expression":[["well run"],["30:38"]],
         "Polarity":"Positive",
         "Intensity":"Standard"
         ....
      }
   ]
}
```

Figure 4: Example of annotated review to illustrate data format.



```
{
  "sent_id": "opener_en/kaf/hotel/english00214_f731285c2d232cf15e5cdae66ab186b1-6",
  "text": "The first floor 24 hr bar was well run and no matter what time of day
           there was always a waiter on hand to serve .",
  "opinions": [
    {
      "Source": [[],[]],
      "Target": [["bar"],["22:25"]],
      "Polar_expression": [["well run"],["30:38"]],
      "Polarity": "Positive"
      ....
    }
  ]
}
```

Figure 5: Example of discrepancy between model target prediction and annotation ("24 hr bar" vs "bar"). While the prediction does not match the label, the parsed target is still correct in the sentiment analysis.

opinion but more closely resembles a fact about the hotel.

4.3 Limitations

Our project faced challenges due to the limited hardware resources available to us. GPU memory constraints restricted our exploration of hops beyond 4, hindering our ability to investigate deeper neural networks and their effectiveness in extracting information from longer and more complex reviews.

Given our resources, the average training time for a single model exceeded 8 hours. Consequently, computational time posed a significant limitation in the exploration of the model performance.

5 Conclusion

In this project, we introduce a system designed for inferring (holder, target, expression, polarity) quadruples from documents using a single model. The proposed model demonstrates compelling performance when compared to the more intricate and resource-intensive model from which it was initially derived. Our single model avoids error propagation that might plague pipeline systems for this extraction task.

We believe the gap in performance between the model and the best models in the SemEval-2022 Task 10 could be narrowed with additional hardware resources to explore deeper networks to disentangle the opinion on longer reviews.

Acknowledgements

We would like to thank the ISCAS team for providing an excellent repository which served as a foundation for our study.

References

Rodrigo Agerri, Montse Cuadros, Sean Gaines, and German Rigau. 2013. Opener: Open polarity enhanced named entity recognition. *Sociedad Española para el Procesamiento del Lenguaje Natural*, 51(90):215–218.

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. 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 3387–3402, Online. Association for Computational Linguistics.

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States. Association for Computational Linguistics.

Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Xinyu Lu, Mengjie Ren, Yaojie Lu, and Hongyu Lin. 2022. ISCAS at SemEval-2022 task 10: An extraction-validation pipeline for structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1305–1312, Seattle, United States. Association for Computational Linguistics.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2576–2585, Online. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

6 Appendix

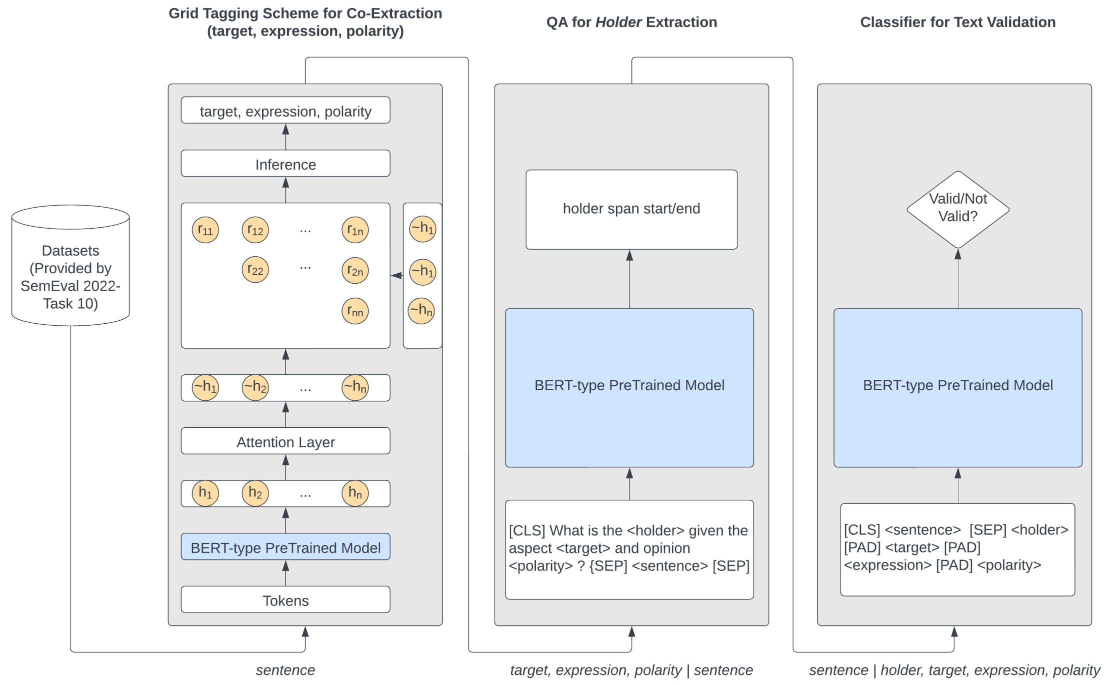

Grid Tagging Scheme for Co-Extraction
(target, expression, polarity)
QA for *Holder* Extraction
Classifier for Text Validation
Datasets (Provided by SemEval 2022-Task 10)
target, expression, polarity
Inference
r_{11} r_{12} ... r_{1n}
r_{22} ... r_{2n}
r_{nn}
$\sim h_1$ $\sim h_1$ $\sim h_n$
$\sim h_1$ $\sim h_2$ ... $\sim h_n$
Attention Layer
h_1 h_2 ... h_n
BERT-type PreTrained Model
Tokens
holder span start/end
BERT-type PreTrained Model
[CLS] What is the <holder> given the aspect <target> and opinion <polarity> ? {SEP] <sentence> [SEP]
Valid/Not Valid?
BERT-type PreTrained Model
[CLS] <sentence> [SEP] <holder> [PAD] <target> [PAD] <expression> [PAD] <polarity>
sentence
target, expression, polarity | sentence
sentence | holder, target, expression, polarity

Figure 6: ISCAS Original GTS Extraction / QA / Validation Pipeline

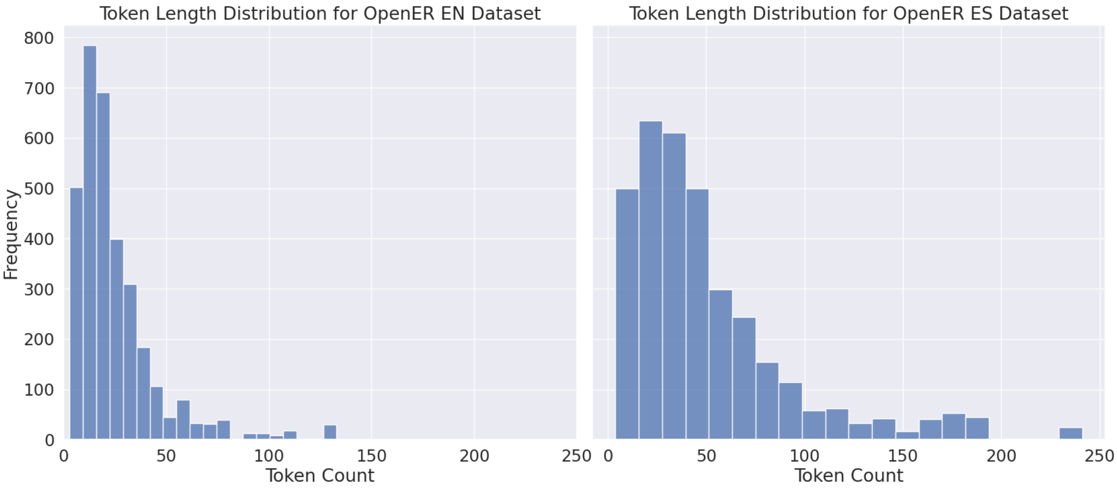


Figure 7: Distribution of Token Count by Reviews