CAISA Qualification Report
Aspect Extraction (AE)

Mohammad Mehdi Balouchi
Professor: Dr. Lucie Fleck
Supervisor: Vahid Sadiri Javadi

University of Bonn, Germany
s6mobalo@uni-bonn.de

Abstract. The goal of this qualification test is to use an Aspect Extraction (AE) model to automatically identify and extract aspects or features mentioned in a given text. As Aspect Extraction is a sub-task of Aspect Based Sentiment Analysis (ABSA), After an extensive investigation on the matter, we determined that the SemEval-2014 Task 4 laptop reviews dataset was the most crucial one to employ [4], then we chose the best performed pre-trained model called InstructABSA which is an instruction learning paradigm for Aspect-Based Sentiment Analysis (ABSA) subtasks.[6]. After running and evaluating the model on the labeled dataset, we used an unlabled dataset from amazon product reviews for the cell phones and accessories category [1], moreover, we extracted the aspects for this unlabeled dataset and exported the result as a csv file.

Keywords: NLP, Aspect Extraction, Sentiment Analysis, Aspect Based Sentiment Analysis, ABSA

1 Introduction

Early work in sentiment analysis mainly aimed to detect the overall polarity (e.g., positive or negative) of a given text or text span [2]. However, the need for a more fine-grained approach, such as aspect-based (or 'feature-based') sentiment analysis (ABSA), soon became apparent [8]. For example, laptop reviews not only express the overall sentiment about a specific model but also sentiments relating to its specific aspects, such as the hardware, software, price, etc. Subsequently, a review may convey opposing sentiments [4] and the aspect extraction is the first step to achieve ABSA goal. The first task of this qualification test is to have a runnable python program which uses a language model to extract the aspects of a text, this an example for the AE task:

Text: "The camera on this phone is excellent, but the battery life is disappointing."

Aspects: ["camera", "battery life"]

Moreover, our python program has the ability to run and evaluate the model for a labeled datasets with the same format as SemEval-2014 Task 4 dataset, which we will elaborate the format on the related section. Another requirement for this task which is successfully accomplished is running the model for an unlabeled dataset and exporting the result in a csv file. This qualification test is consists of the following **subtasks:**

1. You should find a labeled dataset for the task
2. You can find a pre-trained model (e.g. from GitHub) for the dataset or train a model on the dataset.
3. You should know how the model is being trained: What are the input and output of the model?
4. You need to find an unlabeled dataset
5. You should use your model for the new dataset to extract the aspects
6. Save the results in any format you prefer (e.g., CSV or JSON).
7. Write a report for all the steps you have done (which is this document)

2 Background

One of the most important previous research projects on this topic is SemEval 2014 task 4. SemEval is a series of international natural language processing (NLP) research workshops whose mission is to advance the current state of the art in semantic analysis. Task 4 of SemEval is based on laptop and restaurant reviews and consists of four subtasks. The First subtask (SB1) is **Aspect term extraction** [4], which is similar to our qualification test.

3 Datasets

To complete the subtasks 1 and 4, we had to find a labeled and unlabeled dataset. The labeled dataset should consist of product reviews of users with human-annotated aspects; there isn't any requirement for the aspects' polarities; however, as the selected dataset is from SemEval-2014 ABSA Task 4, the polarities are also included in the dataset. For the unlabeled dataset, we chose the Amazon product reviews for the "Cell phones and Accessories" category. We wrote a parser for both labeled (xml parser) and unlabeled (json parser) datasets, and for easier computation, we converted the datasets to the pandas dataframe format with the "text" of the review and a list of corresponding "labels" as their aspects.

- **Labeled Dataset:**
  The datasets of the ABSA task were provided in an XML format (see Fig. 1) . They are available with a non-commercial, no-redistribution license through META-SHARE, a repository devoted to the sharing and dissemination of language resources [3].

```xml
<sentence id="2339">
    <text>I charge it at night and skip taking the cord with me because of the good battery life.</text>
    <aspectTerms>
        <aspectTerm term="cord" polarity="neutral" from="41" to="45"/>
        <aspectTerm term="battery life" polarity="positive" from="74" to="86"/>
    </aspectTerms>
</sentence>
<sentence id="812">
    <text>I bought a HP Pavilion DV4-1222nr laptop and have had so many problems with the computer.</text>
</sentence>
```

Fig. 1. Train dataset with .xml format example. This is the Laptop reviews dataset with the annotated aspects.

We used the laptops dataset contains 3845 English sentences extracted from laptop customer reviews. Human annotators tagged the aspect terms; 3045 sentences were used for training and 800 for testing (evaluation). In the training set, there are 2373 annotated aspects, of which 1048 are unique. We note that laptop reviews often evaluate each laptop as a whole rather than expressing opinions about particular aspects. Furthermore, when they express opinions about particular aspects, they often do so by using adjectives that refer **implicitly** to aspects (e.g., ‘expensive’, ‘heavy’), rather than using **explicit** aspect terms (e.g., ‘cost’, ‘weight’); the annotators were instructed to tag only explicit aspect terms, not adjectives implicitly referring to aspects [4].

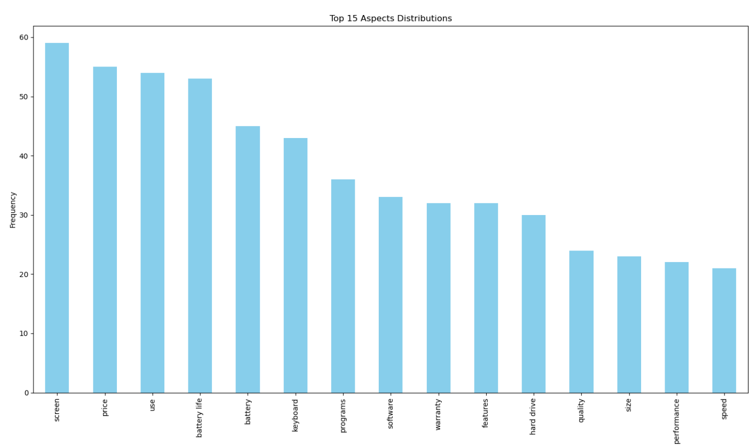

Fig. 2. Top 15 rank of the labeled dataset aspects distributions.

– **Unlabeled Dataset:**
As our unlabeled dataset we used the Amazon Review Data (2018), The pro-

vided the amazon user reviews based on different categories (233.1 million reviews). There are raw review data and 5-core reviews, which 5-core is the subset of the data in which all users and items have at least 5 reviews (75.26 million reviews in total). We chose the "Cell Phones and Accessories" category for the 5-core version (1,128,437 reviews), which is the closest category to our pre-trained model on laptop reviews. The format of this dataset is one review per line in a json format (see Fig 3).


```json
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "vote": 5,
  "style": {
    "Format:": "Hardcover"
  },
  "reviewText": "I bought this for my husband who plays the piano.  He is having a wonderful time
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```


Fig. 3. Unlabeled Amazon Review Data (2018) with .json format example.

Here we are reporting some statistics about our labeled and unlabeled datasets (see table 1). We calculated the averge number of tokens for all the reviews, maximum and minimum number of tokens that exists in reviews.

Table 1. Datasets statistics

Dataset	#Reviews	Average #tokens	Max #tokens	Min #tokens
train_labeled	3048	16.83	83	2
test_labeled	800	14.76	71	3
amazon unlabeled	1127672	57.50	6686	0

4 Model Selection

There are several models available for the ABSA task. we used the paperswithcode.com website for comparing the existed models. These F1 score comparison is for the Aspect Extraction on SemEval 2014 Task 4 Sub Task 2 for the Laptop dataset. Based on the evaluation comparison (see table 3). We chose the **InstructABSA** model which is an instruction learning paradigm for Aspect-Based Sentiment Analysis [6].

Table 2. Models' performance comparison on Aspect Extraction on SemEval 2014 Task 4 Sub Task 2 for Laptop dataset

Rank	Model	F1	Year
1	**InstructABSA**	**92.30**	**2023**
2	ACE	87.4	2020
3	PH-SUM	86.09	2020
4	BAT	85.57	2020
5	Wei et al.	82.7	2020

5 Model Description (InstructABSA)

InstructABSA is an instruction learning model for aspect based sentiment analysis. Their approach involves further instruction tuning of the Tk-Instruct model [7]. Tk-INSTRUCT, is a generative model for transforming task inputs given declarative in-context instructions (task definition or kshot examples). It is built by multi-task training of the T5 model [5] over all the task instructions in their training set, and is evaluated on unseen tasks in the test set. In InstructABSA model they add instruction prompts specific to the downstream ABSA subtasks in the form of task definitions, followed by positive, negative, and neutral examples [6]. The instruction used for the pre-trained model which is fine-tunned for the SemEval 2014 Laptop dataset consists of a definition + 2 positive examples + 2 negative examples + 2 neutral examples (see Fig 4). The InstructABSA model generates the prompts based on the top ranked aspects distributions, moreover, they analyse the effect of instruction tuning based on task definition manipulation and also task examples manipulation.

```
# definition + 2 positive examples + 2 negative examples + 2 neutral examples
BOS = """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: I charge it at night and skip taking the cord with me because of the good battery life.
    output: battery life
    Positive example 2-
    input: I even got my teenage son one, because of the features that it offers, like, iChat, Photobooth, garage band and more!.
    output: features, iChat, Photobooth, garage band
    Negative example 1-
    input: Speaking of the browser, it too has problems.
    output: browser
    Negative example 2-
    input: The keyboard is too slick.
    output: keyboard
    Neutral example 1-
    input: I took it back for an Asus and same thing- blue screen which required me to remove the battery to reset.
    output: battery
    Neutral example 2-
    input: Nightly my computer defrags itself and runs a virus scan.
    output: virus scan
    Now complete the following example-
    input: """
```

Fig. 4. Instruction prompt used in InstructABSA for the AE task for SemEval 2014 Laptop dataset

We used the pre-trained model of InstructABSA finetuned for the Aspect Term Extraction (ATE) subtask. The above prompt is prepended onto each input review. Finally it outputs the related aspects of a text (review) and "noaspect-term" when it does not recognize any aspect.

6 Evaluation

For evaluating our result, we used the proposed evaluation metrics in SemEval 2014. They proposed to first create two sets G and S. G is the set of all Gold (true) aspects from the dataset and S is the set of all aspect terms returned by our model. Then we calculate the precision and recall based on the following formulas:

$$P = \frac{|S \cap G|}{|S|} \tag{1}$$

$$R = \frac{|S \cap G|}{|G|} \tag{2}$$

Furthermore, we calculate the F1 score:

$$F1 = \frac{2.P.R}{P + R} \tag{3}$$

Here are the evaluation reports of the model that we executed with our python program:

Table 3. InstructABSA evaluation report on the test labeled dataset

Precision	**Recall**	**F1**
87.55%	**86.08%**	**86.81 %**

7 Conclusion

In this qualification test we wrote python program in order to do the Aspect Extraction tasks for the product reviews. We selected train and test labeled dataset from the SemEval 2014 task laptop reviews, moreover, we selected an unlabeled dataset from the Amazon product reviews. Then we chose the InstructABSA pre-trained model and make it runnable for extracting the aspects of a single review text and also for a dataset (xml; similar to the laptops format). We evaluated the model on the test dataset and reported precision, recall and F1 based on the SemEval 2014 task 4 evaluation metrics. It is possible extend our word to the ABSA (Aspect Based Sentiment Analysis) task as well.

References

1. Ni, J., Li, J., McAuley, J.: Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 188–197. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1018, https://aclanthology.org/D19-1018
2. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of EMNLP. pp. 79–86. Philadelphia, Pennsylvania, USA (2002)
3. Piperidis, S.: The meta-share language resources sharing infrastructure: Principles, challenges, solutions. In: Proceedings of LREC-2012. pp. 36–42. Istanbul, Turkey (2012)
4. Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., 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. Association for Computational Linguistics, Dublin, Ireland (2014)
5. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research (JMLR) (2020)
6. Scaria, K., Gupta, H., Sawant, S.A., Mishra, S., Baral, C.: Instructabsa: Instruction learning for aspect based sentiment analysis (2023)
7. Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Arunkumar, A., Ashok, A., Dhanasekaran, A.S., Naik, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H.G., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Patel, M., Pal, K.K., Moradshahi, M., Parmar, M., Purohit, M., Varshney, N., Kaza, P.R., Verma, P., Puri, R.S., Karia, R., Sampat, S.K., Doshi, S., Mishra, S., Reddy, S., Patro, S., Dixit, T., Shen, X., Baral, C., Choi, Y., Smith, N.A., Hajishirzi, H., Khashabi, D.: Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks (2022)
8. Zhang, L., Liu, B.: Sentiment analysis and opinion mining. In: Encyclopedia of Machine Learning and Data Mining (2012)