Challenges for Open-domain Targeted Sentiment Analysis

Anonymous ACL submission

Abstract

Since previous studies on open-domain targeted sentiment analysis are limited in dataset domain variety and sentence level, we propose a novel dataset consisting of 6,013 human-labeled data to extend the data domains in topics of interest and document level. Furthermore, we offer a nested target annotation schema to extract the complete sentiment information in documents, boosting the practicality and effectiveness of open-domain targeted sentiment analysis. Moreover, we leverage the pre-trained model BART in a sequence-to-sequence generation method for the task. Benchmark results show that there exists large room for improvement of open-domain targeted sentiment analysis. Meanwhile, experiments have shown that challenges remain in the effective use of open-domain data, long documents, the complexity of target structure, and domain gaps.

1 Introduction

Open-domain targeted sentiment analysis refers to the task of extracting entities and sentiment polarities (e.g., positive, negative, neutral) towards them in free texts (Mitchell et al., 2013) (Figure 1). It has received much research attention due to wide applications to market prediction, recommendation system, product selection, public opinion surveillance. For example, a business might be interested in monitoring the mentioning of itself or its products and services from all media sources, and an investment fund can be interested in learning the sentiment towards a range of open-ended topics that can potentially be influential to market volatilities. Ideally, the task requires algorithms to process open-domain texts from different genres such as news, reports and tweets. For each domain, topics and opinion expressions can be highly different.

As shown in Figure 1 (a), existing research on open-domain targeted sentiment analysis has focused on a sentence-level setting (Mitchell et al., 2013), where different models have been proposed to extract or tag text spans as the mentioned targets, assigning sentiment polarity labels (i.e., positive, negative and neutral) on each extracted span. Both pipeline methods (Mitchell et al., 2013; Zhang et al., 2015; Hu et al., 2019) and joint methods (Mitchell et al., 2013; Zhang et al., 2015; Ma et al., 2018; Li et al., 2019a; Zhou et al., 2019; Song et al., 2019; Pingili and Li, 2020; Hu et al., 2019) have been considered, with the former taking separate models for opinion target extraction and target sentiment classification, and the latter using a single model for solving both subtasks. The current state-of-the-art results (Luo et al., 2020) has been achieved by using pre-trained model BERT (Kenton et al., 2019).

Existing work, however, is limited in several aspects. First, it is constrained by the use of relatively small datasets from Mitchell et al. (2013) and Pontiki et al. (2014, 2015, 2016), which are confined to the restaurant review, laptop review and twitter domains. As a consequence, a strong performance on the benchmarks does not necessarily reflect strong

---

The price of food is high in that Italian restaurant, but its service is good.

<table>
<thead>
<tr>
<th>Traditional</th>
<th>Our Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>price - Negative</td>
<td>Italian restaurant - food - price - Negative</td>
</tr>
<tr>
<td>Food - Negative</td>
<td>Italian restaurant - service - Positive</td>
</tr>
<tr>
<td>Service - Positive</td>
<td>Italian restaurant - Mixed</td>
</tr>
</tbody>
</table>

(a) Sentence-level Example

Bought but didn’t use for months. When I finally did decide to use it, I took it to Hawaii with me to be able to charge my phone in the rental car while visiting. Worked for the first two days then the it stopped charging. Seller for the charger will not work with me.

<table>
<thead>
<tr>
<th>Our Work</th>
<th>charger - USB connection - Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>charger - Seller - Negative</td>
</tr>
<tr>
<td></td>
<td>charger - Negative</td>
</tr>
</tbody>
</table>

(b) Multi-sentence Example

Figure 1: Traditional open-domain targeted sentiment analysis (Traditional in the figure) and our work.
work extracts the target, aspect and sentiment at the document-level, and the sentiment model representations over diverse text domains are three-tuple failing to indicate agreement in each domain (F1); The average number of sentence (#S), tokens (#Tok), targets (#AT) in each domain; micro-F1 scores of annotator performance in open-domain texts (Orbach et al., 2021). Given the existence of pre-trained language model representations over diverse text domains (Kenton et al., 2019; Radford et al., 2019; Lewis et al., 2020), it is timely to investigate open-domain targeted sentiment analysis in more practical and realistic settings.

Second, existing work considers open-domain targeted sentiment analysis only at the sentence level. However, text sources in the open domain are typically in the form of documents, such as a piece of news, or a product review. Sentence-level sentiment analysis fails to give accurate information due to lack of co-reference and discourse knowledge. Take the simple sentence “It is quite useful in helping me with the housework.” from the dataset of Mitchell et al. (2013) for example, the gold-standard target entity is represented by the span “it”. However, significant post-processing can be necessary to correctly identify the true sentiment polarity on the target entity, which involves co-reference resolution and mention-level polarity information integration.

Third, complex relations are not fully extracted in the previous work, which just extracts opinion targets separately. For example, ‘The price of the food is high in that Italian restaurant’, the relationships of price, food and Italian restaurant are not implied in the previous datasets. Although some work extracts the target, aspect and sentiment at the same time (Yang et al., 2018; Saeidi et al., 2016), it is still limited in the extensibility, having restricted the schema of target expression e.g. food-price-Negative which is three-tuple failing to indicate Italian restaurant.

To address the above issues, we consider open-domain targeted sentiment analysis at the document level with a variety of text domains. A contrast between our dataset and traditional open-domain targeted sentiment analysis is shown in Figure 1. In particular, for increasing diversity, our data are sourced from six different domains with three different linguistic genres. To address the limitation on span-based target representation, we define the problem of open-domain targeted sentiment as a fully end-to-end task, where the input is a document and the output is a list of mentioned target entities with their sentiment polarities.

While pre-trained sequence-to-sequence models are useful for solving our task, results show that there exists much need for further improvement. Challenges exist in the effective use of open-domain data, long documents, the complexity of target structure, and domain gaps. To our knowledge, we are the first to consider the open-domain targeted sentiment analysis in a document-level setting. We will release our code and dataset later.

2 Dataset

Our proposed dataset contains six domains, including book reviews, clothing reviews, restaurant reviews, hotel reviews, financial news and social media data (PhraseBank). The details of data sources are shown in Appendix B.

2.1 Annotation Schema

The document-level sentiment task enables us to comprehensively extract (1) relations among entities and aspects; (2) the inference of implicit entities and aspects. Considering that targets can have fine-grained levels of specificity (e.g., restaurant-food-price), we denote sentiment targets with tuples, where all the targets and their relations are extracted in a nested data structure. To allow better document-level representation and avoid noise, we adopt the [Positive, Negative, Mixed] sentiment schema (Orbach et al., 2021). The task form is shown Figure 1, and our schema has the advantages of comprehensibility, extensibility and specificity.
2.2 Annotation Procedure

The procedure of annotation is shown in Figure 2. Each domain is distributed three different annotators, who are trained before making annotation. The data is divided into the annotation and validation sections – the former is allocated to one of the annotators, and the latter is annotated by all the three annotators. After annotation, we calculate the F1 score to check annotation precision on the validation section. If the F1 score does not reach an acceptable level, we discuss about the issues and revise the annotation guidelines when necessary, and the data are re-annotated. If the F1 score reaches an acceptable level, the data are re-checked by one more individual. The details of the final annotation rules are shown in Appendix A.

Considering the complexity of the nested target structure, we use a loose-match score replacing the exact-match score in the calculation of the F1 score, which is also used in our experimental evaluation. The exact-match score means that each labeled target is assigned correct score 1.0 only if all the components and the sentiment are the same with the golden text. But in loose-match score for each target if the sentiment is correct, we calculate the ratio of overlapped nests in labels and the golden text, and if the ratio reaches acceptable levels, we assign it with corresponding scores. The loose-match score is reasonable for the annotation because the components of nested targets tend to have similar sentiments. For example in Figure 1 (a), in Italian restaurant - food - price - Negative, the target components food and Italian restaurant also tend to have negative polarities for high price. The acceptable levels we set 0.5 and 0.66 with the corresponding score 0.5, 1.0.

2.3 Analysis and Statistics

Table 1 shows statistics in each domain of our dataset. First, the numbers of documents are roughly the same for each domain, with all domains having more than 900 documents. Second, the average number of sentences is the lowest in PhraseBank domain which is one feature of the PhraseBank, and it is the largest in News domain. The average number of targets is the largest in Restaurant reviews implying the difficulty in this domain is the largest. The numbers of targets in the different number of target nests (last 3 columns in 1) show that most of the targets are 1-nest and 2-nest, some are 3-nest and few are 4-nest (which is neglected in Table 1). Third, label imbalance exists in the dataset, with positive sentiments being the dominant. We did not deliberately control the label distribution, to keep it as close to practical situations as feasible (similar to Pontiki et al. (2014, 2015, 2016)).

3 Approach

In our schema, the nested opinion targets are in a structure that involves the relations of each component and inference of implicit targets, which can be challenging for traditional structured prediction models (Mitchell et al., 2013; Zhang et al., 2015). Neural sequence-to-sequence modeling provides a useful solution (Vinyals et al., 2015), and we take BART (Lewis et al., 2020) as the sequence-to-sequence framework, which is a denoising autoencoder for pre-training sequence-to-sequence models based on Transformer (Vaswani et al., 2017). BART has shown to be particularly effective in tasks of text summarization, machine translation, information retrieval and sequence generation (Lewis et al., 2020; Liu et al., 2020b; Chen and Song, 2021; Liu et al., 2021; Yan et al., 2021).

3.1 Model

We consider both the joint task of open-domain targeted sentiment analysis and its subtasks. Formally our model takes \( X = [x_1, x_2, ..., x_n] \) as inputs, and output a target sequence \( Y = [y_0, y_1, ..., y_m] \) where \( y_0 \) is the start token of the sentence. For target sentiment classification, the output is \( Y \) a sentiment polarity.
3.1.1 Opinion Target Extraction
For opinion target extraction, the target sequence \([y_1, ..., y_m]\) is a target list \([t_1, t_2, ..., t_l]\). Each element is \(t_j = [e_b, ..., e_i, ..., e_e]\) where \(e_b, e_e\) is the beginning and ending token of each target respectively, and \(e_i\) is the token to separate the nest structure of targets, e.g. \([e_b, \text{restaurant}, e_i, \text{food}, e_e, \text{positive}, e_se]\).

3.1.2 Target Sentiment Classification
For target sentiment classification, we set a target set for each document \(T = \{t_1, t_2, ..., t_{|L|}\}\) where \(|L|\) is the number of targets for each document in the dataset and the sentiment polarity set \(P = \{p_1, p_2, ..., p_{|C|}\}\) where \(|C|\) is the number of sentiment polarity in the task. Each element \(t_j = [e_b, ..., e_i, ..., e_e]\) is in the same format mentioned above. Similar to (Liu et al., 2021), we create the templates \(T_{i,p_k} = [w_1, w_2, ..., w_l] = [t_j + p_k, e_se]\) (e.g. \([e_b, \text{restaurant}, e_i, \text{food}, e_e, \text{positive}, e_se]\)). For a given target set, we can obtain a list of templates \(T_j = [T_{ij,p_1}, T_{ij,p_2}, ..., T_{ij,p_{|C|}}]\). Then we feed the template sets into fine-tuned pre-trained generative language model to sign a score for each template \(T_{ij,p_k} = [w_1, w_2, ..., w_l]\), formulate as

\[
f(T_{ij,p_k}) = \sum_{i=1}^{l} \log P(w_i|w_{1,i-1}, X) \tag{1}
\]

We choose the sentiment polarity of the target \(t_j\) with the largest score.

3.1.3 Open-domain Targeted Sentiment Analysis
For open-domain targeted sentiment analysis, the target sequence \([y_1, ..., y_m]\) is a target list \([t_1, t_2, ..., t_l]\). Each element is \(t_j = [e_b, ..., e_i, ..., e_e, s_j, e_se]\), where \(e_b, e_e, e_se\) are the beginning, ending tokens of each target, and ending token of sentiment respectively. \(e_i\) is to separate the nest structure of targets, \(s_j\) is the sentiment towards this target such as \([e_b, \text{restaurant}, e_i, \text{food}, e_e, \text{positive}, e_se]\).

3.1.4 Training
In opinion target extraction and open-domain targeted sentiment analysis, the gold texts are given directly as \(Y\). For target sentiment classification, gold texts are generated for each target with a gold polarity which we use \(Y\) to represent as well.

Given a sequence input \(X\), we feed the input \(X\) into BART encoder to obtain the hidden states:

\[
h_{\text{encoder}} = \text{BARTEncoder}(X) \tag{2}
\]

At the \(i\)th step of the BART decoder, the generated output tokens \(y_{1:i-1}\) are taken as inputs to yield a representation

\[
h_{\text{decoder}} = \text{BARTDecoder}(h_{\text{encoder}}, y_{1:i-1}) \tag{3}
\]

The loss function for training instance \((X, Y)\) is formulated as

\[
\mathcal{L} = -\sum_{i=1}^{m} \log P(y_i|y_{1,i-1}, X) \tag{4}
\]

4 Experiments
We conduct experiments for verifying the influence of the open-domain data, the document length, the complex target structure and the model structure in open-domain targeted sentiment analysis.

4.1 Experimental Settings
We perform experiments using the official pre-trained BART model provided by Huggingface\(^1\). The max sequence length for inputs is 512, and the max sequence length for output generation

\(^1\)https://huggingface.co/facebook/bart-base
We carry out experiments over the single-domain and multi-domain settings. The performance of the pipeline model, we train the model on a single domain and the data in each domain, respectively. Then, we carry out experiments over the single-domain setting by training the model on a single domain and test the model on the corresponding test data.

**Multi-domain and single-domain settings.** We first mix up the data on the six domains and fine-tune the BART model over a multi-domain setting before testing the trained model on the mixed data and the data in each domain, respectively. Then, we carry out experiments over the single-domain setting, by training the model on a single domain and test the model on the corresponding test data.

**Test on complex nested target structure.** For exploring the influence of complex nested target structure, we try to mix the dataset and split it w.r.t. the number of target nests. The statistics of the number of targets with different numbers of nests in each domain is shown in Table 1 (last 4 columns). Then we train and test the model on each data split with the number of target nests. The statistics of the number of target nests in each domain is shown in Table 1 (last 4 columns). Then we train and test the model on each data split with the number of target nests.

**Out-of-domain test.** Models for open-domain targeted sentiment analysis are expected to learn sufficient knowledge about various domains and be applied to unseen domains for open-domain requirements. We design 5-1 (1-1) out-of-domain tests, indicating that we use training data on five (one) domains to train the model, and test the model on another domain.

**Pipeline model.** In order to evaluate the performance of the pipeline model, we train the model of opinion target extraction and target sentiment classification on each domain separately and test on the corresponding test data by concatenating the two models together.

### 4.2 Overall Results

First, according to the results of multi-domain setting in Table 4, the loose-match evaluation score (F1 = 31.25) is relatively higher than the exact match evaluation score (F1 = 21.98). The results of BART-Loose (F1 = 31.25) provide evidence that there exist much room for improvement in open-domain targeted sentiment analysis comparing with the F1 score reported by the previous traditional work (Hu et al., 2019) where the F1 scores of LAPTOP, REST, Twitter are 68.06, 57.69 and 74.92, respectively.

Second, the results of the multi-domain setting trained BART model of each domain are shown in Table 2 (first seven columns). The performance of open-domain targeted sentiment analysis on Books (31.90), Restaurant (20.82) and News (14.52) domains are relatively the weakest. This could be impacted by different factors, including the influence of documents, domains, and target structure, which are analyzed in Section 4.3, 4.4, and 4.5.

Third, it is worth noting that the average recall values (36.66 and 28.89) for opinion target extraction and open-domain targeted sentiment analysis are all higher than the precision (53.70 and 42.21).
4.3 Influence of Document-level Inputs

We are interested in understanding the influence of documents for open-domain targeted sentiment analysis, which can be characterized by the average number of tokens or sentences. In particular, we illustrate the relation between the document length and the performance on the multi-domain setting (the illustration on single-domain setting is similar) in Figure 4. The Figure shows the performance of the model on open-domain targeted sentiment analysis and opinion target extraction shows a strong correlation to the average number of tokens or sentences, which is one characteristic in the document-level task. With the increase of tokens or sentences, the performance of open-domain targeted sentiment analysis and opinion target extraction decreases significantly. But for target sentiment classification, the performance does not have such an obvious relation (Figure 4 (c)(d)) as shown before. This implies that the model can be negatively affected by the document length for open-domain targeted sentiment analysis.

4.4 Influence of Complex Target Structure

According to the results shown in Table 5, the F1 scores of the 1-nest, 2-nest, and 3-nest settings are 38.31 29.59 and 23.9 showing that the number of target nests negatively affects the performance. The F1 score in the 3-nest target setting is 14.41 lower than that in 1-nest targets experiment. It implies another reason why the performance of Restaurant (with a large number of 2-nest targets (2566)) is
Figure 5: Comparisons between out-of-domain tests and the multi-domain setting. ● symbol for F1 score of the multi-domain setting, ▲ symbol for F1 score of 5-1 out-of-domain test and ■ symbol for F1 score of 1-1 out-of-domain test.

Table 8: Results of pipeline model on the single-domain setting.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books</td>
<td>35.46</td>
<td>29.94</td>
<td>32.46</td>
</tr>
<tr>
<td>Clothing</td>
<td>45.00</td>
<td>41.44</td>
<td>43.14</td>
</tr>
<tr>
<td>Restaurant</td>
<td>35.60</td>
<td>26.31</td>
<td>30.25</td>
</tr>
<tr>
<td>Hotel</td>
<td>59.76</td>
<td>41.82</td>
<td>49.20</td>
</tr>
<tr>
<td>News</td>
<td>17.85</td>
<td>11.29</td>
<td>13.83</td>
</tr>
<tr>
<td>PhraseBank</td>
<td>58.60</td>
<td>51.57</td>
<td>54.86</td>
</tr>
<tr>
<td><strong>Avg</strong></td>
<td><strong>42.04</strong></td>
<td><strong>33.72</strong></td>
<td><strong>37.29</strong></td>
</tr>
</tbody>
</table>

4.5 Influence of Domain

The results of 5-1 out-of-domain test are shown in Table 6. In particular, the average F1 scores is 17.22, which is 16.75 lower than that on the multi-domain setting. The performance decay implies the generalization performance of the model on our dataset is weak, for the reason huge gaps exist among the domains. Then the results of 1-1 out-of-domain test are shown in Table 7. The average F1 scores of 1-1 out-of-domain test is 14.59, which is 20.38 lower than that on the multi-domain setting, also lower than that on the 5-1 setting. It suggests open-domain data can help to boost the performance of generalization. The visualization of results in the 5-1 test, 1-1 test and the multi-domain setting is shown in Figure 5. The performance in News domain (1.98 and 1.73 in 5-1 and 1-1 tests) is especially low, that the model can scarcely learn useful knowledge from other domains for news domain. Note that the results on 1-1 out-of-domain test are better than that on 5-1 test in Hotel (18.39-17.69) and PhraseBank (34.18-28.76), which implies there exist great differences among the domain gaps and more open-domain data does not mean better-trained models.

4.6 Pipeline vs Joint Models

Different from the observation of Mitchell et al. (2013), Zhang et al. (2015) and Hu et al. (2019), the average F1 score of the pipeline model (37.29) is better than the joint model (34.27). Better results of pipeline models (comparing with joint models) lie in the domains Restaurant (30.25-19.08), Hotel (39.20-37.12) and News (13.83-12.91). We notice the performance of the joint model is strongly related to the average number of targets in the dataset (Figure 6). With the increase of the average number of targets, the performance of the joint model becomes worse than the pipeline model. In the domains that the average number of targets is small (Books (2.50), Clothing (1.67), PhraseBank (1.23)), joint models perform better than pipeline models. Conversely, in the domains that the average number of targets is large (Restaurant (5.03), Hotel (3.33), News (2.91)), pipeline models have better performance.

4.7 Case Study

Table 9 shows two qualitative cases sampled from single-domain setting. As observed in the first case, the model output a partially correct answer (strap#Negative), but the information of the relation between shoe and strap is not extracted. Although the words ‘They fit almost perfectly’ and
‘it’s way too loose’ express positive and negative sentiments for the target shoe, but it is not extracted, which means that the model fails to infer the representation of ‘They’ and ‘it’. In the second case, Valerie’s place — stairs #Negative fails to be extracted when the model facing a relative large number of targets.

5 Related Work

Open-domain targeted sentiment analysis can be divided into two sub-tasks, namely, the opinion target extraction and target sentiment classification. Traditionally, the sub-tasks are solved separately (Lafferty et al., 2001; Shu et al., 2017; Zhang et al., 2016; Ren et al., 2016; Wang et al., 2017; Chen et al., 2017; Fan et al., 2018; Song et al., 2019), which can be pipelined together to solve the open-domain targeted sentiment analysis. The joint task of open-domain targeted sentiment analysis is modeled as an end-to-end span extraction problem (Zhou et al., 2019; Hu et al., 2019) or span tagging problem: tagging as {B, I, E, S} - {POS, NEG, NEU} and O (Mitchell et al., 2013; Zhang et al., 2015; Ma et al., 2018; Li et al., 2019a; Song et al., 2019; Pingili and Li, 2020). Recent work compares pipeline model and joint model (Mitchell et al., 2013; Zhang et al., 2015; Hu et al., 2019), finding that the pipeline model can achieve better performance.

Previous studies mainly conduct experiments on three datasets: (1) LAPTOP, product reviews from the laptop domain in SemEval 2014 challenge (Pontiki et al., 2014); (2) TWITTER, comprised by the tweets collected by Mitchell (Mitchell et al., 2013); (3) REST, a union of restaurant reviews in SemEval 2014, 2015, or 2016 (Pontiki et al., 2014, 2015, 2016). Some work also tries to propose datasets in news domain (Hamborg et al., 2021; Hamborg and Donnay, 2021) which are mainly on the political spectrum. To evaluate previous models’ ability to solve open-domain targeted sentiment analysis in various domains, Orbach et al. (2021) constructs a new evaluation dataset in extensive domains finding that there is ample room for improvement on this challenging new dataset.

Aspect-based sentiment analysis is a similar work, which aims to extract the aspect term and then identify its sentiment orientation, like (Li et al., 2019b; Chen et al., 2020; Wang et al., 2019; Chen and Qian, 2020; Liu et al., 2020a). The task needs to find the aspects related to the elements in a given aspect category set. But for open-domain targeted sentiment analysis, no pre-defined aspect categories are given. For example in LAPTOP dataset, ‘But the performance of Mac Mini is a huge disappointment.’ For the target ‘Mac Mini’ is not in the focused aspect categories, thus it is not labeled and only ‘performance’ is labeled. Some work tries to extract the target, aspect and sentiment at the same time (Yang et al., 2018; Saeidi et al., 2016), while it limits the extensibility. Meanwhile, document-level aspect-based sentiment analysis task is also studied in (Chen et al., 2020; Wang et al., 2019) to alleviate the information deficiency problem for the implicit targets (pronouns).

6 Conclusion

In this study, we propose a challenging dataset for open-domain targeted sentiment analysis to overcome the limitations of data domain and sentence level. By using the dataset, we expect to boost the effectiveness and practicality of this task. Benchmark performance using BART has demonstrated that the challenges exist in the effective use of open-domain data, long documents, the complexity of target structure and domain gaps.

Table 9: Case Study. The symbols ‘—–’, ‘#’ and $e_{se}$ represent the split, ending tokens of target and the ending token of sentiment, respectively. The beginning token of targets is neglected here for simplicity.
7 Ethical Statement

We honor the ACL Code of Ethics. No private data or non-public information was used in this work. All annotators have received labor fees corresponding to the amount of their annotated instances.

References


Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016. Improving twitter sentiment classification using topic-enriched multi-prototype word embeddings. In Thirtieth AAAI conference on artificial intelligence.


A Appendix: Rules for Annotation

A.1 Target Candidates and sentiment Annotation

General Instructions.
In this task you will review a set of documents. Your goal is to identify the nested items in the documents that have a sentiment expressed to them.

Steps
1. Read the documents thoroughly and carefully.
2. Identify the items that have a sentiment expressed to them.
3. Mark each item by the form of nested target structure connected by ‘–’ and for each nested target choose the expressed sentiment:
   (a). Positive: the expressed sentiment is positive.
   (b). Negative: the expressed sentiment is negative.
   (c). Mixed: the expressed sentiment is both positive and negative.
4. If there is no item with a sentiment expressed towards them, proceed to the next document.

Rules and Tips
1. The nest target structures are labeled as they appear in the document, even though they have overlapping parts (see example 2).
2. If the target of pronoun (it, this, that, etc.) could not be inferred from the whole text, the pronoun will be a target, but it will not be considered as a part of nested target structure (see example 2).
3. The sentiment should be expressed towards the marked items, it cannot come from with the marked item (see example 3).
4. Unfactual content will not be marked in conditional or subjunctive sentences (see example 5).
5. Verbs will not serve as targets even though there exist sentiment words towards them (see example 6).
6. “the” cannot be a part of a marked item. (see example 7).

A.2 Examples

1. Basics

Example 1.1: The food is good.

Answer: food # Positive
Explanation: The word good expresses a positive sentiment towards food.

Example 1.2: The food is awful.
Answer: food # Negative
Explanation: The word awful expresses a negative sentiment towards food.

Example 1.3: The food is tasty but expensive.
Answer: food # Mixed
Explanation: The word good expresses a positive sentiment towards food while the word awful expresses a negative sentiment towards food. So the correct sentiment to food is mixed.

Example 1.4: The restaurant is near downtown.
Answer: Nothing should be selected, for there is no sentiment expressed.

2. Nested target structure

Example 2.1: Good charger and is perfect because it also has a USB connection. Also love that it is original material it works like that too giving a quick charge when I need it.
Answer: charger # Positive
charger - USB connection # Positive
charger - material # Positive
charger - charge # Positive
Explanation: The word good expresses a positive sentiment towards charger, and the word perfect expresses a positive sentiment to the USB connection of charger. Meanwhile, the next sentence has positive sentiments towards material and charge separately, and they can be inferred to be a part of the charge.

Example 2.2: It charges my phone quickly and the cord is super long.
Answer: It # Positive
cord # Positive
Explanation: The word quickly expresses a positive sentiment to the target it while it cannot be inferred what it represents, then it is marked. For cord, although we can know cord is a part of it, but it will not be considered to be marked in the nested target structure.

Example 2.3: The food was served good for a meal in the Italian restaurant, but the atmosphere was awful.
Answer: Italian restaurant - food # Positive
Italian restaurant - atmosphere # Negative
Italian restaurant # Mixed
Explanation: The word good expresses a positive sentiment to the target food and food is a part of the Italian restaurant, meanwhile for the item food has marked, the duplicate item meal will not be marked. Then word awful expressed a negative opinion towards the atmosphere of the Italian restaurant. Further, from the two nested target items, the sentiment of Italian restaurant can be inferred to be mixed.

3. Sentiment location

Example 3.1: I love this great car.
Answer: car # Positive
Explanation: Both words love and great expresses positive sentiment towards car, so car is marked, but not great car is marked.

4. Long-document examples

Example 4.1: Could not power my S2 phone. The LG charger I was using had no problem but I needed a second charger. I thought buying an Official Samsung charger would be the best route to go. With nothing running on my phone except Waze and Audible (my usual combo when driving) the battery icon showed charging on AC, but was losing power at the rate of 5% per hour. On a long trip I was forced to turn the phone completely off for a few hours to get it to charge. In fairness it could have been a defective unit but I won’t be wasting time trying another of this model. The company has been very accommodating in the return. The return has been smooth and I WOULD buy from them again.
Answer: LG charger # Positive
Official Samsung charger # Negative
company # Positive
company–return # Positive

Example 4.3: My wife liked my Nokia 3650 so much that she switched chips with me and is carrying it. My favorite features:1. Speaker Phone. Nice when driving or multitasking. Good audible range. I slip it in my shirt pocket and speak into the air. Works great!2. Display is very good for its size. The camera takes 640 x 480 color images. I bought a 32 meg card to increase storage. I recently used the phone as my principle camera on vacation to the Smokies. Worked great.3. Contacts is a nice feature that can pull your chip's phone numbers and store them. Just add email addresses and you can send the camera pics to any email via the multimedia option. Disadvantages: The blue lighted round keyboard. In low light it is hard to see. This can be a problem when text-messaging or adding contact details. I’m buying a 2nd phone which will be another Nokia 3650. (...):)
Answer: Nokia 3650 # Mixed
Nokia 3650–Speaker # Positive
Nokia 3650–Speaker–audible range # Positive
Nokia 3650–display–size # Positive
Nokia 3650–camera # Positive
Nokia 3650–contact # Positive
Nokia 3650–multimedia option # Positive
Nokia 3650–keyboard # Negative

5. Unfactual content will not be marked in conditional or subjunctive sentences

Example 5.1: For example, if the Asia Pacific market does not grow as anticipated, our results could suffer.
Answer: Nothing should be selected, for the sentence is a conditional sentence.

6. Verbs not for targets

Example 6.1: Works well.
Answer: Nothing should be selected, for verbs will not be targets. It is normal to be marked in the ABSA work, for they can be aspects of the items.

7. “the” cannot be a part of a marked item

Example 7.1: The food is awful.
Answer: food # Negative
Error: The food # Negative

8. Idioms

Example 8.1: The laptop's performance was in the middle of the pack, but so is its price.
Answer: None

B Appendix: Data Source

Our proposed dataset contains six domains, including books reviews, clothing reviews, restaurant reviews, hotel reviews, financial news and social media data.
B.1 Dataset Sources

Raw document data are from several datasets or collected by ourselves and they are used for annotation inputs. The details are as follows:

1. **Books and Clothing.** The reviews of books and clothing are from ². The annotated data contains 986 book reviews and 928 clothing reviews which are randomly selected from the downloaded dataset. We used the data of books domain and clothing domain of 5-core version in this data source.

2. **Restaurant.** Restaurant reviews are in Boston, collected by Yelp (April 17, 2021).³ The annotated data contains 940 reviews which are randomly selected from the downloaded dataset (only restaurant reviews remain).

3. **Hotels.** Hotel reviews are in Boston, collected by AirBnb (February 19, 2021).⁴ The annotated data contains 1029 reviews which are randomly selected from the downloaded dataset.

4. **Social Media.** A random sample of 1194 sentences was chosen to represent the overall social media database⁵. Annotators were asked to consider the sentiment of sentences from the viewpoint of an investor only.

5. **Business News.** Our business news dataset was collected from Reuters⁶ and Bloomberg⁷ containing 936 news. In particular, Reuters News was collected from March 2021 to April 2021, resulting in 498 instances. While Bloomberg News was collected over the period from October 2006 to November 2013, resulting in 438 samples.

³https://www.yelp.com/dataset/download
⁴http://insideairbnb.com/get-the-data.html
⁵https://huggingface.co/datasets/financial_phrasebank
⁷https://github.com/philipperemy/financial-news-dataset