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 variances.

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

Traditional	price - Negative Food - Negative Service - Positive Italian restaurant - Mixed
Our Work	Italian restaurant - food - price - Negative Italian restaurant - food - Negative Italian restaurant - service - Positive Italian restaurant - Mixed

(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.

Our Work	charger - USB connection - Negative charger - Seller - Negative charger - Negative

(b) Multi-sentence Example

Figure 1: Traditional open-domain targeted sentiment analysis (Traditional in the figure) and our work.

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 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 neces-

Domain	#Doc	#T	#P	#N	#M	F1	#S	#Tok	#AT	#1-n	#2-n	#3-n
Books	986	2,470	1,624	542	304	59.06	7.59	109.10	2.50	1,465	988	17
Clothing	928	1,555	1034	299	222	60.51	4.54	44.12	1.67	1,166	385	4
Restaurant	940	4,739	3,457	828	454	57.44	10.08	116.63	5.03	1,943	2,566	221
Hotel	1,029	3,436	3,165	154	117	72.07	5.24	55.63	3.33	1,408	1,795	231
News	936	2,725	1,358	1,254	113	75.34	12.53	175.72	2.91	2,053	618	53
PhraseBank	1,194	1,481	1,006	464	11	75.04	1.00	23.30	1.23	918	541	49

Table 1: Details for our proposed datasets, include the number of documents (#Doc) and targets (#T) in each domain, the number of Positive (#P), Negative (#N), Mixed (#M) sentiment labels, micro-F1 scores of annotator agreement (F1 for micro F1 score, henceforth), the average number of sentence (#S), tokens (#Tok), targets (#AT) and the number of 1-nest targets (#1-n), 2-nest targets (#2-n) and 3-nest targets (#3-n) in each domain.

sarily reflect strong performance in open-domain texts (Orbach et al., 2021). Recent availability of representation models pre-trained on diverse text domains (Kenton et al., 2019; Radford et al., 2019; Lewis et al., 2020) allows us to investigate open-domain targeted sentiment 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 models fail 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 is a large gap for further improvement. Challenges exist in the effective use of open-domain data, long documents, the complexity of target structure, and domain variances. To our knowledge, we are the first to consider the open-domain targeted sentiment analysis in the 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 task form is shown Figure 1, which shows a balance between comprehensibility, extensibility and specificity. 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).

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.

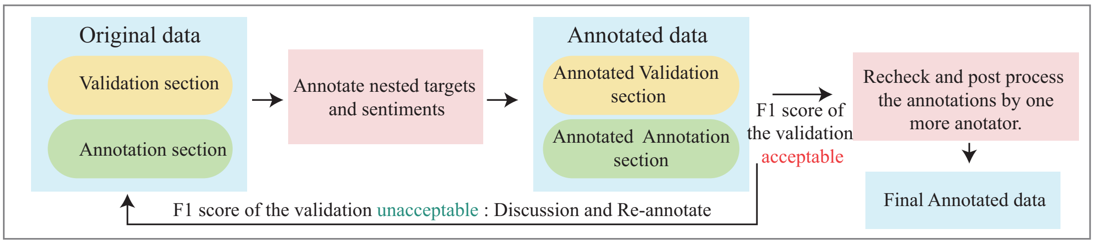

Figure 2: Annotation procedure of our proposed dataset.

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 at least two annotators. After annotation, we calculate the average micro F1 score of each two annotators to check annotation agreements on the validation section. The F1 score is calculated in phrase level for the reason we consider the relations of target components in the evaluation procedure, similar to Kim and Klinger (2018). 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 rechecked 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 chosen 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 the 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 smallest in the **PhraseBank** domain which is one feature of the **PhraseBank**, and the largest in the **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 $Y_t = [y_0, y_1, ..., y_m]$ where y_0 is the start token for BART. For target sentiment classification, the output is Y_s a polarity in an given text template.

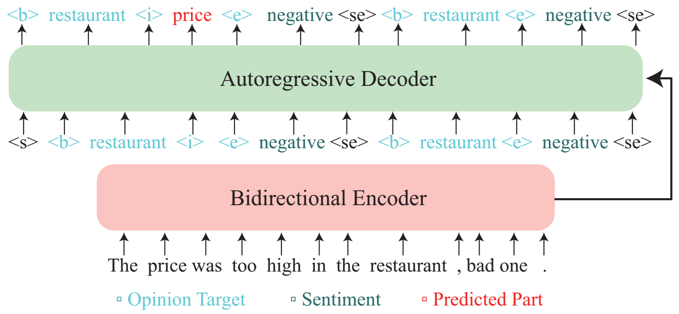

Figure 3: Pre-trained model for generation method in open-domain targeted sentiment analysis task.

3.1.1 Opinion Target Extraction

For opinion target extraction, the target sequence $[y_1, ..., y_m]$ (not includes the beginning token for BART) 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 are the beginning and ending token of each target respectively, and e_i is the token to separate the nest structure of targets. For instance, given the input '*The food in this restaurant is awful*', the output is $[e_b,\ restaurant,\ e_i,\ food,\ e_e,\ e_b,\ restaurant,\ e_e]$.

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 $\mathbf{T}_{t_j,p_k} = w_1, w_2, ..., w_l = [t_j + p_k, e_{se}]$ (e.g. $[e_b,\ restaurant,\ e_i,\ food,\ e_e,\ positive,\ e_{se}]$). For a given target set, we can obtain a list of templates $\mathbf{T}_{t_j} = [T_{t_j,p_1}, T_{t_j,p_2}, ..., T_{t_j,p_{|C|}}]$, and feed the template sets into fine-tuned pre-trained generative language model to assign a score to each template $\mathbf{T}_{t_j,p_k} = w_1, w_2, ..., w_l$:

$$f(\mathbf{T}_{t_j,p_k}) = \sum_{i=1}^{l} logP(w_i|w_{1,i-1}, X) \quad (1)$$

We choose the sentiment polarity with the largest score for the target t_j.

3.1.3 Open-domain Targeted Sentiment Analysis

For open-domain targeted sentiment analysis, the target sequence $[y_1, ..., y_m]$ (not includes the beginning token for BART) 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. For example, given the input '*The food is too awful*', the model output is $[e_b,\ food,\ e_e,\ negative,\ e_{se}]$.

3.1.4 Training

In opinion target extraction and open-domain targeted sentiment analysis, the gold outputs are given directly as a token list $Y_t = [y_0, y_1, ..., y_m]$ where y_0 denotes the start token for BART. For target sentiment classification, gold texts are generated for each target with a gold polarity, which we use a token token sequence Y_s to represent.

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

$$\mathbf{h}^{encoder} = BARTEncoder(X) \quad (2)$$

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

$$\mathbf{h}_i^{decoder} = BARTDecoder(h^{encoder}, y_{1:i-1}) \quad (3)$$

The loss function for the training instance (X, Y_t) or (X, Y_s) is formulated as

$$\mathcal{L} = -\sum_{i=1}^{m} logP(y_i|y_{1,i-1}, X) \quad (4)$$

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.

Domain	OTE			TSC	OTSA			OTSA-Single		
	Precision	Recall	F1	Precision	Precision	Recall	F1	Precision	Recall	F1
Books	56.84	38.12	45.63	73.85	40.65	26.25	31.90	43.02	29.17	34.76
Clothing	62.93	47.20	53.94	83.55	49.36	38.32	43.14	60.67	41.66	49.40
Restaurant	47.11	25.46	33.05	83.26	32.00	15.44	20.82	35.98	12.99	19.08
Hotel	68.85	44.14	53.79	95.69	50.39	29.38	37.12	47.67	26.64	34.17
News	23.16	10.93	14.85	69.94	20.23	11.33	14.52	18.57	9.90	12.91
PhraseBank	63.28	54.10	58.32	91.48	60.67	52.62	56.35	58.92	52.05	55.27
Avg	53.70	36.66	43.26	82.96	42.21	28.89	33.98	44.13	28.73	34.27

Table 2: Experimental results (OTE for opinion target extraction task, TSC for target sentiment classification task and OTSA for results of open-domain targeted sentiment analysis on the multi-domain setting; OTSA-Single for results of open-domain targeted sentiment analysis on the single-domain setting).

4.1 Experimental Settings

We perform experiments using the official pre-trained BART model provided by Huggingface[1]. The maximum input sequence length is 512, and the maximum output sequence length is 100. We split our dataset into training/validation/testing sets in the same ratio of 7:1:2 for all tasks. The best model configuration is selected according to the highest performance on the validation set. In particular, the batch size 4, learning rate is initialized as 1e-4, our model is trained for 20 epochs. The experiments include:

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 datasets and split the data 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). We train and test the model on each data split of different numbers of nests (1-nest, 2-nest and 3-nest) respectively.

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, using training data on five (one) domains to train the model, and testing 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 model pipeline.

4.2 Overall Results

First, the loose-match evaluation scores of the multi-domain setting experiment on test mixed data are precision 41.40, recall 25.10, F1 31.25, relatively higher than the exact match evaluation score (precision 19.13, recall 17.66, F1 21.98). The values of loose-match evaluation scores 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. Meanwhile, the F1 score of Transformer model on mixed test data is only 3.76, which indicates the significance of using pre-trained models for the task.

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 due to different factors including the size of documents, domains, and target structure, which are analyzed in Section 4.3, 4.4, and 4.5, respectively.

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 lower than the precision (53.70 and 42.21). It suggests that the model tends to output more correct targets and sentiments, but fails to identify all the targets and sentiments. Then, by comparing the results on opinion target extraction and target sentiment classification, the precision of the latter task (82.96) is strongly better than the former (53.70),

[1]https://huggingface.co/facebook/bart-base

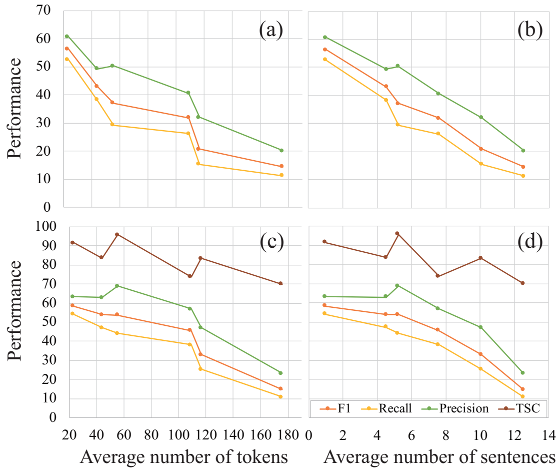

Figure 4: The relation between the document length and the performance on the multi-domain setting. (a) (b) for open-domain targeted sentiment analysis; (c) (d) for opinion target extraction and target sentiment classification.

	Precision	Recall	F1
1-nest	47.45	**32.12**	**38.31**
2-nest	**50.16**	21.09	29.69
3-nest	29.72	20.00	23.90

Table 3: Results on different numbers of nests.

which implies the difficulty is extracting targets.

The results of the single-domain setting are shown in Table 2 (last 3 columns). The average F1 score of open-domain targeted sentiment analysis on the single domain setting is 34.27, better than that of the multi-domain setting (33.98). Overall, open-domain data do not help improve the performance of the model. The worse results for the multi-domain setting (comparing with single-domain setting) are on **Books** (31.90-34.76) and **Clothing** (43.14-49.40), which implies that no useful information could be obtained from other domains for these domains. But for **Restaurant** (20.82-19.08), **Hotel** (37.12-34.17), **News** (14.52-12.91) and **PhraseBank** (56.35-55.27), open-domain data can help boost the model performance. More effective use of open-domain data requires further research.

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 results show that the performance of the model on open-domain targeted sentiment analysis and opinion target extraction has 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.

Domain	Precision	Recall	F1
Books	29.30	12.68	17.69
Clothing	32.47	20.72	25.29
Restaurant	23.29	7.98	11.89
Hotel	27.78	13.00	17.69
News	3.84	1.33	1.98
PhraseBank	33.33	25.30	28.76
Avg	25.00	13.50	17.22

Table 4: Out-of-domain test results (5-1 experiments).

Domain	Precision	Recall	F1
P->B	13.86	6.14	8.50
B->C	16.02	15.13	15.56
C->R	7.19	1.97	3.09
R->H	28.99	13.58	18.49
H->N	1.87	1.62	1.73
N->P	38.07	31.02	34.18
Avg	17.67	11.08	13.59

Table 5: Out-of-domain test results (1-1 experiments) (P for PhraseBank, B for Books, C for Clothing, R for Restaurant, H for Hotel and N for News).

4.4 Influence of Complex Target Structure

According to the results shown in Table 3, 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 weak. Nested targets are challenging to identify which requires more inference for the relations between target components for open-domain targeted sentiment analysis.

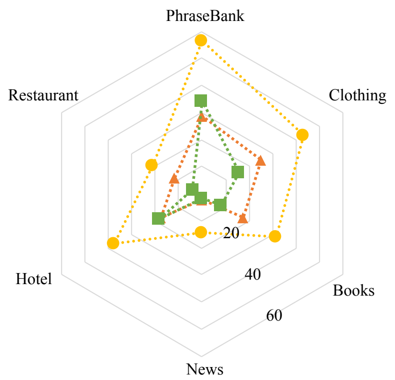

Figure 5: Comparisons between out-of-domain tests and the multi-domain setting. ● symbol for F1 score of the multi-domain setting, ▲ symbol for F1 score of 5-1 out-of-domain test and ■ symbol for F1 score of 1-1 out-of-domain test.

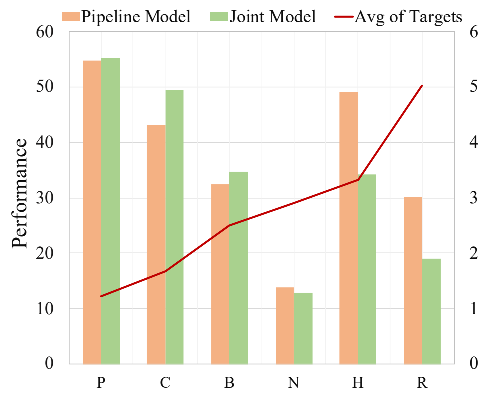

Figure 6: Comparisons between performance of pipeline model and joint model on the single-domain setting (P for PhraseBank, B for Books, C for Clothing, R for Restaurant, H for Hotel and N for News).

Domain	Precision	Recall	F1
Books	35.46	29.94	32.46
Clothing	45.00	41.44	43.14
Restaurant	35.60	26.31	30.25
Hotel	59.76	41.82	49.20
News	17.85	11.29	13.83
PhraseBank	58.60	51.57	54.86
Avg	42.04	33.72	37.29

Table 6: Single-domain setting results of pipeline model.

4.5 Influence of Domain

The results of 5-1 out-of-domain test are shown in Table 4. 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, due to the fact that much difference exists between the domains. The results of 1-1 out-of-domain test are shown in Table 5. 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 on the **News** domain (1.98 and 1.73 in 5-1 and 1-1 tests) is especially low, that the model can hardly 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 that more open-domain data does not always lead to 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 performer 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. The phenomenon may be due to the complexity of the generation content, i.e. with the increase of the length of outputs, it becomes harder to generate correct texts for open-domain targeted sentiment analysis, but opinion target extraction is relatively easier.

4.7 Case Study

Table 7 shows two qualitative cases from the single-domain setting. As observed in the first case, the model outputs a partially correct answer (*strap#Negative*), but the information of the relation between *shoe* and *strap* is not extracted. Al-

Context	Gold Labels	Output
I've ordered similar character shoes from other manufacturers and, as long as I size up. They fit almost perfectly... perhaps a tad big but a 7 would probably have been too snug. My dissatisfaction is with the strap. Even at the tightest supplied hole, it 's way too loose.	shoes — strap # Negative e_{se} shoes # Mixed e_{se}	strap # Negative e_{se}
Valerie 's place is spotless with a wonderful kitchen. The only thing that might be difficult for some is the need to climb 2 flights of stairs to access the bedroom. I would stay here again without hesitation.	Valerie 's place — stairs # Negative e_{se} Valerie 's place # Mixed e_{se} Valerie 's place — kitchen # Positive e_{se}	Valerie 's place # Mixed e_{se} Valerie 's place — kitchen # Positive e_{se}

Table 7: 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.

though the words '*They fit almost perfectly*' and '*it's way too loose*' express sentiments for the target *shoe*, it is not extracted, which means that the model fails to infer the anaphora of '*They*' and '*it*'. In the second case, *Valerie 's place — stairs #Negative* fails to be extracted when the model faces 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. As a solution to the challenging joint and pipeline tasks, we considered a single unified baselines using a seq2seq pre-trained language model, which is close to real-world practical settings. Benchmark performance demonstrated that the task is very difficult even given the current pre-trained technologies, and challenges exist in the effective use of open-domain data, long documents, the complexity of target structure and domain variances.

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

Peng. Chen, Z. Sun, L. Bing, and Y. Wei. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Xiao Chen, Changlong Sun, Jingjing Wang, Shoushan Li, Luo Si, Min Zhang, and Guodong Zhou. 2020. Aspect sentiment classification with document-level sentiment preference modeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3677, Online. Association for Computational Linguistics.

Yisong Chen and Qing Song. 2021. News text summarization method based on bart-textrank model. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 5, pages 2005–2010.

Zhuang Chen and Tieyun Qian. 2020. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3694, Online. Association for Computational Linguistics.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3433–3442.

Felix Hamborg and Karsten Donnay. 2021. NewsMTSC: A dataset for (multi-)target-dependent sentiment classification in political news articles. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 1663–1675.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2021. Towards Target-Dependent Sentiment Classification in News Articles. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12646 LNCS:156–166.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 537–546.

Jacob Devlin Kenton, Chang Ming-Wei, and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

John D Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019a. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6714–6721.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019b. Exploiting bert for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41.

Jian Liu, Zhiyang Teng, Leyang Cui, Hanmeng Liu, and Yue Zhang. 2021. Solving aspect category sentiment analysis as a text generation task. *arXiv preprint arXiv:2110.07310*.

Shu Liu, Wei Li, Yunfang Wu, Qi Su, and Xu Sun. 2020a. Jointly modeling aspect and sentiment with dynamic heterogeneous graph neural networks. *arXiv preprint arXiv:2004.06427*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Huaishao Luo, Lei Ji, Tianrui Li, Daxin Jiang, and Nan Duan. 2020. Grace: Gradient harmonized and cascaded labeling for aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 54–64.

Dehong Ma, Sujian Li, and Houfeng Wang. 2018. Joint learning for targeted sentiment analysis. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 4737–4742.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654.

Matan Orbach, Orith Toledo-Ronen, Artem Spector, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Yaso: A targeted sentiment analysis evaluation dataset for open-domain reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9154–9173.

Shashipal Reddy Pingili and Longzhuang Li. 2020. Target-based sentiment analysis using a bert embedded model. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1124–1128.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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*.

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

Lei Shu, Hu Xu, and Bing Liu. 2017. Lifelong learning crf for supervised aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 148–154.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Targeted Sentiment Classification with Attentional Encoder Network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11730 LNCS:93–103.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. 2015. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*.

Jingjing Wang, Changlong Sun, Shoushan Li, Jiancheng Wang, Luo Si, Min Zhang, Xiaozhong Liu, and Guodong Zhou. 2019. Human-like decision making: Document-level aspect sentiment classification via hierarchical reinforcement learning.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3316–3322. AAAI Press.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429, Online. Association for Computational Linguistics.

Jun Yang, Runqi Yang, Chongjun Wang, and Junyuan Xie. 2018. Multi-entity aspect-based sentiment analysis with context, entity and aspect memory. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 6029–6036.

Meishan Zhang, Yue Zhang, and Duy Tin Vo. 2015. Neural networks for open domain targeted sentiment. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, (September):612–621.

Meishan Zhang, Yue Zhang, and Duy Tin Vo. 2016. Gated neural networks for targeted sentiment analysis. *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, pages 3087–3093.

Yan Zhou, Longtao Huang, Tao Guo, Jizhong Han, and Songlin Hu. 2019. A span-based joint model for opinion target extraction and target sentiment classification. *IJCAI International Joint Conference on Artificial Intelligence*, 2019-Augus:5485–5491.

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

Explanation: A sentiment may be conveyed with an idiom – be sure you understand the meaning of an input sentence before answering. When unsure, look up potential idioms online. in the middle of the pack does not convey a positive nor a negative sentiment, and certainly not both (so the answer is not "mixed" as well).

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 database[5]. Annotators were asked to consider the sentiment of sentences from the view point of an investor only.
5. **Business News.** Our business news dataset was collected from Reuters[6] and Bloomberg[7] 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.

[2] https://nijianmo.github.io/amazon/index.html
[3] https://www.yelp.com/dataset/download
[4] http://insideairbnb.com/get-the-data.html
[5] https://huggingface.co/datasets/financial_phrasebank
[6] https://www.reuters.com/news/
[7] https://github.com/philipperemy/financial-news-dataset