YASO: A Targeted Sentiment Analysis Evaluation Dataset for Open-Domain Reviews

Matan Orbach, Orith Toledo-Ronen, Artem Spector,
Ranit Aharonov, Yoav Katz, Noam Slonim
IBM Research
{matano, oritht, artems, katz, noams}@il.ibm.com
ranit.aharonov2@ibm.com

Abstract

Current TSA evaluation in a cross-domain setup is restricted to the small set of review domains available in existing datasets. Such an evaluation is limited, and may not reflect true performance on sites like Amazon or Yelp that host diverse reviews from many domains. To address this gap, we present YASO – a new TSA evaluation dataset of open-domain user reviews. YASO contains 2215 English sentences from dozens of review domains, annotated with target terms and their sentiment. Our analysis verifies the reliability of these annotations, and explores the characteristics of the collected data. Benchmark results using five contemporary TSA systems show there is ample room for improvement on this challenging new dataset. YASO is available at github.com/IBM/yaso-tsa.

1 Introduction

Targeted Sentiment Analysis (TSA) is the task of identifying the sentiment expressed towards single words or phrases in texts. For example, given the sentence *"it's a useful **dataset** with a complex **download procedure**"* the desired output is identifying **dataset** and **download procedure**, with a positive and negative sentiments expressed towards them, respectively. Our focus in this work is on TSA of user reviews data in English.

Till recently, typical TSA evaluation was *in-domain*, for example, by training on labeled restaurant reviews and testing on restaurant reviews. New works (e.g. Rietzler et al. (2020)) began considering a *cross-domain* setup, training models on labeled data from one or more domains (e.g., restaurant reviews) and evaluating on others (e.g., laptop reviews). For many domains, such as car or book reviews, TSA data is scarce or non-existent. This suggests that cross-domain experimentation is more realistic, as it aims at training on a small set of labeled domains and producing predictions for reviews from any domain. Naturally, the *evaluation* in this setup should resemble real-world content from sites like Amazon or Yelp that host reviews from dozens or even hundreds of domains.[1]

Existing English TSA datasets do not facilitate such a broad evaluation, as they typically include reviews from a small number of domains. For example, the popular SemEval (SE) datasets created by Pontiki et al. (2014, 2015, 2016) (henceforth SE14, SE15, and SE16, respectively), contain English reviews of restaurants, laptops and hotels (see §2 for a discussion of other existing datasets). To address this gap, we present YASO,[2] a new TSA dataset collected over user reviews taken from four sources: the Yelp and Amazon (Keung et al., 2020) datasets of reviews from those two sites; the Stanford Sentiment Treebank (SST) movie reviews corpus (Socher et al., 2013); and the Opinosis dataset of reviews from over 50 topics (Ganesan et al., 2010). To the best of our knowledge, while these resources have been previously used for sentiment analysis research, they were not annotated and used for *targeted* sentiment analysis. The new YASO evaluation dataset contains 2215 annotated sentences, on par with the size of existing test sets (e.g., one of the largest is the SE14 test set, with 1,600 sentences).

The annotation of open-domain reviews data is different from the annotation of reviews from a small fixed list of domains. Ideally, the labels would include both targets that are explicitly mentioned in the text, as well as aspect *categories* that are implied from it. For example, in *"The restaurant serves good but expensive **food**"* there is a sentiment towards the explicit target **food** as well as towards the implied category **price**. This approach of aspect-based sentiment analysis (Liu, 2012) is implemented in the SE datasets. However, because the categories are domain specific, the annotation

[1] E.g. Yelp has more than 1,200 business categories here.

[2] The name is an acronym of the data sources.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9154–9173
November 7–11, 2021. ©2021 Association for Computational Linguistics

of each new domain in this manner first requires defining a list of relevant categories, for example, *reliability* and *safety* for cars, or *plot* and *photography* for movies. For open-domain reviews, curating these domain-specific categories over many domains, and training annotators to recognize them with per-domain guidelines and examples, is impractical. We therefore restrict our annotation to sentiment-bearing targets that are *explicitly present in the review*, as in the annotation of open-domain tweets by Mitchell et al. (2013).

While some information is lost by this choice, which may prohibit the use of the collected data in some cases, it offers an important advantage: the annotation guidelines can be significantly simplified. This, in turn, allows for the use of crowd workers who can swiftly annotate a desired corpora with no special training. Furthermore, the produced annotations are consistent across all domains, as the guidelines are domain-independent.

TSA annotation in a pre-specified domain may also distinguish between targets that are entities (e.g., a specific restaurant), a part of an entity (e.g., the restaurant's balcony), or an aspect of an entity (e.g., the restaurant's location). For example, Pontiki et al. (2014) use this distinction to exclude targets that represent entities from their annotation. In an open-domain annotation setup, making such a distinction is difficult, since the reviewed entity is not known beforehand.

Consequently, we take a comprehensive approach and annotate all sentiment-bearing targets, including mentions of reviewed entities or their aspects, named entities, pronouns, and so forth. Notably, pronouns are potentially important for the analysis of multi-sentence reviews. For example, given *"I visited the restaurant. **It** was nice."*, identifying the positive sentiment towards **It** allows linking that sentiment to the restaurant, if the coreference is resolved.

Technically, we propose a two-phase annotation scheme. First, each sentence is labeled by five annotators that should identify and mark all *target candidates* – namely, all terms to which sentiment is expressed in the sentence. Next, each target candidate, in the context of its containing sentence, is labeled by several annotators who determine the sentiment expressed towards the candidate – either positive, negative, or mixed (if any).[3] The full scheme is exemplified in Figure 1. We note that this scheme is also applicable to general non-review texts (e.g., tweets or news).

Several analyses are performed on the collected data: (i) its reliability is established through a manual analysis of a sample; (ii) the collected annotations are compared with existing labeled data, when available; (iii) differences from existing datasets are characterized. Lastly, benchmark performance on YASO was established in a cross-domain setup. Five state-of-the-art (SOTA) TSA systems were reproduced, using their available codebases, trained on data from SE14, and applied to predict targets and their sentiments over our annotated texts.

In summary, our main contributions are (i) a new domain-independent annotation scheme for collecting TSA labeled data; (ii) a new evaluation dataset with target and sentiment annotations of 2215 open-domain review sentences, collected using this new scheme; (iii) a detailed analysis of the produced annotations, validating their reliability; and (iv) reporting cross-domain benchmark results on the new dataset for several SOTA baseline systems. All collected data are available online.[4]

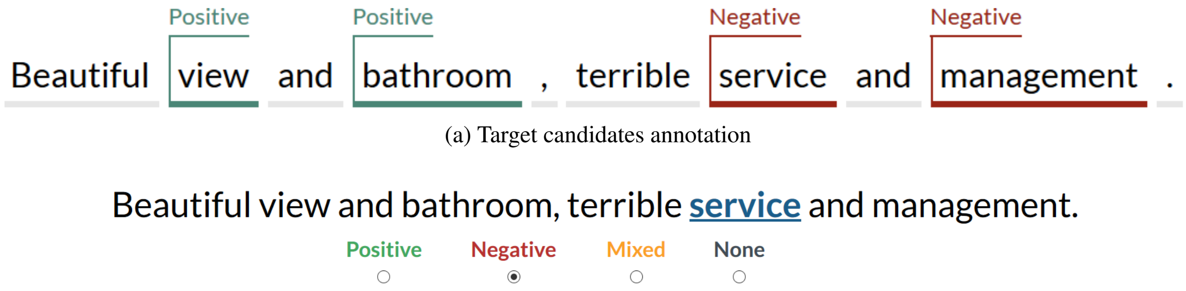
2 Related work

Review datasets The Darmstadt Review Corpora (Toprak et al., 2010) contains annotations of user reviews in two domains – online universities and online services. Later on, SE14 annotated laptop and restaurant reviews (henceforth SE14-L and SE14-R). In SE15 a third domain (hotels) was added, and SE16 expanded the English data for the two original domains (restaurants and laptops). Jiang et al. (2019) created a challenge dataset with multiple targets per-sentence, again within the restaurants domain. Saeidi et al. (2016) annotated opinions from discussions on urban neighbourhoods. Clearly, the diversity of the reviews in these datasets is limited, even when taken together.

Non-review datasets The Multi-Purpose Question Answering dataset (Wiebe et al., 2005) was the first opinion mining corpus with a detailed annotation scheme applied to sentences from news documents. Mitchell et al. (2013) annotated open-domain tweets using an annotation scheme similar to ours, where target candidates were annotated for their sentiment by crowd-workers, yet the annotated terms were limited to automatically detected

[3] Mixed: a positive and a negative sentiment towards one target, e.g., for **car** in *"a beautiful yet unreliable **car**"*.

[4] github.com/IBM/yaso-tsa


Positive
Positive
Negative
Negative
Beautiful view and bathroom , terrible service and management .

(a) Target candidates annotation

Beautiful view and bathroom, terrible service and management.

○ Positive ● Negative ○ Mixed ○ None

(b) Sentiment annotation

Figure 1: The UI of our two-phase annotation scheme (detailed in §3): *Target candidates annotation* (top) allows multiple target candidates to be marked in one sentence. In this phase, aggregated sentiments for candidates identified by a few annotators may be incorrect (see §5 for further analysis). Therefore, marked candidates are passed through a second *sentiment annotation* (bottom) phase, which separately collects their sentiments.

named entities. Other TSA datasets on Twitter data include targets that are either celebrities, products, or companies (Dong et al., 2014), and a multi-target corpus on UK elections data (Wang et al., 2017a). Lastly, Hamborg et al. (2021) annotated named entities for their sentiment within the news domain.

Multilingual Other datasets exist for various languages, such as: Norwegian (Øvrelid et al., 2020), Catalan and Basque (Barnes et al., 2018), Chinese (Yang et al., 2018), Hungarian (Szabó et al., 2016), Hindi (Akhtar et al., 2016), SE16 with multiple languages (Pontiki et al., 2016), Czech (Steinberger et al., 2014) and German (Klinger and Cimiano, 2014).

Annotation Scheme Our annotation scheme is reminiscent of two-phase data collection efforts in other tasks. These typically include an initial phase where annotation candidates are detected, followed by a verification phase that further labels each candidate by multiple annotators. Some examples include the annotation of claims (Levy et al., 2014), evidence (Rinott et al., 2015) or mentions (Mass et al., 2018).

Modeling TSA can be divided into two subtasks: target extraction (TE), focused on identifying all sentiment targets in a given text; and sentiment classification (SC), of determining the sentiment towards a specific candidate target in a given text. TSA systems are either *pipelined* systems running a TE model followed by an SC model (e.g., Karimi et al. (2020)), or *end-to-end* (sometimes called *joint*) systems using a single model for the whole task, which is typically regarded as a sequence labeling problem (Li and Lu, 2019; Li et al., 2019a; Hu et al., 2019; He et al., 2019). Earlier works (Tang et al., 2016a,b; Ruder et al., 2016; Ma et al., 2018; Huang et al., 2018; He et al., 2018) have utilized pre-transformer models (see surveys by Schouten and Frasincar (2015); Zhang et al. (2018)). Recently, focus has shifted to using pre-trained language models (Sun et al., 2019; Song et al., 2019; Zeng et al., 2019; Phan and Ogunbona, 2020). Generalization to unseen domains has also been explored with pre-training that includes domain-specific data (Xu et al., 2019; Rietzler et al., 2020), adds sentiment-related objectives (Tian et al., 2020), or combines instance-based domain adaptation (Gong et al., 2020).

3 Input Data

The input data for the annotation was sampled from the following datasets:

– **Yelp**:[5] A dataset of 8M user reviews discussing more than 200k businesses. The sample included 129 reviews, each containing 3 to 5 sentences with a length of 8 to 50 tokens. The reviews were sentence split, yielding 501 sentences.

– **Amazon**:[6] A dataset in 6 languages with 210k reviews per language (Keung et al., 2020). The English test set was sampled in the same manner as Yelp, yielding 502 sentences from 151 reviews.

– **SST**:[7] A corpus of 11,855 movie review sentences (Socher et al., 2013) originally extracted from Rotten Tomatoes by Pang and Lee (2005). 500 sentences, with a minimum length of 5 tokens, were randomly sampled from its test set.

– **Opinosis**:[8] A corpus of 7,086 user review sen-

[5] www.yelp.com/dataset

[6] registry.opendata.aws/amazon-reviews-ml

[7] nlp.stanford.edu/sentiment

[8] github.com/kavgan/opinosis-summarization

tences from Tripadvisor (hotels), Edmunds (cars), and Amazon (electronics) (Ganesan et al., 2010). Each sentence discusses a topic comprised of a product name and an aspect of the product (e.g. "performance of Toyota Camry”). At least 10 sentences were randomly sampled from each of the 51 topics in the dataset, yielding 512 sentences.

Overall, the input data includes reviews from many domains not previously annotated for TSA, such as books, cars, pet products, kitchens, movies or drugstores. Further examples are detailed in Appendix A.

The annotation input also included 200 randomly sampled sentences from the test sets of SE14-L and SE14-R (100 per domain). Such sentences have an existing annotation of targets and sentiments, which allows a comparison against the results of our proposed annotation scheme (see §5).

4 YASO

Next, we detail the process of creating YASO. An input sentence was first passed through two phases of annotation, followed by several post-processing steps. Figure 2 depicts an overview of that process, as context to the details given below.

4.1 Annotation

Target candidates annotation Each input sentence was tokenized (using `spaCy` by Honnibal and Montani (2017)) and shown to 5 annotators who were asked to mark target candidates by selecting corresponding token sequences within the sentence. Then, they were instructed to identify the sentiment expressed towards the candidate – positive, negative, or mixed (Figure 1a).

This step is recall-oriented, without strict quality control, and some candidates may be detected by only one or two annotators. In such cases, sentiment labels based on annotations from this step alone may be incorrect (see §5 for further analysis).

Selecting multiple non-overlapping target candidates in one sentence was allowed, each with its own sentiment. To avoid clutter and maintain a reasonable number of detected candidates, the selection of overlapping spans was prohibited.

Sentiment annotation To verify the correctness of the target candidates and their sentiments, each candidate was highlighted within its containing sentence, and presented to 7 to 10 annotators who were asked to determine its sentiment (without being shown the sentiment chosen in the first phase). For cases in which an annotator believes a candidate was wrongly identified and has no sentiment expressed towards it, a "none" option was added to the original labels (Figure 1b).

To control the quality of the annotation in this step, test questions with an a priori known answer were interleaved between the regular questions. A per-annotator accuracy was computed on these questions, and under-performers were excluded. Initially, a random sample of targets was labeled by two of the authors, and cases in which they agreed were used as test questions in the first annotation batch. Later batches also included test questions formed from unanimously answered questions in previously completed batches.

All annotations were done using the Appen platform.[9] Overall, 20 annotators took part in the target candidates annotation phase, and 45 annotators worked on the sentiment annotation phase. The guidelines for each phase are given in Appendix B.

4.2 Post-processing

The sentiment label of a candidate was determined by majority vote from its sentiment annotation answers, and the percentage of annotators who chose that majority label is the annotation *confidence*. A threshold t defined on these confidence values (set to 0.7 based on an analysis detailed below) separated the annotations between high-confidence targets (with confidence $\geq t$) and low-confidence targets (with confidence $< t$).

A target candidate was considered as *valid* when annotated with high-confidence with a particular sentiment (i.e., its majority sentiment label was not "none"). The valid targets were clustered by considering overlapping spans as being in the same cluster. Note that non-overlapping targets may be clustered together, for example, if t_1, t_2, t_3 are valid targets, t_1 overlaps t_2 and t_2 overlaps t_3, then all three are in one cluster, regardless of whether t_1 and t_3 overlap. The sentiment of a cluster was set to the majority sentiment of its members.

The clustering is needed for handling overlapping labels when computing recall. For example, given the input *"The food was great"*, and the annotated (positive) targets **The food** and **food**, a system which outputs only one of these targets should be evaluated as achieving full recall. Representing both labels as one cluster allows that (see details in §6). An alternative to our approach is considering

[9] `www.appen.com`

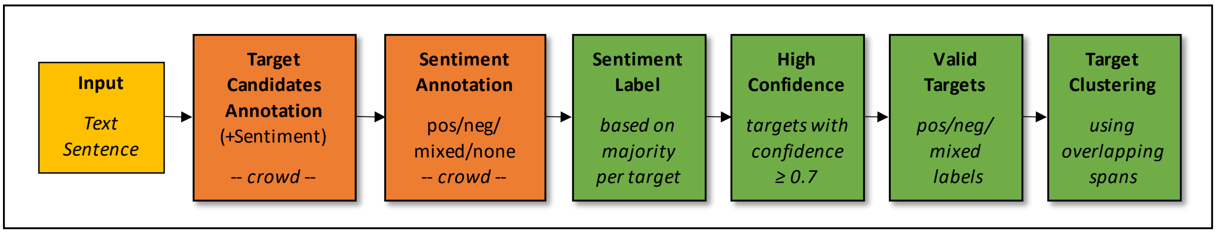


Figure 2: The process for creating YASO, the new TSA evaluation dataset. An input sentence is passed through two phases of annotation (in orange), followed by four post-processing steps (in green).

any prediction that overlaps a label as correct. In this case, continuing the above example, an output of **food** or **The food** alone will have the desired recall of 1. Obviously, this alternative comes with the disadvantage of evaluating outputs with an inaccurate span as correct, e.g., an output of **food was great** will not be evaluated as an error.

4.3 Results

Confidence The per-dataset distribution of the confidence in the annotations is depicted in Figure 3a. For each confidence bin, one of the authors manually annotated a random sample of 30 target candidates for their sentiments, and computed a per-bin annotation error rate (see Table 1). Based on this analysis, the confidence threshold for valid targets was set to 0.7, since under this value the estimated annotation error rate was high. Overall, around 15%-25% of all annotations were considered as low-confidence (light red in Figure 3a).

Bin	$[0.0, 0.7)$	$[0.7, 0.8)$	$[0.8, 0.9)$	$[0.9, 1.0]$
Error	33.3%	10%	3.3%	3.3%

Table 1: The annotation error rate per confidence-bin.

Sentiment labels Observing the distribution of sentiment labels annotated with high-confidence (Figure 3b), hardly any targets were annotated as mixed, and in all datasets (except AMAZON) there were more positive labels than negative ones. As many as 40% of the target candidates may be labeled as not having a sentiment in this phase (grey in Figure 3b), demonstrating the need for the second annotation phase.

Clusters While a cluster may include targets of different sentiments, in practice, cluster members were always annotated with the same sentiment, further supporting the quality of the sentiment annotation. Thus, the sentiment of a cluster is simply the sentiment of its targets.

The distribution of the number of valid targets in each cluster is depicted in Figure 3c. As can be seen, the majority of clusters contain a single target. Out of the 31% of clusters that contain two targets, 70% follow the pattern *"the/this/a/their <T>"* for some term *T*, e.g., *color* and *the color*. The larger clusters of 4 or more targets (2% of all clusters), mostly stem from conjunctions or lists of targets (see examples in Appendix C).

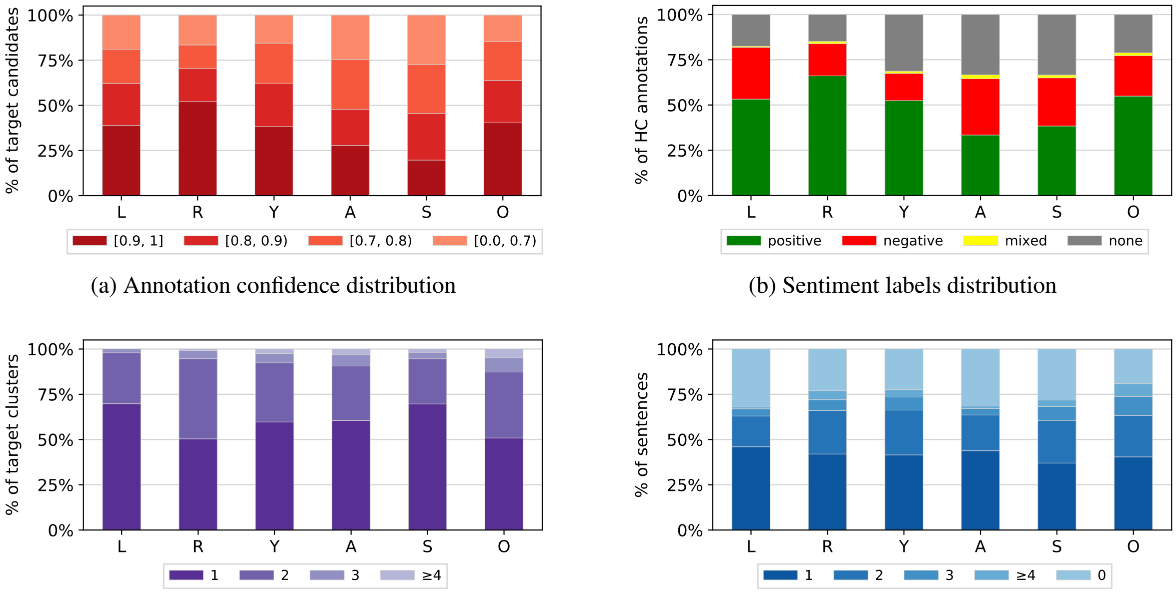
The distribution of the number of clusters identified in each sentence is depicted in Figure 3d. Around 40% of the sentences have one cluster identified within, and as many as 40% have two or more clusters (for OPINOSIS). Between 20% to 35% of the sentences contain no clusters, i.e. no term with a sentiment expressed towards it was detected. Exploring the connection between the number of identified clusters and properties of the annotated sentences (e.g., length) is an interesting direction for future work.

Summary Table 2 summarizes the statistics of the collected data. It also shows the average pairwise inter-annotator agreement, computed with Cohen's Kappa (Cohen, 1960), which was in the range considered as moderate agreement (substantial for SE14-R) by Landis and Koch (1977).

Overall, the YASO dataset contains 2215 sentences and 7415 annotated target candidates. Several annotated sentences are exemplified in Appendix C. To enable further analysis, the dataset includes **all** candidate targets, not just valid ones, each marked with its confidence, sentiment label (including raw annotation counts), and span. YASO is released along with code for performing the post-processing steps described above, and computing the evaluation metrics presented in §6.

5 Analysis

Next, three questions pertaining to the collected data and its annotation scheme are explored.



(a) Annotation confidence distribution

(b) Sentiment labels distribution

(c) Cluster size distribution

(d) Clusters per sentence distribution

Figure 3: Per-dataset statistics showing the distributions of: (a) The confidence in the sentiment annotation of each target candidate; (b) The sentiment labels of targets annotated with high-confidence (HC); (c) The number of valid targets within each cluster; (d) The number of clusters in each annotated sentence. The datasets are marked as: SE14-L (L), SE14-R (R), YELP (Y), AMAZON (A), SST (S) and OPINOSIS (O).

Dataset	#S	#TC	#HC	#VT	#TC	K
SE14-L	100	190	154	127	96	0.54
SE14-R	100	290	242	206	131	0.62
YELP	501	1716	1449	995	655	0.53
AMAZON	502	1540	1161	774	501	0.47
SST	500	1751	1271	846	613	0.41
OPINOSIS	512	1928	1644	1296	763	0.56
Total	2215	7415	5921	4244	2759	-

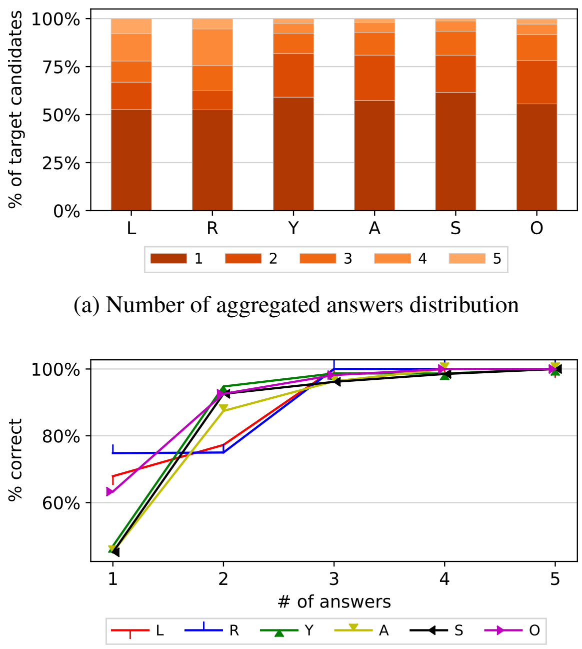
Table 2: Per-dataset annotation statistics: The number of annotated sentences (**#S**) and target candidates annotated within those sentences (**#TC**); The number of targets annotated with high confidence (**#HC**), and as valid targets; (**#VT**); The number of clusters formed from the valid targets (**#TC**); The average pairwise inter-annotator agreement (**K**). See §4.

Is the sentiment annotation phase mandatory? Recall that each sentence in the target candidates annotation phase was shown to 5 annotators who chose candidates and their sentiments. As a result, each candidate has 1 to 5 "first-phase" sentiment answers that can be aggregated by majority vote to a *detection-phase sentiment label*. These can be compared with the sentiment labels from the sentiment annotation phase (which are always based on ≥7 answers).

The distribution of the number of answers arising from the detection-phase labeling is depicted in Figure 4a. In most cases, only one or two answers were available (e.g., in ≥80% of cases for YELP). Figure 4b further details how many of them were correct; for example, those based on one answer for YELP were correct in <50% of cases. In such cases, the sentiment annotation phase is essential for obtaining the correct label. On the other hand, when based on three or more answers, the detection-phase sentiments were correct in ≥96% of cases, for all datasets. Such cases may be exempt from the second sentiment annotation phase, thus reducing costs in future annotation efforts.

What are the differences from SE14? The collected clusters for sentences sampled from SE14 were compared with the SE14 original annotations by pairing each cluster, based solely on its span, with overlapping SE14 annotations (excluding SE14 neutral labels), when available. The sentiments within each pair were compared, and, in most cases, were found to be identical (see Table 3).

Table 3 further shows many clusters are exclusively present in YASO – they do not overlap any SE14 annotation. A manual analysis of such clus-



(a) Number of aggregated answers distribution

(b) Percentage of correct labels

Figure 4: A per-dataset analysis of the detection-phase sentiment labels, showing (a) the distribution of the number of answers that the labels are based on, and (b) how it affects the percentage of correct labels. The datasets are marked as: SE14-L (L), SE14-R (R), YELP (Y), AMAZON (A), SST (S) and OPINOSIS (O).

ters revealed only a few were annotation errors (see Table 4). The others were of one of these categories: (i) **Entities**, such as company/restaurant names; (ii) **Product** terms like *computer* or *restaurant*; (iii) **Other** terms that are not product aspect, such as **decision** in *"I think that was a great **decision** to buy"*; (iv) **Indirect** references, including pronouns, such as **It** in *"**It** was delicious!"*. This difference is expected as such terms are by construction excluded from SE14. In contrast, they are included in YASO since by design it includes all spans people consider as having a sentiment. This makes YASO more complete, while enabling those interested to discard terms as needed for downstream applications. The per-domain frequency of each category, along with additional examples, is given in Table 4.

A similar analysis performed on the 20 targets that were exclusively found in SE14 (i.e., not paired with any of the YASO clusters), showed that 8 cases were SE14 annotation errors, some due to complex expressions with an implicit or unclear sentiment. For example, in *"They're a bit more expensive then typical, but then again, so is their **food**."*, the sentiment of **food** is unclear (and labeled as positive in SE14). From the other 12 cases not paired with any cluster, three were YASO annotation errors (i.e. not found through our annotation scheme), and the rest were annotated but with low-confidence.

Labeled in:	***Both***		***Exclusive***		
Domain	**Ag**	***Dis***	**YASO**	**SE**	**Total**
Laptops	41	5	64	11	121
Restaurants	93	5	38	9	145

Table 3: A comparison of YASO annotations to labels from the SE14 dataset. The sentiment labels of targets labeled in ***both*** datasets may agree (**Ag**) or disagreee (***Dis***). Targets ***exclusively*** present in one of the datasets (**YASO** or **SE**) are further analyzed in §5.

Category	**L**	**R**	**Examples**
Entities	14	6	Apple, iPhone, Culinaria
Product	13	6	laptop, this bar, this place
Other	10	11	process, decision, choice
Indirect	24	11	it, she, this, this one, here
Error	3	4	–

Table 4: A categorization of valid targets in YASO that are not part of SE14, for the laptops (**L**) and restaurants (**R**) domains. The categories are detailed in §5.

What is the recall of the target candidates annotation phase? The last comparison also shows that of the 156 targets[10] annotated in SE14 within the compared sentences, 98% (153) were detected as target candidates, suggesting that our target candidates annotation phase achieved good recall.

6 Benchmark Results

Recall the main purpose of YASO is cross-domain evaluation. The following results were obtained by training on data from SE14 (using its original training sets), and predicting targets over YASO sentences. The results are reported for the full TSA task, and separately for the TE and SC subtasks.

Baselines The following five recently proposed TSA systems were reproduced using their available codebases, and trained on the training set of each of the SE14 domains, yielding ten models overall.
– ***BAT***:[11] (Karimi et al., 2020): A pipelined system

[10] The sum of **Ag**, ***Dis*** and **SE** in Table 3, subtracting the 8 exclusive SE14 annotations manually identified as errors.

[11] `github.com/IMPLabUniPr/BERT-for-ABSA`

with domain-specific language models (Xu et al., 2019) augmented with adversarial data.

– ***LCF***:[12] (Yang et al., 2020): An end-to-end model based on Song et al. (2019), with domain adaptation and a local context focus mechanism.

– ***RACL***:[13] (Chen and Qian, 2020): An end-to-end multi-task learning and relation propagation system. We used the RACL-GloVe variant, based on pre-trained word embeddings.

– ***BERT-E2E***:[14] (Li et al., 2019b): A BERT-based end-to-end sequence labeling system. We used the BERT+Linear architecture, which computes per-token labels using a linear classification layer.

– ***HAST+MCRF***: A pipeline of (i) HAST,[15] a TE system based on capturing aspect detection history and opinion summary (Li et al., 2018); and (ii) MCRF-SA,[16] an SC system utilizing multiple CRF-based structured attention models (Xu et al., 2020a).

Evaluation Metrics As a pre-processing step, any predicted target with a span equal to the span of a target candidate annotated with low-confidence was excluded from the evaluation, since it is unclear what is its true label.

The use of clusters within the evaluation requires an adjustment of the computed recall. Specifically, multiple predicted targets contained within one cluster should be counted once, considering the cluster as one true positive. Explicitly, a predicted target and a cluster are *span-matched*, if the cluster contains a valid target with a span equal to the span of the prediction (an *exact* span match). Similarly, they are *fully-matched* if they are span-matched and their sentiments are the same. Predictions that were not span-matched to any cluster were considered as errors for the TE task (since their span was not annotated as a valid target), and those that were not fully-matched to any cluster were considered as errors for the full task. Using span-matches, precision for the TE task is the percentage of span-matched predictions, and recall is the percentage of span-matched clusters. These metrics are similarly defined for the full task using full-matches.

For SC, evaluation was restricted to predictions that were span-matched to a cluster. For a sentiment label l, precision is the percentage of fully-matched predictions with sentiment l (out of all span-matched predictions with that sentiment); recall is the percentage of fully-matched clusters with sentiment l (out of all span-matched clusters with that sentiment). Macro-F_1 (mF_1) is the average F_1 over the positive and negative sentiment labels (mixed was ignored since it was scarcely in the data, following Chen and Qian (2020)).

Our data release is accompanied by code for computing all the described evaluation metrics.

Results Table 5 presents the results of our evaluation. BAT trained on the restaurants data was the best-performing system for TE and the full TSA tasks, on three of the four datasets (YELP, SST and OPINOSIS). For SC, BERT-E2E was the best model on three datasets. Generally, results for SC were relatively high, while TE results by some models may be very low, typically stemming from low recall. The precision and recall results for each task are further detailed in Appendix D.

Appendix D also details additional results when relaxing the TE evaluation criterion from exact span-matches to overlapping span-matches – where a predicted target and a cluster are span-matched if their spans overlap. While with this relaxed evaluation the TE performance was higher (as expected), the absolute numbers suggest a significant percentage of errors were not simply targets predicted with a misaligned span.

TSA task performance was lowest for SST, perhaps due to its domain of movie reviews, which is furthest of all datasets from the product reviews training data. Interestingly, it was also the dataset with the lowest level of agreement among humans (see Figure 3a).

The choice of the training domain is an important factor for most algorithms. This is notable, for example, in the TE performance obtained for YELP: the gap between training on data from the laptops domain or the restaurants domain is ≥ 20 (in favor of the latter) for all algorithms (except LCF). A likely cause is that the YASO data sampled from YELP has a fair percentage of reviews on food related establishments. Future work may further use YASO to explore the impact of the similarity between the training and test domains, as well as develop new methods that are robust to the choice of the training domain.

[12] github.com/yangheng95/LCF-ATEPC

[13] github.com/NLPWM-WHU/RACL

[14] github.com/lixin4ever/BERT-E2E-ABSA

[15] github.com/lixin4ever/HAST

[16] github.com/xuuuluuu/Aspect-Sentiment-Classification

		Yelp			Amazon			SST			Opinosis		
System	**Train**	**TE**	**SC**	**TSA**	**TE**	**SC**	**TSA**	**TE**	**SC**	**TSA**	**TE**	**SC**	**TSA**
BAT	**Lap.**	27.8	88.0	24.8	34.5	96.3	33.1	8.8	**100.0**	8.8	57.2	92.2	53.6
	Res.	**58.0**	91.6	**54.4**	29.3	89.4	25.6	**34.9**	90.6	**31.9**	**59.1**	91.8	**55.3**
BERT-E2E	**Lap.**	28.2	91.3	26.5	35.5	97.8	**34.5**	12.2	97.4	12.0	56.1	94.0	53.4
	Res.	52.7	**93.4**	49.9	28.6	**98.0**	27.7	9.9	92.5	9.0	50.4	**94.3**	48.0
HAST+MCRF	**Lap.**	16.7	68.3	11.9	21.4	82.0	17.5	2.8	64.9	1.9	34.8	82.2	29.6
	Res.	40.7	88.4	36.5	9.4	95.6	9.0	3.1	67.0	2.2	31.6	87.7	28.0
LCF	**Lap.**	41.0	72.6	33.3	**37.9**	85.0	31.9	17.0	80.4	13.7	54.7	91.1	50.6
	Res.	48.8	84.8	43.7	36.1	87.1	31.0	16.5	75.7	12.8	55.7	86.5	49.4
RACL	**Lap.**	23.0	88.2	20.8	29.0	89.6	25.9	13.2	78.1	10.2	43.2	83.1	37.8
	Res.	44.5	87.9	39.9	22.5	88.9	19.7	7.9	86.3	7.0	43.8	85.0	38.4
Average	**Lap.**	27.3	81.7	23.5	31.7	90.1	28.6	10.8	84.2	9.3	49.2	88.5	45.0
	Res.	48.9	89.2	44.9	25.2	91.8	22.6	14.5	82.4	12.6	48.1	89.1	43.8

Table 5: Benchmark results on YASO with five SOTA systems, trained on data from one SE14 domain (laptops – **Lap.** or restaurants – **Res.**). The reported metric is F_1 for target extraction (**TE**) and the entire task (**TSA**), and macro-F_1 for sentiment classification (**SC**).

7 Conclusion

We collected a new open-domain user reviews TSA evaluation dataset named YASO. Unlike existing review datasets, YASO is not limited to any particular reviews domain, thus providing a broader perspective for cross-domain TSA evaluation. Benchmark results established in such a setup with contemporary TSA systems show there is ample headroom for improvement on YASO.

YASO was annotated using a new scheme for creating TSA labeled data, that can be also applied to non-review texts. The reliability of the annotations obtained by this scheme has been verified through a manual analysis of a sample and a comparison to existing labeled data.

One limitation of our scheme is that aspect categories with a sentiment implied from the reviews were excluded, since their annotation requires prespecifying the domain along with its associated categories. While this may limit research for some applications, the dataset is useful in many real-world use cases. For example, given a brand name, one may query a user reviews corpus for sentences containing it, and analyze the sentiment towards that brand in each sentence along with the sentiment expressed to other terms in these sentences.

Future work may improve upon the presented results by training on multiple domains or datasets, adapting pre-trained models to the target domains in an unsupervised manner (e.g., Rietzler et al. (2020)), exploring various data augmentation techniques, or utilizing multi-task or weak-supervision algorithms. Another interesting direction for further research is annotating opinion terms within the YASO sentences, facilitating their co-extraction with corresponding targets (Wang et al., 2016, 2017b), or as triplets of target term, sentiment, and opinion term (Peng et al., 2020; Xu et al., 2020b).

All benchmark data collected in this work are available online.[17] We hope that these data will facilitate further advancements in the field of targeted sentiment analysis.

Acknowledgments

We wish to thank the anonymous reviewers for their insightful comments, suggestions, and feedback.

References

Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. Aspect based sentiment analysis in Hindi: Resource creation and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2703–2709, Portorož, Slovenia. European Language Resources Association (ELRA).

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of the Eleventh International*

[17] github.com/IBM/yaso-tsa

Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

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. Association for Computational Linguistics.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland. Association for Computational Linguistics.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 340–348. Association for Computational Linguistics.

Chenggong Gong, Jianfei Yu, and Rui Xia. 2020. Unified feature and instance based domain adaptation for aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7035–7045, Online. Association for Computational Linguistics.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2021. Towards target-dependent sentiment classification in news articles. In *Diversity, Divergence, Dialogue*, pages 156–166, Cham. Springer International Publishing.

Ruidan He, Wee Sun Lee, H. Ng, and Daniel Dahlmeier. 2018. Effective attention modeling for aspect-level sentiment classification. In *COLING*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515, Florence, Italy. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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

Binxuan Huang, Yanglan Ou, and Kathleen M. Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In *Social, Cultural, and Behavioral Modeling - 11th International Conference, SBP-BRiMS 2018, Washington, DC, USA, July 10-13, 2018, Proceedings*, volume 10899 of *Lecture Notes in Computer Science*, pages 197–206. Springer.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6280–6285, Hong Kong, China. Association for Computational Linguistics.

Akbar Karimi, Leonardo Rossi, Andrea Prati, and Katharina Full. 2020. Adversarial training for aspect-based sentiment analysis with bert. *arXiv preprint arXiv:2001.11316*.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Roman Klinger and Philipp Cimiano. 2014. The USAGE review corpus for fine grained multi lingual opinion analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2211–2218, Reykjavik, Iceland. European Language Resources Association (ELRA).

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500.

Hao Li and Wei Lu. 2019. Learning explicit and implicit structures for targeted sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. 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, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4194–4200. International Joint Conferences on Artificial Intelligence Organization.

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, Hong Kong, China. Association for Computational Linguistics.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Yukun Ma, Haiyun Peng, and E. Cambria. 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In *AAAI*.

Yosi Mass, Lili Kotlerman, Shachar Mirkin, Elad Venezian, Gera Witzling, and Noam Slonim. 2018. What did you mention? a large scale mention detection benchmark for spoken and written text.

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, Seattle, Washington, USA. Association for Computational Linguistics.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5025–5033, Marseille, France. European Language Resources Association.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8600–8607.

Minh Hieu Phan and Philip O. Ogunbona. 2020. Modelling context and syntactical features for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3211–3220, Online. Association for Computational Linguistics.

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.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France. European Language Resources Association.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. A hierarchical model of reviews for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 999–1005, Austin, Texas. Association for Computational Linguistics.

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

Kim Schouten and Flavius Frasincar. 2015. Survey on aspect-level sentiment analysis. *IEEE Transactions* *on Knowledge and Data Engineering*, 28(3):813–830.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *CoRR*, abs/1902.09314.

Josef Steinberger, Tomáš Brychcín, and Michal Konkol. 2014. Aspect-level sentiment analysis in Czech. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–30, Baltimore, Maryland. Association for Computational Linguistics.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota. Association for Computational Linguistics.

Martina Katalin Szabó, Veronika Vincze, Katalin Ilona Simkó, Viktor Varga, and Viktor Hangya. 2016. A Hungarian sentiment corpus manually annotated at aspect level. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2873–2878, Portorož, Slovenia. European Language Resources Association (ELRA).

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan. The COLING 2016 Organizing Committee.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224, Austin, Texas. Association for Computational Linguistics.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4067–4076, Online. Association for Computational Linguistics.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584, Uppsala, Sweden. Association for Computational Linguistics.

Bo Wang, Maria Liakata, Arkaitz Zubiaga, and Rob Procter. 2017a. TDParse: Multi-target-specific sentiment recognition on Twitter. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 483–493, Valencia, Spain. Association for Computational Linguistics.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. *arXiv preprint arXiv:1603.06679*.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017b. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2):0.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Lu Xu, Lidong Bing, Wei Lu, and Fei Huang. 2020a. Aspect sentiment classification with aspect-specific opinion spans. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3561–3567. Association for Computational Linguistics.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020b. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online. Association for Computational Linguistics.

Heng Yang, Biqing Zeng, JianHao Yang, Youwei Song, and Ruyang Xu. 2020. A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction. *arXiv preprint arXiv:1912.07976*.

Jun Yang, Runqi Yang, Chong-Jun Wang, and Junyuan Xie. 2018. Multi-entity aspect-based sentiment analysis with context, entity and aspect memory. In *AAAI*.

Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9(16).

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

A Annotated Domains

YASO includes, among others, review texts from the following product and business domains: apparel, automotive, baby products, beauty, books, cameras, cars, car washes, cinemas, digital e-books, drugstores, electronics, furniture, food, grocery, home improvement, hotels, industrial supplies, jewelry, kitchen, lawn and garden, luggage, movies, musical instruments, office products, personal computers, pet products, restaurants, shoes, sports, toys, video games, watches, and wireless.

B Annotation Guidelines

B.1 Target Candidates Annotation

Below are the guidelines for the labeling task of detecting potential targets and their sentiment.

General instructions

In this task you will review a set of sentences. Your goal is to identify items in the sentences that have a sentiment expressed towards them.

Steps

1. Read the sentence carefully.
2. Identify items that have a sentiment expressed towards them.
3. Mark each item, and for each selection 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 are no items with a sentiment expressed towards them, proceed to the next sentence.

Rules & Tips

- Select all items in the sentence that have a sentiment expressed towards them.
- It could be that there are several correct overlapping selections. In such cases, it is OK to choose only one of these overlapping selections.
- The sentiment towards a selected item(s) should be expressed from other parts of the sentence, it cannot come from within the selected item (see Example #2 below).
- Under each question is a comments box. Optionally, you can provide question-specific feedback in this box. This may include a rationalization of your choice, a description of an error within the question or the justification of another answer which was also plausible. In general, any relevant feedback would be useful, and will help in improving this task.

Examples

Here are a few example sentences, categorized into several example types. For each sentence, the examples show item(s) which should be selected, and the sentiment expressed towards each such item. Further explanations are provided within the examples, when needed. Please review the examples carefully before starting the task.

1. **Basics**

   Example #1.1: *The food was good.*
   Correct answer: The food was good.
   Explanation: The word *good* expresses a positive sentiment towards *food*.

   Example #1.2: *The food was bad.*
   Correct answer: The food was bad.
   Explanation: The word *bad* expresses a negative sentiment towards *food*.

   Example #1.3: *The food was tasty but expensive.*
   Correct answer: The food was tasty but expensive.
   Explanation: *tasty* expresses a positive sentiment, while *expensive* expresses a negative sentiment, so the correct answer is Mixed.

Input Dataset	Sentence
SE14-R	Although I moved uptown I try to stop in as often as possible for the GREAT [cheap [food]$_{\text{P}}$]$_{\text{P}}$ and to pay the friendly [staff]$_{\text{P}}$ a visit.
SE14-L	A great [college [tool]$_{\text{P}}$]$_{\text{P}}$!
OPINOSIS	The [Waitrose supermarket]$_{\text{P}}$ has many take out food options .
AMAZON	[The protective [seal]$_{\text{N}}$]$_{\text{N}}$ was broken when I received this [item]$_{\text{N}}$ and a large amount of the contents had spilled out of the container into the plastic bag that the item was in.
YELP	[The [wait]$_{\text{N}}$]$_{\text{N}}$ was a little longer than what I prefer, but [the [service]$_{\text{P}}$]$_{\text{P}}$ was kind, [the [food]$_{\text{P}}$]$_{\text{P}}$ was incredible, and [the [Phuket Bucket]$_{\text{P}}$]$_{\text{P}}$ was refreshing on a warm evening.
SST	[The Irwins]$_{\text{P}}$ emerge unscathed , but [the [fictional [footage]$_{\text{N}}$]$_{\text{N}}$]$_{\text{N}}$ is unconvincing and criminally badly [acted]$_{\text{N}}$.

Table 6: Annotation examples from the various input datasets. A target t that has a positive/negative sentiment expressed towards it is marked as $[t]_{\text{P}}$ / $[t]_{\text{N}}$.

Example #1.4: *The food was served.*
Correct answer: Nothing should be selected, since there is no sentiment expressed in the sentence.

2. **Sentiment location**

Example #2.1: *I love this great car.*
Correct answer #1: I love this great car.
Correct answer #2: I love this great car.
Explanation: The word *love* expresses a positive sentiment towards *great car* or *car*.
Note: It is OK to select only one of the above options, since they overlap.

Example #2.2: *I have a great car.*
Correct answer: I have a great car.
Explanation: The word *great* expresses a positive sentiment towards *car*.
Note: Do NOT select the item *great car*, because there is NO sentiment expressed towards *great car* outside of the phrase *great car* itself. The only other information is that *i have a* item, which does not convey a sentiment towards it.

3. **Multiple selections in one sentence**

Example #3.1: *The food was good, but the atmosphere was awful.*
Correct answer: The food was good, but the atmosphere was awful.
Explanation: the word *good* expresses a positive sentiment towards *food*, while the word *awful* expresses a negative sentiment towards *atmosphere*.
Note: Both items should be selected!

Example #3.2: *The camera has excellent lens.*
Correct answer: The camera has excellent lens.
Explanation: The word *excellent* expresses a positive sentiment towards *lens*. • An *excellent lens* is a positive thing for a camera to have, thus expressing a positive sentiment towards *camera*.
Note: Both items should be selected!

Example #3.3: *My new camera has excellent lens, but its price is too high.*
Correct answer: My new camera has excellent lens, but its price is too high.
Explanation: The word *excellent* expresses a positive sentiment towards *lens*, while the words *too high* expresses a negative sentiment towards *price*. There is a positive sentiment towards the camera, due to its *excellent lens*, and also a negative sentiment, because *its*

price is too high, so the sentiment towards *camera* is Mixed.
Note: All three items should be selected. Other acceptable selections with a Mixed sentiment are *new camera* or *My new camera*. Since they overlap, it is OK to select just one of them.

4. **Sentences without any expressed sentiments**
   Below are some examples of sentences without any expressed sentiment in them. For such sentences, nothing should be selected.

   Example #4.1: *Microwave, refrigerator, coffee maker in room.*
   Example #4.2: *I took my Mac to work yesterday.*

5. **Long selected items**
   There is no restriction on the length of a select item, so long as there is an expressed sentiment towards it in the sentence (which does not come from within the marked item).

   Example #5.1: *The food from the Italian restaurant near my office was very good.*
   Correct answer #1: The food from the Italian restaurant near my office was very good.
   Correct answer #2: The food from the Italian restaurant near my office was very good.
   Correct answer #3: The food from the Italian restaurant near my office was very good.
   Correct answer #4: The food from the Italian restaurant near my office was very good.
   Explanation: the words *very good* express a positive sentiment towards
   emphfood.

   Note: It is also a valid choice to select *food* along with its details description: *food from the Italian restaurant near my office*, or add the prefix *The* to the selection (or both). The selection must be a coherent phrase. *food from the* is not a valid selection. Since these selections all overlap, it is OK to select one of them.

B.2 Sentiment Annotation

Below are the guidelines for labeling the sentiment of identified target candidates.

General instructions

In this task you will review a set of sentences, each containing one marked item. Your goal is to determine the sentiment expressed in the sentence towards the marked item.

Steps

1. Read the sentence carefully.

2. Identify the sentiment expressed in the sentence towards the marked item, by selecting one of these four options:

   (a) Positive: the expressed sentiment is **positive**.
   (b) Negative: the expressed sentiment is **negative**.
   (c) Mixed: the expressed sentiment is **both** positive and negative.
   (d) **None**: there is **no sentiment** expressed towards the item.

3. If there are no items with a sentiment expressed towards them, proceed to the next sentence.

Rules & Tips

- The sentiment should be expressed towards the marked item, it cannot come from within the marked item (see Example #2 below).

- A sentence may appear multiple times, each time with one marked item. Different marked items may have different sentiments expressed towards each of them in one sentence (see Example #3 below)

- Under each question is a **comments box**. Optionally, you can provide question-specific feedback in this box. This may include a rationalization of your choice, a description of an error within the question or the justification of another answer which was also plausible. In general, any relevant feedback would be useful, and will help in improving this task.

Input Dataset	Sentence
Yelp	Great [[office staff]$_{\text{P}}$, [[nurse]$_{\text{P}}$ practitioner]$_{\text{P}}$ and [pediatric doctor]$_{\text{P}}$]$_{\text{P}}$.
Amazon	[Her [[office [routine]$_{\text{P}}$]$_{\text{P}}$ and [morning routine]$_{\text{P}}$]$_{\text{P}}$]$_{\text{P}}$ are wonderful.
Opinosis	As of today, I am a bit disappointed in [the [[build]$_{\text{N}}$ [quality]$_{\text{N}}$]$_{\text{N}}$ of [the [car]$_{\text{N}}$]$_{\text{N}}$]$_{\text{N}}$.
Opinosis	[This car]$_{\text{P}}$ is nearly perfect when compared to other cars in this class regarding [[interior dimensions]$_{\text{P}}$, [visibility]$_{\text{P}}$, [exterior styling]$_{\text{P}}$]$_{\text{P}}$, etc .

Table 7: Examples of sentences in which large target clusters were annotated.

Examples

Here are a few examples, each containing a sentence and a marked item, along with the correct answer and further explanations (when needed). Please review the examples carefully before starting the task.

1. **Basics**

   Example #1.1: *The food was good.*
   Answer: Positive

   Example #1.2: *The food was bad.*
   Answer: Negative

   Example #1.3: *The food was tasty but expensive.*
   Answer: Mixed
   Explanation: *tasty* expresses a positive sentiment, while *expensive* expresses a negative sentiment, so the correct answer is **Mixed**.

   Example #1.4: *The food was served.*
   Answer: None

2. **Sentiment location**

   Example #2.1: *I love this great car.*
   Answer: Positive
   Explanation: There is a positive sentiment expressed towards *great car* outside of the marked item *car* – in the statement that *I love* the car.

   Example #2.2: *I love this great car.*
   Answer: Positive
   Explanation: There is a positive sentiment expressed towards *car* outside of the marked item *car* – in the word *great* and the statement that *I love* the car.

   Example #2.3: *I have a great car.*
   Answer: Positive
   Explanation: There is a positive sentiment (*great*) expressed towards *car* outside of the marked item *car*.

3. **Different marked items in one sentence**

   Example #3.1: *The food was good, but the atmosphere was awful.*
   Answer: Positive
   The food was good, but the atmosphere was awful.
   Answer: Negative

   Example #3.2: *The camera has excellent lens.*
   Answer: Positive
   The camera has excellent lens.
   Answer: Positive

   Example #3.3: *My new camera has excellent lens, but its price is too high.*
   Answer: Mixed
   Explanation: There is a positive sentiment towards the camera, due to its *excellent lens*, and also a negative sentiment, because *its price is too high*, so the correct answer is **Mixed**.
   My new camera has excellent lens, but its price is too high.

Answer: Positive
My new camera has excellent lens, but its price is too high.
Answer: Negative

4. **Marked items without a sentiment**

   Below are some examples of marked items without an expressed sentiment in the sentence. In cases where there is a expressed sentiment towards other words in the same sentence, it is exemplified as well.

   Example #4.1: *Microwave, refrigerator, coffee maker in room.*
   Answer: None

   Example #4.2: *Note that they do not serve beer, you must bring your own.*
   Answer: None

   Example #4.3: *The cons are more annoyances that can be lived with.*
   Answer: None
   Explanation: While the marked item contains a negative sentiment, there is no sentiment towards the marked item.

   Example #4.4: *working with Mac is so much easier, so many cool features.*
   Answer: None
   working with Mac is so much easier, so many cool features.
   Answer: Positive
   working with Mac is so much easier, so many cool features.
   Answer: Positive

   Example #4.5: *The battery life is excellent- 6-7 hours without charging.*
   Answer: None
   The battery life is excellent- 6-7 hours without charging.
   Answer: Positive

   Example #4.6: *I wanted a computer that was quiet, fast, and that had overall great performance.*
   Answer: None

5. **"the" can be a part of a marked item**

   I feel a little bit uncomfortable in using the Mac system.
   Answer: Negative
   I feel a little bit uncomfortable in using the Mac system.
   Answer: Negative
   I feel a little bit uncomfortable in using the Mac system.
   Answer: None

6. **Long marked items**

   There is no restriction on the length of a marked item, so long as there is an expressed sentiment towards it in the sentence (which does not come from within the marked item).

   The food from the Italian restaurant near my office was very good.
   Answer: Positive
   The food from the Italian restaurant near my office was very good.
   Answer: Positive
   The food from the Italian restaurant near my office was very good.
   Answer: None

7. **Idioms**

   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.

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

C Annotation Examples

Table 6 presents sentences included in YASO, along with the annotated targets and their corresponding sentiments found within each sentence.

A target t that has a positive sentiment expressed towards it is marked as $[t]_{\mathrm{P}}$. Similarly $[t]_{\mathrm{N}}$ is used for a negative sentiment. For brevity, the examples only show the valid targets annotated within the sentences, hiding any low-confidence annotations or target candidates that were annotated as not having a sentiment in the second annotation phase. As can be seen in the examples, annotated valid targets may overlap, demonstrating the need for the definition of the target clusters.

Table 7 further exemplifies sentences in which a cluster containing more than 4 valid targets were detected.

D Detailed Benchmark Results

In addition to the main benchmark results presented in the paper, Table 8 shows the precision, recall and F_1 for target extraction and the entire task. For sentiment classification, the same metrics are separately reported for the positive and negative sentiment labels, as well as macro-F_1 over these two classes.

Table 9 presents results similar to Table 5 with another TE evaluation criteria, where a predicted target and a cluster are span-matched if their spans overlap. This is a more relaxed evaluation criteria than the one used in the main results (which consider a predicted target and a cluster as span-matched if the cluster contains a target with a span equal to the span of the prediction).

		Task:	***TE***			***SC***							***TSA***		
						Positive			**Negative**						
Dat.	**System**	**Train**	**P**	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$	$\mathbf{mF_1}$	**P**	**R**	$\mathbf{F_1}$
Y	**BAT**	**Lap.**	73.7	17.1	27.8	94.3	94.3	94.3	72.0	94.7	81.8	88.0	65.8	15.3	24.8
		Res.	**76.3**	**46.7**	**58.0**	96.1	96.9	96.5	81.2	92.9	86.7	91.6	**71.6**	**43.8**	**54.4**
	BERT-E2E	**Lap.**	68.6	17.7	28.2	94.9	**98.9**	**96.9**	**88.2**	83.3	85.7	91.3	64.5	16.6	26.5
		Res.	63.9	44.9	52.7	**97.4**	96.1	96.8	84.4	**96.4**	**90.0**	**93.4**	60.4	42.4	49.9
	HAST+ MCRF	**Lap.**	63.0	9.6	16.7	87.5	72.9	79.5	43.5	83.3	57.1	68.3	45.0	6.9	11.9
		Res.	64.9	29.6	40.7	95.8	92.5	94.1	73.1	95.0	82.6	88.4	58.2	26.6	36.5
	LCF	**Lap.**	60.6	31.0	41.0	85.7	91.1	88.3	60.0	53.8	56.8	72.6	49.3	25.2	33.3
		Res.	56.8	42.7	48.8	91.8	95.5	93.6	79.2	73.1	76.0	84.8	50.9	38.3	43.7
	RACL	**Lap.**	58.4	14.4	23.0	92.2	95.9	94.0	82.4	82.4	82.4	88.2	52.8	13.0	20.8
		Res.	59.4	35.6	44.5	94.2	92.6	93.4	77.0	88.7	82.5	87.9	53.3	31.9	39.9
A	**BAT**	**Lap.**	57.1	24.8	34.5	96.4	94.6	95.5	95.7	98.5	97.1	96.3	**54.8**	23.8	33.1
		Res.	**61.9**	19.2	29.3	84.5	**100**	91.6	96.0	80.0	87.3	89.4	54.2	16.8	25.6
	BERT-E2E	**Lap.**	50.6	27.3	35.5	95.6	98.5	97.0	98.6	98.6	98.6	97.8	49.1	**26.5**	**34.5**
		Res.	53.0	19.6	28.6	94.1	100	97.0	**100**	97.9	**98.9**	**98.0**	51.4	19.0	27.7
	HAST+ MCRF	**Lap.**	46.1	14.0	21.4	85.7	81.1	83.3	77.1	84.4	80.6	82.0	37.5	11.4	17.5
		Res.	48.1	5.2	9.4	**100**	94.4	**97.1**	88.9	**100**	94.1	95.6	46.3	5.0	9.0
	LCF	**Lap.**	47.4	**31.5**	**37.9**	80.9	90.0	85.2	88.4	81.3	84.7	85.0	39.9	26.5	31.9
		Res.	46.0	29.7	36.1	81.0	93.2	86.6	92.3	83.3	87.6	87.1	39.5	25.5	31.0
	RACL	**Lap.**	51.8	20.2	29.0	88.7	94.8	91.7	89.7	85.4	87.5	89.6	46.2	18.0	25.9
		Res.	49.0	14.6	22.5	89.5	89.5	89.5	85.7	90.9	88.2	88.9	43.0	12.8	19.7
S	**BAT**	**Lap.**	**64.4**	4.7	8.8	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**64.4**	4.7	8.8
		Res.	61.0	**24.5**	**34.9**	90.3	100	94.9	96.2	78.1	86.2	90.6	55.7	**22.3**	**31.9**
	BERT-E2E	**Lap.**	57.5	6.9	12.2	96.3	100	98.1	100	93.8	96.8	97.4	56.2	6.7	12.0
		Res.	46.6	5.5	9.9	90.5	95.0	92.7	92.3	92.3	92.3	92.5	42.5	5.1	9.0
	HAST+ MCRF	**Lap.**	33.3	1.5	2.8	100	57.1	72.7	40.0	100	57.1	64.9	22.2	1.0	1.9
		Res.	41.7	1.6	3.1	100	62.5	76.9	40.0	100	57.1	67.0	29.2	1.1	2.2
	LCF	**Lap.**	32.3	11.6	17.0	89.5	77.3	82.9	69.7	88.5	78.0	80.4	25.9	9.3	13.7
		Res.	35.7	10.8	16.5	79.1	85.0	81.9	73.9	65.4	69.4	75.7	27.6	8.3	12.8
	RACL	**Lap.**	37.7	8.0	13.2	81.5	78.6	80.0	72.7	80.0	76.2	78.1	29.2	6.2	10.2
		Res.	28.0	4.6	7.9	90.5	95.0	92.7	85.7	75.0	80.0	86.3	25.0	4.1	7.0
O	**BAT**	**Lap.**	64.0	51.8	57.2	95.7	97.3	96.5	87.5	88.4	88.0	92.2	60.0	48.5	53.6
		Res.	**72.3**	49.9	**59.1**	95.2	**98.7**	96.9	87.3	86.1	86.7	91.8	**67.7**	46.8	**55.3**
	BERT-E2E	**Lap.**	60.9	52.0	56.1	96.1	98.3	97.2	92.3	89.4	90.8	94.0	58.0	**49.5**	53.4
		Res.	62.8	42.1	50.4	**97.1**	97.9	**97.5**	89.9	**92.2**	**91.0**	**94.3**	59.9	40.1	48.0
	HAST+ MCRF	**Lap.**	48.4	27.1	34.8	93.6	85.6	89.4	67.2	84.9	75.0	82.2	41.1	23.1	29.6
		Res.	58.0	21.8	31.6	93.8	91.4	92.6	77.4	89.1	82.8	87.7	51.4	19.3	28.0
	LCF	**Lap.**	56.1	53.3	54.7	92.4	97.7	94.9	**93.5**	81.9	87.3	91.1	51.9	49.4	50.6
		Res.	57.0	**54.4**	55.7	93.4	92.4	92.9	76.3	84.5	80.2	86.5	50.5	48.2	49.4
	RACL	**Lap.**	50.7	37.6	43.2	91.0	94.0	92.4	75.4	72.1	73.7	83.1	44.3	32.9	37.8
		Res.	56.1	35.9	43.8	93.9	91.5	92.7	71.8	83.6	77.2	85.0	49.2	31.5	38.4

Table 8: Detailed benchmark results on YASO with five SOTA systems, trained on data from one SE14 domain (laptops – **Lap.** or restaurants – **Res.**). The reported metrics are precision (**P**), recall (**R**) and $\mathbf{F_1}$ for target extraction (***TE***) and the entire task (***TSA***). For sentiment classification (***SC***), the same metrics are separately reported for the positive and negative sentiment labels, as well as macro-F_1 ($\mathbf{mF_1}$) over these two classes. The datasets (**Dat.**) are marked as: Yelp (Y), Amazon (A), SST (S) and Opinosis (O).

		YELP			AMAZON			SST			OPINOSIS		
System	**Train**	**TE**	**SC**	**TSA**	**TE**	**SC**	**TSA**	**TE**	**SC**	**TSA**	**TE**	**SC**	**TSA**
BAT	**Lap.**	32.3	88.7	29.1	45.1	94.0	42.3	11.5	97.2	11.2	68.9	92.2	64.0
	Res.	**70.6**	91.5	**66.3**	41.4	90.8	36.6	**48.0**	88.7	**42.8**	**73.5**	92.0	**68.5**
BERT-E2E	**Lap.**	32.7	90.6	30.8	47.2	95.6	**44.9**	15.7	**98.1**	15.4	68.8	92.4	64.2
	Res.	65.0	**91.6**	60.7	39.0	**97.7**	37.8	13.7	93.6	12.5	63.0	**94.3**	59.6
HAST+MCRF	**Lap.**	20.1	71.3	15.1	25.9	79.8	20.4	4.4	62.6	2.8	47.0	80.7	38.7
	Res.	53.2	85.5	46.7	13.3	93.8	12.6	4.1	67.5	2.8	42.8	84.5	36.5
LCF	**Lap.**	51.7	73.4	42.2	**52.2**	82.6	42.8	28.9	77.7	22.2	70.5	88.5	63.3
	Res.	65.7	86.4	59.4	49.7	84.6	41.5	25.2	69.7	17.7	69.2	85.7	60.5
RACL	**Lap.**	29.8	89.0	27.4	35.7	87.8	31.1	19.8	73.1	14.2	59.1	81.9	50.1
	Res.	59.6	86.0	52.6	32.2	88.5	28.2	15.1	68.8	10.6	62.1	85.4	53.9
Average	**Lap.**	33.3	82.6	28.9	41.2	88.0	36.3	16.1	81.7	13.2	62.9	87.2	56.1
	Res.	62.8	88.2	57.2	35.1	91.1	31.3	21.2	77.7	17.3	62.1	88.4	55.8

Table 9: Benchmark results on YASO using overlapping span-matches instead of exact span-matches. This table is similar to Table 5: it presents results from five SOTA systems, trained on data from one SE14 domain (laptops – **Lap.** or restaurants – **Res.**). The reported metric is F_1 for target extraction (**TE**) and the entire task (**TSA**), and macro-F_1 for sentiment classification (**SC**).