SemEval-2015 Task 12: Aspect Based Sentiment Analysis

Maria Pontiki*, Dimitrios Galanis*, Haris Papageorgiou*,
Suresh Manandhar[±], Ion Androutsopoulos[◊]*

*Institute for Language and Speech Processing, Athena R.C., Athens, Greece
[±] Dept. of Computer Science, University of York, UK
[◊] Dept. of Informatics, Athens University of Economics and Business, Greece
{mpontiki, galanisd, xaris} @ilsp.gr
suresh@cs.york.ac.uk
ion@aueb.gr

Abstract

SemEval-2015 Task 12, a continuation of SemEval-2014 Task 4, aimed to foster research beyond sentence- or text-level sentiment classification towards Aspect Based Sentiment Analysis. The goal is to identify opinions expressed about specific entities (e.g., laptops) and their aspects (e.g., price). The task provided manually annotated reviews in three domains (restaurants, laptops and hotels), and a common evaluation procedure. It attracted 93 submissions from 16 teams.

1 Introduction and Related Work

The rise of e-commerce, as a new shopping and marketing channel, has led to an upsurge of review sites for a variety of services and products. In this context, Aspect Based Sentiment Analysis (ABSA) -i.e., mining opinions from text about specific entities and their aspects- can help consumers decide what to purchase and businesses to better monitor their reputation and understand the needs of the market (Pavlopoulos 2014). Given a target of interest (e.g., Apple Mac mini), an ABSA method can summarize the content of the respective reviews in an aspect-sentiment table like the one in Fig 1. Some review sites also generate such tables based on customer ratings, but usually only for a limited set of predefined aspects and not from free-text reviews.

Several ABSA methods have been proposed for various domains, like consumer electronics (Hu and Liu {2004a, 2004b}), restaurants (Ganu et al., 2009) and movies (Thet et al., 2010). The available methods can be divided into those that adopt domain-independent solutions (Lin and He, 2009), and those that use domain-specific knowledge to improve their results (Thet et al., 2010). Typically, most methods treat aspect extraction and sentiment classification separately (Brody and Elhadad, 2010), but there are also approaches that model the two problems jointly (Jo and Oh, 2011).

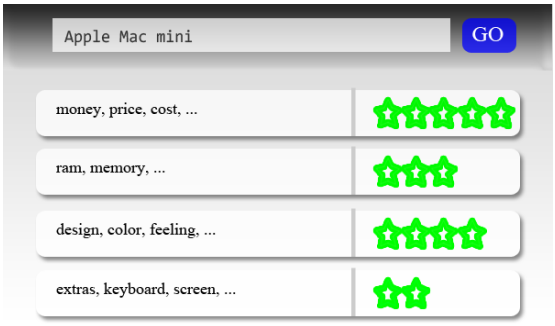

Figure 1. Table summarizing the average sentiment for each aspect of an entity.

Publicly available ABSA datasets adopt different annotation schemes for different subtasks and languages (Pavlopoulos 2014). For example, the datasets of McAuley et al. (2012) provide aspects and respective ratings at the review level (i.e., aspects and ratings associated with entire reviews, not particular sentences)[1] about Beers, Pubs, Toys and Games, and Audiobooks. The reviews are obtained from sites that allow users to evaluate a product not only in terms of its overall quality, but also focusing on specific predefined aspects (e.g. "smell" and "taste" for Beers, "fun" and "educational value" for Toys and Games). The IGGSA Shared Tasks on German Sentiment Analysis (Ruppenhofer et al., 2014) provided human annotated datasets of political speeches (STEPS task)

[1] A subset of the datasets has been annotated with aspects at the sentence level.

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 486–495,
Denver, Colorado, June 4-5, 2015.

and reviews about products (StAR task) like coffee machines and washers. The StAR task focused on the extraction of evaluative phrases (e.g., "*bad*") and aspect expressions (e.g., "*washer*"). The STEPS dataset includes annotations for evaluative phrases, opinion targets, and the corresponding sources (opinion holders). The extraction of opinion targets and holders has also been addressed in the context of the Multilingual Opinion Analysis Task (Seki et al., 2007; Seki et al., 2008; Seki et al., 2010) and the Sentiment Slot Filling[2] Task of the Knowledge Base Population Track (Mitchell, 2013). However, these tasks deal with the identification of opinion targets in general, not in the context of ABSA.

SemEval-2014 Task 4 (SE-ABSA14) provided datasets annotated with aspect terms (e.g., "hard disk", "pizza") and their polarity for laptop and restaurant reviews, as well as coarser aspect categories (e.g., PRICE) and their polarity only for restaurants[3] (Pontiki et al., 2014). The task attracted 165 submissions from 32 teams that experimented with a variety of features (e.g., based on n-grams, parse trees, named entities, word clusters), techniques (e.g., rule-based, supervised and unsupervised learning), and resources (e.g., sentiment lexica, Wikipedia, WordNet). The participants obtained higher scores in the restaurants domain. The laptops domain proved to be harder involving more entities (e.g., hardware and software components) and complex concepts (e.g., usability, portability) that are often discussed implicitly in the text. The SE-ABSA14 task set-up has been adopted for the creation of aspect-level sentiment datasets in other languages, like Czech (Steinberger et al., 2014).

SemEval-2015 Task 12 (SE-ABSA15) built upon SE-ABSA14 and consolidated its subtasks (aspect category extraction, aspect term extraction, polarity classification) into a principled unified framework (described in Section 2). In addition, SE-ABSA15 included an aspect level polarity classification subtask for the hotels domain in which no training data were provided (out-of-domain ABSA). The annotation schema and the provided datasets are described in Section 3. The evaluation measures and the baseline methods are described in Section 4, while the evaluation scores and the main characteristics of the developed systems are presented in Section 5. The paper concludes with a general assessment of the task.

2 Task Set-Up

2.1 ABSA Framework: From SE-ABSA14 to SE-ABSA15

In SE-ABSA14, given a sentence from a user review about a target entity *e* (e.g., a laptop), the goal was to identify all aspects (explicit terms or categories) and the corresponding polarities. Following Liu (2006) & Zhang and Liu (2014), an aspect (term or category) indicated: (a) a part/component of *e* (e.g., battery), (b) an attribute of *e* (e.g., price), or (c) an attribute of a part/component of *e* (e.g., battery life). In SE-ABSA15, an aspect category is defined as a combination of an entity type E and an attribute type A. This definition of aspect makes more explicit the difference between entities and the particular facets that are being evaluated. E can be the reviewed entity *e* itself (e.g., laptop), a part/component of it (e.g., battery or customer support), or another relevant entity (e.g., the manufacturer of *e*), while A is a particular attribute (e.g., durability, quality) of E. E and A are concept names (classes) from a given domain ontology and do not necessarily occur as terms in a sentence. For example, in "*They sent it back with a huge crack in it and it still didn't work; and that was the fourth time I've sent it to them to get fixed*" the reviewer is evaluating the *quality* (A) of the *customer support* (E) without explicitly mentioning it.

In contrast to SE-ABSA14, in the current framework aspect terms correspond to explicit mentions of the entities E (e.g., service, pizza) or attributes A (e.g., price, quality). However, only the extraction of the explicit mentions of E is required (see Section 2.2). Another difference is that the datasets of SE-ABSA15 consist of whole reviews, not isolated sentences. Correctly identifying the E, A pairs of a sentence and their polarities often requires examining a wider part or the whole review.

In this setting, the ABSA problem has been formalized into a principled unified framework in which all the identified constituents of the expressed opinions (i.e., opinion target expressions, aspects and sentiment polarities) meet a set of guidelines/specifications and are linked to each other within tuples. The extracted tuples directly

[2] http://www.nist.gov/tac/2014/KBP/Sentiment/index.html

[3] The SE-ABSA14 inventory of categories for the restaurants domain is similar to the one of Ganu et al. (2009).

reflect the intended meaning of the texts and, thus, can be used to generate structured aspect-based opinion summaries from user reviews in realistic applications (e.g., review sites).

2.2 Task Description

SE-ABSA15 consisted of the following subtasks. Participants were free to choose the subtasks, slots and domains they wished to participate in.

Subtask 1: In-domain ABSA. Given a review text about a laptop or restaurant, identify all the opinion tuples with the following types (tuple slots) of information:

Slot 1: **Aspect Category**. The goal is to identify every entity E and attribute A pair towards which an opinion is expressed in the given text. E and A should be chosen from predefined inventories of entity types (e.g., LAPTOP, MOUSE, RESTAURANT, FOOD) and attribute labels (e.g., DESIGN, PRICE, QUALITY). The E, A inventories for each domain are described in section 3.

Slot 2: Opinion Target Expression (**OTE**). The task is to extract the OTE, i.e., the linguistic expression used in the given text to refer to the reviewed entity E of each E#A pair. The OTE is defined by its starting and ending offsets. When there is no explicit mention of the entity, the slot takes the value "NULL". The identification of Slot 2 values was required only in the restaurants domain.

Slot 3: Sentiment Polarity. Each identified E#A pair has to be assigned one of the following polarity labels: positive, negative, neutral (mildly positive or mildly negative sentiment).

Two examples of opinion tuples with Slot 1-3 values from the restaurants domain are shown below. Such tuples can be used to generate aspect-sentiment tables like the one of Fig 1.

a. *The food was delicious but do not come here on an empty stomach.* →
{category= "FOOD#QUALITY", target= "food", from: "4", to: "8", polarity= "positive"},
{category= "FOOD#STYLE_OPTIONS"[4], target = "food", from: "4", to: "8", polarity= "negative"}

b. *Prices are in line.* →
{category: "RESTAURANT#PRICES", target= "NULL", from: "-", to: "-", polarity: "neutral"}

[4] Opinions evaluating the food quantity (e.g. portions size) are assigned the label "FOOD#STYLE_OPTIONS".

Subtask 2: Out-of-domain ABSA. In this subtask, participants had the opportunity to test their systems in a previously unseen domain (hotel reviews) for which no training data was made available. The gold annotations for Slots 1 and 2 were provided and the teams had to return the sentiment polarity values (Slot 3).

3 Datasets and Annotation

3.1 Data Collection

Datasets for three domains (laptops, restaurants, hotels) were provided; consult Table 1 for more information.

	Laptops	Restaurants	Hotels
	Training data		
Review texts	277	254	-
Sentences	1739	1315	-
	Test data		
Review texts	173	96	30
Sentences	761	685	266

Table 1. Datasets provided for ABSA.

Note that in the domain of hotels no training data were provided (Out-of-Domain ABSA).

3.2 Annotation Schema and Guidelines

Given a review text about a laptop, a restaurant or a hotel, the task of the annotators was to identify opinions expressed towards specific entities and their attributes and to assign the respective aspect category (Slot 1) and polarity (Slot 3) labels. The category (E#A) values had to be chosen from predefined inventories of entities and attributes for each domain; the inventories were described in detail in the respective annotation guidelines[5]. In particular, the entity E could be assigned 22 possible labels for the laptops domain (e.g., LAPTOP, SOFTWARE, SUPPORT), 6 labels for the restaurants domain (e.g., RESTAURANT, FOOD), and 7 labels for the hotels domain (e.g., HOTEL, ROOMS). The attribute A could be assigned 9 possible labels for the laptops domain (e.g., USABILITY), 5 labels for the restaurants domain (e.g., QUALITY), and 8 labels for the hotels domain (e.g., COMFORT). The

[5] The detailed annotation guidelines are available at: http://alt.qcri.org/semeval2015/task12/index.php?id=data-and-tools

full inventories of the aspect category labels for each domain are provided below in appendices A-C. Quite often reviews contain opinions towards entities that are not directly related to the entity being reviewed, for example, restaurants/hotels that the reviewer has visited in the past, other laptops or products (and their components) of the same or a competitive brand. Such entities as well as comparative opinions are considered to be out of the scope of SE-ABSA15. In these cases, no opinion annotations were provided.

The {E#A, polarity} annotations had to be assigned at the sentence level taking into account the context of the whole review. For example, "*Laptop still did not work, blue screen within a week...*" (Previous sentence: "*Horrible customer support-they lost my laptop for a month-got it back 3 months later*") had to be assigned a negative opinion about the customer support, not about the operation of the laptop, as implied by the previous sentence. Similarly, in "*I was so happy with my new Mac.*" (Next sentences: "*For two months... Then the hard drive failed."*), even though the reviewer says how happy he/she was with the laptop, he/she is expressing a negative opinion.

For the polarity slot the possible values were: positive, negative, and neutral. Contrary to SE-ABSA14, the "neutral" label applies only to mildly positive or mildly negative sentiment, thus it does not indicate objectivity (e.g., "*Food was okay, nothing great*." → {FOOD#QUALITY, "*Food*", neutral}). Another difference is that this year the "conflict" label was not used, since –due to the adopted fine-grained aspect classification schema– it is very rare to encounter (in a sentence) both a positive and a negative opinion about the same attribute A of an entity E. In the few cases where this happened, the dominant sentiment was chosen (e.g., "*The OS takes some getting used to but the learning curve is so worth it*!" → {OS#USABILITY, positive}).

For the restaurants and the hotels domain the annotators also had to tag the OTE (explicit mention) for each identified entity E (Slot 2). Such mentions can be named entities (e.g., "The Four Seasons"), common nouns (e.g., "place", "steak", "bed") or multi-word terms (e.g., "vitello alla marsala", "conference/banquet room"). Similarly to SE-ABSA14, the identified OTEs were annotated as they appeared, even if misspelled. When an evaluated entity E was only implicitly inferred or referred to (e.g., through pronouns), the OTE slot was assigned the value "NULL" (e.g. "*Everything was wonderful*." → {RESTAURANT#GENERAL, NULL, positive}).

In the laptops domain we did not provide OTE annotations, since most entities are instantiated through a limited set of expressions (e.g., MEMORY: "memory", "ram", CPU: "processing power", "processor", "cpu") as opposed to the restaurants domain, where for example, the entity "FOOD" is instantiated through a variety of food types and dishes (e.g. "pizza", "Lobster Cobb Salad"). Furthermore, LAPTOP, which is the majority category label in laptops (see Section 3.3), is instantiated mostly through pronominal mentions, while the explicit mentions are limited to nouns like laptop, computer, product, etc.

3.3 Annotation Process and Statistics

Each dataset was annotated by a linguist (annotator A) using BRAT (Stenetorp et al., 2012), a web-based annotation tool, which was configured appropriately for the needs of the task. Then, one of the organizers (annotator B) validated/inspected the resulting annotations. When B was not confident or disagreed with A, a decision was made collaboratively between them and a third annotator. The main disagreements encountered during the annotation process are summarized below:

Slot 1. In the laptops domain the main difficulty was that in some negative evaluations the annotators were unsure about the actual problem/target. For example, in "*Sometimes the screen even goes black on this computer*", the black screen may be related to the graphics, the laptop operation (e.g., motherboard issue) or the screen itself. The decision for such cases was to assign the E#A pair that reflected what the reviewer is saying and not the possible interpretations that a technician would give. So, if someone reports screen issues without providing further details, then the opinion is considered to be about the screen[6]. Another issue was when an attribute could be inferred from an explicitly evaluated attribute. For example, DESIGN affects USABILITY (e.g., *"With the switch being at the top you need to memorize the key combination*

[6] "Blue screen" is an exception since it is well-known that it refers to the laptop operation.

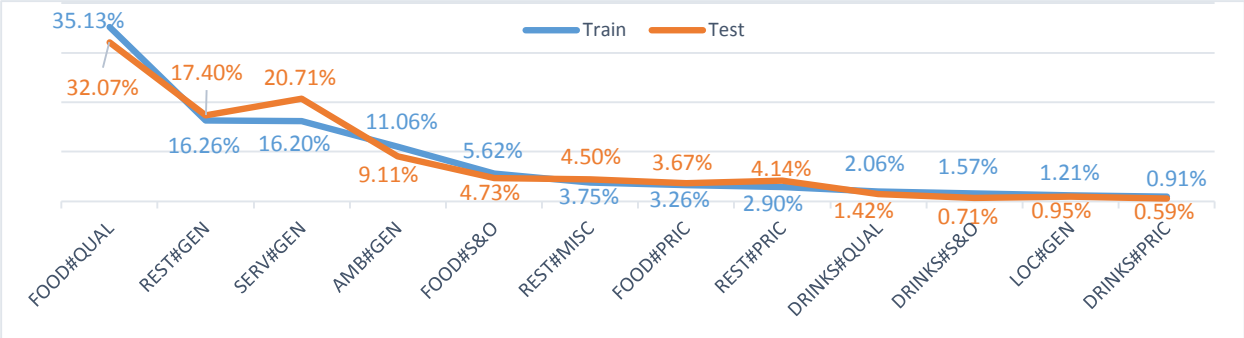

Figure 2. Aspect category (E#A) distribution in the restaurants domain. REST = restaurant, SERV = service, AMB = ambience, LOC = location, GEN=general, PRIC = price, S&O = style&options, MISC= miscellaneous

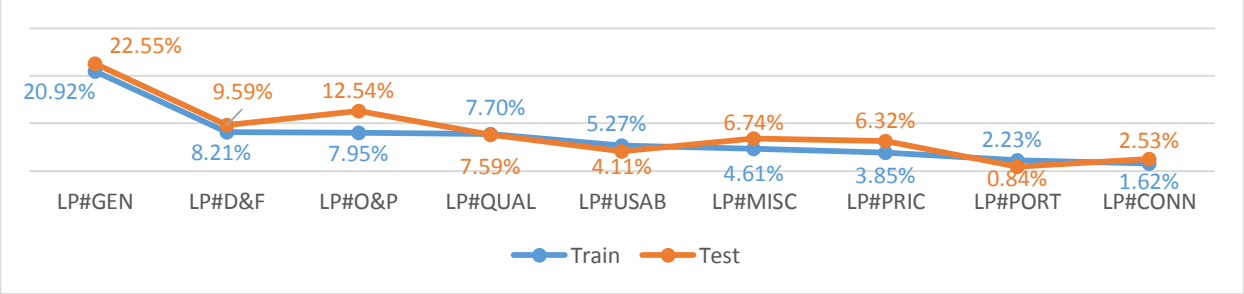

Figure 3. LAPTOP#ATTRIBUTE categories distribution in the laptops domain. LP= laptop, O&P= operation &performance, QUAL= quality, D&F= design &features, USAB=usability, CONN=connectivity, PORT=portability.

rather than just flicking a switch"). In such cases annotators assigned both attribute labels. The annotation in the restaurants domain was easier, due to the less fine-grained schema. A common problem was that (as in SE-ABSA14) the distinction between the GENERAL and MISCELLANEOUS and between the RESTAURANT and AMBIENCE labels was not always clear.

Slot 2. The annotators found it easier to identify explicit references to the target entities as opposed to the more general aspect terms of SE-ABSA14. However, the problem of distinguishing aspect terms when they appear in conjunctions or disjunctions remains. In this case the maximal phrase (e.g. the entire conjunction or disjunction) is annotated (e.g. *"Greek or Cypriot dishes"* instead of *"Greek dishes*", "*Cypriot dishes*").

Slot 3. Most cases in which the annotators had difficulty deciding the correct polarity label fall into one of the following categories: (a) *Change of sentiment over time*. Some reviewers tend to start their review by saying how excited they were at first (e.g., with the laptop) and continue by reporting problems or negative evaluations. (b) *Negative fact vs. positive opinion*. Some reviewers do mention particular deficiencies of a laptop or a restaurant saying, however, at the same time that they do not bother (e.g., *"Overheats but put a pillow and problem solved!"*). (c) *Mildly positive and negative sentiments are both denoted by the "neutral" label*. In some cases the annotators reported that it would be helpful to have a more fine-grained schema (e.g., "negative", "somewhat negative", "neutral", "somewhat positive", "positive"). Finally, in some cases it is difficult to decide a polarity label without knowing the reviewer's intention (e.g., *"50% of the food was very good"*).

The annotation process resulted in 5,761 opinion tuples in total that correspond to more than 15,000 label assignments (E, A, OTE, polarity); consult Table 2 for more information.

Laptops			
	training	test	total
{E#A, polarity}	1974	949	2923
Restaurants			
	training	test	total
{E#A, OTE, polarity}	1654	845	2499
Hotels			
	training	test	total
{E#A, OTE, polarity}	-	339	339

Table 2. Number of tuples annotated per dataset.

The distribution of the category annotations in the restaurants domain (Fig. 2) is similar across the training and test set. In the laptops domain, 81 E,A combinations (different pairs) were annotated in the training set and 58 in the test set. LAPTOP is the majority entity class in both sets; 62.36% in training, 72.81% in test data. Figure 3 presents the distribution for all the attributes of the LAPTOP entity in the training and test sets. Again, the category distributions are similar. The remaining 37.64% of the annotations in the laptops training data correspond to 72 categories with frequencies ranging from 6.53% to 0.05%. In the test set, the remaining 27.19% of the annotations correspond to 49 categories with frequencies from 2.32 % to 0.11%.

Regarding polarity, positive is the majority class in all domains (Table 3). The polarity distribution is balanced in the laptops domain, while in the restaurants domain there is a significant imbalance between the positive and negative classes across the training and the test sets.

	positive	negative	neutral
RS-TR	72.43%	24.36%	3.20%
RS-TE	53.72%	40.96%	5.32%
LP-TR	55.87%	38.75%	5.36%
LP-TE	57%	34.66%	8.32%
HT-TE	71.68%	24.77%	3.53%

Table 3. Polarity distribution per domain (RS-restaurants, LP-laptops, HT-hotels). TR and TE indicate the training and test sets.

3.4 Datasets Format and Availability

The datasets[7] of the SE-ABSA15 task were provided in an XML format. They are available under a non-commercial, no redistribution license through META-SHARE[8], a repository devoted to the sharing and dissemination of language resources (Piperidis, 2012).

4 Evaluation Measures and Baselines

Similarly to SE-ABSA14, the evaluation ran in two phases. In Phase A, the participants were asked to return the {category, OTE} tuples for the restaurants domain and only the category slot (Slot1) for the laptops domain. Subsequently, in Phase B, the participants were given the gold annotations for the reviews of Phase A and they were asked to return the polarity (Slot3). Each participating team was allowed to submit up to two runs per slot and domain in each phase; one constrained (C), where only the provided training data could be used, and one unconstrained (U), where other resources (e.g., publicly available lexica) and additional data of any kind could be used for training. In the latter case, the teams had to report the resources they used. To evaluate aspect category (Slot1) and OTE extraction (Slot2) in Phase A, we used the F-1 measure. To evaluate sentiment polarity (Slot 3) in Phase B, we used accuracy. Furthermore, we implemented and provided three baselines (see below) for the respective slots.

4.1 Evaluation Measures

Slot 1: F-1 scores are calculated by comparing the category annotations that a system returned (for all the sentences) to the gold category annotations (using micro-averaging). These category annotations are extracted from the values of Slot 1 (category). Duplicate occurrences of categories (for the same sentence) are ignored.

Slot 2: F-1 scores are calculated by comparing the targets that a system returned (for all the sentences) to the corresponding gold targets (using micro-averaging). The targets are extracted using their starting and ending offsets. The calculation for each sentence considers only distinct targets and discards NULL targets, since they do not correspond to explicit mentions.

Slot 1&2 (jointly): Again F-1 scores are calculated by comparing the {category, OTE} tuples of a system to the gold ones (using micro-averaging).

Slot 3: To evaluate sentiment polarity detection in Phase B, we calculated the accuracy of each system, defined as the number of correctly predicted polarity labels of aspect categories, divided by the total number of aspect categories. Recall that we use the gold aspect categories in Phase B.

4.2 Baselines

Slot 1: For category (E#A) extraction, a Support Vector Machine (SVM) with a linear kernel was trained. In particular, *n* unigram features are extracted from the respective sentence of each tuple that is encountered in the training data. The category value (e.g., SERVICE#GENERAL) of the tuple is

[7] The data are available at http://metashare.ilsp.gr:8080/.

[8] META-SHARE (http: //www.metashare.org/) was implemented in the framework of the META-NET Network of Excellence (http://www.meta-net.eu/).

used as the correct label of the feature vector. Similarly, for each test sentence *s*, a feature vector is built and the trained SVM is used to predict the probabilities of assigning each possible category to *s* (e.g., {SERVICE#GENERAL, 0.2}, {RESTAURANT#GENERAL, 0.4}. Then, a threshold[9] t is used to decide which of the categories will be assigned[10] to *s*. As features, we use the 1,000 most frequent unigrams of the training data excluding stop-words.

Slot 2: The baseline uses the training reviews to create for each category *c* (e.g., SERVICE#GENERAL) a list of OTEs (e.g., SERVICE#GENERAL → {"staff", "waiter"}). These are extracted from the (training) opinion tuples whose category value is *c*. Then, given a test sentence *s* and an assigned category *c*, the baseline finds in *s* the first occurrence of each OTE of *c*'s list. The OTE slot is filled with the first of the target occurrences found in *s*. If no target occurrences are found, the slot is assigned the value NULL.

Slot 3: For polarity prediction we trained a SVM classifier with a linear kernel. Again, as in Slot 1, *n* unigram features are extracted from the respective sentence of each tuple of the training data. In addition, an integer-valued feature[11] that indicates the category of the tuple is used. The correct label for the extracted training feature vector is the corresponding polarity value (e.g., positive). Then, for each tuple {category, OTE} of a test sentence *s*, a feature vector is built and it is classified using the trained SVM. Furthermore, for Slot 3 we also used a majority baseline that assigns the most frequent polarity (in the training data) to all test tuples.

The baseline systems and evaluation scripts are available for download as a single zip from the SE-ABSA15 website[12]. They are implemented in Java and can be used via a Linux shell script. The baselines use the LibSVM package[13] (Chang and Lin, 2011) for SVM training and prediction. The scores of the baselines in the test datasets are presented in Tables 4–8 along with the system scores.

[9] The threshold *t* was tuned on a subset of the training data (for each domain) using a trial and error approach.

[10] We use the –b 1 option of LibSVM to obtain probabilities.

[11] Each category (E#A pair) has been assigned a distinct integer value.

[12] http://alt.qcri.org/semeval2015/task12/index.php?id=data-and-tools

[13] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

5 Evaluation Results

In total, the task attracted 92 submissions from 16 teams. The evaluation results per phase and slot are presented below. For the teams that submitted more than one unconstrained runs per slot and domain, we included in the tables only the run with the highest score.

5.1 Results of Phase A

The aspect category identification slot attracted 6 teams for the laptops dataset and 9 teams for the restaurants dataset (consult Table 4). As expected, the systems achieved significantly higher scores (+12%) in the restaurants domain since in this domain the classification schema is less fine-grained; it contains 6 entity types and 5 attribute classes that result in 12 possible combinations, as opposed to the laptops domain where the 22 entities and 9 attribute labels give rise to more than 80 combinations. The best F-1 scores in both domains, 50.86% for laptops and 62.68% for restaurants, were achieved by the unconstrained submission of the NLANGP team, which modeled aspect category extraction as a multiclass classification problem with features based on n-grams, parsing, and word clusters learnt from Amazon and Yelp data (for laptops and restaurants, respectively). The system of Sentiue (scores: 50% on laptops, 54.10% on restaurants) used a separate MaxEnt classifier with bag-of-word-like features (e.g. words, lemmas) for each entity and for each attribute. Subsequently, heuristics are applied to the output of the classifiers to determine which categories will be assigned to each sentence.

Laptops		Restaurants	
Team	F1	Team	F1
NLANGP	50.86*	NLANGP	62.68*
Sentiue	50.00*	NLANGP	61.94
IHS-RD.	49.59	UMDuluthC	57.19
NLANGP	49.06	UMDuluthT	57.19
TJUdeM	46.49	SIEL	57.14*
UFRGS	44.95	Sentiue	54.10*
UFRGS	44.73*	LT3	53.67*
V3	24.94*	TJUdeM	52.44*
		UFRGS	52.09*
		UFRGS	51.88
		IHS-RD.	49.87
		IHS-RD.	49.16
		V3	41.85*

Baseline	**48.06**	Baseline	**51.32**

Table 4. F-1 scores for aspect category extraction (slot 1). * indicate unconstrained systems.

The OTE slot, which was used only in the restaurants domain, attracted 14 teams; consult Table 5. The best F1 score (70.05%) was achieved by the unconstrained submission of EliXa that addressed the problem using an averaged perceptron with a BIO tagging scheme. The features EliXa used included n-grams, token classes, n-gram prefixes and suffixes, and word clusters learnt from additional data (Yelp for Brown and Clark clusters; Wikipedia for word2vec clusters). Similarly, NLANGP (67.11%) was based on a Conditional Random Fields (CRF) model with features based on word strings, head words (obtained from parse trees), name lists (e.g. extracted using frequency), and Brown clusters.

Restaurants			
Team	F1	Team	F1
EliXa	70.05*	UMDuluthC	50.36
NLANGP	67.11*	UMDuluthT	50.36
IHS-RD.	63.12	LT3	49.97*
Lsislif	62.22	UFRGS	49.32*
NLANGP	61.49	V3	45.67*
wnlp	57.63	Sentiue	39.82*
SIEL	53.38*	CU-BDDA	36.01
TJUdeM	52.44*	CU-BDDA	33.86*
Baseline		**48.06**	

Table 5. Results for OTE extraction (slot 2). * indicate unconstrained systems.

Finally, as expected, the scores are significantly lower when systems have to link the extracted OTEs to the relevant aspect categories (Slot1&2 jointly). As shown in Table 6, the best F-1 score (42.90%) was achieved by the NLANGP team that simply combined the output for each slot to construct the corresponding tuples.

Restaurants			
Team	F1	Team	F1
NLANGP	42.90*	LT3	35.50*
IHS-RD.	42.72	UFRGS	34.87*
IHS-RD.	41.96	UMDuluthC	32.59
NLANGP	39.81	UMDuluthT	32.59
TJUdeM	37.15*	Sentiue	31.20*
Baseline		**34.44**	

Table 6. Results for Slot1&2. * indicate unconstrained systems.

5.2 Results of Phase B

The sentiment polarity slot attracted 10 teams for the laptops and 12 teams for the restaurants domain (see Table 7). The best accuracy scores in both domains, 79.34% for laptops and 78.69% for restaurants, were achieved by Sentiue with a MaxEnt classifier along with features based on n-grams, POS tagging, lemmatization, negation words and publicly available sentiment lexica (MPQA, Bing Liu's lexicon, AFINN). The system of ECNU (scores: 78.29% laptops, 78.10% restaurants) used features based on n-grams, PMI scores, POS tags, parse trees, negation words and scores based on 7 sentiment lexica. The lsislif team (77.87% laptops, 75.50% restaurants) relied on a logistic regression model (Liblinear) with various features: syntactic (e.g., unigrams, negation), semantic (Brown dictionary), sentiment (e.g., MPQA, SentiWordnet).

Laptops		Restaurants	
Team	Acc.	Team	Acc.
Sentiue	79.34*	Sentiue	78.69*
ECNU	78.29	ECNU	78.10*
Lsislif	77.87	Lsislif	75.50
ECNU	74.49*	LT3	75.02*
LT3	73.76*	UFRGS	71.71
TJUdeM	73.23*	Wnlp	71.36
EliXa	72.91*	UMDuluthC	71.12
Wnlp	72.07	EliXa	70.05*
EliXa	71.54	ECNU	69.82
V3	68.38*	V3	69.46*
UFRGS	67.33	TJUdeM	68.87*
SINAI	65.85	EliXa	67.33
SINAI	51.84*	SINAI	60.71*
		SIEL	70.76*
SVM+ BOW Baseline	**69.96**	SVM+ BOW Baseline	**63.55**
Majority Baseline	**57.00**	Majority Baseline	**53.72**

Table 7. Accuracy scores for slot 3 (polarity extraction). * indicate unconstrained systems. The evaluated run of SIEL team was submitted after the deadline had expired, but before the release of the gold polarity labels.

Most teams performed (slightly) better in the laptops domain. This is probably due to the fact that in the restaurants domain the positive polarity is significantly more frequent in the training than in the test data, which may have led to biased models. Nevertheless, most system scores indicate robustness across the two domains, with Sentiue

achieving the most stable performance: 79.34% in laptops and 78.69% in restaurants.

A similar score was obtained also by Sentiue in the hidden domain (78.76%). The (hidden) hotels domain (subtask 2) attracted 9 teams. Lsislif achieved the best score based on a Liblinear model developed for the restaurants domain. LT3 achieved the second best score (80.53%) with an SVM model trained on the restaurants training data. The model used features based on unigrams, sentiment lexica (by Bing Liu, General Inquirer) and PMI scores learnt from TripAdvisor data. The team of EliXa (79.64%) used a multiclass SVM and features based on word clusters, lemmas, n-grams, POS tagging, and well known sentiment lexica. The system of Sentiue (78.76%) is somewhat similar; it uses BOW, POS tags, lemmas, and sentiment lexica. The results of some systems (LT3, EliXa, V3) suggest that the hidden domain was easier, but other systems (e.g., ECNU, wnlp) achieved significantly lower scores in the hidden domain, compared to the in-domain ABSA scores.

Hotels			
Team	Acc.	Team	Acc.
lsislif	85.84	V3	71.09*
LT3	80.53*	UFRGS	65.78
EliXa	79.64*	SINAI	63.71*
sentiue	78.76*	Wnlp	55.45
EliXa	74.92	UMDuluthC	71.38
Majority Baseline		**71.68**	

Table 8. Accuracy scores for slot 3 (polarity extraction). * indicate unconstrained systems. The evaluated run of UMDuluthC team was submitted after the deadline had expired but before the release of the gold polarity labels.

6 Conclusions

The SE-ABSA15 task is a continuation of SE-ABSA14 task. The SE-ABSA15 task provided a new definition of aspect –that makes explicit the difference between entities and the particular facets that are being evaluated- within a new principled, unified ABSA framework and output representation, which may be used in realistic applications (e.g., review sites). We also provided benchmark datasets containing manually annotated reviews from three domains (restaurants, laptops, hotels) and baselines for the respective SE-ABSA15 slots. The task attracted 93 submissions from 16 teams that were evaluated in three slots: aspect categories, opinion target expressions, and polarity classification. Future work includes applying the new framework and annotation schema to other languages (e.g., Spanish, Greek) and enhancing it with information about topics or events, opinion holders, and annotations for linguistic phenomena like metaphor and irony.

Acknowledgments

We thank Konstantina Papanikolaou, who carried out a critical part of the annotation process, Thomas Keefe for his help during the initial phases of the annotation process, Juli Bakagianni for her support on the META-SHARE platform and John Pavlopoulos for his valuable contribution in shaping the SE-ABSA tasks. Maria Pontiki & Haris Papageorgiou were supported by the POLYTROPON (KRIPIS-GSRT, MIS: 448306) project.

References

Bing Liu. 2006. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2006 and 2011: Springer.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In Proceedings of NAACL, pages 804–812, Los Angeles, California.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27.

Gayatree Ganu, Noemie Elhadad, and Amelie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In Proceedings of WebDB, Providence, Rhode Island, USA.

Minqing Hu and Bing Liu. 2004a. Mining and summarizing customer reviews. In Proceedings of KDD, pages 168–177, Seattle, WA, USA.

Minqing Hu and Bing Liu. 2004b. Mining opinion features in customer reviews. In Proceedings of AAAI, pages 755–760, San Jose, California.

Yohan Jo, and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In Proceedings of the fourth ACM international conference on Web search and data mining, pages 815-824.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, pages 375–384.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In Proceedings of the 12th IEEE Inter-

national Conference on Data Mining, ICDM '12, pages 1020–1025, Brussels, Belgium.

Margaret Mitchell. 2013. Overview of the TAC2013 Knowledge Base Population Evaluation: English Sentiment Slot Filling. In Proceedings of the Text Analysis Conference (TAC), Gaithersburg, MD, USA.

John Pavlopoulos. 2014. *Aspect based sentiment analysis*. PhD thesis, Dept. of Informatics, Athens University of Economics and Business, Greece.

Stelios Piperidis. 2012. The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In Proceedings of LREC-2012, pages 36–42, Istanbul, Turkey.

Maria Pontiki, Dimitrios 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.

Josef Ruppenhofer, Roman Klinger, Julia Maria Struß, Jonathan Sonntag, and Michael Wiegand. 2014. IGGSA Shared Tasks on German Sentiment Analysis (GESTALT). In Workshop Proceedings of the 12th Edition of the KONVENS Conference, pages 164-173.

Yohei Seki, David Kirk Evans, Lun-Wei Ku, Hsin-Hsi. Chen, Noriko Kando, and Chin-Yew Lin. 2007. Overview of opinion analysis pilot task at ntcir-6. In Proceedings of NTCIR-6 Workshop Meeting, pages 265–278.

Yohei Seki, David Kirk Evans, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, and Noriko Kando. 2008. Overview of multilingual opinion analysis task at NTCIR-7. In Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access, pages 185–203.

Yohei Seki, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, and Noriko Kando. 2010. Overview of Multilingual Opinion Analysis Task at NTCIR-8: A Step Toward Cross Lingual Opinion Analysis. In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, pages 209–220.

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*, Association for Computational Linguistics, pages 24–30, Baltimore, Maryland.

Pontus Stenetorp, Sampo Pyysalo, Goran Topi´c, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In Proceedings of EACL, pages 102–107, Avignon, France.

Tun Thura Thet, Jin-Cheon Na, and Christopher S. G. Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. J. Information Science, 36(6):823–848.

Lei Zhang and Bing Liu. 2014. Aspect and Entity Extraction for Opinion Mining", book chapter in Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenges, and Opportunities, Springer, 2014.

Appendix A. Laptop Aspect Categories

Entity Labels	
1. LAPTOP	13. BATTERY
2. DISPLAY	14. GRAPHICS
3. KEYBOARD	15. HARD DISK
4. MOUSE	16. MULTIMEDIA DEVICES
5. MOTHERBOARD	17. HARDWARE
6. CPU	18. SOFTWARE
7. FANS& COOLING	19. OS
8. PORTS	20. WARRANTY
9. MEMORY	21. SHIPPING
10. POWER SUPPLY	22. SUPPORT
11. OPTICAL DRIVES	23. COMPANY
Attribute Labels	
A. GENERAL	E. USABILITY
B. PRICE	F. DESIGN& FEATURES
C. QUALITY	G. PORTABILITY
D. OPERATION& PERFORMANCE	H. CONNECTIVITY
	I. MISCELLANEOUS

Appendix B. Restaurant Aspect Categories

Entity Labels	**Attribute Labels**
1. RESTAURANT	A. GENERAL
2. FOOD	B. PRICES
3. DRINKS	C. QUALITY
4. AMBIENCE	D. STYLE & OPTIONS
5. SERVICE	E. MISCELLANEOUS
6. LOCATION	

Appendix C. Hotel Aspect Categories

Entity Labels	**Attribute Labels**
1. HOTEL	A. GENERAL
2. ROOMS	B. PRICE
3. FACILITIES	C. COMFORT
4. ROOM AMENITIES	D. CLEANLINESS
5. SERVICE	E. QUALITY
6. LOCATION	F. DESIGN & FEATURES
7. FOOD & DRINKS	G. STYLE & OPTIONS
	H. MISCELLANEOUS