ABSITA: Aspect-based Sentiment Analysis at EVALITA 2018 - Submission Guidelines

Pierpaolo Basile
University of Bari
pierpaolo.basile@uniba.it

Valerio Basile
Sapienza University of Rome
basile@di.uniroma1.it

Danilo Croce
University of Rome "Tor Vergata"
croce@info.uniroma2.it

Marco Polignano
University of Bari
marco.polignano@uniba.it

September 10, 2018

1 Task Description

We approach Aspect-based Sentiment Analysis as a cascade of two subtasks:

1. **Aspect Category Detection** (ACD)

2. **Aspect Category Polarity** (ACP)

For example, let us consider the sentence describing an hotel:

$$I \text{ } servizi \text{ } igienici \text{ } sono \text{ } puliti \text{ } e \text{ } il \text{ } personale \text{ } cordiale \text{ } e \text{ } disponibile.$$
$$(\textit{Toilets are clean but the staff is not friendly nor helpful.}) \tag{1}$$

In the ACD task one or more "aspect categories" evoked in a sentence are identified, e.g. the `pulizia` (`cleanliness`) and `staff` categories in sentence 1. In the **Aspect Category Polarity** (ACP) task, the polarity of each expressed category is recognized, e.g. a `positive` category polarity is expressed concerning the `pulizia` category while it is `negative` if considering the `staff` category.

In our evaluation framework, the set of aspect categories is known and given to the participants, so the ACD task can be seen as a multi-class, non-exclusive classification task where each input text has to be classified as evoking or not each aspect category. The systems will be asked to return a binary vector where each dimension corresponds to an aspect category and the values 0 (false) and 1 (true) indicate whether each aspect has been detected in the text. Table 1 shows examples of annotation for the ACD task with four aspects.

Text	Cleanliness	Staff	Comfort	Location
I servizi igienici sono puliti e il personale cordiale e disponibile	1	1	0	0
La posizione è molto comoda per il treno e la metro.	0	0	0	1
Ottima la disponibilit√† del personale, e la struttura della stanza	0	1	1	0

Table 1: Examples of categories detection ACD.

For the ACP task, the input is the review text paired with the set of aspects identified in the text by the ACD subtask, and the goal is to assign polarity labels to each of the aspect category. Two binary polarity labels are expected for each aspect: POS an NEG, indicating a positive and

negative sentiment expressed towards a specific aspect, respectively. Note that the two labels are not mutually exclusive: in addition to the annotation of *positive* aspects (POS:true, NEG:false) and *negative* aspects (POS:false, NEG:true), there can be aspects with no polarity, or *neutral* polarity (POS:false, NEG:false). This is also the default polarity annotation for the aspects that are not detected in a text. Finally, the polarity of an aspect can be *mixed* (POS:true, NEG:true), in cases where both sentiments are expressed towards a certain aspect in a text. Table 2 summarizes the possible annotations with examples.

Text	Aspect	POS	NEG
Il bagno andrebbe ristrutturato	Cleanliness	0	0
Camera pulita e spaziosa.	Cleanliness	1	0
Pulizia della camera non eccelsa.	Cleanliness	0	1
Il bagno era pulito ma lasciava un po a desiderare	Cleanliness	1	1

Table 2: Examples of polarity annotations with respect to the *cleanliness* aspect.

2 Annotation Process and Dataset format

In order to obtain a complete dataset for ABSA, we annotated the sentences from the hotel reviews according to seven aspects:

1. *pulizia* (cleanliness)

2. *comfort*

3. *servizi* (amenities)

4. *staff*

5. rapporto *qualita-prezzo* (value)

6. *wifi* (wireless Internet connection)

7. *posizione* (location)

For each aspect, the polarity (positive, negative) of its mention has been annotated. The positive and negative polarities are annotated independently, thus for each aspect four sentiment classes are possible: *positive* (positive=yes, negative=no), *negative* (positive=no, negative=yes), *neutral* (positive=no, negative=no), *mixed* (positive=yes, negative=yes).

Please note that the special topic "other" has been added for completeness, to annotate sentences with opinions on aspects not among the seven considered by the task. **The aspect "other" is provided additionally and it will be not part of the evaluation of results provided for the task.**

Four different annotators have been involved in the task in parallel, i.e., dividing the dataset into four equal subsets of about **2500 sentences each**. Furthermore, a portion of the dataset made of 250 sentences was annotated by all four annotators. We computed the calculated the inter-annotator agreement per-class over this subset and reported a variation between 85% and 100% (percentage of sentences for which all annotators agreed). Incomplete, irrelevant, and incomprehensible sentences have been discarded from the dataset during the annotation.

The data format used is CSV with UTF-8 encoding and semicolon as separator. The first attribute is the id of the review. It is important to note that in booking.com the order of positive and negative sentences is strictly defined and this can make easier the task. To overcome this issues, we randomly assign for each sentence a new position in the review. As a consequence, the final positional id showed in the data file will not reflect the real order of the sentences in the review. The text of the sentence will be provided at the end of the line and delimited by ", preceded by three binary values for each aspect indicating respectively: the presence of the aspect in the sentence ($aspectX\_presence$:0/1), the positive polarity for that aspect ($aspectX\_pos$:0/1) and finally the

```
sentence_id; aspect1_presence; aspect1_pos; aspect1_neg; ...; sentence
201606240;0;0;0;0;0;0;0;0;0;0;0;0;1;1;0;0;0;0;1;1;0;"Considerato il prezzo e per una sola notte,
va benissimo come punto di appoggio."
201606241;1;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;"Almeno i servizi igienici andrebbero rivisti
e migliorati nella pulizia."
201606242;0;0;0;1;0;1;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;"La struttura purtroppo √® vecchia e ci
vorrebbero dei lavori di ammodernamento."
```

Figure 1: Sample of the annotated dataset in CSV format.

negative polarity ($aspectX\_neg$:0/1). Figure 1 shows an example of the annotated dataset in the proposed format.

The *_presence field for the ACD task is computed as the logic inclusive OR of the respective *_positive and *_negative fields.

3 Evaluation and baselines

We evaluate the ACD and ACP subtasks separately by comparing the classifications provided by the participant systems to the gold standard annotations of the test set. For the ACD task, we compute Precision, Recall and F_1-score defined as: $F1_a = \frac{2P_a R_a}{P_a + R_a}$, where Precision ($P_a$) and Recall ($R_a$) are defined as: $P_a = \frac{|S_a \cap G_a|}{|S_a|}$; $R_a = \frac{|S_a \cap G_a|}{|G_a|}$. Here S_a is the set of aspect category annotations that a system returned for all the test sentences, and G_a is the set of the gold (correct) aspect category annotations. For instance, if a review is labeled in the gold standard with the two aspects $G_a = \{\text{CLEANLINESS}, \text{STAFF}\}$, and the system predicts the two aspects $S_a = \{\text{CLEANLINESS}, \text{COMFORT}\}$, we have that $|S_a \cap G_a| = 1$, $|G_a| = 2$ and $|S_a| = 2$ so that $P_a = \frac{1}{2}$, $R_a = \frac{1}{2}$ and $F1_a = \frac{1}{2}$. For the ACD task the baseline will be computed by considering a system which assigns the most frequent aspect category (estimated over the training set) to each sentence.

For the ACP task we will evaluate the entire chain, thus considering both the aspect categories detected in the sentences together with their corresponding polarity, in the form of $(aspect, polarity)$ pairs. We again compute Precision, Recall and F_1-score now defined as $F1_p = \frac{2P_p R_p}{P_p + R_p}$. Precision ($P_p$) and Recall ($R_p$) are defined as $P_p = \frac{|S_p \cap G_p|}{|S_p|}$; $R_p = \frac{|S_p \cap G_p|}{|G_p|}$, where S_p is the set of $(aspect, polarity)$ pairs that a system returned for all the test sentences, and G_a is the set of the gold (correct) pairs annotations. For instance, if a review is labeled in the gold standard with the pairs $G_p = \{(\text{CLEANLINESS}, POS), (\text{STAFF}, POS)\}$, and the system predicts the three pairs $S_p = \{(\text{CLEANLINESS}, POS), (\text{CLEANLINESS}, NEG), (\text{COMFORT}, POS)\}$, we have that $|S_p \cap G_p| = 1$, $|G_2| = 2$ and $|S_p| = 3$ so that $P_a = \frac{1}{3}$, $R_a = \frac{1}{2}$ and $F1_a = 0.28$.

For the ACP task, the baseline will be computed by considering a system which assigns the most frequent $(aspect, polarity)$ pair (estimated over the training set) to each sentence.

We will produce separate rankings for the tasks, based on the F_1 scores. Participants who submit only the result of the ACD task will appear in the first ranking only.

4 Submission guide

The participants may choose to submit the results obtained for the ACD subtask only or them obtained for the complete ACP task. In both the cases, the submissions must follow the full **CSV format of the training set**. **The textual content of the sentence could be omitted, along with the values in the columns of the aspect "other".** The submission format is the same for the systems that participates in the ACD task only and those that participates to both tasks. For the ACD-only participants, the *_positive and *_negative fields can be left blank.

Submissions for the **ACD** task must have valued only columns which refers to sentence categories.

On the contrary, results submitted for the **ACP** task must have assessed the polarity columns for a given sentence (POS:false, NEG:false) only if the aspect is detected and annotated with "1" in the category column.

Once your system has produced the results for the task over the test set, you have to follow the next instructions for completing your submission:

1. Create a CSV file which respect the following filename format:
   absita_taskname_teamName_systemID.csv

2. Send the file to the following email addresses: absita.evalita@gmail.com and marco.polignano@uniba.it, using the subject "[ABSITA] Submission: taskName - teamName", where the "taskName" is one of ACD and ACP

Participants are allowed to change their submission only 1 time. For each sub-task, a separate official ranking will be provided.

The submission deadline is defined for the **September 16th 2018**.

5 Contacts

For any information about the ABSITA tasks you can send an email to absita.evalita@gmail.com or start a new topic on our google group mailing list: absita2018@googlegroups.com