Overview of the EVALITA 2018 Aspect-based Sentiment Analysis task (ABSITA)

Pierpaolo Basile, Valerio Basile, Danilo Croce, Marco Polignano

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

UNIVERSITÀ DEGLI STUDI DI TORINO

Università di Roma Tor Vergata

EVALITA
Evaluation of NLP and Speech Tools for Italian

User’s opinion

Booking.com

«Posizione e comodità ai mezzi.Personale cortesissimo.»

R **Rita**
Italia
7,0

Arredo camere, un po' vetusto. Assenza di mini-bar interno o anche semplice macchinetta caffè/bevande calde, abbastanza anomalo in una struttura simile!

L **Luca**
Italia
4,3

I croissant erano buoni tipici di Roma! Grande varietà di cibo, cappuccino ottimo! Grande varietà di cioccolata Perugina, marmellate, succhi di frutta.

R **Roserica**
Svizzera
7,6

Location and Staff Positive

Comfort Negative

Service Positive

Motivation

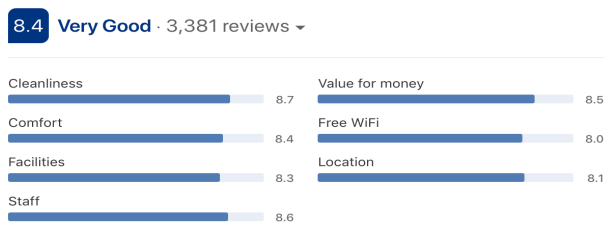
Users no more passive

Amazon, TripAdvisor or Booking.com allow people to express their opinions on items and services, such as **hotels** and **restaurants.**

Sentiment Analysis

Aspect-based Sentiment Analysis (**ABSA**) is an evolution of Sentiment Analysis that aims at capturing the **aspect-level opinions expressed in natural language texts**

Application in real domains



Relevant Research Topic

The task was repeated in SemEval 2015 and 2016, aiming to facilitate more in-depth research

Task Description

At a glance

Participants are asked to **detect** within sentences (expressing opinions about accommodation services) some of the **aspects considered by the writer**.

For each detected aspect, participants are asked to detect a specific **polarity class**

The set of considered aspects is: PULIZIA (cleanliness), COMFORT, SERVIZI (amenties), STAFF, QUALITA-PREZZO (value), WIFI and POSIZIONE (location).

Aspect Category Detection (ACD)

In the ACD task, one or more **"aspect categories"** evoked in a sentence have to be identified, e.g. the posizione (location).

COMFORT ✗ STAFF ✗ LOCATION ✓ VALUE ✗ ...

Aspect Polarity Detection (ACP)

Each **category aspect** detected in the ACD task have to be annotated with **polarity label**: **POS** (positive) , **NEG** (negative), also in a not exclusive way (**Mixed**)

LOCATION POS ✓ LOCATION NEG ✗

DATASET

Booking.com

The data source chosen for creating the ***ABSITA datasets*** is the popular website ***booking.com***

We extracted the **textual reviews** in the Italian language, labeled on the website with one of the 8 considered aspects. We collect in total **4,121** distinct reviews.

Annotation Strategy

The reviews have been manually checked to verify the annotation and to add missing links between sentences and aspects

- We started by annotating 250 randomly chosen sentences observing an **inter-annotators agreement** rating of **94.4%** average
- In order to complete the annotation, we assigned different 1,000 reviews to each annotator that correspond to **2,500 sentences on average**

Each annotator received a **uniformly balanced distribution** of positive and negative aspects. We annotated in total more than 10,000 sentences.

DATASET: Statistics

Released datasets:

Trial set:	**30 sentences**	*0.34%*
Training set:	**6,337 sentences**	*69.75%*
Test set:	**2,718 sentences**	*29.91%*

Splitting percentage:

```
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    ..."
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   ..."
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 \'e vecchia e ci vorrebbero ..."
```

Dataset	clean_pos	comf_pos	amen_pos	staff_pos	value_pos	wifi_pos	loca_pos
Trial set	2	8	6	3	1	1	5
Training set	504	978	948	937	169	43	1,184
Test set	193	474	388	411	94	18	526

Dataset	clean_neg	comf_neg	amen_neg	staff_neg	value_neg	wifi_neg	loca_neg
Trial set	1	2	3	1	1	0	1
Training set	383	1,433	920	283	251	86	163
Test set	196	666	426	131	126	52	103

Evaluation and baselines

Evaluation protocol:

We evaluate the **ACD and ACP subtasks separately.** The **baseline** is computed by considering a system which assigns the **most frequent (aspect, polarity) pair** estimated over the training set to each sentence. This pair is equal to **"comfort : negative"**

ACD TASK

We calculate the micro Precision (**P_a**), Recall (**R_a**) and F1-score (**$F1_a$**):

$$P_a = \frac{|S_a \cap G_a|}{|S_a|} \qquad R_a = \frac{|S_a \cap G_a|}{|G_a|} \qquad F1_a = \frac{2P_aR_a}{P_a+R_a}$$

Where **S_a** is the set of labels returned for each sentence and **G_a** the set of the gold (correct) aspect category annotations.

As an example:

$$S_a = \{\text{CLEANLINESS}, \text{COMFORT}\} \qquad G_a = \{\text{CLEANLINESS}, \text{STAFF}\}$$

$$P_a = \frac{1}{2} \qquad R_a = \frac{1}{2} \qquad F1_a = \frac{1}{2}$$

ACP TASK

We calculate the micro Precision (**P_b**), Recall (**R_b**) and F1-score (**$F1_b$**) considering both the **aspect** categories detected in the sentences together with their corresponding **polarity**.

Where **S_a** is the set of labels returned for each sentence and **G_p** the set of the gold (correct) aspect category annotations.

As an example:

$$G_p = \{(\text{CLEANLINESS}, POS), (\text{STAFF}, POS)\}$$

$$S_a = \{(\text{CLEANLINESS}, POS), (\text{CLEANLINESS}, NEG), (\text{COMFORT}, POS)\}$$

$$P_a = \frac{1}{3} \qquad R_a = \frac{1}{2} \qquad F1_a = 0.28$$

Results

Partecipants

- **7** teams
- **11** partecipants

- *20 total runs*
- 12 runs for ACD
- 8 runs for ACP

- Of the 7 teams who participated to the ACD task, 5 teams also participated to the ACP task.

ACD Partecipants

Systems	Micro-P	Micro-R	Micro-F1
italiaNLP_1	0.8397	**0.7837**	**0.8108**
gw2017_1	0.8713	0.7504	0.8063
gw2017_2	0.8697	0.7481	0.8043
X2Check_gs	0.8626	0.7519	0.8035
UNIPV	0.8819	0.7378	0.8035
X2Check_w	**0.8980**	0.6937	0.7827
italiaNLP_2	0.8658	0.697	0.7723
SeleneBianco	0.7902	0.7181	0.7524
VENSES_1	0.6232	0.6093	0.6162
VENSES_2	0.6164	0.6134	0.6149
ilc_2	0.5443	0.5418	0.5431
ilc_1	0.6213	0.433	0.5104
mfc	**0.4111**	**0.2866**	**0.3377**

ACP Partecipants

Systems	Micro-P	Micro-R	Micro-F1
italiaNLP_1	0.8264	0.7161	**0.7673**
UNIPV	**0.8612**	0.6562	0.7449
gw2017_2	0.7472	**0.7186**	0.7326
gw2017_1	0.7387	0.7206	0.7295
italiaNLP_2	0.8735	0.5649	0.6861
SeleneBianco	0.6869	0.5409	0.6052
ilc_2	0.4123	0.3125	0.3555
ilc_1	0.5452	0.2511	0.3439
mfc baseline	**0.2451**	**0.1681**	**0.1994**

Submitted systems

- **5 systems** (*ItaliaNLP*, *gw2017*, *X2Check*, *UNIPV*, *SeleneBianco*) are based **on supervised machine learning** and **3 systems** (*ItaliaNLP*, *gw2017*, *UNIPV*) employ **deep learning** (in particular LTSM networks, often in their bi-directional variant).

- Pre-trained word embeddings are used as word representations by ***UNIPV*** and ***gw2017***. ***ItaliaNLP*** employs word embedding created from the ItWaC corpus (Baroni et al., 2009) and corpus extracted from **Booking.com**.

- ***ItaliaNLP***, ***VENSES*** and ***X2Check*** *used* **pre-existing NLP pipelines**. Other systems make use of off-the-shelf NLP tools such as SpaCy (***gw2017***, ***UNIPV***) and Freeling (***SeleneBianco***).

- Additional resources used by the systems often include **domain-specific or affective lexicons**. ***ItaliaNLP*** employed the MPQA affective lexicon. ***UNIPV*** system makes use of the affective lexicon for Italian developed in the framework of the OpeNER project

- All runs submitted can be considered **"constrained runs"**, the systems were trained on the provided data set only

Consideration

- The results obtained by the teams **largely outperform the baseline** demonstrating the efficacy of the solutions proposed and the **affordability** of all the two tasks

- The results obtained for the **ACD** task show a **small range of variability**: top results are concentrated around a F1 score value of 0.80

- The values of **precision** and **recall** show **higher variability**, indicating significant difference among the proposed approaches

- Good results have also been obtained using **rule-based systems**, even though they suffer from generalization issues and need to be tailored on the set of sentences to classify

Conclusion

Good Results

Systems based on Machine Learning strategies performed very well on the task and they largely outperforming baselines

Italia_NLP

The system is first classified in both the two subtask: **ACD** and **ACP**

Relevant outcomes

The results achieved by the systems strongly supports the state of the art of ABSA for the Italian language

Systems Details Available

More details about the implementation of the systems that participated in the task can be found in their specific reports

Future Directions

The decision to use additional resources as lexicons in conjunction with semantic word embeddings have been demonstrated to be successful…

Extra resources

The definition of new lexicons and resources for supporting the task in the Italian language is an exciting future research direction

ABSITA

Aspect-based Sentiment Analysis at EVALITA 2018

http://sag.art.uniroma2.it/absita/

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

Valerio Basile
University of Turin
basile@di.unito.it

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

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