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

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

Valerio Basile
University of Turin
basile@di.uniroma1.it

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

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

Abstract

English. ABSITA is the Aspect-based Sentiment Analysis task at EVALITA 2018 (Caselli et al., 2018). This task aimed to foster research in the field of aspect-based sentiment analysis within the Italian language: the goal is to identify the aspects of given target entities and the sentiment expressed for each aspect. Two subtasks are defined, namely Aspect Category Detection (ACD) and Aspect Category Polarity (ACP). In total, 20 runs were submitted by 7 teams comprising 11 total individual participants. The best system achieved a micro F1-score of 0.810 for ACD and 0.767 for ACP.

Italiano. *ABSITA è l’esercizio di valutazione di aspect-based sentiment analysis di EVALITA 2018 (Caselli et al., 2018). Il compito ha l’obiettivo di promuovere la ricerca nel campo della sentiment analysis per lingua italiana: ai partecipanti è stato richiesto di identificare gli aspetti rilevanti per le entitá fornite come input e la sentiment espressa per ognuno di essi. In particolare abbiamo definito come sottotask l’Aspect Category Detection (ACD) e l’Aspect Category Polarity (ACP). In totale, sono state presentate 20 soluzioni di 7 team composti in totale da 11 singoli partecipanti. Il miglior sistema ha ottenuto un punteggio di micro F1 di 0,810 per ACD e 0,767 per ACP.*

1 Introduction

In recent years, many websites started offering a high level interaction with users, who are no more a passive audience, but can actively produce new content. For instance, platforms like Amazon[1] or TripAdvisor[2] allow people to express their opinions on products, such as food, electronic items, clothes, and services, such as hotels and restaurants.

In such a social context, *Sentiment Analysis* (SA) is the task of automatically extract subjective opinions from a text. In its most basic form, a SA system takes in input a text written in natural language and assign it a label indicating whether the text is expressing a positive or negative sentiment, or neither (neutral, or objective, text). However, reviews are often quite detailed in expressing the reviewer’s opinion on several aspects of the target entity. *Aspect-based Sentiment Analysis* (ABSA) is an evolution of Sentiment Analysis that aims at capturing the aspect-level opinions expressed in natural language texts (Liu, 2007).

At the international level, ABSA was introduced as a shared task at SemEval, the most prominent evaluation campaign in the Natural Language Processing field, in 2014 (SE-ABSA14), providing a benchmark dataset of reviews in English (Pontiki et al., 2014). Datasets of computer laptops and restaurant reviews were annotated with aspect terms (both fine-grained, e.g. ”hard disk”, ”pizza”, and coarse-grained, e.g., ”food”) and their polarity (positive or negative).

The task was repeated in SemEval 2015 (SE-ABSA15) and 2016 (SE-ABSA16), aiming to facilitate more in-depth research by providing a new ABSA framework to investigate the relations between the identified constituents of the expressed opinions and growing up to include languages other than English and different domains (Pontiki et al., 2015; Pontiki et al., 2016).

ABSITA (Aspect-based Sentiment Analysis on Italian) aims at providing a similar evaluation with respect to texts in Italian. In a nutshell, partic-

[1] http://www.amazon.com

[2] http://www.tripadvisor.com

ipants are asked to detect within sentences (expressing opinions about accommodation services) some of the aspects considered by the writer. These aspects belongs to a close set ranging from the cleanliness of the room to the price of the accommodation. Moreover, for each detected aspect, participants are asked to detect a specific polarity class, expressing appreciation or criticism towards it.

During the organization of the task, we collected a dataset composed of more than 9,000 sentences and we annotated them with aspects and polarity labels. During the task, 20 runs were submitted by 7 teams comprising 11 individual participants.

In the rest of the paper Section 2 provides a detailed definition of the task. Section 3 describes the dataset made available in the evaluation campaign, while Section 4 reports the official evaluation measures. In Section 5 and 6, the results obtained by the participants are reported and discussed, respectively. Finally, Section 7 derives the conclusions.

2 Definition of the task

In ABSITA, Aspect-based Sentiment Analysis is decomposed as a cascade of two subtasks: **Aspect Category Detection** (ACD) and **Aspect Category Polarity** (ACP). For example, let us consider the sentence describing an hotel:

I servizi igienici sono puliti e il personale cordiale e disponibile. (*Toilets are clean but the staff is not friendly nor helpful.*)

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 2. 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 participant systems are 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.

For the ACP task, the input is the review text paired with the set of aspects identified in the text within 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.

The participants could choose to submit only the results of the ACD subtask, or both tasks. In the latter case, the output of the ACD task is used as input for the ACP. As a constraint on the results submitted for the ACP task, the polarity of an aspect for a given sentence can be different than (`POS:false`, `NEG:false`) only if the aspect is detected in the ACD step.

3 Dataset

The data source chosen for creating the ABSITA datasets is the popular website *booking.com*[3]. The platform allows users to share their opinions about hotels visited through a positive/negative textual review and a fine-grain rating system that can be used for assigning a score to each different aspect: cleanliness, comfort, facilities, staff, value for money, free/paid WiFi, location. Therefore, the website provides a large number of reviews in many languages.

We extracted the textual reviews in Italian, labeled on the website with one of the eighth considered aspects. The dataset contains reviews left by users for hotels situated in several main Italian cities such as Rome, Milan, Naples, Turin, Bari, and more. We split the reviews into groups of sentences which describe the positive and the negative characteristics of the selected hotel. The reviews have been collected between the 16th and

[3] `https://www.booking.com`

Sentence	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 disponibilitá del personale, e la struttura della stanza	0	1	1	0

Table 1: Examples of categories detection ACD.

Sentence	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.

the 17th of April 2018 using Scrapy[4], a Python web crawler. We collect in total 4,121 distinct reviews in Italian language.

The reviews have been manually checked to verify the annotation of the aspects provided by booking.com, and to add missing links between sentences and aspects. We started by annotating a small portion of the whole dataset split by sentences (250 randomly chosen sentences) using four annotators (the task organizers) in order to check the agreement of the annotation. For the ACD task, we asked the annotators to answer straightforward questions in the form of "Is aspect X mentioned in the sentence Y?" (Tab. 1).

The set of italian aspects is the direct translation of those booking.com: PULIZIA (cleanliness), COMFORT, SERVIZI (amenities), STAFF, QUALITA-PREZZO (value), WIFI (wireless Internet connection) and POSIZIONE (location). Similarly, for the ACP subtask, the annotation is performed at sentence level, but with the set of aspects already provided by the ACD annotation, and checkboxes to indicate positive and negative polarity of each aspect (Tab. 2). The result of the pilot annotation has been used to compute an inter-annotator agreement measure, in order to understand if it was possible to allow annotators to work independently each other on a different set of sentences. We found agreement ranging from 82.8% to 100% with an average value of 94.4% obtained counting the number of sentences annotated with the same label by all the annotators.

In order to complete the annotation, we assigned different 1,000 reviews to each annotator (about 2,500 sentences on average). We split the dataset among the annotators so that each of them received a uniformly balanced distribution of positive and negative aspects, based on the scores provided by the original review platform. Incomplete, irrelevant, and incomprehensible sentences have been discarded from the dataset during the annotation. At the end of the annotation process, we obtained the gold standard dataset with the associations among sentence, sentiment and aspect. The entire annotation process took a few weeks to complete. The positive and negative polarities are annotated independently, thus for each aspect the four sentiment combination discussed in Section 2 are possible: *positive*, *negative* , *neutral* and *mixed*. The resulting classes are: *cleanliness_positive, cleanliness_negative, comfort_positive, comfort_negative, amenities_positive, amenities_negative, staff_positive, staff_negative, value_positive, value_negative, wifi_positive, wifi_negative, location_positive, location_negative, other_positive, other_negative*. For each aspect, the sentiment is encoded in two classes:

- *negative* = (**_positive = 0, *_negative = 1*)
- *positive* = (**_positive = 1, *_negative = 0*)
- *neutral* = (**_positive = 0, *_negative = 0*)
- *mixed* = (**_positive = 1, *_negative = 1*)

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 is not part of the evaluation of results provided for the task.

We released the data in Comma-separated Value format (CSV) with UTF-8 encoding and semicolon as separator. The first attribute is the id of the review. Note that in booking.com the order of positive and negative sentences is strictly defined and this can make too easy the task. To overcome

[4] https://scrapy.org

Dataset	Description	#Sentences
Trial set	Trial dataset containing a small set of features used for checking the format of the file format	*30* *0.34% of Total*
Training set	The dataset contains sentences provided for training. They have been selected using a random stratification of the whole dataset.	*6,337* *69.75% of Total*
Test set	The dataset contains sentences provided for testing. They contains sentences without the annotations of aspects.	*2,718* *29.91% of Total*

Table 3: List of datasets released for the ABSITA task at EVALITA 2018.

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

Table 4: Distribution of the sentences in the datasets among the aspects and polarities.

this issue, we randomly assign for each sentence a new position in the review. As a consequence, the final positional id showed in the data file do not reflect the real order of the sentences in the review. The text of the sentence is provided at the end of the line and delimited by ”. It is preceded by three binary values for each aspect indicating respectively: the presence in the sentence ($aspectX\_presence$:0/1), the positive polarity for that aspect ($aspectX\_pos$:0/1) and finally the negative polarity ($aspectX\_neg$:0/1). Fig. 1 shows an example of the annotated dataset in the proposed format.

The list of the datasets released for the task is provided in Tab. 3 and the distribution of the sentences among aspects and polarity is provided in Tab. 4. The subdivision adopted for it is respectively 0.34%, 69.75%, 29,91% for trial, training and test data. The datasets can be freely downloaded from `http://sag.art.uniroma2.it/absita/` and reused in non-commercial projects and researches. After the submission deadline, we also distributed the gold standard test set and evaluation script.

4 Evaluation measures 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: $F1_a = \frac{2P_aR_a}{P_a+R_a}$, 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 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 $F1_p = \frac{2P_pR_p}{P_p+R_p}$. 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 $|S_p \cap G_p| = 1$, $|G_p| = 2$ 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 is computed by considering a system which assigns the most fre-

```
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 ..."
```

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

System	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	0.6970	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	0.4330	0.5104
mfc baseline	0.4111	0.2866	0.3377

Table 5: Results of the submissions for the ACD subtask.

System	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

Table 6: Results of the submissions for the ACP subtask.

quent $(aspect, polarity)$ pair (estimated over the training set) to each sentence.

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

5 Results

We received submissions from several teams that participated in past editions of EVALITA, in particular to the SENTIPOLC (Sentiment Polarity Classification (Barbieri et al., 2016)) and NEEL-it (Named Entity Recognition (Basile et al., 2016)), but also some new entries in the community. In total, 20 runs were submitted by 7 teams comprising 11 individual participants. The task allowed participant teams to send up to 2 submissions from each team. In particular, 12 runs were submitted to ACD task and 8 runs to the ACP task.

We also provide the result of a baseline system that assigns to each instance the most frequent class in each task, i.e., the aspect (COMFORT) and polarity (*positive*) for that aspect, according to the frequency of classes in the training set. The results of the submissions for the two tasks, and the baseline (namely mfc baseline), are reported in Tab. 5 and Tab. 6. Of the seven teams who participated to the ACD task, five teams also participated to the ACP task.

The results obtained by the teams largely outperform the baseline demonstrating the efficacy of the solutions proposed and the affordability of all two tasks. The results obtained for the ACD task (Tab. 5) show a small range of variability, at least in the first part of the ranking (the top results are concentrated around a F1 score value of 0.80). On the contrary, the values of precision and recall show higher variability, indicating significant difference among the proposed approaches.

6 Discussion

The teams of the ABSITA challenge have been invited to describe their solution in a technical report and to fill in a questionnaire, in order to gain an insight on their approaches and to support their replicability. Five systems (*ItaliaNLP*, *gw2017*, *X2Check*, *UNIPV*, *SeleneBianco*) are based on supervised machine learning, that is, all the systems for which we have access to the implementation details, with the exception of *VENSES*, which is a rule-based unsupervised system. Among the system that use supervised approaches, three systems (*ItaliaNLP*, *gw2017*, *UNIPV*) employ deep learning (in particular LTSM networks, often in their bi-directional variant).

All runs submitted can be considered ”constrained runs”, that is, the systems were trained on the provided data set only.

Besides additional training data, some systems employ different kind of external resources. Among these, pre-trained word embeddings are used as word representations by *UNIPV* (Fasttext[5]) and *gw2017* (word embeddings provided by the SpaCy framework[6]). The system of *ItaliaNLP* employs word embedding created from the ItWaC corpus (Baroni et al., 2009) and corpus extracted from Booking.com.

Some of the systems are ABSA extensions built on top of custom or pre-existing NLP pipelines. This is the case for *ItaliaNLP*, *VENSES* and *X2Check*. Other systems make use of off-the-shelf NLP tools for preprocessing the data, such as SpaCy (*gw2017*, *UNIPV*) and Freeling[7] (*SeleneBianco*).

Finally, additional resources used by the systems often include domain-specific or affective lexicons. *ItaliaNLP* employed the MPQA affective lexicon (Wilson et al., 2005), and further developed an affective lexicon from a large corpus of tweets by distant supervision. The *UNIPV* system makes use of the affective lexicon for Italian developed in the framework of the OpeNER project[8].

In the ACD task, the precision of the second ranked system (*gw2017*) is significantly higher than that of the first system (*ItaliaNLP*), although the latter ranks at the top because of a higher recall. This unbalance between precision and recall is mainly due to the high number of aspect that can be assigned at the same time to a sentence: a system returning too many aspects is exposed to low precision but higher recall, while a more conservative system would achieve the opposite situation. Further details about the systems developed for the task can be found in the technical reports of the partecipants: ItaliaNLP (Cimino et al., 2018), *UNIPV* (Nicola, 2018), *VENSES* (Delmonte, 2018), *X2Check* (Di Rosa and Durante, 2018), *gw2017* (Bennici and Portocarrero, 2018)

7 Conclusion

The large availability of user-generated contents over the Web that characterizes the current tendencies of virtually sharing opinions with others has promoted the diffusion of platforms able to analyze and reuse them for personalized services. A challenging task is the analysis of the users' opinions about a product, service or topic of discussion. In particular, the ABSA (Aspect-based Sentiment Analysis) task concerns the association of a polarity (positive, negative, neutral/objective) to the piece of the sentence that refers to an aspect of interest. In ABSITA, we proposes to automatically extract users' opinions about aspects in hotel rewievs. The complexity of the task has been successfully faced by the solutions submitted to the task. Systems that used supervised machine learning approaches, based on semantic and morphosyntactic features representation of textual contents, demonstrate encouraging performances in the task. 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. The decision to use additional resources as additional lexicons in conjunction with semantic word embeddings have been demonstrated to be successful. More details about the implementation of the systems that participated in the task can be found in their specific reports. In conclusion, we consider the ABSITA 2018 task a success and an improvement of state of the art for the ABSA task in the Italian language.

References

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the Evalita 2016 SENTIment POLarity Classification Task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, Naples, Italy, December.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43:209–226.

P. Basile, A. Caputo, A.L. Gentile, and G. Rizzo. 2016. Overview of the evalita 2016 named entity recognition and linking in italian tweets (neel-it) task. In *5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*, Napoli, Italia, 12/2016.

Mauro Bennici and Xileny Seijas Portocarrero. 2018. Ensemble for aspect-based sentiment analysis. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

[5] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

[6] https://spacy.io/

[7] http://nlp.lsi.upc.edu/freeling/node/

[8] https://github.com/opener-project/\\VU-sentiment-lexicon

Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. Evalita 2018: Overview of the 6th evaluation campaign of natural language processing and speech tools for italian. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. 2018. Multi-task Learning in Deep Neural Networks at EVALITA 2018. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Rodolfo Delmonte. 2018. Itvenses - a symbolic system for aspect-based sentiment analysis. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Emanuele Di Rosa and Alberto Durante. 2018. Aspect-based sentiment analysis: X2check at absita 2018. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Bing Liu. 2007. *Web data mining*. Springer.

Giancarlo Nicola. 2018. Bidirectional attentional lstm for aspect based sentiment analysis on italian. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

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, August. Association for Computational Linguistics and Dublin City University.

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, June. Association for Computational Linguistics.

Maria Pontiki, Dimitrios 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, Evgeny 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 '16, San Diego, California, June. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.