ATE_ABSITA @ EVALITA2020: Overview of the Aspect Term Extraction and Aspect-based Sentiment Analysis Task

Lorenzo De Mattei
University of Pisa
Ist. di Ling. Comp.
"Antonio Zampolli"
Pisa ItaliaNLP Lab
lorenzo.demattei@di.unipi.it

Graziella De Martino
University of Bari A. Moro
Dept. Computer Science
E. Orabona 4, Bari
graziella.demartino@uniba.it

Andrea Iovine
University of Bari A. Moro
Dept. Computer Science
E. Orabona 4, Bari
SWAP Research Group
andrea.iovine@uniba.it

Alessio Miaschi
University of Pisa
Ist. di Ling. Comp.
"Antonio Zampolli"
Pisa ItaliaNLP Lab
alessio.miaschi@phd.unipi.it

Marco Polignano
University of Bari A. Moro
Dept. Computer Science
E. Orabona 4, Bari
SWAP Research Group
marco.polignano@uniba.it

Giulia Rambelli
University of Pisa
Coling Lab Pisa
Aix-Marseille University
giulia.rambelli@phd.unipi.it

Abstract

Over the last years, the rise of novel sentiment analysis techniques to assess aspect-based opinions on product reviews has become a key component for providing valuable insights to both consumers and businesses. To this extent, we propose ATE_ABSITA: the EVALITA 2020 shared task on Aspect Term Extraction and Aspect-Based Sentiment Analysis. In particular, we approach the task as a cascade of three subtasks: Aspect Term Extraction (ATE), Aspect-based Sentiment Analysis (ABSA) and Sentiment Analysis (SA). Therefore, we invited participants to submit systems designed to automatically identify the "aspect term" in each review and to predict the sentiment expressed for each aspect, along with the sentiment of the entire review. The task received broad interest, with 27 teams registered and more than 45 participants. However, only three teams submitted their working systems. The results obtained underline the task's difficulty, but they also show how it is possible to deal with it using innovative approaches and models. Indeed, two of them are based on large pre-trained language models as typical in the current state of the art for the English language. (de Mattei et al., 2020)

1 Introduction and motivation

Leaving comments and reviews on the Web has become a common practice for users to express their opinions about products, experiences, and more. Thus, companies need to deal with increasingly large amounts of textual data, which can be useful to identify their products' strengths and weaknesses. However, the automatic analysis of reviews poses numerous problems related to its processing. First of all, reviewers often use informal language, with a wide variety of colloquialisms and contractions, which make review analysis through lexicon-based techniques difficult. Second, automatically identifying aspects of the product within a sentence is not easy, due to the intrinsic subjectivity in the definition of "aspect". These issues have already been addressed in the area of Text Mining and Sentiment Analysis. Recently, the sentiment analysis and opinion mining tasks have seen a surge in interest, thanks to the large quantity of data available for analysis and the new natural language processing techniques based on language models such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2019). Thus, we proposed the ATE_ABSITA: the EVALITA 2020 (Basile et al., 2020) shared task on Aspect Term Extraction and Aspect-Based Sentiment Analysis.

Sentiment Analysis (or *Opinion Mining*) is the task of identifying what the user thinks about a particular element. It often takes the form of a classification task with the purpose of annotating a portion of text with a positive, negative, or neutral label. *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). Very often, the ABSA task is performed on a set of aspects defined a priori, limiting its applicability in the real scenario. *Aspect Term Extraction* (ATE) is the task of identifying "aspect term" in a text without knowing a priori the list that contains it. According to the literature definition, a term/phrase is considered as an aspect when it co-occurs with some "opinion words" that indicate a sentiment polarity on it (Pontiki et al., 2016a).

At the international level, SemEval, the most prominent evaluation campaign in the Natural Language Processing field, provided in 2014 SE-ABSA14 (Pontiki et al., 2014) a benchmark dataset of reviews in the English language for the ABSA task. Given a set of sentences with pre-identified entities (e.g., *restaurants*), the task was about identifying the aspect term occurring in the sentences and returning a list containing all the distinct aspect term. It was then asked for all retrieved aspect term to determine whether the polarity of each of them was positive, negative, neutral, or conflict. The same task was replicated in 2015, 2016, consolidating the four subtasks of SE-ABSA14 (Pontiki et al., 2014) within a unified framework. Besides, SE-ABSA15 (Pontiki et al., 2016b) included an out-of-domain ABSA subtask, involving test data from a domain unknown to the participants.

ABSA is not a novel task at EVALITA. A first edition was proposed at EVALITA 2018 by (Basile et al., 2018). The task was subdivided into two subtasks: Aspect Category Detection (ACD) and Aspect Category Polarity (ACP). The first was about the identification of categories mentioned into the review, knowing the categories a priori. The latter was about the detection of the polarity of the opinion of the user about the previous detected categories. However, it bears some similarities with at least other two tasks from the previous editions of the campaign. SENTIPOLC (Basile et al., 2014), featured in the 2014 and 2016 editions of EVALITA, is a shared task on the polarity classification of social media content. The other is NEEL-it (Basile et al., 2016), held at EVALITA 2016. NEEL-it is the task of Named Entity Recognition and Linking, that is, the task of identifying the spans of an input text that refer to named entities, and linking them to entries in a knowledge base, e.g., pages of Wikipedia.

aspect term	Positive	Negative
mantenere la temperatura	1	0
costruzione	1	0

Table 1: Examples of Aspect-Based Sentiment Analysis annotations.

2 Definition of the Task

We define the ATE_ABSITA task as a cascade of three subtasks: **Aspect Term Extraction** (ATE), **Aspect-based Sentiment Analysis** (ABSA), **Sentiment Analysis** (SA).

For example, let us consider the sentence describing a review of a metallic bottle:

La borraccia termica svolge egregiamente il proprio compito di mantenere la temperatura, calda o fredda che sia. La costruzione è ottimale e ben rifinita. Acquisto straconsigliato!

The thermal water bottle does its job very well to keep the temperature, whether hot or cold. The construction is optimal and well finished. Purchase highly recommended!

In the **Aspect Term Extraction** (ATE) task, one or more "aspect term" mentioned in a sentence are identified, e.g. `mantenere la temperatura` (`keep the temperature`) and `costruzione` (`construction`) in the sentence. Given a sequence $X = x, ..., x_T$ of T words, the ATE task can be formulated as a token/word level sequence labeling problem to predict an aspect label sequence $Y = y_1, ..., y_T$, where each y_i comes from a finite label set $Y = B, I, O$ which describes the possible aspect labels (begin, inside, outside). An example of ATE annotation is provided in Fig. 1.

In the **Aspect-based Sentiment Analysis** (ABSA) task, the polarity of each expressed aspect is recognized, e.g. a `positive` category polarity is expressed concerning the `mantenere la temperatura` aspect. 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 *mixed polarity* (POS:true, NEG:true), or *neutral* polarity (POS:false, NEG:false). An example ot ABSA annotation is showed in Tab. 1.

In the **Sentiment Analysis** (SA) task, the polarity of the review is provided. In particular, we

La	borraccia	termica	svolge	egregiamente	il	proprio
O	O	O	O	O	O	O

compito	di	mantenere	la	temperatura,	*calda*	o
O	O	**B**	**I**	**I**	O	O

fretta	che	sia.	La	costruzione	è	ottimale
O	O	O	O	B	O	O

e	ben	rifinita.	Acquisto	straconsigliato!
O	O	O	B	O

Figure 1: Result of the ATE annotation.

decided to use the score left by the user at the item as the polarity value. It is defined as an integer number within the $[1..5]$ range. An example is provided in Tab. 2.

Review	**Score**
La borraccia termica svolge egregiamente il proprio compito di mantenere la temperatura,calda o fretta che sia. La costruzione è ottimale e ben rifinita. Acquisto straconsigliato!	5

Table 2: Example of Sentiment Analysis polarity annotation on the whole sentence.

In the ATE task here described, the set of aspects is not defined in advance, and the task itself is formalized as a Sequence Labeling task. The ABSA task can, instead, be formalized as a multi-class classification task. Finally, the Sentiment Analysis is considered as a regression task. For each review, participants will be asked to return a vector of aspects, a vector of aspect:polarity pairs, and a review:score pair. Two binary polarity labels are expected for each aspect: POS and NEG, indicating a positive and negative sentiment expressed towards a specific aspect, respectively. The participants may choose to submit each of the three subtasks independently.

3 Dataset

The data source chosen for creating the datasets is an eCommerce platform famous worldwide. The platform allows users to share their opinions about the items that they bought through a textual review and a final score of satisfaction. Therefore, the website provides a large number of reviews in many languages, including Italian (Fig. 2). We have collected 4364 real user reviews, written in the Italian language, involving 23 products. The training, dev and test sets will be randomly generated in the following ratios: 70% training, 2.5% dev, 27.5% test set. This means that the test set will be not out-of-domain. The items cover very different domains of use. In particular, the existing objects refer to: SD Memory Cards, Irons, Water Bottles, Action Cameras, Razors, Phones, Printer Cartridges, Coffee Capsules, Backpacks, Hair Dryers, 2 different Movies, 2 different Books, Toy Phones, Car Light bulbs, Sweatshirts, Boots, Fans, Storage Chest, Shoe Cabinets, Personal Digital Assistants, TV streaming boxes/sticks. A portion of the collected data has been **manually annotated** by three different subjects. Then, we measured the inter-annotator agreement metric as the value of quality of all the annotations. In particular, we obtained a score of 73.2% over 100 reviews. Thanks to the good score, we decided to continue the annotation process by annotating each review individually (i.e. one annotator per review). At the end of the annotation process, we obtained the gold annotated dataset. We randomly split the gold dataset to create a training/validation/test partition of it.

We do not provide any unique ID that could be used to retrieve more information about the writers. Consequently, we do not violate copyrights and/or we do not have privacy issues. Furthermore, in order to avoid harming the interests of the manufacturers, we do not disclose any information about the specific items for which the reviews have been issued.

The data format used is NDJSON [1] with UTF-8 encoding and newline as delimiter. Note that some reviews may not contain any aspect, but the final review score is always available. An example of

[1] http://ndjson.org/

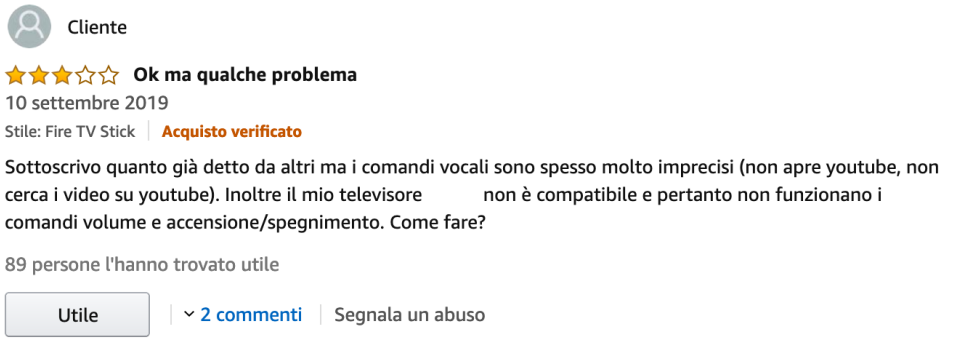

Figure 2: Example of a review about a TV streaming box/stick.


```
{"sentence":"L'attore...e le musiche indimenticabili", "id_sentence":"4c0b","score":5,
"polarities":[[0,0],[1,0]], "aspects_position":[[2,8],[16,23]], "aspects":["attore","musiche"]}
{"sentence":"Schermo guasto dopo appena due settimane,...","id_sentence":"4e1671","score":1,
"polarities":[[0,1]],"aspects_position":[[0,7]],"aspects":["Schermo"]}
{"sentence":"Ottimo telefono belle foto","id_sentence":"4eca9d08","score":4,"polarities":[[1,0]],
"aspects_position":[[22,26]],"aspects":["foto"]}
```


Figure 3: Example of NDJSON dataset records.

annotated data is provided in the code reported in Fig. 3.

4 Annotation Schema

This section describes the protocol that will be used to annotate the datasets for the three subtasks. The objective of this protocol is to get a reasonably objective definition of the characteristics of an aspect term. Due to the highly subjective nature of aspects, it does not encompass all conceivable aspect term. We define an **aspect term** as:

(a) An **attribute** (characteristic, property, feature, quality) of the object itself; (b) a tangible or abstract **part of the object**, for which an opinion can be inferred from the review; (c) the **activities** that the object is able (or not able) to perform; (d) the object's **ability to be suitable** for certain categories of people.

Judgment can be assigned in three ways: 1. *Directly*: the aspect term occurs with an opinion term (i.e., "la **durata della batteria** è ottima"); 2. *Indirectly*: the judgment about the product is transitive to a quality or part of the object. In other words, if an opinion is expressed about the object itself, and it is then stated for which characteristic the judgment is applied, these characteristics are annotated as an aspect term (i.e., "questo telefono è ottimo, soprattutto per la **durata della batteria**"); 3. *Deductible*: the opinion is not expressed directly but it is inferable from the review or from the knowledge of the reference domain.

The aspect term must represent the product characteristics, but it cannot represent a concept that is larger than the product itself. An aspect term **does not identify opinions** regarding elements external to the object, such as: (a) The shipment (it is not an intrinsic property of the object); (b) the company that produced them, the series to which the product belongs or other products with which the object is compared; (c) the elements that refer to the action of purchasing the item; (d) the elements that refer to the customer care. Moreover, in the case of aspect term composed of several words, all the words that make up the aspect term must be contiguous. In case they are separated by one or more words that are not part of the aspect term, the whole expression is discarded. More details and example of annotations are available on the task website[2].

5 Evaluation measures and baselines

We evaluate the three subtasks (ATE, ABSA and SA) separately by comparing the results obtained by the participant systems on the gold standard annotations of the test set.

For the ATE task, we compute Precision, Recall

[2] http://www.di.uniba.it/~swap/ate_absita/examples.html

and F_1-score defined as:

$$F1_a = \frac{2P_aR_a}{P_a + R_a} \quad (1)$$

In order to account for both exact and partial matches of aspect term, we define Precision (P_a) and Recall (R_a) as:

$$P_a = \frac{|S_a \cap G_a| + 0.5 * |PAR_a|}{|S_a|}$$
$$R_a = \frac{|S_a \cap G_a| + 0.5 * |PAR_a|}{|G_a|} \quad (2)$$

Here, S_a is the set of aspect term annotations that a system returned for all the test sentences, G_a is the set of the gold (correct) aspect term annotations and PAR_a is the set of partial matches (predicted and gold aspect term have some overlapping text). For instance, if a review is labeled in the gold standard with the two aspect term $G_a = \{$*costruzione*, *mantenere la temperatura*$\}$, and the system predicts the two aspects $S_a = \{$*costruzione*, *temperatura*$\}$, we have that $|S_a \cap G_a| = 1$, $|PAR_a| = 1$, $|G_a| = 2$ and $|S_a| = 2$, so that $P_a = \frac{1.5}{2} = 0.75$, $R_a = \frac{1.5}{2} = 0.75$ and $F1_a = \frac{1.5}{2} = 0.75$. For the ATE task, we considered a simple baseline approach which considers every name entity as an aspect term. The algorithm is based on a Name Entity Recognition (NER) annotation obtained through the SpaCy [3] tool on the Italian model 'it_core_news_sm'. The implementation of the baseline on the training set is available as a Python3 Notebook on our website.

For the ABSA task (Task 2), we evaluate the entire chain, thus considering both the aspect term detected in the sentences together with their corresponding polarities, in the form of $(aspect, polarity)$ pairs. We again compute Precision (P_p), Recall (R_p) and F_1-score ($F1_p$) defined as following:

$$F1_p = \frac{2P_pR_p}{P_p + R_p} \quad (3)$$

$$P_p = \frac{|S_p \cap G_p| + 0.5 * |PAR_p|}{|S_p|}$$
$$R_p = \frac{|S_p \cap G_p| + 0.5 * |PAR_p|}{|G_p|} \quad (4)$$

Where S_p is the set of $(aspect, polarity)$ pairs that a system returned for all the test sentences, G_a is the set of the gold (correct) pairs annotations and PAR_p is the set of $(aspect, polarity)$ pairs with a partial match. For instance, if a review is labeled in the gold standard with the pairs:
$G_p = \{$(*mantenere la temperatura*, POS), (*costruzione*, POS)$\}$,
and the system predicts the three pairs
$S_p = \{$(*temperatura*, NEG), (*costruzione*, POS), (*acquisto*, POS)$\}$,
we have that $|S_p \cap G_p| = 1$, $|PAR_p| = 0$, $|G_p| = 2$ and $|S_p| = 3$ so that $P_p = \frac{1}{3}$, $R_p = \frac{1}{2}$ and $F1_p = 0.4$. As a baseline for the ABSA task, we decided to assign the most frequent polarity class (i.e. the positive one) to each aspect found by the baseline strategy for Task 1.

To evaluate the SA task (Task 3), we compute the Root Mean Squared Error ($RMSE_w$) between the scores predicted by the participant systems and those found in the gold dataset. For this task, we employed three different baselines. The first predicts the most frequent value in the training set: 5. The second predicts the average value of the scores found on the training set (4.46299). The third one uses AlBERTo (Polignano et al., 2019) as an approach to develop a Regression task.

6 Task statistics

The task has generated great interest in the scientific community. We obtained 27 registered teams, for a total of 45 separate participants. Nevertheless, the difficulty of the task discouraged many of them. At the end of the evaluation phase, we obtained 8 submissions from 3 different teams.

7 Submitted systems

The three teams participating in the task are the following:

- **A2C** (Rosa and Durante, 2020): the team is composed of two members of the App2Check company, who developed a classification model based on state-of-the-art language models. In particular, they investigate the ATE task through the use of four different configurations of language models: 1. Native Italian pre-trained language models, with no specific NER fine-tuning and 3. with NER fine-tuning; 2. Multilingual pre-trained language model, with no specific NER fine-tuning and 4. with NER fine-tuning. For the first and the third configuration, they considered dbmdz/bert-base-

[3] https://spacy.io/

italian-xxl-uncased[4] and GilBERTo[5]. For the second configuration, they considered two implementations of RoBERTa: xml-roberta-large3 (Conneau et al., 2019), xml- roberta-base4 (Liu et al., 2019), and multilingual BERT (Pires et al., 2019). The xlm RoBERTa Large multilingual model was chosen as the competition model. The ABSA task has been performed by fine-tuning a multilingual BERT model in order to assign the polarity label to each portion of text that contains at least one previously detected aspect. Similarly, the SA task has been approached using a multilingual BERT model on a 1 to 5 sentiment scale. The system submitted by the A2C team obtained the best results overall.

- **SentNa** (Francesco Mele and Vettigli, 2020): the authors proposed a hybrid model that joins rule-based and machine learning methodologies in order to combine their respective advantages. The main idea for dealing with the ATE task is to identify a set of plausible aspects via some predefined rules. Then, a classifier is used to filter out the wrong candidates. The rules are defined on POS-Tagging patterns. The authors defined a set of about 3000 rules. The sentiment analysis problem has been solved by building the features representing the text using n-grams, and adding a set of features annotated in SenticNet (Cambria et al., 2010). Then, a regressor composed of 800 Decision Trees with 4 layers has been trained using Gradient Boosting. The final prediction is computed by averaging the output of each tree.

- **ghostwriter19** (Bennici, 2020): the team composed of one member of the YouAre-MyGuide Company proposes a solution based on mixing transfer learning, zero-shot learning (Brown et al., 2020), and ONNX[6], in order to access the power of BERT while using limited resources. In order to deal with the ATE and ABSA tasks, the author uses the AlBERTo (Polignano et al., 2019) language model and an auto training system such as Ktrain[7] for fine-tuning the system. At this point, the model has been exported with ONNX in maximum compatibility mode with the original. The optimization options have been set to a minimum for CPU usage. The performances have remained unchanged, but the speed of inference has significantly improved. For the sentiment analysis task, the author uses a zero-shot learning strategy as a way to make predictions without prior training. In particular, he reuses the embedding of AlBERTo for encoding the sentence and a Bi-LSTM as classification model to predicting a class from 1 to 5.

[4] https://github.com/dbmdz/berts

[5] https://github.com/idb-ita/GilBERTo

[6] https://microsoft.github.io/onnxruntime/

Table 3: Final results obtained by the participants for the ATE sub-task.

Pos.	Team Name	F1 score
1	**A2C**	**0.68222**
2	ghostwriter19	0.53986
3	SentNa	0.34027
4	*Baseline-Name Entities*	*0.2556*

Table 4: Final results obtained by the participants for the ABSA sub-task.

Pos.	Team Name	F1 score
1	**A2C**	**0.61878**
2	ghostwriter19	0.49935
3	SentNa	0.28632
4	*Baseline-Positive pol.*	*0.20000*

8 Discussion of results

The results in tables from 3-5 show the optimal performances of the system developed by the **A2C** team, which obtained first place in all three sub-tasks. The use of pre-trained language models has proven to be the winning strategy. In particular, the differences between the results of **A2C** and **ghostwriter19** show how a large RoBERTa model can strongly outperform a smaller language model such as AlBERTo, even though the latter has been specifically trained on the Italian language. This result was expected, since the ALBERTo baseline also obtained low results. We hypothesize that the difference in style between the tweets that were used to train ALBERTo and the reviews contained in this dataset are a significant factor in the low applicability of this model. Additionally, the results obtained by the A2C system also show that pre-

[7] https://github.com/amaiya/ktrain

Table 5: Final results obtained by the participants for the SA sub-task.

Pos.	Team Name	RMSE
1	**A2C**	**0.66458**
2	SentNa	0.79533
3	ghostwriter19	0.81394
4	*Baseline-Average Score*	*1.00409*
5	*Baseline-AlBERTo*	*1.08063*
6	*Baseline-Most Freq.*	*1.12822*

training the language model for the Named Entity Recognition (NER) task is also useful for identifying aspect term. This is due to the fact that aspect term share some properties with named entities. For example, they are often configured as a noun, an adjective, or a combination of both.

The results obtained by **SentNa** are also interesting. Their model, which is based on decision trees, has obtained a good final score for the SA task. This confirms the findings obtained in earlier Sentiment Analysis tasks in Italian campaigns such as EVALITA, which already demonstrated that techniques such as Decision Trees, Random Forests, and SVD can be effective solutions to this task. Nevertheless, the **SentNa** system demonstrates that an enriched encoding of the sentences, including lexical features such as polarity value, attention, pleasantness, and sensitivity of its composing n-grams, can support a more accurate prediction of the whole sentence polarity.

9 Conclusion

In the ATE_ABSITA task at EVALITA 2020, we focused the attention of research groups that work on computational linguistics for the Italian language on the problem of analyzing user reviews. Specifically, we subdivided the problem into three parts: Aspect Term Extraction (ATE), Aspect-Based Sentiment Analysis (ABSA), Sentence Sentiment Analysis (SA). In the ATE task, the goal was to identify one or more "aspect term" discussed in the review. The second task was about identifying the sentiment evoked by the user while talking about a specific aspect (ABSA). Finally, we asked participants to identify the polarity associated with the entire review (SA). The dataset we released has been collected from a world-famous eCommerce platform. In particular, we extracted and **manually annotated** 4364 real user reviews, written in the Italian language, about 23 different products. Although the results obtained by the systems that participated in the task are very close to those available in the English language literature, the F1 scores for the ATE and ABSA subtasks demonstrate its complexity. It is evident that an F1 score of about 0.60 generates a non-negligible margin of error of prediction. The diversity in terms, linguistic expressions, and in the physical characteristics of the products themselves makes the automatic extraction of "aspect term" a task that is far from being resolved. This complexity can also explain the low number of participants. It is easy to see a substantial discrepancy between the number of people enrolled in the task and those who have proposed a solution for it. In our opinion, this is caused by the difficulty in addressing the problem with the current natural language analysis techniques. However, this also means that there is still a wide margin for improvement in this area, and that this problem can be addressed again in the next edition of EVALITA. We firmly believe that extracting fine-grained opinions from user reviews can be a great asset for improving products, processes, and software systems.

References

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the evalita 2014 sentiment polarity classification task.

Pierpaolo Basile, Annalina Caputo, Anna Lisa Gentile, and Giuseppe Rizzo. 2016. Overview of the evalita 2016 named entity recognition and linking in italian tweets (neel-it) task. In *of the Final Workshop 7 December 2016, Naples*, page 40.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Pierpaolo Basile et al. 2018. O verview of the evalita 2018 aspect-based sentiment analysis task (absita). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:10.

Mauro Bennici. 2020. ghostwriter19 @ ATE_ABSITA: Zero-Shot and ONNX to speed up BERT on sentiment analysis tasks at EVALITA 2020. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. 2010. Senticnet: A publicly available semantic resource for opinion mining. In *AAAI fall symposium: commonsense knowledge*, volume 10. Citeseer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Lorenzo de Mattei, Graziella de Martino, Andrea Iovine, Alessio Miaschi, Marco Polignano, and Giulia Rambelli. 2020. ATE_ABSITA@EVALITA2020: Overview of the Aspect Term Extraction and Aspect-based Sentiment Analysis Task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Antonio Sorgente Francesco Mele and Giuseppe Vettigli. 2020. SentNA@ATE_ABSITA: Sentiment Analysis of customer reviews using Boosted Trees with lexical and lexicon-based features. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019). CEUR*.

Maria Pontiki et al. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation*.

Maria Pontiki et al. 2016a. Semeval-2016 task 5: Aspect based sentiment analysis. In *10th International Workshop on Semantic Evaluation (SemEval 2016)*.

Maria Pontiki et al. 2016b. 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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Emanuele Di Rosa and Alberto Durante. 2020. App2Check@ATE_ABSITA 2020: Aspect Term Extraction and Aspect-based Sentiment Analysis. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.