Author's personal copy
IEEE Transactions on Affective Computing
DOI: 10.1109/TAFFC.2018.2884015

Sentiment Polarity Classification at EVALITA: Lessons Learned and Open Challenges

Valerio Basile, Nicole Novielli, Danilo Croce, Francesco Barbieri, Malvina Nissim, Viviana Patti

***Abstract*—Sentiment analysis in social media is a popular task attracting the interest of the research community, also in recent evaluation campaigns of natural language processing tasks in several languages. We report on our experience in the organization of SENTIPOLC (SENTIment POLarity Classification Task), a shared task on sentiment classification of Italian tweets, proposed for the first time in 2014 within the Evalita evaluation campaign. We present the datasets – which include an enriched annotation scheme for dealing with the impact of figurative language on polarity – the evaluation methodology, and discuss the approaches and results of participating systems. We also offer a reflection on the open challenges of state-of-the-art systems for sentiment analysis of microblogging in Italian, as they emerge from a qualitative analysis of misclassified tweets. Finally, we provide an evaluation of the resources we have created, and share the lessons learned by running this task for two consecutive editions.**

***Index Terms*—Sentiment Analysis, Irony Detection, Social Media Analysis, Evaluation.**

I. Introduction

Sentiment Analysis (SA) on social media, namely detecting whether a message is polarised towards a positive or negative sentiment, is by now an established task of Natural Language Processing (NLP). Solid and growing interest is reflected in the surge of published articles in the area of Affective Computing [15] and in the rising popularity of SA tasks at SemEval [29], where they by now constitute a whole track, attracting the highest number of participants in the last years [36], [35], [28], [34]. Even though this popularity is also motivated by the targeted language (mostly English at SemEval), evaluation campaigns for other languages have recently attracted the attention of the research community. Examples are DEFT@TALN/RECITAL 2017 for French [10], with a special focus on sentiment analysis and figurative language, and StanceCat@Ibereval2017 for Spanish and Catalan, with a special focus on the finer grained task of stance detection [38].

A similar picture emerges from the latest editions of EVALITA[1], the evaluation initiative for language technology on Italian, where we introduced the SENTIment POLarity Classification Task (SENTIPOLC) for the first time in 2014 [8] and replicated it in 2016 [1]. Both editions registered the highest number of participating teams within EVALITA [6].

While the task attracting the largest number of participants is the classification of the *polarity* of a tweet, some related tasks are also deemed important by the community, or are recently gaining traction [4]. Among these, we find *subjectivity* detection, i.e., to detect whether a tweet is *subjective* or is merely reporting some fact, and the analysis of figurative language, including *irony*. Subjectivity, polarity, and irony detection form the three tasks of the SENTIPOLC campaigns, both in its original 2014 version, and in the 2016 rerun. In particular, the 2016 edition of SENTIPOLC featured a few innovations with respect to the original 2014 edition. These include a new annotation layer with two fields that express *literal polarity*, to provide insights into the polarity shifts in the presence of figurative language, and a test set intentionally focused on a different domain than the training set in order to test the generalization ability of the systems, in line with what observed by Basile et al. [3].

One of the key aspects of an evaluation campaign is defining how to build an annotated resource to be used as benchmark. Indeed, the evaluation of systems performing sentiment analysis, subjectivity detection, figurative language analysis, and related tasks typically involves a substantial amount of textual data paired with human judgements on their affective content. While this evaluation usually follows the steps of most evaluation task in NLP - i.e., direct comparison with a manually annotated dataset and computation of correlation metrics between systems and human judgments - evaluating sentiment and emotions is particularly challenging, as label assignment proves tough even for humans. This is demonstrated by the relatively low inter-rater agreement achieved by human judges on affective datasets [26]. Thus, interest has recently surfaced towards producing higher quality gold standard datasets in the area of affective computing [14], [18].

Besides improving the quality of gold standard annotation, collecting datasets that are adequate in size is an important factor too, especially because producing manual annotations is an expensive and time-consuming activity. Crowdsourcing platforms such as Crowdflower[2] are becoming the standard method for collecting large quantities of manually annotated data for Artificial Intelligence and Machine Learning research, for training of supervised systems, and for evaluation purposes. This is the case for sentiment analysis too, as reflected by the use of Crowdflower in the production of the gold standard data of the SA task at SemEval 2016 [28]. In SENTIPOLC 2016, a portion of the data was also annotated with crowdsourcing techniques, rather than entirely by experts as in the 2014 edition. This has led to several observations on the quality of the data, and on the theoretical description of the task itself.

In this paper we offer a retrospective on our experience in

University of Turin, Italy basile@di.unito.it
University of Bari "A. Moro", Italy nicole.novielli@uniba.it
University of Rome "Tor Vergata", Italy croce@info.uniroma2.it
Pompeu Fabra University, Spain francesco.barbieri@upf.edu
University of Groningen, The Netherlands m.nissim@rug.nl
University of Turin, Italy patti@di.unito.it

[1] http://www.evalita.it/

[2] Now Figure Eight: http://www.figure-eight.com/

the organization of SENTIPOLC, aiming at discussing how the quality and size of an annotated resource impact the results of a sentiment analysis evaluation campaign. Furthermore, we discuss the findings of a qualitative investigation of tweets misclassified by the top three scoring systems for the polarity classification task in SENTIPOLC. The contributions of this paper can be summarized as follows: (1) We report and discuss the combined results of the two evaluation campaigns of SENTIPOLC 2014 and 2016; (2) We offer a reflection on the limitations of state-of-the-art systems for sentiment analysis in Italian; (3) We provide an *a posteriori* evaluation of the resources created for the shared tasks and the methodologies for their acquisition.

With respect to goal (1), we provide a detailed analysis of the shared tasks by focusing on the approaches and paradigms adopted by the participants. In particular, after a formal description of the tasks given in Section II, we provide a summary of the results achieved in both editions in Section V, together with a comparison of the solutions presented by the participants. The results obtained from the shared task evaluation supported a deep analysis of the resources acquired so far (goal (3)).

With respect to goal (2), we leverage the results achieved by the best systems to highlight the inherent challenges of the tasks (see Section VI). Specifically, we performed an error analysis on the tweets for which the three top scoring systems from the 2016 edition of the task provided a wrong prediction, in order to identify open challenges in sentiment analysis of Italian tweets. By discussing and sharing the findings of such analysis we hope to encourage the community to address the limitations of state-of-the-art systems.

In Section VII, we present a series of empirical tests aimed at evaluating the impact of different methodologies in creating gold standard data for sentiment analysis. In particular, we considered the interaction between crowd and expert annotation (described in detail in Section III-B), and evaluated its impact on the quality of the gold-standard. Finally, in Section VIII, we reflect upon the experience of running two editions of the SENTIPOLC shared task, drawing a number of useful lessons for the future of sentiment analysis evaluation.

II. The SENTIPOLC Challenge

Sentiment analysis is by now an established task at international campaigns. SENTIPOLC is unique in a few respects. First, the focus of the shared task is on Italian, and it is the only existing sentiment challenge for this language. Second, the sentiment annotation layer is imposed over a dataset which is partly annotated for three other tasks, namely: POS tagging, Linked Named Entities, and Event Factuality [7]. This allows for the joint modeling of various tasks and for easier testing of end-to-end systems. Third, the annotation scheme we employ is more informative than standard ones (also those used at SemEval). Indeed, each category allows for a presence or absence value, thereby letting positive and negative be non-mutually exclusive, and producing innovative combinations, especially in conjunction with the subjectivity layer (see Section III-B for details).

From the start, a particular emphasis has been given to the combinations which allow to mark the presence of irony in tweets, adding a further annotation layer beyond sentiment polarity. Relying on this new layer, SENTIPOLC was the first shared task focusing on sentiment analysis in social media which included a pilot independent task on irony detection, both in the 2014 and 2016 editions. Additionally, in 2016, we have added a layer that specifies the *literal polarity* of an ironic tweet, which in combination with the irony annotation can provide information over the mechanisms that underlie strategies for irony, such as polarity reversal [12].

Interest around the use of non literal language is becoming popular also in other evaluation campaigns. Task 11 at SemEval 2015 [21] was concerned with figurative language in Twitter, but rather than as figurative/literal classification task, it was designed as a polarity detection task in tweets that were already known to be rich in figurative language, as they had been selected and annotated as such. At SemEval 2017, two of the five SA tasks were organized around humor-related topics, but only very recently in 2018 SemEval featured a task on irony detection in English tweets [40]. Finally, the battery of related tasks proposed for French at DEFT@TALN/RECITAL2017 [10] is also reflecting the influence of the SENTIPOLC's experience, where related tasks on polarity classification and irony detection are studied in a joint setting.

A. Task Description

The SENTIPOLC campaign in both the 2014 and the 2016 editions is organized around the three following tasks.

Task 1 - subjectivity classification: a system must decide whether a given message is subjective or objective [31]. Subjectivity classification is often considered a preliminary step necessary to perform sentiment analysis [16].

Task 2 - polarity classification: a system must decide whether a given message is of positive, negative, neutral or mixed sentiment. Differently from most SA tasks (chiefly the SemEval tasks) in our data positive and negative polarities are *not* mutually exclusive and each is annotated as a binary category. A tweet can thus be at the same time positive *and* negative, yielding a mixed polarity, or also neither positive nor negative, meaning it is a subjective statement with neutral polarity, in accordance with [42] (see Section III).

Task 3 - irony detection: a system must decide whether a given message is ironic or not. Twitter communications include a high percentage of ironic messages [17], [24], [23], [33], and platforms monitoring the sentiment in Twitter messages experienced the phenomenon of wrong polarity classification in ironic messages [12], [22]. Indeed, ironic devices in a text can work as unexpected "polarity reversers" (one says something "good" to mean something "bad"), thus undermining the systems' accuracy. In this sense, though not including a specific task on its detection, we have added an annotation layer of *literal polarity* (see Section III-B) which could be potentially used by systems, and also allows us to observe patterns of irony.

The three tasks are meant to be independent: for example, a team could take part in the polarity classification task (Task 2) without tackling Task 1.

III. Gold Standard Creation

In this section we describe how we collected and manually annotated the gold standard for the SENTIPOLC campaigns. Figure 1 provides an overview of the full SENTIPOLC gold standard, with a breakdown of its components from different data sources and methodologies, as detailed in the following sections.

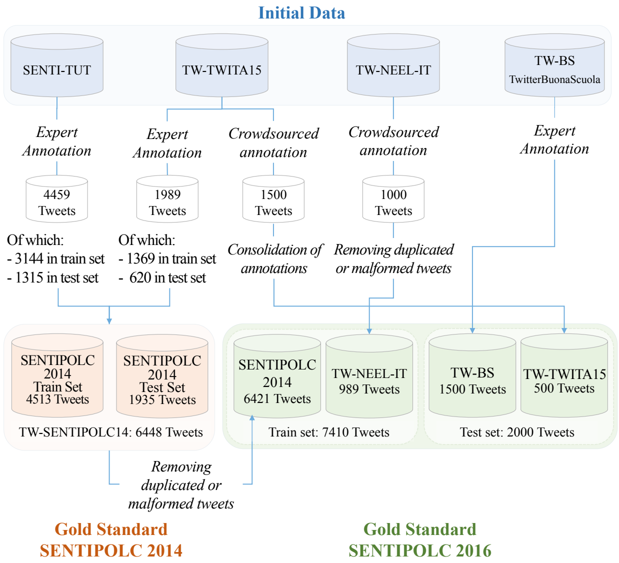


Fig. 1. Creating the gold standard through manual annotation

A. Development and Test Data

The gold standard released for the shared task includes data from different sources. Specifically, the SENTIPOLC 2014 gold standard was created through manual annotation of tweets extracted from two datasets, namely the SENTI-TUT [11] [12] and TWITA 2015 (TW-TWITA15 [9]) collections. To build the SENTIPOLC 2016 gold standard, we re-used the whole SENTIPOLC 2014 dataset, and also added new tweets derived from different resources previously developed for Italian. The dataset composition has been designed in cooperation with the other EVALITA 2016 tasks, in particular the Named Entity rEcognition and Linking in Italian Tweets shared task (NEEL-IT, [5]). Specifically, a portion of the data overlaps with data from NEEL-IT [5], PoSTWITA [13] and FacTA [25]. The multiple layers of annotation on the shared data are intended as a first step towards the long-term goal of enabling participants to develop end-to-end systems from entity linking to entity-based sentiment analysis [3] (see the final report of the EVALITA 2016 evaluation campaign [6] for details).

Both training and test data developed for the 2014 edition of the shared task were included as training data in the 2016 release. Summarizing, the data that we used for the SENTIPOLC 2016 shared task is a collection of tweets which is partially derived from two existing corpora, namely SENTIPOLC 2014 (TW-SENTIPOLC14, 6421 tweets) [8], and TWitterBuonaScuola (TW-BS) [37], from which we selected 1500 tweets. Furthermore, two new sets have been annotated from scratch following the SENTIPOLC 2016 annotation scheme: the first one consists of 1500 tweets selected from TWITA (TW-TWITA15 [9]); the second one consists of 1000 (989 after eliminating malformed tweets) tweets collected in the context of the NEEL-IT shared task (TW-NEELIT [5]).

The tweets in the datasets are marked with a "topic" tag. The training data includes both a *political* collection of tweets and a *generic* collection of tweets. The former has been extracted exploiting specific keywords and hashtags marking political topics ($topic = 1$ in the dataset), while the latter is composed of random tweets on any topic ($topic = 0$). The test material includes tweets from the TW-BS corpus, that were extracted with a specific *socio-political* topic (via hashtags and keywords related to #labuonascuola, different from the ones used to collect the training material). To mark the fact that such tweets focus on a different topic they have been marked with $topic = 2$. While SENTIPOLC does not include any task that takes the "topic" information into account, we release it in case participants want to make use of it.

The annotation scheme of SENTIPOLC 2014 included six fields indicating the manual annotation of the tweet subjectivity (`subj`), its positive (`opos`) and negative (`oneg`) polarity classification, and the presence of irony (`iro`). In SENTIPOLC 2016 this scheme has been enriched with two new fields encoding the *literal* positive (`lpos`) and negative (`lneg`) polarity of tweets. Even if SENTIPOLC does not include any task involving the classification of literal polarity, this information is provided to enable participants to reason about the possible polarity inversion due to the use of figurative language. Table I summarizes the allowed combinations[3].

TABLE I
Combinations of values allowed by our annotation scheme. Order of fields: subj,opos,oneg,iro,lpos,lneg

pattern	description
0,0,0,0,0,0	objective
1,0,0,0,0,0	subj., neutral polarity, no irony
1,1,0,0,1,0	subj., positive polarity, no irony
1,0,1,0,0,1	subj., negative polarity, no irony
1,1,1,0,1,1	subj., both positive and negative polarity (mixed), no irony
1,1,0,1,1,0	subj., positive polarity, ironic twist
1,1,0,1,0,1	subj., positive polarity, ironic twist, negative literal polarity
1,0,1,1,0,1	subj., negative polarity, ironic twist
1,0,1,1,1,0	subj., negative polarity, ironic twist, positive literal polarity
1,1,0,1,0,0	subj., positive polarity, ironic twist, neutral literal polarity
1,0,1,1,0,0	subj., negative polarity, ironic twist, neutral literal polarity
1,1,0,1,1,1	subj., positive polarity, ironic twist, mixed literal polarity
1,0,1,1,1,1	subj., negative polarity, ironic twist, mixed literal polarity

B. Manual Annotation

We followed a mixed approach including both expert annotation and, as a novelty in the 2016 edition, crowdsourcing. For the 2016 edition, annotations from existing corpora (TW-BS and TW-SENTIPOLC14) were revised and finalized via a procedure which involved a group of six expert raters in order

[3] For more details see: http://www.di.unito.it/~tutreeb/sentipolc-evalita16/sentipolc-guidelines2016UPDATED130916.pdf

to make annotation compliant with the SENTIPOLC 2016 updated scheme. Data from NEEL-IT and TWITA15 were annotated from scratch using CrowdFlower. Both the training and the test sets included a mixture of data annotated by the experts and the crowd. In particular, the whole TW-SENTIPOLC14 was included in the development data, while TW-BS was included in the test data. An additional set of 500 crowd-sourced tweets was added to the test set, after a manual check and re-assessment (see below: *Crowdsourced data: consolidation of annotations*). This set also contains the 300 tweets used as test data in the PoSTWITA, NEEL-IT-it and FactA EVALITA 2016 shared tasks.

TW-SENTIPOLC14. Data from the previous evaluation campaign did not include any distinction between literal and overall polarity. Therefore, the old tags `pos` and `neg` were automatically mapped into the new labels `opos` and `oneg`, respectively, which indicate overall polarity. Then, we had to extend the annotation to provide labels for positive and negative literal polarity. In case of tweets without irony, literal polarity values were implied from the overall polarity. For ironic tweets, instead, i.e. $\texttt{iro} = 1$ (806 tweets), we resorted to manual annotation: for each tweet, two independent annotations have been provided for the literal polarity dimension. While for other languages, like English[4] and Dutch [39], ironic tweets can be easily collected by exploiting the presence of specific hashtags (such as #sarcasm), for Italian this isn't quite possible as users do not employ such specific hashtags to mark ironic tweets explicitly. Moreover, we wanted to preserve a natural distribution of irony in the dataset, and extracting ironic tweets directly would not allow for this. At this stage, the two annotators were found in agreement on 53.8% of the tweets. In a second round, a third independent annotation was provided to solve the disagreement. The final label was assigned by majority vote on each field independently. With three annotators, this procedure ensures unambiguous results.

TW-BS. The TW-BS section of the dataset had been previously annotated for polarity and irony[5]. The original TW-BS annotation scheme, however, did not provide any separate annotation for overall and literal polarity. The tags POS, NEG, MIXED and NONE, HUMPOS, HUMNEG in TW-BS were automatically mapped in the following values for the SENTIPOLC's `subj`, `opos`, `oneg`, `iro`, `lpos` and `lneg` annotation fields: POS ⇒ 110010; NEG ⇒ 101001; MIXED ⇒ 111011; NONE ⇒ 000000 ; HUMPOS ⇒ 1101??; HUMNEG ⇒ 1011??. For the last two cases, i.e. where `iro=1`, the same manual annotation procedure described above was applied to obtain literal polarity values: two independent annotations were provided (with agreement on 60.5% of the tweets), and a third annotation was added in a second round in cases of disagreement. Just as with the TW-SENTIPOLC14 set, the final label assignment was done by majority vote.

TW-TWITA15 and TW-NEEL-IT. For these new datasets, all fields were annotated from scratch using CrowdFlower (CF), a crowdsourcing platform which has also been recently used for a similar annotation task [28]. CF enables quality control of the annotations across a number of dimensions, also by employing test questions to find and exclude unreliable annotators. We gave the users a series of guidelines in Italian, including a list of examples of tweets with their annotation according to the SENTIPOLC scheme. The guidelines also contained an explanation of the combinations of values allowed in the annotation schema for the rest of the dataset although in practice these constraints were not enforced in the CF interface. As requested by the platform, we provided a restricted set of "correct" answers to test the reliability of the users. This step proved to be challenging, since in many cases the annotation of at least one dimension is not clear cut. We required to collect at least three independent judgments for each tweet. The total cost of the crowdsourcing has been 55 USD and we collected 9517 judgments in total from 65 workers. We adopted the default CF settings for assigning the majority label (relative majority). The CF reported average confidence (i.e., a measure combining inter-rater agreement and reliability of the contributor) is 0.79 for subjectivity, 0.89 for positive polarity (0.90 for literal positivity), 0.91 for negative polarity (0.93 for literal negativity) and 0.92 for irony. While such scores appear high, they are skewed towards the over-assignment of the "0" label for basically all of classes (see below for further comments on this). Percentage agreement on the assignment of "1" is much lower (ranging from 0.70 to 0.77).[6] On the basis of such observations, we operated a few revisions on the crowd-collected data.

Crowdsourced data: consolidation of annotations. Despite having provided the workers with guidelines, we identified a few cases of value combinations that were not allowed in our annotation scheme, e.g., ironic or polarised tweets (positive, negative or mixed) which were not marked as subjective.

Moreover, we applied a further manual check of crowd-sourced data stimulated by the following observations. When comparing the distributions of values (0,1) for each label in the expert-annotated and crowdsourced data, we observed, as mentioned above, that while the assignment of 1s constituted from 28 to 40% of all assignments for the `opos`/`pos`/ `oneg`/`neg` labels, and about 68% for the subjectivity label in the expert annotation, figures were much lower for the crowdsourced data, with percentages as low as 6 (`neg`), 9 (`pos`), 11 (`oneg`), and 17 (`opos`), and under 50% for subj.[7] This could be an indication of a more conservative interpretation of sentiment on the part of the crowd (note that 0 is also the default value), possibly also due to too few examples in the guidelines, and in any case to the intrinsic subjectivity of the task. On such a basis, we decided to add two more expert annotations to the crowd-annotated test-set adopting the same protocol used by all expert annotators for the rest of the corpus. We assigned the final label for this data based on majority voting from *crowd*, *expert1*, and *expert2*. This does not erase the contribution of the crowd, but hopefully maximises consistency with the guidelines in order to provide a solid evaluation benchmark for this task.

[4] https://competitions.codalab.org/competitions/17468#learn_the_details-data-annotation

[5] For the annotation process and inter-annotator agreement see [37].

[6] This would be taken into account if using Fleiss' Kappa, which is unsuitable in this context due to the varying number of annotators per instance.

[7] The annotation of the presence of irony shows less distance, with 12% in the training set and 8% in the crowd-annotated test set.

C. Format and Distribution

We provided participants with a collection of 7,410 tweets, with IDs and annotations concerning all three SENTIPOLC's subtasks: subjectivity classification (`subj`), polarity classification (`opos`,`oneg`) and irony detection (`iro`), including the two additional fields with respect to SENTIPOLC 2014, namely `lpos` and `lneg`.

The development data include for each tweet the manual annotation for the `subj`, `opos`, `oneg`, `iro`, `lpos` and `lneg` fields, according to the format explained above. Instead, the blind version of the test data, which consists of 2000 tweets, only contains values for the `idtwitter` and `text` fields. The literal polarity might be predicted and used by participants to provide the final classification of the items in the test set, however this should be specified in the submission phase. In addition to the 2000 instances of the official test set, we provided the participants with extra 1000 tweets, without making the difference between the two sets known or explicit. This additional set was entirely annotated via crowdsourcing and did not undergo any expert check, differently than the 500 cases that were instead re-evaluated and included in the official 2000 tweets. One of the aims in the organization of the SENTIPOLC challenge was also evaluating the feasibility of acquiring crowd-annotated data for our tasks. We used the output of the participant on this dataset to gain some insights on the quality of the gold standard when annotated via different means. These experiments are reported in Section VII.

IV. Evaluation metrics

Task 1: subjectivity classification. Systems are evaluated on the assignment of a 0 or 1 value to the subjectivity field. A response is considered plainly correct or wrong when compared to the gold standard annotation. We compute precision (p), recall (r) and F-score (F) for each class, i.e. subjective ((`subj` = 1) and objective (`subj`=0, referred as `Obj` hereinafter).
The overall F-score is the average of the F-scores for subjective and objective classes.

Task 2: polarity classification. Our coding system allows for four combinations of `opos` and `oneg` values: 10 (positive polarity), 01 (negative polarity), 11 (mixed polarity), 00 (no polarity). Accordingly, we evaluate positive and negative polarity independently by computing precision, recall and F-score for both classes (0 and 1). The F-score for the two polarity classes is the average of the F-scores of the respective pairs. Finally, the overall F-score for Task 2 is given by the average of the F-scores of the two polarities.

Task 3: irony detection. Systems are evaluated on their assignment of a 0 or 1 value to the irony field. A response is considered fully correct or wrong when compared to the gold standard annotation. We measure precision, recall and F-score for each class (ironic,non-ironic), similarly to the Task 1, but with different targeted classes. The overall F-score is the average of the F-scores for ironic and non-ironic classes.

Informal evaluation of literal polarity classification. Our coding system allows for four combinations of positive (`lpos`) and negative (`lneg`) values for literal polarity, namely: 10: positive literal polarity; 01: negative literal polarity; 11: mixed literal polarity; 00: no polarity. SENTIPOLC does not include any task that explicitly takes into account the evaluation of literal polarity classification. However, participants could find it useful in developing their system, and might learn to predict it. Therefore, they could choose to submit also this information to receive an informal evaluation of the performance on these two fields, following the same evaluation criteria adopted for Task 2. The performance on the literal polarity classification does not affect the final ranks for the three SENTIPOLC tasks.

V. Results and Methodologies

This section reports an overview of the teams that participated to the two editions of SENTIPOLC 2014 (Table II(a)) and SENTIPOLC 2016 (Table II(b)). It allows to survey most of the approaches used by the NLP community for sentiment analysis of Italian. In particular, during SENTIPOLC 2014, 11 teams from four different countries participated in at least one of the three tasks. These numbers increased in SENTIPOLC 2016, i.e., 13 teams from 6 different countries.

TABLE II
Teams participating to SENTIPOLC 2014 and 2016

(a) SENTIPOLC 2014

team	institution	tasks
CoLingLab (IT)	CoLing Lab, University of Pisa	T2
fbkshelldkm (IT)	Fondazione Bruno Kessler (FBK-IRST)	T1,T2,T3
ficlit+cs@unibo (IT)	FICLIT-University of Bologna	T1,T2
IRADABE (ES/FR)	U Politecnica de Valencia / U Paris 13	T1,T2,T3
Italianlp-wafi (IT)	ItaliaNLP Lab, ILC (CNR)	T2
itgetaruns (IT)	Ca' Foscari University, Venice	T1,T2,T3
mind (IT)	University of Milano-Bicocca	T1,T2,T3
SVMSLU (BY)	Minsk State Linguistic University	T1,T2,T3
uniba2930 (IT)	CS, University of Bari	T1,T2
UNITOR (IT)	University of Roma Tor Vergata	T1,T2,T3
UPFtaln (ES)	TALN, Universitat Pompeu Fabra	T1,T2,T3

(b) SENTIPOLC 2016

team	institution	tasks
ADAPT (IE)	Adapt Centre	T1,T2,T3
CoLingLab (IT)	CoLingLab, University of Pisa	T2
CoMoDI (IT)	FICLIT, University of Bologna	T3
INGEOTEC (MX)	CentroGEO/INFOTEC, CONACyT	T1,T2
IntIntUniba (IT)	University of Bari	T2
IRADABE (ES,FR)	U. Politecnica de Valencia, and U. de Paris	T1,T2,T3
ItaliaNLP (IT)	ItaliaNLP Lab, ILC (CNR)	T1,T2,T3
samskara (IT)	LARI Lab, ILC CNR	T1,T2
SwissCheese (CH)	Zurich University of Applied Sciences	T1,T2,T3
tweet2check (IT)	Finsa s.p.a.	T1,T2,T3
UniBO (IT)	University of Bologna	T1,T2
UniPI (IT)	University of Pisa	T1,T2
Unitor (IT)	University of Roma, Tor Vergata	T1,T2,T3

For each task, we distinguish between *constrained* and *unconstrained* runs. For the constrained runs, the teams had to use the provided development data only, while for unconstrained runs the teams could use additional data for training. Each team had to submit at least a constrained run. We produced a single-ranking table for each subtask, where unconstrained runs are properly marked. Notice that we only use the final F-score for global scoring and ranking. Detailed scores for all classes and tasks for SENTIPOLC

2016 are available on the competition website[8] (similarly for the 2014 edition, see the Appendix in [8]). For each task, we ran a majority class baseline to set a lower-bound for performance as it is standardly done in sentiment analysis evaluation campaigns to deal with class imbalance, and to provide a reference benchmark of a dummy system to compare all developed models [27]. In the tables it is always reported as *Baseline*. When comparing results between the two years, it is important to note that SENTIPOLC 2016 was a bit harder, as the test test was extracted from texts of different topics than the ones present in the training set[9].

Task 1: subjectivity classification. Table III(a) shows results for the subjectivity classification task of SENTIPOLC 2014, which attracted 12 total submissions from 9 teams. The highest F-score was achieved by uniba2930 at 0.7140. Results of SENTIPOLC 2016 are reported in Table III(b) (19 total submissions from 10 different teams). The highest F-score is achieved by Unitor at 0.7444, improving the SENTIPOLC 2014 best results of more than 3 points. Also the average F1 of the systems is better, suggesting that the systems participating to SENTIPOLC 2016 were significantly better than the previous shared task.

Task 2: polarity classification. Table IV(a) shows results for the polarity classification task of SENTIPOLC 2014 (14 submissions from 11 teams) and Table IV(b) shows the results for the same task of SENTIPOLC 2016. The highest F-score of SENTIPOLC 2014 (uniba2930 0.6771) is slightly lower than the highest score of SENTIPOLC 2016 (SwissCheese 0.6828). It is interesting to highlight *SwissCheese* was the top-scoring team also at the 'twin task' for English at Semeval2016-Task4 [28].

Task 3: irony detection Table V(a) shows results for the irony detection task (9 submissions from 7 teams). The highest F-score was achieved by UNITOR at 0.5959, four points higher than the best score of SENTIPOLC 2016 (tweet2check 0.5412, Table V(b)). While all systems score above the baseline, many are close to it, highlighting the task complexity.

TABLE III
TASK 1 - RESULTS : F-SCORES FOR CONSTRAINED ".C" AND UNCONSTRAINED RUNS ".U". AMENDED RUNS ARE MARKED WITH * .

(a) Task 1 - SENTIPOLC 2014

System	Obj	Subj	F
uniba2930.c	**0.6005**	**0.8275**	**0.7140**
UNITOR.u	**0.5762**	0.8032	**0.6897**
uniba2930.u	0.5553	**0.8232**	0.6892
UNITOR.c	0.5819	0.7923	0.6871
IRADABE.c	0.5344	0.8067	0.6706
UPFtaln.c	0.4868	0.8127	0.6497
IRADABE.u	0.5750	0.7178	0.6464
ficlit+cs@unibo.c	0.4480	0.7464	0.5972
mind.c	0.5031	0.6770	0.5901
SVMSLU.c	0.4200	0.7451	0.5825
fbkshelldkm.c	0.4424	0.6761	0.5593
itagetaruns.c	0.3237	0.7211	0.5224
Baseline	0.0000	0.8010	0.4005

(b) Task 1 - SENTIPOLC 2016

System	Obj	Subj	F
Unitor.1.u	**0.6784**	**0.8105**	**0.7444**
Unitor.2.u	0.6723	0.7979	0.7351
samskara.1.c	0.6555	0.7814	**0.7184**
ItaliaNLP.2.c	0.6733	0.7535	0.7134
IRADABE.2.c	0.6671	0.7539	0.7105
INGEOTEC.1.c	0.6623	0.7550	0.7086
Unitor.c	0.6499	0.7590	0.7044
UniPI.1/2.c	**0.6741**	0.7133	0.6937
UniPI.1/2.u	0.6741	0.7133	0.6937
ItaliaNLP.1.c	0.6178	0.7350	0.6764
ADAPT.c	0.5646	0.7343	0.6495
IRADABE.1.c	0.6345	0.6139	0.6242
tweet2check16.c	0.4915	0.7557	0.6236
tweet2check14.c	0.3854	**0.7832**	0.5843
tweet2check14.u	0.3653	0.7940	0.5797
UniBO.1.c	0.5997	0.5296	0.5647
UniBO.2.c	0.5904	0.5201	0.5552
Baseline	0.0000	0.7897	0.3949
*SwissCheese.c	0.6536	0.7748	0.7142
*tweet2check16.u	0.4814	0.7820	0.6317

A. Main outcomes

We compare the participating systems according to the following main dimensions: classification framework (approaches, algorithms, features), tweet representation strategy, exploitation of further Twitter annotated data for training, exploitation of available resources (e.g. sentiment lexicons, NLP tools, etc.), and issues about the interdependency of tasks in case of systems participating in several subtasks.

1) Outcomes from SENTIPOLC 2014: **Tweet representation schema.** As noticed also in the context of similar evaluation campaigns for the English language [30], [36], most systems used supervised learning (uniba2930, mind, IRADABE, UNITOR, UPFtaln, SVMSLU, Italianlp-wafi, CoLingLab, fbkshelldkm). The most popular algorithm was Support Vector Machines, but also Decision Trees, Naive Bayes, K-Nearest Neighbors were used. As mentioned, one team experimented with a co-training approach, too. A variety of features was used, including word-based, syntactic and semantic (mostly lexicon-based) features. The best team in Task 1 and Task 2, uniba2930, specifically mentions that in leave-one-out experiments, (distributional) semantic features were adopted and they appear to contribute the most. uniba2930 is also the only team that explicitly reports using the topic information as a feature, for their constrained runs. The best team in Task 3, UNITOR, employs two sets of features explicitly tailored for the detection of irony, based on emoticons/punctuation and a vector space model to identify words that are out of context. Typical Twitter features were also generally used, such as emoticons, links, usernames, hashtags. Two participants did not adopt a learning approach. ficlit+cs@unibo developed a system based on a sentiment lexicon that uses the polarity of each word in the tweet and the idea of "polarity intensifiers". A syntactic parser was also used to account for polarity inversion cases such as negations. itgetaruns was the only system solely based on deep linguistic analysis exploiting rhetorical relations and pragmatic insights.

Exploitation of additional data for training. Most partic-

[8] http://di.unito.it/sentipolc16

[9] Some teams (Italianlp-wafi in 2014, SwissCheese and tweet2check in 2016) reported conversion errors from their internal format to the official one or similar formal mistakes. The resubmitted amended runs are shown in the tables marked by the * symbol. The team name Italianlp-wafi is also referred to elsewhere as Itanlp-wafi.

TABLE IV
Task 2 - SENTIPOLC Results: F-scores for constrained ”.c” and unconstrained runs ”.u”. Amended runs are marked with * .

(a) Task 2 - SENTIPOLC 2014

System	Pos	Neg	F
uniba2930.c	**0.6752**	**0.6789**	**0.6771**
uniba2930.u	0.6622	**0.6655**	**0.6638**
UNITOR.u	**0.6673**	0.6419	0.6546
IRADABE.c	0.6196	0.6498	0.6347
CoLingLab.c	0.6352	0.6271	0.6312
UNITOR.c	0.6277	0.6321	0.6299
IRADABE.u	0.6058	0.6157	0.6108
UPFtaln.c	0.6079	0.6019	0.6049
SVMSLU.c	0.6153	0.5899	0.6026
ficlit+cs@unibo.c	0.5940	0.6019	0.5980
fbkshelldkm.c	0.5556	0.5695	0.5626
mind.c	0.5293	0.5390	0.5342
itagetaruns.c	0.5021	0.5341	0.5181
Italianlp-wafi.c	0.5159	0.5013	0.5086
Baseline	0.3977	0.3459	0.3718
**Italianlp-wafi.c*	0.6697	0.6576	0.6637

(b) Task 2 - SENTIPOLC 2016

System	Pos	Neg	F
UniPI.2.c	**0.6850**	0.6426	**0.6638**
Unitor.1.u	0.6354	**0.6885**	**0.6620**
Unitor.2.u	0.6312	0.6838	0.6575
ItaliaNLP.1.c	0.6265	**0.6743**	0.6504
IRADABE.2.c	0.6426	0.6480	0.6453
ItaliaNLP.2.c	0.6395	0.6469	0.6432
UniPI.1.u	**0.6699**	0.6146	0.6422
UniPI.1.c	0.6766	0.6002	0.6384
Unitor.c	0.6279	0.6486	0.6382
UniBO.1.c	0.6708	0.6026	0.6367
IntIntUniba.c	0.6189	0.6372	0.6281
IntIntUniba.u	0.6141	0.6348	0.6245
UniBO.2.c	0.6589	0.5892	0.6241
UniPI.2.u	0.6586	0.5654	0.6120
CoLingLab.c	0.5619	0.6579	0.6099
IRADABE.1.c	0.6081	0.6111	0.6096
INGEOTEC.1.u	0.5944	0.6205	0.6075
INGEOTEC.2.c	0.6414	0.5694	0.6054
ADAPT.c	0.5632	0.6461	0.6046
IntIntUniba.c	0.5779	0.6296	0.6037
tweet2check16.c	0.6153	0.5878	0.6016
tweet2check14.u	0.5585	0.6300	0.5943
tweet2check14.c	0.5660	0.6034	0.5847
samskara.1.c	0.5198	0.6168	0.5683
Baseline	0.4518	0.3808	0.4163
*SwissCheese.c	0.6529	0.7128	0.6828
*tweet2check16.u	0.6528	0.6373	0.6450

TABLE V
Task 3 - Irony Detection: F-scores for constrained ”.c” and unconstrained runs ”.u”. Amended runs are marked with * .

(a) Task 3 - SENTIPOLC 2014

System	Non-Iro	Iro	F
UNITOR.u	**0.8345**	**0.3573**	**0.5959**
UNITOR.c	**0.7963**	**0.3554**	**0.5759**
IRADABE.u	0.7983	0.3044	0.5513
IRADABE.c	0.8371	0.2459	0.5415
SVMSLU.c	0.8254	0.2533	0.5394
itagetaruns.c	0.8257	0.1602	0.4929
mind.c	0.7344	0.2197	0.4771
fbkshelldkm.c	0.8328	0.1086	0.4707
UPFtaln.c	0.8842	0.0532	0.4687
baseline	0.8882	0.0000	0.4441

(b) Task 3 - SENTIPOLC 2016

System	Non-Iro	Iro	F
tweet2check16.c	0.9115	0.1710	**0.5412**
CoMoDI.c	0.8993	0.1509	0.5251
tweet2check14.c	0.9166	0.1159	0.5162
IRADABE.2.c	0.9241	0.1026	0.5133
ItaliaNLP.1.c	0.9359	0.0625	0.4992
ADAPT.c	0.8042	**0.1879**	0.4961
IRADABE.1.c	0.9259	0.0484	0.4872
Unitor.2.u	**0.9372**	**0.0248**	**0.4810**
Unitor.c	0.9358	0.0163	0.4761
Unitor.1.u	0.9373	0.0084	0.4728
ItaliaNLP.2.c	**0.9367**	0.0083	0.4725
Baseline	0.9376	0.000	0.4688
*SwissCheese.c	0.9355	0.1367	0.5361

ipants restricted themselves to the provided data and submitted constrained systems. Only three teams submitted unconstrained runs, and apart from UNITOR, results are worse than those obtained by the constrained runs. We believe this situation is triggered by the current lack of sentiment-annotated, available large datasets for Italian. Additionally, what might be available is not necessarily annotated according to the same principles adopted in SENTIPOLC. Interestingly, uniba2930 attempted acquiring more training data via co-training. They trained two SVM models on SentiDevSet, each with a separate feature set, and then used them to label a large amount of acquired unlabelled data progressively adding training instances to one another's training set, and re-training. No significant improvement was observed, due to the noise introduced by the automatically labelled training instances.

External Resources. Almost all participants relied on various sentiment lexicons. At least six teams (uniba2930, UPFtaln, fbkshelldkm, ficlit+cs@unibo, UNITOR, IRADABE) used information from SentiWordNet [19], either using the mapping of SentiWordNet to Italian given by the Sentix lexicon [9] or otherwise alternative resources. Several other lexica and dictionaries were used, either natively in Italian or translated from English (e.g. AFINN, Hu-Liu lexicon, Whissel's Dictionary). Native tools for Italian (which were allowed also for unconstrained runs) were used for pre-processing, such as tokenisers, POS-taggers, and parsers.

2) Outcomes from SENTIPOLC 2016: **Tweet representation schemas.** Almost all teams adopted (i) traditional manual feature engineering or (ii) distributional models (i.e. word embeddings) to represent tweets. The teams adopting the strategy (i) make use of traditional feature modeling, using specific features that encode word-based, syntactic and semantic (mostly lexicon-based) features. In addition, micro-blogging specific features such as emoticons and hashtags are also adopted, for example by ColingLab, INGEOTEC or CoMoDi. Deep learning methods adopted by some teams, such as UniPI and SwissCheese required to model individual tweets through geometrical representation of tweets, i.e. vectors. Words from individual tweets are represented through word embeddings, mostly derived by using the Word2Vec tool or similar approaches. Unitor extends this representation with additional features derived from Distributional Polarity Lexicons. In addition, some teams (e.g. ColingLab) adopted Topic Models to represent tweets. Samskara also used feature modelling with a communicative and pragmatic value. CoMoDi is one of the few systems that investigated irony-specific features.

Exploitation of additional data for training. Some teams submitted unconstrained results, as they used additional Twitter annotated data for training their systems. In particular, UniPI used a silver standard corpus made of more than 1M tweets to pre-train a Convolutional Neural Network (CNN); this corpus is annotated using a polarity lexicon and specific polarised words. Also Unitor used external tweets to pre-train their CNN. This corpus is made of the contexts of the tweets populating the training material and automatically annotated using the classifier trained only over the training material, in a semi-supervised fashion. Moreover, Unitor used distant supervision to label a set of tweets used for the acquisition of their so-called Distribution Polarity Lexicon. Distant supervision is also adopted by INGEOTEC to extend the training material for the their SVM classifier.

External Resources. The majority of teams used external resources, such as lexicons specific for Sentiment Analysis tasks. Some teams used already existing lexicons, such as Samskara, ItaliaNLP, CoLingLab, or CoMoDi, while others created their own task specific resources, such as Unitor, IRADABE, CoLingLab.

Multi-task learning Among the systems participating in more than one SENTIPOLC task, SwissCheese and Unitor designed systems that exploit the task interdependency. In particular, SwissCheese trained one CNN for all the tasks simultaneously, by joining the labels. The results of their experiments indicate that the multi-task CNN outperforms the single-task CNN. Unitor made the training step dependent on the subtask, e.g. considering only subjective tweets when training the Polarity Classifier. However it is difficult to assess the contribution of cross-task information based only on the experimental results obtained by the single teams.

B. Comparing SENTIPOLC 2014 and 2016

The majority of participants in SENTIPOLC 2016 adopted learning methods already investigated in SENTIPOLC 2014; in particular, Support Vector Machine (SVM) is the most adopted learning algorithm. The SVM is generally based on specific linguistic/semantic feature engineering.

The main difference between 2014 and 2016 approaches are the deep learning systems, that were used only in 2016. In particular, Convolutional Neural Networks (CNN) have been investigated in 2016 by a few teams, following the same line as the international community on applying deep learning techniques to sentiment-related tasks [32]. Moreover, multi-task learning was introduced in 2016, and one team learned to classify subjectivity, polarity and irony at the same time. One participant adopted a rule based approach in combination with a rich set of linguistic cues dedicated to irony detection.

In the irony task, performances drop significantly in SENTIPOLC 2016. An explanation for this could be that unlike SENTIPOLC 2014, at this edition the topics in the train and in the test sets are different, and it has been shown that systems might be modeling topic rather than irony [2]. This evidence suggests that examples are probably not sufficient to generalise over the structure of ironic tweets. We plan to run further experiments on this issue, including a larger and more balanced dataset of ironic tweets in future campaigns.

Although evaluated over different data, we see that in 2016 best systems show better, albeit comparable, performance for subjectivity with respect to systems of 2014, and outperform them for polarity (if we consider late submissions).

VI. Error Analysis

To get some deeper insight on the difficulties inherent to the polarity detection task (Task 2), we manually examined cases where the three systems obtaining the top ranks on Task 2 of SENTIPOLC 2016 yielded the wrong predictions. In particular, we selected the subset of the tweets in the test set on which all the three systems predicted a wrong label for `opos` or `oneg` (or both). The resulting set (the *hard cases* set, HC set henceforth) is composed of 495 tweets. Each tweet in the HC set was individually annotated with possible causes of errors by at least one of task's organizers, and the results were collectively discussed to identify potential reasons and error patterns. In the following, we report and discuss notable error classes resulting from our analysis. For each class, we indicate the percentage of misclassified tweets belonging to it. In 39 cases (8%), multiple error categories were selected because of the co-occurrence of difficulties that can be responsible for misclassification.

Implicit sentiment polarity and polarity inferred from contextual knowledge (35%) - We observed that in 172 cases users do not express their sentiment, mood or personal opinions in an explicit manner. However, an evaluation of it could nonetheless be inferred by human annotators by relying on *common sense knowledge*, *world knowledge*, and shared contextual knowledge in general. In the literature, a gap exists in this sense and studies dedicated to implicit expressions of sentiment, such as the context-aware model of sentiment proposed in [41], are limited. Also, systems at SENTIPOLC unavoidably focused mainly on the detection of explicit sentiment, as they often relied on domain independent affective lexicons. As emerged from our analysis, such approaches do not allow to deal with cases where the expression of a negative or positive evaluation towards an entity is not accompanied by the presence of explicitly polarized lexical clues, like in the following tweet from the HC set: *(ITA) "@matteorenzi 'la buona scuola' pare 'l'opera nazionale balilla' - (EN) " @matteorenzi 'la buona scuola' reminds me of the 'opera nazionale balilla'"*, whose polarity cannot be understood without word knowledge about the recent and historical Italian political context ("la buona scuola"/"*the good school*" is the name of the school reform proposed by the Renzi's government; "opera nazionale balilla" was an Italian fascist youth organization).

Some misclassified tweets, where contextual knowledge is required for a correct interpretation of the overall sentiment, are cases that do not express a generic sentiment or opinion but rather a finer-grained stance towards a specific target (especially in the #labuonascuola sub-corpus), as in the tweet:

(ITA) "A me non risultano quegli errori che dite #labuonascuola Controllate anche voi [URL]" - (EN) "I didn't find the errors you are talking about #labuonascuola Please double check here".

In this example, there are no explicit lexical clues that could support an automatic sentiment analyzer in assigning the correct positive polarity to the tweet. Indeed, the positive evaluation is related to the recognition of a positive stance towards an Italian political reform and it is implicit: in a conversational context where a reform is under discussion, the absence of mistakes or typos assumes a positive connotation.

Along this line, we also observed that some cases are even more complex, since the contextual knowledge needed is related to a 'belief framework', which can be recognised as shared by a narrower group of people (but not by most people), as in the following example:

(ITA) "@tuttoprof nel #labuonascuola si studia #informatica x sviluppare #pensierocomputazionale: cultura x tutti [URL] @dskutz"- (EN) " @tuttoprof in #labuonascuola we study to enhance computational thinking: culture for all".

The positive connotation of *pensiero computazionale* ('computational thinking') can be inferred only when recognizing that the author of the post belongs to a community of social media people who usually show a positive orientation towards integrating computational thinking into the school curricula.

Figurative language devices (20%) - In a 97 tweets of the HC set we observed figurative language [20], such as sarcasm as in *(ITA) "Una buona scuola per un mondo buono. Firmato Mulino Bianco*[10]" - *(EN) "A good school for a good world. By Mulino Bianco"*, metaphors as in *(ITA)"@SteGiannini @FusacchiA @matteorenzi @pdnetwork @SenatoriPD la 'buona scuola' perché prevedete di mangiarci sopra!! #sfplm85bis"*[11], and a remarkable presence of rhetorical questions as in *(ITA) "@matteorenzi dove sta questa buona scuola? Dove" - (EN) "@matteorenzi where is this good school? Where"*, which can be difficult for the systems to properly comprehend. Much figurative language is based on conventions, such as idiomatic expressions and proverbs.

Twitter language: noisy text, hashtags, and expressive signals (18%) - In 90 misclassified tweets we observed the presence of noisy texts (misspellings, abbreviations, new words), expressive signals such as elongated words as well as emojis, or evaluative hashtags, especially multi-word hashtags. Elongated words ("*Raga stasera tutti in piazzaaaaaaa*" - untranslatable) and emojis must be taken care of during preprocessing to preserve the information they carry, which is often crucial to disambiguate polarity.

As for hashtags, they are employed by Twitter users to accomplish different linguistic functions, thus enabling embedding of metadiscourse in social media communication [44]. In a sentiment analysis setting, a relevant function that can be played by the hashtag is the one of expressing an evaluative metacomment construing a sort of stance, which sometimes alone determine the polarity of an entire post (e.g. #cattivascuola (*#badschool*) in *(ITA) "Il paradosso del sorite e la consultazione su #labuonascuola #cattivascuola [URL]" - (EN) "The sorites paradox and the consultation on #goodschool #badschool [URL]"*. See also the multi-word hashtags '#TuttiInGalera' ('#AllInJail') and '#ersistema' ('thesystem', with reference to the political system) in the following tweet from the HC set:

(ITA) "*#AndreaColletti #M5S: #Riforma della #prescrizione [URL] #Incalza #TuttiInGalera #ersistema #terradeifuochi*" - (EN) "*#AndreaColletti #M5S: #Statute of #limitations [URL] #Advances #AllInJail #thesystem #terradeifuochi*".

The multi-word nature of these hashtags makes even harder to interpret their meaning in terms of sentiment polarity.

Colloquial expressions and specific jargons (12%) - 57 misclassified tweets contain colloquial expressions (e.g., "è un pacco"/"*it's a scam*"), rare words (e.g., "imbecille"/"*stupid*"), dialectal expression (idiomatic expression especially from the dialect, e.g. "sei nà sola"/"*you're a fraud*"), slurs, slang words (e.g. "Quanto è gnocco Rollo?'/'"*How handsome is Rollo?*") words belonging to specific jargons (e.g., soccer jargon).

Subjective neutral and mixed cases (12%) - In general, mixed cases were misclassified. One hypothesis is that most of the systems assigned a sort of 'prevailing' polarity to the tweet, without recognising the presence of signals for the coexistence of both the positive and negative polarity. Similarly, systems often misclassified 'subjective neutral tweets', but we have to take into account that classes in the training set were unbalanced (reflecting the situation in the real data). In particular, there were probably too few examples of neutral subjectivity for systems to be able to generalize well. The HC set includes 54 cases of misclassified 'subjective neutral tweets' (on a total of 219 tweets of this nature in the test set). We observed that the systems were attributing a polarity to such tweets in almost all cases (58), which means that they recognized subjectivity, but they were not able to discriminate the neutrality w.r.t. the polarity.

Relationship between factual information and polarity (9%) - Among misclassified tweets, 45 involve a 'polar fact' [43], i.e. factual information (such as news), where the event reported usually invokes for most people a positive or negative feeling. In *(ITA) " #Trieste respinge #labuonascuola renziana, 10 ottobre sciopero e corteo #scuola..." - (EN) " #Trieste reject #labuonascuola promoted by Renzi, strike on October 10 with demonstration at #school...".* the 'polar fact' is that (most people know that) strikes and demonstrations related to a reform are evidence of a negative attitude towards it. Cases like this are extremely challenging for a text-based classifier if the system does not embed any rule or strategy to deal with the pragmatic context.

Use of negation or adversative conjunction (4%) - Negation is also often present in 20 misclassified tweets, which suggests that the top systems were not provided with an appropriate and effective way of calculating and representing the role of negation in sentiment polarity classification. Moreover, many mixed misclassified tweets contained adversative (like "ma"/"*but*") or concessive conjunctions like "anche se"/"*even if*", see for instance the following example:

(ITA) "nuove energie per la #scuola, anche se manca un ripensamento dei cicli [URL] @LuigiBerlinguer su #labuonascuola" - (EN) "New energy for the school, even if a restyling of the cycles would still be needed http://t.co/jEpbRdLgff @LuigiBerlinguer on #labuonascuola"

[10] Mulino Bianco is a brand for cookies that is famous in Italy for its commercials depicting the "perfect family" stereotype

[11] *mangiare sopra* ('to eat on something') means to profit off something in questionable ways.

In particular, adversative conjunctions are used to connect two clauses with opposite meanings, and the sentiment orientations of these two clauses are usually different from each other. This could be an interesting feature to exploit to discriminate cases of mixed polarity.

Lastly, we checked the distribution of labels in the HC as compared to the whole test set. The figures are reported in Table VI.

TABLE VI
DISTRIBUTIONS OF THE LABELS IN THE HC SET, COMPARED TO THE DISTRIBUTIONS IN THE TEST SET.

	Test set	**Hard cases**
subj (0/1)	695/1,305 (34.7%/65.2%)	43/452 (8.6%/91.3%)
opos (0/1)	1,648/352 (82.4%/17.6%)	357/138 (72.1%/27.8%)
oneg (0/1)	1,230/770 (61.5%/38.5%)	201/294 (40.6%/59.3%)
iro (0/1)	1,765/235 (88.2%/11.7%)	416/79 (84.0%/15.9%)

Among the noticeable findings, we can observe how the hard cases set presents a higher rate of ironic content (15.9% of the tweets vs. 11.7% in the original test set) and an even higher rate of subjective content (91.3% vs. 65.2% in the test set). Moreover, the polarity of the hard cases set is slightly less unbalanced, with a ratio of `opos` labels to `oneg` labels in this subset of 0.469 against 0.457 in the test set.

Finally, we also exploited the information contained in the gold standard about literal polarity, and compared it with the intended polarity of the message. We found that 429 tweets (86.6%) in HC have the same literal and intended polarity, while in the test set this is true for 1,826 tweets (91.3%).

VII. QUALITY OF THE GOLD STANDARD

As detailed in Section III-B, the gold standard for the SENTIPOLC 2016 edition has mixed origin, including annotations done by the crowd, and parts corrected by experts. How much does this mixed nature affect the quality of the dataset and its use in the evaluation of the systems' performance? In order to answer such questions, we ran validation experiments focusing on two aspects: the impact of the quality of the gold standard on the evaluation of system performance, and the homogeneity of the gold standard, i.e. its internal consistency.

As a first test, we assessed the difference in the quality derived from the expert revision of a portion of the gold standard annotated by the crowd (Section III-B), for a total of 500 instances. We call this set *ER500* when the labels are the expert revised ones, and *C500* when the labels are the original ones assigned by the crowd. We computed a measure of inter-rater agreement between the three systems who achieved the top scores on the polarity detection task of SENTIPOLC 2016 and the gold standard. The considered systems are UniPI.2 (constrained run), Unitor.1 and Unitor.2 (unconstrained runs). We computed the Fleiss' Kappa, considering the gold standard set and the systems as they were independent annotators, one time using the crowdsourced labels, and a second time replacing them with the expert revised labels. The results, broken down by label, are reported in Table VII.

TABLE VII
INTER-RATER AGREEMENT (FLEISS' KAPPA) BETWEEN THE TOP THREE SYSTEMS AND THE CROWDSOURCED GOLD STANDARD (*C500* SET) VS. EXPERT REVISED GOLD STANDARD (*ER500*).

Set	**`subj`**	**`opos`**	**`oneg`**	**`iro`**
C500	0.611	0.518	0.453	0.024
ER500	0.657	0.564	0.524	-0.002

The results of the experiment show that the inter-rater agreement is higher when computed on the expert revised part of the gold standard dataset, and lower when the annotation is provided by the crowd. This result suggests that the quality of the crowdsourced data can be improved by employing experts to re-annotate part of it, although this could raise a scalability issue, see VIII. Note also that the effect may be due to the new annotation being simply more consistent with the training set because annotated by the same experts.

Internal consistency of the gold standard is further evaluated via a second experiment that we ran in order to test the ability of the gold standard to withstand statistical noise. Our assumption is that a good quality gold standard should be able to provide the same evaluation scores even if only subsets of it are used, i.e., the full datasets should exhibit similar properties to its parts. Is our dataset of mixed origin eventually composed homogeneously, at least in terms of system evaluation? In other words: if we test systems on slices of the gold standard, even if their annotation comes from different sources, do we get approximately the same results?

We thus devised the following experiment: given our gold standard set G to test, we divided it randomly into two halves G_1 and G_2. We ran the evaluation of the submissions to the original task using both halves of the gold standard independently, producing two sets of F-measure scores R_1 and R_2. Finally, we computed the statistical correlation (Pearson's) between the F-measures: $c(G) = Pearson(R_1, R_2)$. In this experiment we focused again on the results obtained by the systems on Task 2 (polarity detection). We repeated the meta-evaluation experiment on several datasets, summarized in Table VIII, including also the dataset of 1000 additional tweets annotated by the crowd (C) that were released to the participants together with the test set but were not considered for the official evaluation and ranking (see Section III-C). We obtained the following results: $c(G) = 0.994$, $c(C) = 0.989$, $c(ER500) = 0.983$, $c(C500) = 0.984$. Generally speaking, the high correlation of the F-measures indicates that the gold standard is robust against random sampling, implying that it has been annotated following the guidelines in a coherent way throughout the dataset. The slightly lower correlation scores that we observe on the datasets that include crowd-based annotation suggest that mixing different sources of annotation might need to be supervised closely to ensure internal coherence. Nevertheless, at least in our case, the impact on system evaluation seems to be minimal. Indeed, the almost zero difference between $c(ER500)$ and $c(C500)$ indicates that the manual correction of part of the dataset has basically no impact on the final evaluation outcome. This may be due to the small size of the manually corrected subset or to the high level of noise in the original data that exceeds the potential gain in the quality of the annotation. As a final reflection, we can summarize the findings of these additional experiments aimed at evaluating the quality of the gold standard and its impact on evaluation as follows: Does it matter that not all annotations come from the same source and not all annotators are expert? We do observe some difference as manual correction of crowd-annotated data seems to improve quality, but in terms of system evaluation, keeping in a portion of crowd-annotated, potentially more noisy data, appears to be rather irrelevant.

TABLE VIII
DATASETS USED AS GOLD STANDARDS IN THE EXPERIMENTS.

Name(label)	Description	Size
Gold 2000 (G)	Official test set used for the evaluation of SENTIPOLC 2016	2000
Crowd (C)	Portion of the test set annotated with Crowd Flower, not used for the official evaluation	1000
Expert Revised ($ER500$)	Subset of G that have been expert revised, with the revised labels	500
Crowdsourced ($C500$)	Subset of G that have been expert revised, with the original labels	500

VIII. Closing Remarks

SENTIPOLC has been successful in bringing together the sentiment analysis community towards the analysis of social media in the Italian language. Running two editions in the span of three years allowed us to depict a significant picture of the state of the art of this area of language technologies.

The overwhelming trend in terms of approaches to the task is clearly the use of supervised machine learning techniques. Among them, word embeddings proved to improve the performance of the systems across the boards regardless of the particular learning framework employed. Nevertheless, the absolute figures emerging from the evaluation suggest that we are still far from having solved sentiment analysis on Italian social media texts. However, from a historical perspective, we found that the performance of the systems on the subjectivity and polarity detection tasks is increasing over time.

One of the challenges emerged particularly from the 2016 edition, where we intentionally provided a test set focusing on a domain that was absent from the training set. As a consequence, some of the systems saw a substantial drop in performance with respect to the evaluation on the development set, due to a lack of generalization power of their learning architecture.

The experience of running two editions of SENTIPOLC, involving gathering annotated data sets and analyzing the results of the participating systems, allowed us to take a deeper look at how the gold standard is produced and how the methodology for its creation interacts with the results of the shared task. In particular, we obtained an empirical confirmation that crowdsourcing is a valid alternative to expert annotation as a way of producing large-scale high-quality data sets for evaluation of sentiment analysis. However, we also observed that mixing different sources of annotation on the same dataset might require close supervision in order to ensure internal coherence of data annotation. On the one hand, expert revision of crowdsourced annotation ensures high-quality data for benchmarking sentiment analysis-related tasks. On the other hand, this two-step annotation may suffer from scalability issues. This is a problem that is not frequently addressed but we believe definitely deserves further study and serious consideration in the creation of benchmark resources for sentiment analysis.

In conclusion, in this paper we presented a meta-evaluation of sentiment analysis of social media, which was made possible by the SENTIPOLC initiative. The data and observations we collected highlighted a number of critical points, which will be considered for the organization of future shared tasks.

References

[1] Barbieri, F., Basile, V., Croce, D., Nissim, M., Novielli, N., Patti, V.: Overview of the evalita 2016 sentiment polarity classification task. In: Proc. of CLiC-it & EVALITA 2016, *CEUR Workshop Proceedings*, vol. 1749. CEUR-WS.org (2016)

[2] Barbieri, F., Ronzano, F., Saggion, H.: How Topic Biases Your Results? A Case Study of Sentiment Analysis and Irony Detection in Italian. In: RANLP, Recent Advances in Natural Language Processing (2015)

[3] Basile, P., Basile, V., Nissim, M., Novielli, N.: Deep tweets: from entity linking to sentiment analysis. In: Proc. of CLiC-it 2015, p. 41 (2015)

[4] Basile, P., Basile, V., Nissim, M., Novielli, N., Patti, V.: Sentiment analysis of microblogging data. In: Encyclopedia of Social Network Analysis and Mining, pp. 1–17. Springer (2017)

[5] Basile, P., Caputo, A., Gentile, A.L., Rizzo, G.: Overview of the EVALITA 2016 Named Entity rEcognition and Linking in Italian Tweets (NEEL-IT) Task. In: Proc. of of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. aAcademia University Press (2016)

[6] Basile, P., Cutugno, F., Nissim, M., Patti, V., Sprugnoli, R.: EVALITA 2016: Overview of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In: Proc. of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. aAcademia University Press (2016)

[7] Basile, P., Cutugno, F., Nissim, M., Patti, V., Sprugnoli, R.: EVALITA Goes Social: Tasks, Data, and Community at the 2016 Edition. Italian Journal of Computational Linguistics (IJCoL) **3**(1) (2017)

[8] Basile, V., Bolioli, A., Nissim, M., Patti, V., Rosso, P.: Overview of the Evalita 2014 SENTIment POLarity Classification Task. In: Proc. of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian, pp. 50–57. Pisa University Press, Pisa, Italy (2014)

[9] Basile, V., Nissim, M.: Sentiment analysis on Italian tweets. In: Proc. of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 100–107. NAACL (2013)

[10] Benamara, F., Grouin, C., Karoui, J., Moriceau, V., Robba, I. (eds.): Analyse d'opinion et langage figuratif dans des tweets: présentation et résultats du Défi Fouille de Textes DEFT2017. TALN 2017 (2017)

[11] Bosco, C., Allisio, L., Mussa, V., Patti, V., Ruffo, G., Sanguinetti, M., Sulis, E.: Detecting happiness in italian tweets: Towards an evaluation dataset for sentiment analysis in Felicittà. In: Proc. of ESSLOD 2014, LREC 2014, pp. 56–63. Reykjavik, Iceland (2014)

[12] Bosco, C., Patti, V., Bolioli, A.: Developing Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT. IEEE Intelligent Systems, Special Issue on Knowledge-based Approaches to Content-level Sentiment Analysis **28**(2), 55–63 (2013)

[13] Bosco, C., Tamburini, F., Bolioli, A., Mazzei, A.: Overview of the EVALITA 2016 Part Of Speech on TWitter for ITAlian Task. In: Proc. of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. aAcademia University Press (2016)

[14] Burmania, A., Parthasarathy, S., Busso, C.: Increasing the reliability of crowdsourcing evaluations using online quality assessment. IEEE Transactions on Affective Computing **7**(4), 374–388 (2016)

[15] Cambria, E.: Affective computing and sentiment analysis. IEEE Intelligent Systems **31**(2), 102–107 (2016). DOI 10.1109/MIS.2016.31

[16] Chaturvedi, I., Cambria, E., Welsch, R.E., Herrera, F.: Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. Information Fusion **44**, 65–77 (2018)

[17] Davidov, D., Tsur, O., Rappoport, A.: Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In: Proc. of the 14th Conf. on Computational Natural Language Learning, CoNLL '10, pp. 107–116. Uppsala, Sweden (2010)

[18] D'Mello, S.K.: On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability. IEEE Transactions on Affective Computing **7**(2), 136–149 (2016)

[19] Esuli, A., Baccianella, S., Sebastiani, F.: SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Proc. of LREC'10. ELRA (2010)

[20] Garavelli, B.: Il parlar figurato: Manualetto di figure retoriche. Universale Laterza. Editori Laterza (2014)

[21] Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Barnden, J., Reyes, A.: Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 470–478 (2015)

[22] Ghosh, A., Li, G., Veale, T., Rosso, P., Shutova, E., Reyes, A., Barnden, J.: Semeval-2015 task 11: Sentiment analysis of figurative language in Twitter. In: Proc. of the 9th Int. Workshop on Semantic Evaluation (SemEval 2015), pp. 470–475. Denver, Colorado, USA (2015)

[23] González-Ibáñez, R., Muresan, S., Wacholder, N.: Identifying sarcasm in twitter: A closer look. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT '11, pp. 581–586. Portland, Oregon (2011)

[24] Hao, Y., Veale, T.: An ironic fist in a velvet glove: Creative misrepresentation in the construction of ironic similes. Minds Mach. **20**(4), 635–650 (2010)

[25] Minard, A.L., Speranza, M., Caselli, T.: The EVALITA 2016 Event Factuality Annotation Task (FactA). In: Proc. of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. aAcademia University Press (2016)

[26] Mozetič, I., Grčar, M., Smailović, J.: Multilingual twitter sentiment classification: The role of human annotators. PloS one **11**(5) (2016)

[27] Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., Wilson, T.: Semeval-2013 task 2: Sentiment analysis in twitter (2013)

[28] Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V.: SemEval-2016 Task 4: Sentiment Analysis in Twitter. In: Proc. of the 10th Int. Workshop on Semantic Evaluation (SemEval-2016), pp. 1–18. San Diego, California (2016)

[29] Nakov, P., Rosenthal, S., Kiritchenko, S., Mohammad, S.M., Kozareva, Z., Ritter, A., Stoyanov, V., Zhu, X.: Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. Language Resources and Evaluation **50**(1), 35–65 (2016)

[30] Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., Wilson, T.: Semeval-2013 task 2: Sentiment analysis in Twitter. In: Proc. of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 312–320. ACL (2013)

[31] Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and trends in information retrieval **2**(1-2), 1–135 (2008)

[32] Poria, S., Cambria, E., Hazarika, D., Vij, P.: A deeper look into sarcastic tweets using deep convolutional neural networks. In: COLING, pp. 1601–1612. ACL (2016)

[33] Reyes, A., Rosso, P.: On the difficulty of automatically detecting irony: Beyond a simple case of negation. Knowledge and Information Systems **40**(3), 595–614 (2014)

[34] Rosenthal, S., Farra, N., Nakov, P.: Semeval-2017 task 4: Sentiment analysis in twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 502–518 (2017)

[35] Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S.M., Ritter, A., Stoyanov, V.: SemEval-2015 Task 10: Sentiment Analysis in Twitter. In: Proc of. the 9th Int. Workshop on Semantic Evaluation, SemEval '2015. Denver, Colorado (2015)

[36] Rosenthal, S., Ritter, A., Nakov, P., Stoyanov, V.: SemEval-2014 Task 9: Sentiment Analysis in Twitter. In: Proc. of the 8th Int. Workshop on Semantic Evaluation (SemEval 2014), pp. 73–80. Dublin, Ireland (2014)

[37] Stranisci, M., Bosco, C., Farías, D.I.H., Patti, V.: Annotating sentiment and irony in the online italian political debate on #labuonascuola. In: Proc. of the LREC 2016, pp. 2892–2899. ELRA (2016)

[38] Taulé, M., Martí, M.A., Rangel Pardo, F.M., Rosso, P., Bosco, C., Patti, V.: Overview of the task on stance and gender detection in tweets on catalan independence. In: Proc. of the 2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with SEPLN 2017, *CEUR Workshop Proceedings*, vol. 1881. CEUR-WS.org, Murcia, Spain (2017)

[39] Van Hee, C., Lefever, E., Hoste, V.: Exploring the realization of irony in twitter data. In: LREC (2016)

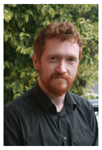
[40] Van Hee, C., Lefever, E., Hoste, V.: SemEval-2018 Task 3: Irony Detection in English Tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation, pp. 39–50. Association for Computational Linguistics (2018)

[41] Vanzo, A., Croce, D., Basili, R.: A context-based model for sentiment analysis in twitter. In: COLING 2014, 25th International Conference on Computational Linguistics, Proc. of the Conference: Technical Papers, August 23-29, 2014,Dublin, Ireland, pp. 2345–2354 (2014)

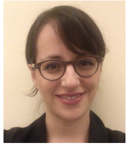
[42] Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language Res Eval **1**(2) (2005)

[43] Wilson, T.: Annotating subjective content in meetings. In: Proc. of LREC'08. ELRA, Marrakech, Morocco (2008)

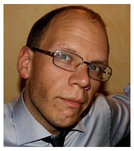
[44] Zappavigna, M.: Searchable talk: the linguistic functions of hashtags. Social Semiotics **25**(3), 274–291 (2015)



Valerio Basile is a postdoc researcher at University of Turin. His work spans across several areas such as: formal representations of meaning, linguistic annotation, natural language generation, commonsense knowledge, semantic parsing and sentiment analysis. He is one of the creators of TWITA, the large-scale collection of Italian tweets, and an organizer of both editions of the Sentiment Polarity Classification task on Italian (SENTIPOLC).



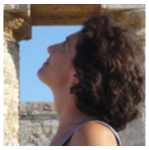
Nicole Novielli is an Assistant Professor at the University of Bari. Her research interests lie at the intersection of software engineering and affective computing with a specific focus on sentiment analysis of social media and developers communication traces. In 2016, she started the workshop series on Emotion Awareness in Software Engineering co-located with the International Conference on Software Engineering (ICSE).



Danilo Croce is an Assistant Professor at the University of Roma, Tor Vergata. His expertise concerns theoretical and applied Machine Learning in the areas of Natural Language Processing, Information Retrieval and Data Mining. He is author of more than 70 publications on international journals, proceedings of international Conferences and Workshops. He acts regularly as a reviewer for the most important International Conferences with a yearly schedule (ACL, COLING and AAAI).

Francesco Barbieri is a postdoctoral researcher in Natural Language Processing at Universitat Pompeu Fabra (UPF). He worked on automatic text classification of social media content, focusing on irony detection and in the last two years on emojis. He also participated to the sentiment in figurative language task (Task 11) of SemEval 2015, presenting the second best system and co-organised the Multilingual Emoji Prediction Task at SemEval 2018.

Malvina Nissim is Associate Professor in Language Technology at the University of Groningen. She has experience in sentiment analysis and author identification and profiling. Before joining the University of Groningen, she was a tenured researcher at the University of Bologna (2006-2014), and a post-doc at the University of Edinburgh (2001-2005), and the National Research Council in Rome (2005-2006).



Viviana Patti is Associate Professor at University of Turin. Her recent research interests include sentiment analysis, irony detection and hate speech detection. She has been guest editor of the special issue on Emotion and Sentiment in Social and Expressive Media in the IPM Journal (2016) and of the special section on Affect and Interaction in Agent-based Systems and Social Media in ACM TOIT (2017). She led the development of Twitter corpora for sentiment analysis in Italian, English and Spanish, exploited in international evaluation campaigns.