Overview of the Evalita 2016 SENTIment POLarity Classification Task

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

Overview of the Evalita 2016 SENTIment POLarity Classification Task

Francesco Barbieri
Pompeu Fabra University
Spain
francesco.barbieri@upf.edu

Valerio Basile
Université Côte d'Azur,
Inria, CNRS, I3S
France
valerio.basile@inria.fr

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

Malvina Nissim
University of Groningen
The Netherlands
m.nissim@rug.nl

Nicole Novielli
University of Bari "A. Moro"
Italy
nicole.novielli@uniba.it

Viviana Patti
University of Torino
Italy
patti@di.unito.it

Abstract

English. The SENTIment POLarity Classification Task 2016 (SENTIPOLC), is a rerun of the shared task on sentiment classification at the message level on Italian tweets proposed for the first time in 2014 for the Evalita evaluation campaign. It includes three subtasks: *subjectivity classification*, *polarity classification*, and *irony detection*. In 2016 SENTIPOLC has been again the most participated EVALITA task with a total of 57 submitted runs from 13 different teams. We present the datasets – which includes an enriched annotation scheme for dealing with the impact on polarity of a figurative use of language – the evaluation methodology, and discuss results and participating systems.

Italiano. *Descriviamo modalità e risultati della seconda edizione della campagna di valutazione di sistemi di sentiment analysis (SENTIment POLarity Classification Task), proposta nel contesto di "EVALITA 2016: Evaluation of NLP and Speech Tools for Italian". In SENTIPOLC è stata valutata la capacità dei sistemi di riconoscere diversi aspetti del sentiment espresso nei messaggi Twitter in lingua italiana, con un'articolazione in tre sottotask:* subjectivity classification*,* polarity classification *e* irony detection*. La campagna ha suscitato nuovamente grande interesse, con un totale di* 57 *run inviati da* 13 *gruppi di partecipanti.*

1 Introduction

Sentiment classification on Twitter, namely detecting whether a tweet is polarised towards a positive or negative sentiment, is by now an established task. Such solid and growing interest is reflected in the fact that the Sentiment Analysis tasks at SemEval (where they constitute now a whole track) have attracted the highest number of participants in the last years (Rosenthal et al., 2014; Rosenthal et al., 2015; Nakov et al., 2016), and so it has been for the latest Evalita campaign, where a sentiment classification task (SENTIPOLC 2014) was introduced for the first time (Basile et al., 2014).

In addition to detecting the *polarity* of a tweet, it is also deemed important to detect whether a tweet is *subjective* or is merely reporting some fact, and whether some form of figurative mechanism, chiefly *irony*, is also present. Subjectivity, polarity, and irony detection form the three tasks of the SENTIPOLC 2016 campaign, which is a rerun of SENTIPOLC 2014.

Innovations with respect to SENTIPOLC 2014 While the three tasks are the same as those organised within SENTIPOLC 2014, we want to highlight the innovations that we have included in this year's edition. First, we have introduced two new annotation fields which express *literal polarity*, to provide insights into the mechanisms behind polarity shifts in the presence of figurative usage. Second, the test data is still drawn from Twitter, but it is composed of a portion of random tweets and a portion of tweets selected via keywords, which do not exactly match the selection procedure that led to the creation of the training set. This was intentionally done to observe the portability of supervised systems, in line with what observed in (Basile et al., 2015). Third, a portion of the data was annotated via Crowdflower rather than by experts. This has led to several observations on the quality of the data, and on the theoretical description of the task itself. Fourth, a portion

of the test data overlaps with the test data from three other tasks at Evalita 2016, namely PoST-WITA (Bosco et al., 2016), NEEL-IT (Basile et al., 2016a), and FactA (Minard et al., 2016). This was meant to produce a layered annotated dataset where end-to-end systems that address a variety of tasks can be fully developed and tested.

2 Task description

As in SENTIPOLC 2014, we have three tasks.

Task 1: Subjectivity Classification: *a system must decide whether a given message is subjective or objective* (Bruce and Wiebe, 1999; Pang and Lee, 2008).

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) and in accordance with (Basile et al., 2014), 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.[1] Section 3 provides further explanation and examples.

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 (Davidov et al., 2010; Hao and Veale, 2010; González-Ibáñez et al., 2011; Reyes et al., 2013; Reyes and Rosso, 2014), and platforms monitoring the sentiment in Twitter messages experienced the phenomenon of wrong polarity classification in ironic messages (Bosco et al., 2013; Ghosh et al., 2015). Indeed, ironic devices in a text can work as unexpected "polarity reversers" (one says something "good" to mean something "bad"), thus undermining 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 3.2) 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 without tackling Task 1.

[1] In accordance with (Wiebe et al., 2005).

3 Development and Test Data

Data released for the shared task comes from different datasets. We re-used the whole SENTIPOLC 2014 dataset, and also added new tweets derived from different datasets previously developed for Italian. The dataset composition has been designed in cooperation with other Evalita 2016 tasks, in particular the Named Entity rEcognition and Linking in Italian Tweets shared task (NEEL-IT, Basile et al. (2016a)). The multiple layers of annotation 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 (Basile et al., 2015). A portion of the data overlaps with data from NEEL-IT (Basile et al., 2016a), PoSTWITA (Bosco et al., 2016) and FacTA (Minard et al., 2016). See (Basile et al., 2016b) for details.

3.1 Corpora Description

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 are using for this shared task is a collection of tweets which is partially derived from two existing corpora, namely Sentipolc 2014 (TW-SENTIPOLC14, 6421 tweets) (Basile et al., 2014), and TWitterBuonaScuola (TW-BS) (Stranisci et al., 2016), 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 a set of 1500 tweets selected from the TWITA 2015 collection (TW-TWITA15, Basile and Nissim (2013)), the second one consists of 1000 (reduced to 989 after eliminating malformed tweets) tweets collected in the context of the NEEL-IT shared task (TW-NEELIT, Basile et al. (2016a)). The subsets of data extracted from existing corpora (TW-SENTIPOLC14 and TW-BS) have been revised according to the new annotation guidelines specifically devised for this task (see Section 3.3 for details).

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 which takes the "topic" information into account, we release it in case participants want to make use of it.

3.2 Annotation Scheme

Six fields contain values related to manual annotation are: `subj`, `opos`, `oneg`, `iro`, `lpos`, `lneg`.

The annotation scheme applied in SENTIPOLC 2014 has been enriched with two new fields, `lpos` and `lneg`, which encode the *literal* positive and negative polarity of tweets, respectively. Even if SENTIPOLC does not include any task which involves the actual 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 in ironic tweets. Indeed, in the presence of a figurative reading, the literal polarity of a tweet might differ from the intended overall polarity of the text (expressed by `opos` and `oneg`). Please note the following issues about our annotation scheme:

- An objective tweet will not have any polarity nor irony, thus if `subj` $= 0$, then `opos` $= 0$, `oneg` $= 0$, `iro` $= 0$, `lpos` $= 0$, and `lneg` $= 0$.
- A subjective, non ironic, tweet can exhibit at the same time *overall* positive *and* negative polarity (mixed polarity), thus `opos` $= 1$ and `oneg` $= 1$ can co-exist. Mixed *literal* polarity might also be observed, so that `lpos` $= 1$ and `lneg` $= 1$ can co-exist, and this is true for both non-ironic and ironic tweets.
- A subjective, non ironic, tweet can exhibit no specific polarity and be neutral but with a subjective flavor, thus `subj` $= 1$ and `opos` $= 0$, `oneg` $= 0$. Neutral *literal* polarity might also be observed, so that `lpos` $= 0$ and `lneg` $= 0$ is a possible combination; this is true for both non-ironic and ironic tweets.
- An ironic tweet is always subjective and it must have one defined polarity, so that `iro` $= 1$ cannot be combined with `opos` and `oneg` having the same value. However, mixed or neutral literal polarity could be observed for ironic tweets. Therefore, `iro` $= 1$, `lpos` $= 0$, and `lneg` $= 0$ can co-exist, as well as `iro` $= 1$, `lpos` $= 1$, and `lneg` $= 1$.
- For subjective tweets without irony (`iro` $= 0$), the overall (`opos` and `oneg`) and the literal (`lpos` and `lneg`) polarities are always annotated consistently, i.e. `opos` $=$ `lpos` and `oneg` $=$ `lneg`. Note that in such cases the literal polarity is implied automatically from the overall polarity and not annotated manually. The manual annotation of literal polarity only concerns tweets with `iro` $= 1$.

Table 1 summarises the allowed combinations.

3.3 Annotation procedure

Annotations for data from existing corpora (TW-BS and TW-SENTIPOLC14) have been revised and completed by exploiting an annotation procedure which involved a group of six expert annotators, in order to make them compliant to the SENTIPOLC 2016 annotation scheme. Data from NEEL-IT and TWITA15 was annotated from scratch using CrowdFlower. Both training and test data included a mixture of data annotated by experts and crowd. In particular, the whole TW-SENTIPOLC14 has been included in the development data release, while TW-BS was included in the test data release. Moreover, a set of 500 tweets from crowdsourced data was included in the test set, after a manual check and re-assessment (see below: *Crowdsourced data: consolidation of annotations*). This set 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 didn't 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. `iro` $= 1$ (806 tweets), we resorted to manual annotation: for each tweet, two independent annotations have been provided for the literal polarity dimension. The inter-annotator agreement at this stage was $\kappa = 0.538$. In a second round, a third independent annotation was provided to solve the disagreement. The final label

Table 1: Combinations of values allowed by our annotation scheme

subj	opos	oneg	iro	lpos	lneg	**description and *explanatory tweet in Italian***
0	0	0	0	0	0	objective *l'articolo di Roberto Ciccarelli dal manifesto di oggi* `http://fb.me/1BQVy5WAk`
1	0	0	0	0	0	subjective with neutral polarity and no irony *Primo passaggio alla #strabrollo ma secondo me non era un iscritto*
1	1	0	0	1	0	subjective with positive polarity and no irony *splendida foto di Fabrizio, pluri cliccata nei siti internazionali di Photo Natura* `http://t.co/GWoZqbxAuS`
1	0	1	0	0	1	subjective with negative polarity and no irony *Monti, ripensaci: l'inutile Torino-Lione inguaia l'Italia: Tav, appello a Mario Monti da Mercalli, Cicconi, Pont...* `http://t.co/3CazKS7Y`
1	1	1	0	1	1	subjective with both positive and negative polarity (mixed polarity) and no irony *Dati negativi da Confindustria che spera nel nuovo governo Monti. Castiglione: "Avanti con le riforme"* `http://t.co/kIKnbFY7`
1	1	0	1	1	0	subjective with positive polarity, and an ironic twist *Questo governo Monti dei paschi di Siena sta cominciando a carburare; speriamo bene...*
1	1	0	1	0	1	subjective with positive polarity, an ironic twist, and negative literal polarity *Non riesco a trovare nani e ballerine nel governo Monti. Ci deve essere un errore! :)*
1	0	1	1	0	1	subjective with negative polarity, and an ironic twist *Calderoli: Governo Monti? Banda Bassotti ..infatti loro erano quelli della Magliana.. #FullMonti #fuoritutti #piazzapulita*
1	0	1	1	1	0	subjective with negative polarity, an ironic twist, and positive literal polarity *Ho molta fiducia nel nuovo Governo Monti. Più o meno la stessa che ripongo in mia madre che tenta di inviare un'email.*
1	1	0	1	0	0	subjective with positive polarity, an ironic twist, and neutral literal polarity *Il vecchio governo paragonato al governo #monti sembra il cast di un film di lino banfi e Renzo montagnani rispetto ad uno di scorsese*
1	0	1	1	0	0	subjective with negative polarity, an ironic twist, and neutral literal polarity *arriva Mario #Monti: pronti a mettere tutti il grembiulino?*
1	1	0	1	1	1	subjective with positive polarity, an ironic twist, and mixed literal polarity *Non aspettare che il Governo Monti prenda anche i tuoi regali di Natale... Corri da noi, e potrai trovare IDEE REGALO a partire da 10e...*
1	0	1	1	1	1	subjective with negative polarity, an ironic twist, and mixed literal polarity *applauso freddissimo al Senato per Mario Monti. Ottimo.*

was assigned by majority vote on each field independently. With three annotators, this procedure ensures an unambiguous result for every tweet.

TW-BS The TW-BS section of the dataset had been previously annotated for polarity and irony[2]. 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 $\Rightarrow$ 110010; NEG $\Rightarrow$ 101001; MIXED $\Rightarrow$ 111011; NONE $\Rightarrow$ 000000[3]; HUMPOS $\Rightarrow$ 1101??; HUMNEG $\Rightarrow$ 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 (inter-annotator agreement $\kappa = 0.605$), 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 on each field.

TW-TWITA15 and TW-NEEL-IT For these new datasets, all fields were annotated from scratch using CrowdFlower (CF)[4], a crowdsourcing platform which has also been recently used for a similar annotation task (Nakov et al., 2016). 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 and their annotation according to the SENTIPOLC scheme. The guidelines also contained an explanation of the rules we followed for the annotation of the rest of the dataset, although in practice these constraints were not enforced in the CF

[2]For the annotation process and inter-annotator agreement see (Stranisci et al., 2016)

[3]Two independent annotators reconsidered the set of tweets tagged by NONE in order to distinguish the few cases of subjective, neutral, not-ironic tweets, i.e. 100000, as the original TW-BS scheme did not allow such finer distinction. The inter-annotator agreement on this task was measured as $\kappa = 0.841$ and a third independent annotation was used to solve the few cases of disagreement.

[4]`http://www.crowdflower.com/`

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., inter-rater agreement) 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).[5] On the basis of such observations and on a first analysis of the resulting combinations, 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. We automatically fixed the annotation for such cases, in order to release datasets of only tweets annotated with labels consistent with the SENTIPOLC's annotation scheme.[6]

Moreover, we applied a further manual check of crowdsourced data stimulated by the following observations. When comparing the distributions of values (0,1) for each label in both training and crowdsourced test 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, 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 basis, we decided to add two more expert annotations to the crowd-annotated test-set, and take the majority vote 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.

Table 2: Distribution of value combinations

combination						dev	test
subj	opos	oneg	iro	lpos	lneg		
0	0	0	0	0	0	2,312	695
1	0	0	0	0	0	504	219
1	0	1	0	0	1	1,798	520
1	0	1	1	0	0	210	73
1	0	1	1	0	1	225	53
1	0	1	1	1	0	239	66
1	0	1	1	1	1	71	22
1	1	0	0	1	0	1,488	295
1	1	0	1	0	0	29	3
1	1	0	1	0	1	22	4
1	1	0	1	1	0	62	8
1	1	0	1	1	1	10	6
1	1	1	0	1	1	440	36
total						7,410	2,000

3.4 Format and Distribution

We provided participants we a single development set, which consists of 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 final data format of the distribution is as follows: "`id`", "`subj`, "`opos`", "`oneg`", "`iro`", "`lpos`", "`lneg`", "`top`", "`text`".

The development data includes 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. In other words, the development data contains the six columns manually annotated,

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

[6] In particular, for CF data we applied two automatic transformations for restoring consistency of configurations of annotated values in cases where we observed a violation of the scheme: when at least a value 1 is present in the fields `opos`, `oneg`, `iro`, `lpos`, or `lneg`, we set the field `subj` accordingly: `subj`=0 ⇒ `subj`=1; when `iro`=0, the literal polarity value is overwritten by the overall polarity value.

[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.

while the test data will contain values only in the first (`idtwitter`) and last two columns (`top` and `text`). 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. The distribution of combinations in both development and test data is given in Table 2.

4 Evaluation

Task1: 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 (`subj`, `obj`):

$$p_{class} = \frac{\#correct_{class}}{\#assigned_{class}} \quad r_{class} = \frac{\#correct_{class}}{\#total_{class}}$$
$$F_{class} = 2\frac{p_{class}\, r_{class}}{p_{class} + r_{class}}$$

The overall F-score will be the average of the F-scores for subjective and objective classes.

Task2: 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):

$$p^{pos}_{class} = \frac{\#correct^{pos}_{class}}{\#assigned^{pos}_{class}} \quad r^{pos}_{class} = \frac{\#correct^{pos}_{class}}{\#total^{pos}_{class}}$$
$$p^{neg}_{class} = \frac{\#correct^{neg}_{class}}{\#assigned^{neg}_{class}} \quad r^{neg}_{class} = \frac{\#correct^{neg}_{class}}{\#total^{neg}_{class}}$$
$$F^{pos}_{class} = 2\frac{p^{pos}_{class}\, r^{pos}_{class}}{p^{pos}_{class} + r^{pos}_{class}} \quad F^{neg}_{class} = 2\frac{p^{neg}_{class}\, r^{neg}_{class}}{p^{neg}_{class} + r^{neg}_{class}}$$

The F-score for the two polarity classes is the average of the F-scores of the respective pairs:

$$F^{pos} = \frac{(F^{pos}_0 + F^{pos}_1)}{2} \quad F^{neg} = \frac{(F^{neg}_0 + F^{neg}_1)}{2}$$

Finally, the overall F-score for Task 2 is given by the average of the F-scores of the two polarities.

Task3: 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 Task1, but with different targeted classes. The overall F-score will be 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 will not affect in any way the final ranks for the three SENTIPOLC tasks.

5 Participants and Results

A total of 13 teams from 6 different countries participated in at least one of the three tasks of SENTIPOLC. Table 3 provides an overview of the teams, their affiliation, their country (C) and the tasks they took part in.

Table 3: Teams participating to SENTIPOLC 2016

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

Almost all teams participated to both subjectivity and polarity classification subtasks. Each team had to submit at least a constrained run. Furthermore, teams were allowed to submit up to four runs (2 constrained and 2 unconstrained) in

case they implemented different systems. Overall we have 19, 26, 12 submitted runs for the subjectivity, polarity, and irony detection tasks, respectively. In particular, three teams (UniPI, Unitor and tweet2check) participated with both a constrained and an unconstrained runs on the both the subjectivity and polarity subtasks. Unconstrained runs were submitted to the polarity subtask only by IntIntUniba.SentiPy and INGEOTEC.B4MSA. Differently from SENTIPOLC 2014, unconstrained systems performed better than constrained ones, with the only exception of UniPI, whose constrained system ranked first for the polarity classification subtask.

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. However, systems that are ranked midway might have excelled in precision for a given class or scored very bad in recall for another.[8]

For each task, we ran a majority class baseline to set a lower-bound for performance. In the tables it is always reported as *Baseline*.

5.1 Task1: subjectivity classification

Table 4 shows results for the subjectivity classification task, which attracted 19 total submissions from 10 different teams. The highest F-score is achieved by `Unitor` at 0.7444, which is also the best unconstrained performance. Among the constrained systems, the best F-score is achieved by samskara with $F = 0.7184$. All participating systems show an improvement over the baseline.

5.2 Task2: polarity classification

Table 5 shows results for polarity classification, the most popular subtask with 26 submissions from 12 teams. The highest F-score is achieved by UniPi at 0.6638, which is also the best score among the constrained runs. As for unconstrained runs, the best performance is achieved by Unitor with $F = 0.6620$. All participating systems show an improvement over the baseline.[9]

[8] Detailed scores for all classes and tasks are available at `http://www.di.unito.it/~tutreeb/sentipolc-evalita16/index.html`

[9] After the deadline, SwissCheese and tweet2check reported about a conversion error from their internal format to the official one. The resubmitted amended runs are shown in the table (marked by the * symbol), but the official ranking was not revised.

Table 4: Task 1: F-scores for constrained ".c" and unconstrained runs ".u". After the deadline, two teams reported about a conversion error from their internal format to the official one. The resubmitted amended runs are marked with *.

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_late	0.6536	0.7748	0.7142
*tweet2check16.u_late	0.4814	0.7820	0.6317

5.3 Task3: irony detection

Table 6 shows results for the irony detection task, which attracted 12 submissions from 7 teams. The highest F-score was achieved by tweet2check at 0.5412 (constrained run). The only unconstrained run was submitted by Unitor achieving 0.4810 as F-score. While all participating systems show an improvement over the baseline (F = 0.4688), many systems score very close to it, highlighting the complexity of the task.

6 Discussion

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.

Since we did not receive details about the systems adopted by some participants, i.e., tweet2check, ADAPT and UniBO, we are not including them in the following discussion. We consider however tweet2check's results in the discussion regarding irony detection.

Approaches based on Convolutional Neural

Table 5: Task 2: F-scores for constrained ".c" and unconstrained runs ".u". Amended runs are marked with * .

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_late	0.6529	0.7128	0.6828
*tweet2check16.u_late	0.6528	0.6373	0.6450

Table 6: Task 3: F-scores for constrained ".c" and unconstrained runs ".u". Amended runs are marked with *.

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_late	0.9355	0.1367	0.5361

Networks (CNN) have been investigated at SENTIPOLC this year for the first time by a few teams. Most of the other teams 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 over specific linguistic/semantic feature engineering, as discussed for example by ItaliaNLP, IRADABE, INGEOTEC or ColingLab. Other methods have been also used, as a Bayesian approach by samskara (achieving good results in polarity recognition) combined with linguistically motivated feature modelling. CoMoDi is the only participant that adopted a rule based approach in combination with a rich set of linguistic cues dedicated to irony detection.

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, as presented in SENTIPOLC 2014, 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 the 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.

Issues about the interdependency of tasks. Among the systems participating in more than one task, SwissCheese and Unitor designed systems that exploit the interdependency of specific subtasks. 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.

Irony detection. As also observed at SENTIPOLC 2014, irony detection appears truly challenging, as even the best performing system submitted by Tweet2Check ($F = 0.5412$) shows a low recall of 0.1710. We also observe that the performances of the supervised system developed by Tweet2Check and CoMoDi's rule-based approach, specifically tailored for irony detection, are very similar (Table 6).

While results seem to suggest that irony detection is the most difficult task, its complexity does not depend (only) on the inner structure of irony, but also on unbalanced data distribution (1 out of 7 examples is ironic in the training set). The classifiers are thus biased towards the non-irony class, and tend to retrieve all the non-ironic examples (high recall in the class non-irony) instead of actually modelling irony. If we measure the number of correctly predicted examples instead of the average of the two classes, the systems perform well (micro F1 of best system is 0.82).

Moreover, performance for irony detection drops significantly compared to SENTIPOLC 2014. 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 modelling topic rather than irony (Barbieri et al., 2015). 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.

7 Closing Remarks

All systems, except CoMoDI, exploited machine learning techniques in a supervised setting. Two main strategies emerged. One involves using linguistically principled approaches to represent tweets and provide the learning framework with valuable information to converge to good results. The other exploits state-of-the-art learning frameworks in combination with word embedding methods over large-scale corpora of tweets. On balance, the last approach achieved better results in the final ranks. However, with F-scores of 0.744 (unconstrained) and 0.7184 (constrained) in *subjectivity recognition* and 0.6638 (constrained) and 0.6620 (unconstrained) in *polarity recognition*, we are still far from having solved sentiment analysis on Twitter. For the future, we envisage the definition of novel approaches, for example by combining neural network-based learning with a linguistic-aware choice of features.

Besides modelling choices, *data* also matters. At this campaign we intentionally designed a test set with a sampling procedure that was close but not identical to that adopted for the training set (focusing again on political debates but on a different topic), so as to have a means to test the generalisation power of the systems (Basile et al., 2015). A couple of teams indeed reported substantial drops from the development to the official test set (e.g. IRADABE), and we plan to further investigate this aspect in future work. Overall, results confirm that sentiment analysis of micro-blogging is challenging, mostly due to the subjective nature of the phenomenon, and it's reflected in the inter-annotator agreement (Section 3.3). Crowdsourced data for this task also proved to be not entirely reliable, but this requires a finer-grained analysis on the collected data, and further experiments including a stricter implementation of the guidelines.

Although evaluated over different data, we see that this year's best systems show better, albeit comparable, performance for subjectivity with respect to 2014's systems, and outperform them for polarity (if we consider late submissions). For a proper evaluation across the various editions, we propose the use of a progress set for the next edition, as already done in the SemEval campaign.

References

Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2015. How Topic Biases Your Results? A Case Study of Sentiment Analysis and Irony Detection in Italian. In *RANLP, Recent Advances in Natural Language Processing*.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In *Proc. of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, Georgia, June.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proc. of EVALITA 2014*, pages 50–57, Pisa, Italy. Pisa University Press.

Pierpaolo Basile, Valerio Basile, Malvina Nissim, and Nicole Novielli. 2015. Deep tweets: from entity linking to sentiment analysis. In *Proc. of CLiC-it 2015*.

Pierpaolo Basile, Annalina Caputo, Anna Lisa Gentile, and Giuseppe Rizzo. 2016a. Overview of the EVALITA 2016 Named Entity rEcognition and Linking in Italian Tweets (NEEL-IT) Task. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Pierpaolo Basile, Franco Cutugno, Malvina Nissim, Viviana Patti, and Rachele Sprugnoli. 2016b. EVALITA 2016: Overview of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. 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.

Cristina Bosco, Fabio Tamburini, Andrea Bolioli, and Alessandro Mazzei. 2016. Overview of the EVALITA 2016 Part Of Speech on TWitter for ITAlian Task. In In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Rebecca F. Bruce and Janyce M. Wiebe. 1999. Recognizing Subjectivity: A Case Study in Manual Tagging. *Nat. Lang. Eng.*, 5(2):187–205, June.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proc. of CoNLL '10*, pages 107–116.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, Antonio Reyes, and Jhon Barnden. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proc. of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–475.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proc. of ACL–HLT '11*, pages 581–586, Portland, Oregon.

Yanfen Hao and Tony Veale. 2010. An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. *Minds Mach.*, 20(4):635–650.

Anne-Lyse Minard, Manuela Speranza, and Tommaso Caselli. 2016. The EVALITA 2016 Event Factuality Annotation Task (FactA). In Pierpaolo Basile, Anna Corazza, Franco Cutugno, Simonetta Montemagni, Malvina Nissim, Viviana Patti, Giovanni Semeraro and Rachele Sprugnoli, editors, *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. Associazione Italiana di Linguistica Computazionale (AILC).

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proc. of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18.

Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, January.

Antonio Reyes and Paolo Rosso. 2014. On the difficulty of automatically detecting irony: Beyond a simple case of negation. *Knowl. Inf. Syst.*, 40(3):595–614.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Lang. Resour. Eval.*, 47(1):239–268, March.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment Analysis in Twitter. In *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland, August.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 Task 10: Sentiment Analysis in Twitter. In *Proc. of the 9th International Workshop on Semantic Evaluation*, SemEval '2015.

Marco Stranisci, Cristina Bosco, Delia Iraz Hernndez Faras, and Viviana Patti. 2016. Annotating sentiment and irony in the online italian political debate on #labuonascuola. In *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2892–2899. ELRA.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2).