Evalita 2016 Sentipolc Task
Task Guidelines

Valerio Basile[1], Francesco Barbieri[2], Danilo Croce[3], Malvina Nissim[4], Nicole Novielli[5], and Viviana Patti[6]

[1]INRIA, France
[2]Pompeu Fabra University, Barcelona, Spain
[3]University of Rome "Tor Vergata", Rome, Italy
[4]University of Groningen, Groningen, The Netherlands
[5]University of Bari "A. Moro", Bari, Italy
[6]University of Torino, Italy

Contents

1 Task description

The main goal of Sentipolc is the sentiment analysis (SA) at message level of Italian tweets. The task is divided into three sub-tasks with an increasing level of complexity. Participants may choose to participate in one or more sub-tasks.

- **Task 1: Subjectivity Classification:** given a message, decide whether the message is subjective or objective.
- **Task 2: Polarity Classification:** given a message, decide whether the message is of positive, negative, neutral or mixed sentiment (i.e. conveying both a positive and negative sentiment).
- **Task 3: Irony Detection:** given a message, decide whether the message is ironic or not.

2 Development and Test Data

2.1 Corpora Description

The training data include 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).

NEW

With respect to the training data, the test material includes an additional topic. A tweet that was explicitly extracted with a socio-political topic (via specific hashtags and keywords different from the one used to collect the training material) will have topic = 2, otherwise topic = 0.

While Sentipolc does not include any task which takes this distinction into account, each tweet is marked with a political or non-political tag. In case participants want to make use of this information they are obviously free to do so, but should remember to mention this in the final description of their system.

We provide now a development set that participants can use to build their systems, while test set will be released in September 2016 (see Section 2.2 for details and the Sentipolc Web Page for updates: `http://di.unito.it/sentipolc16`).

2.2 Format and Distribution

A single development set will be provided, SentiDevSet henceforth. In particular, the distribution consists of a set of 7,410 tweets, with IDs and annotations concerning all three Sentipolc's subtasks: subjectivity classification (`subj`), polarity classification (`opos` and `oneg`) and irony detection (`iro`).

Notice that, with respect to the annotation adopted in Sentipolc 2014 [1], two additional fields (namely `lpos` and `lneg`) are reported to capture the *literal* polarity exhibited by a tweet. While Sentipolc does not include any task which takes the classification of literal polarity into account, `lpos` and `lneg` encode respectively the literal positive and negative polarity of tweets. This information is provided to enable participant to reason about the possible polarity inversion due to the use of figurative language in ironic tweets, thus the existing `lpos` and `lneg` fields refer to literal polarity of a tweet, which might differ from the intended overall polarity of the text expressed by `opos` and `oneg`.

It will be possible to to download the data at the following address: `http://www.di.unito.it/~tutreeb/sentipolc-evalita16/data.html`. We will provide soon also a web interface based on the use of RESTful Web API technology at `http://www.di.unito.it/~tutreeb/sentipolc-evalita16/tweets.html`.

The data format is as follows:

```
"idtwitter","subj","opos","oneg","iro","lpos","lneg", "top", "text"
```

idtwitter	Twitter status ID it is used by the API to fetch the actual tweet.
subj	Subjectivity: possible values are 0 and 1. A subjective tweet will have `subj` = 1; an objective tweet `subj` = 0.
opos	Positive *overall* polarity: possible values are 0 and 1. A tweet exhibiting positive polarity will have `opos` = 1; a tweet without positive polarity will have `opos` = 0.
oneg	Negative *overall* polarity: possible values are 0 and 1. A tweet exhibiting negative polarity will have `neg` = 1; a tweet without negative polarity will have `neg` = 0.
iro	Irony: possible values are 0 and 1. A tweet with an ironic twist will have `iro` = 1, otherwise `iro` = 0.
lpos	Positive *literal* polarity: possible values are 0 and 1. A tweet exhibiting positive *literal* polarity will have `pos` = 1; tweet without positive *literal* polarity will have `pos` = 0.
lneg	Negative *literal* polarity: possible values are 0 and 1. A tweet exhibiting negative *literal* polarity will have `neg` = 1; tweet without negative *literal* polarity will have `neg` = 0.
top	Topic: possible values are 0,1, and 2. A tweet that was explicitly extracted with a political topic (via specific hashtags and keywords) will have top = 1, A tweet that was explicitly extracted with a socio-political topic (using different hashtags and keywords) will have top = 2, otherwise top = 0.
text	Twitter message: this column is filled with the actual tweet's text.

The fields that contain values related to manual annotation are: `subj`, `opos`, `oneg`, `iro`, `lpos`, `lneg`. 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 just neutral but with a clear 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, and 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, that is tweets for which `iro` = 0, the overall polarity (`opos` and `oneg`) and the literal polarity (`lpos` and `lneg`) 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.

To sum up, the combinations in Table 1 are allowed in our annotation scheme.

Table 1: Combinations of values allowed by our annotation scheme

	`subj`	`opos`	`oneg`	`iro`	`lpos`	`lneg`	**description**
(a)	**0**	0	0	0	0	0	an objective tweet
(b)	**1**	0	0	0	0	0	a subjective tweet with neutral polarity and no irony
(c)	**1**	1	0	0	1	0	a subjective tweet with positive polarity and no irony
(d)	**1**	0	1	0	0	1	a subjective tweet with negative polarity and no irony
(e)	**1**	1	1	0	1	1	a subjective tweet with both positive and negative polarity (mixed polarity) and no irony
(f)	**1**	1	0	1	1	0	a subjective tweet with positive polarity, and an ironic twist
(g)	**1**	1	0	1	0	1	a subjective tweet with positive polarity, an ironic twist, and negative literal polarity
(h)	**1**	0	1	1	0	1	a subjective tweet with negative polarity, and an ironic twist
(i)	**1**	0	1	1	1	0	a subjective tweet with negative polarity, an ironic twist, and positive literal polarity
(j)	**1**	1	0	1	0	0	a subjective tweet with positive polarity, an ironic twist, and neutral literal polarity
(k)	**1**	0	1	1	0	0	a subjective tweet with negative polarity, an ironic twist, and neutral literal polarity
(l)	**1**	1	0	1	1	1	a subjective tweet with positive polarity, an ironic twist, and mixed literal polarity
(m)	**1**	0	1	1	1	1	a subjective tweet with negative polarity, an ironic twist, and mixed literal polarity

Examples for each combinations are provided in the Appendix. You can refer to the letters for matching examples in the Appendix and combinations in the Table.

The version of the data of the SentiDevSet 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 data for the test set (SentiTestSet henceforth) will only contain values for the `idtwitter` and `text` fields. In other words, the development data contains the six columns manually annotated, while the test data will contain values only in the first (`idtwitter`) and last (`text`) columns. The literal polarity might be predicted and used by participant to provide the final classification of the items in the test set, however this should be specified in the submission phase.

3 Submission format

Results for all tasks should be submitted in a plain text file with comma-separated fields. The format of the run files submitted by participants must be as follows:

```
"idtwitter","subj","opos","oneg","iro","lpos", "lneg", "top"
```

In the following, we report an example of a what a submitted run should look like. You can see in blue the values you will have to fill, and in black the values you have to include as inherited from SentiTestSet. The field "lpos" and "lneg" (in green, 5th and 6th columns after the `idtwitter`), denoting the literal polarity of the tweet, could also bee included in case you want to take part to the informal evaluation of the literal polarity classification (see the 'Evaluation' Section for details).

```
"<idtwitter>","0","0","0","0","0","0","0"
"<idtwitter>","1","0","1","1","0","0","0"
"<idtwitter>","1","0","0","0","0","0","0"
...
```

Specifically, submitted runs must contain one tweet per line including the original values provided in SentiTestSet for what concerns the `idtwitter` field, plus the annotations for the fields which are relevant w.r.t. the chosen task(s). In particular:

- **Task 1 - Subjectivity Classification:** we will consider relevant annotations for this task 0 or 1 values under the `subj` field (1st column after the `idtwitter`)
- **Task 2 - Polarity Classification:** we will consider relevant annotations for this task 0 or 1 values under the `pos` and `neg` fields (2nd and 3rd column after the `idtwitter`)
- **Task 3 - Irony detection:** we will consider relevant annotations for this task 0 or 1 values under the `iro` field (4th column after the `idtwitter`)

Note that each line should **NOT** include the tweet's text in your submission, as you can see from the example box above.

```
"<idtwitter>","","","","","0"
```

Number and types of runs For each task, we distinguish between constrained and unconstrained runs:

- for a **constrained** run teams must use the provided training data only; other resources, such as lexicons are allowed; however, it is not allowed to use additional training data in the form of tweets or sentences with sentiment annotations;
- for an **unconstrained** run teams can use additional data for training, e.g., additional tweets annotated for sentiment.

NEW

Please note that we will provide two saparate ranks for the constrained and unconstrained runs.

Important: if you take part in a given task, you **must** submit a constrained run, while the unconstrained one is optional. Each team may perform up to two submissions for the constrained run. Similarly, we allow a maximum of two submissions for the unconstrained run per team.

NEW

In the true spirit of research, we allow more than one submission for each run to encourage participants to experiment novel approaches as well as more traditional ones. However, we strongly recommend participants to submit two runs only if they implement substantially different approaches. If you rather want to fine-tune the performance of your system by varying the features included in your classifier, we suggest submitting only one run and to describe the feature engineering activity in the final report.

Notice that even if you take part in more than one task, your results will have to be included in **one file only** per run. For Task 1 only the second column (`subj` field) will be evaluated, for Task 2 columns 3 and 4 (`opos` and `oneg` fields), and for Task 3 column 5 (`iro` field). Thus, if a team decides to take part in all three tasks with both a constrained and an unconstrained setting, they will submit a total of six runs contained in **two files**: one including the constrained runs for all tasks, the other the unconstrained runs for all tasks. Teams will be asked to report what resources they have used for each run.

4 How to submit your runs

Once you have run your system over the test data you have downloaded, you will have to send it to us following these recommendations:

- choose a team name and name the files containing your runs in the following way:

  NEW

  - `sentipolc16.teamName.systemID.c` for the constrained runs
  - `sentipolc16.teamName.systemID.u` for the unconstrained runs

- send all relevant files to the following email address: `francesco.barbieri@upf.edu` using the subject "sentipolc — nameTeam"

- in the body of the email please specify all resources you used in the unconstrained run

Please note that SentiTestSet consists of 3000 tweets, so you can double check that your files are complete by verifying that you have the correct number of lines.

5 Evaluation

5.1 Task1: subjectivity classification

Systems will be evaluated on their assignment of a 0 or 1 value to the subjectivity field. A response will thus be considered plainly correct or wrong when compared to the gold standard annotation. We precision, recall and F-score for each class (`subj,obj`):

$$precision_{class} = \frac{\#correct\_class}{\#assigned\_class}$$

$$recall_{class} = \frac{\#correct\_class}{\#total\_class}$$

$$F_{class} = 2\frac{precision_{class} recall_{class}}{precision_{class} + recall_{class}}$$

The overall F-score will be the average of the F-scores for subjective and objective classes: $(F_{subj} + F_{obj})/2$

5.2 Task2: polarity classification

Our coding system allows for four combinations of `positive` (`opos`) and `negative` (`oneg`) values (see Guidelines for details), namely:

- 10: positive polarity
- 01: negative polarity
- 11: mixed polarity
- 00: no polarity

Accordingly with our scheme, we allow for *partial scoring* of system answers. Thus, each class (`opos,oneg`) will be evaluated independently via F-score, and the final score per tweet will be given by the average of the single F-scores. We chose to represent the final score as the F-score average of `opos` and `oneg` in accordance with SemEval's scoring system [3]. In SemEval it is only done over the whole corpus as there is only one possible class out of three that can be assigned to a tweet.

The per-tweet F-scores will be eventually averaged to get an overall score of the system. For example, the system in Table 2 over the shown corpus of nine tweets would get $F = 0.55$. Over a corpus of size n, the final F for a given system will be:

$$F = \frac{1}{n} \sum_{i=1}^{n} F_i$$

This corresponds to averaging F_{opos} and F_{oneg} over the whole corpus, as it's done in SemEval $((F_{pos} + F_{neg})/2)$, but as given above it's easier to see that we perform (partial) scoring of each tweet.

Table 2: Scoring per tweet, Task 2

gold	system	positive			negative			final score
		prec	rec	F	prec	rec	F	F_{tweet}
01	01	1.0	1.0	1.0	1.0	1.0	1.0	1.0
11	01	0.0	0.0	0.0	1.0	1.0	1.0	0.5
10	00	0.0	0.0	0.0	1.0	1.0	1.0	0.5
00	01	1.0	1.0	1.0	0.0	0.0	0.0	0.5
11	00	0.0	0.0	0.0	0.0	0.0	0.0	0.0
01	10	0.0	0.0	0.0	0.0	0.0	0.0	0.0
00	00	1.0	1.0	1.0	1.0	1.0	1.0	1.0
00	10	0.0	0.0	0.0	1.0	1.0	1.0	0.5
10	10	1.0	1.0	1.0	1.0	1.0	1.0	1.0

5.3 Task3: irony detection

Systems will be evaluated on their assignment of a 0 or 1 value to the irony field. A response will thus be considered fully correct or wrong when compared to the gold standard annotation. We will measure precision, recall and F-score for each class (`ironic,non-ironic`):

$$precision_{class} = \frac{\#correct\_class}{\#assigned\_class}$$

$$recall_{class} = \frac{\#correct\_class}{\#total\_class}$$

$$F_{class} = 2\frac{precision_{class} recall_{class}}{precision_{class} + recall_{class}}$$

The overall F-score will be the average of the F-scores for ironic and non-ironic classes: $(F_{ironic} + F_{non-ironic})/2$

5.4 Informal evaluation of literal polarity classification

Our coding system allows for four combinations of `positive` and `negative` values for literal polarity expressed by the text. The fields we used for coding this information are `lpos` and `lneg`. (see sections above for details), namely:

- 10: positive literal polarity
- 01: negative literal polarity
- 11: mixed literal polarity
- 00: neutral literal polarity

Sentipolc does not include any task that explicitly takes into account evaluation of literal polarity classification. However, it might be useful for participants to develop their system by also considering information about the literal polarity expressed by the text of a tweet. Participants can 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 about polarity classification. However, the performance on the literal polarity classification will not affect in any way the final ranks for the three Sentipolc task.

6 Final remarks

In order to download the data, we provide a web interface for downloading the tweet's text on the fly (see Section 2.2).

If you have any questions or problems, please start a topic on the googlegroups mailing list (`sentipolc-evalita2016@googlegroups.com`).

Sentipolc@Evalita2016 Web Page: `http://di.unito.it/sentipolc16`.

References

[1] V. Basile, A. Bolioli, M. Nissim, V. Patti, and P. Rosso. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *In Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA14)*, pages 5057, Pisa, Italy. Pisa University Press.

[2] V. Basile and M. Nissim. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

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

[4] C. Bosco, V. Patti, and A. Bolioli. 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.

Appendix: Examples of possible combinations

For a wordy explanation of classes and columns, please refer to Table 1. Reference letters match.

	subj	opos	oneg	iro	lpos	lneg	example
(a)	0	0	0	0	0	0	l'articolo di Roberto Ciccarelli dal manifesto di oggi http://fb.me/1BQVy5WAk
(b)	1	0	0	0	0	0	Primo passaggio alla #strabrollo ma secondo me non era un iscritto
(c)	1	1	0	0	1	0	splendida foto di Fabrizio, pluri cliccata nei siti internazionali di Photo Natura http://t.co/GWoZqbxAuS
(d)	1	0	1	0	0	1	Monti, ripensaci: l'inutile Torino-Lione inguaia l'Italia: Tav, appello a Mario Monti da Mercalli, Cicconi, Pont... http://t.co/3CazKS7Y
(e)	1	1	1	0	1	1	Dati negativi da Confindustria che spera nel nuovo governo Monti. Castiglione: "Avanti con le riforme" http://t.co/kIKnbFY7
(f)	1	1	0	1	1	0	Questo governo Monti dei paschi di Siena sta cominciando a carburare; speriamo bene...
(g)	1	1	0	1	0	1	Non riesco a trovare nani e ballerine nel governo Monti. Ci deve essere un errore! :)
(h)	1	0	1	1	0	1	Calderoli: Governo Monti? Banda Bassotti ..infatti loro erano quelli della Magliana.. #FullMonti #fuoritutti #piazzapulita
(i)	1	0	1	1	1	0	Ho molta fiducia nel nuovo Governo Monti. Più o meno la stessa che ripongo in mia madre che tenta di inviare un'email.
(j)	1	1	0	1	0	0	Il vecchio governo paragonato al governo #monti sembra il cast di un film di lino banfi e Renzo montagnani rispetto ad uno di scorsese
(k)	1	0	1	1	0	0	arriva Mario #Monti: pronti a mettere tutti il grembiulino?
(l)	1	1	0	1	1	1	Non aspettare che il Governo Monti prenda anche i tuoi regali di Natale... Corri da noi, e potrai trovare IDEE REGALO a partire da 10*e*...
(m)	1	0	1	1	1	1	applauso freddissimo al Senato per Mario Monti. Ottimo.