Evalita 2014 Sentipolc Task
Task Guidelines

Valerio Basile[1], Andrea Bolioli[2], Malvina Nissim[3], Viviana Patti[4], and Paolo Rosso[5]

[1]University of Groningen, The Netherlands
[2]CELI, Torino, Italy
[3]FICLIT, University of Bologna, Italy
[4]Dipartimento di Informatica, University of Torino, Italy
[5]NLEL, Universitat Politècnica de València, Spain

Contents

1 Task description

The main goal of Sentipolc is sentiment analysis (SA) at message level on 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. The first two are standard SA tasks, whereas the third one is a pilot task aimed at studying the presence of irony in tweets.

- **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 (Pilot Task):** Irony Detection: Given a message, decide whether the message is ironic or not.

2 Development and Test Data

2.1 Corpora Description

There are two main components of the data: a *generic* collection of tweets and a *political* collection of tweets. The latter has been extracted exploiting specific keywords and hashtags marking political topics, while the latter is composed of random tweets on any topic. 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 2014 (see Section 2.2 for details).

2.2 Format and Distribution

A single development set will be provided, SentiDevSet henceforth. In particular, the distribution consists of a set of 4,513 twitter status IDs, with annotations concerning all three Sentipolc's subtasks: subjectivity classification, polarity classification and irony detection.

In order to address privacy concerns and in compliance with Twitter's terms, rather than releasing the original tweet's text, we are providing a web interface based on the use of RESTful Web API technology to download the text: `http://www.di.unito.it/~tutreeb/sentipolc-evalita14/tweet.html`.

The data format is as follows:

```
"idtwitter","subj","pos","neg","iro","top","text"
```

where the field `text` is to be filled using the procedure available on the website mentioned above. (Notice that in cases where the tweet is no longer available, the `text` field is filled by the string: "Tweet Not Available", rather than by the text of the tweet.) The meaning of each field is described below:

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.
pos	Positive polarity: possible values are 0 and 1. A tweet exhibiting positive polarity will have `pos` = 1; a tweet without positive polarity will have `pos` = 0.
neg	Negative 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.
top	Topic: possible values are 0 and 1. A tweet that was explicitly extracted with a political topic (via specific hashtags and keywords) will have `top` = 1, otherwise `top` = 0.
text	Twitter message: this column will be filled with the actual tweet's text once you follow the population procedure provided by our web interface

The fields that contain values related to manual annotation are: `subj`, `pos`, `neg`, `iro`. Please note the following issues about our annotation scheme:

- An objective tweet will not have any polarity nor irony, thus if `subj` = 0, then `pos` = 0, `neg` = 0, and `iro` = 0.
- A subjective tweet can exhibit at the same time positive *and* negative polarity (mixed polarity), thus `pos` = 1 and `neg` = 1 can co-exist.
- A subjective tweet can exhibit no specific polarity and be just neutral but with a clear subjective flavor, thus `subj` = 1 and `pos` = 0 and `neg` = 0 is a possible combination.
- An ironic tweet is always subjective and it must have one defined polarity, so that `iro` = 1 cannot be combined with `pos` and `neg` having the same value.

To sum up, the following are the combinations that are allowed in our annotation scheme:

Table 1: Combinations of values allowed by other annotation scheme

subj	pos	neg	iro	**description**
0	0	0	0	an objective tweet
1	0	0	0	a subjective tweet with neutral polarity and no irony
1	1	0	0	a subjective tweet with positive polarity and no irony
1	0	1	0	a subjective tweet with negative polarity and no irony
1	1	1	0	a subjective tweet with both positive and negative polarity (mixed polarity) and no irony
1	1	0	1	a subjective tweet with positive polarity, and an ironic twist
1	0	1	1	a subjective tweet with negative polarity, and an ironic twist

Examples for each combinations are provided in the Appendix.

The version of the data of the SentiDevSet includes for each tweet the manual annotation for the `subj`, `pos`, `neg` and `iro` 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 `top` fields. In other words, the development data contains the first six columns annotated, while the test data will contain values only in the first (id) and last (topic) columns. In both cases, the `idtwitter` will allow to fetch the Twitter message.

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","pos","neg","iro","top"
```

This is 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:

```
"<idtwitter>","0","0","0","0","0"
"<idtwitter>","1","0","1","1","1"
"<idtwitter>","1","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` and `top` fields, 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`)

Number and types of runs For each task, each team may submit two runs:

- a **constrained** run - using 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;
- an **unconstrained** run - using additional data for training, e.g., additional tweets annotated for sentiment.

Important: if you take part in a given task, you **must** submit a constrained run, while the unconstrained one is optional. Notice that even if you take part in more than one task, your results will have to be included in **one file only** per type of run. For Task 1 only the second column (`subj` field) will be evaluated, for Task 2 columns 3 and 4 (`pos` and `neg` 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 Evaluation

4.1 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: $(F_{subj} + F_{obj})/2$

4.2 Task2: polarity classification

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

- 10: positive polarity
- 01: negative polarity
- 11: mixed polarity

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

- 00: no polarity

Accordingly with our scheme, we allow for *partial scoring* of system answers. Thus, each class (`pos,neg`) 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 `pos` and `neg` in accordance with SemEval's scoring system [1]. 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_{pos} and F_{neg} 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.

4.3 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 Final remarks

Due to Twitter's privacy policy we cannot provide tweets directly, but only the twitter status ID referring to them. You will have to download them yourself. For your convenience, we provide a web interface for downloading the tweet's text on the fly (see Section 2.2). However, notice that the text could be not available anymore for various reasons: Twitter users can delete their own posts anytime; their accounts can be temporarily suspended or deactivated. As a consequence, it is possible that the number of the available messages in the development dataset will vary over time. In order to deal with this issue, at submission time participants will be asked to equip their runs with the information about the number of tweets actually exploited from SentiDevSet.

If you have any questions or problems, please start a topic on the googlegroups mailing list (`sentipolc-evalita2014@googlegroups.com`).

References

[1] 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.

Appendix: Examples of possible combinations

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

0	0	0	0	l'articolo di Roberto Ciccarelli dal manifesto di oggi `http://fb.me/1BQVy5WAk`
1	0	0	0	Primo passaggio alla #strabrollo ma secondo me non era un iscritto
1	1	0	0	splendida foto di Fabrizio, pluri cliccata nei siti internazionali di Photo Natura `http://t.co/GWoZqbxAuS`
1	0	1	0	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	Dati negativi da Confindustria che spera nel nuovo governo Monti. Castiglione: "Avanti con le riforme" `http://t.co/kIKnbFY7`
1	1	0	1	Letta: sicuramente non farò parte del governo Monti . e siamo un passo avanti. #finecorsa
1	0	1	1	Botta di ottimismo a #lInfedele: Governo Monti, o la va o la spacca.