10.12871/clicit201429

Overview of the Evalita 2014 SENTIment POLarity Classification Task

Valerio Basile
University of Groningen
v.basile@rug.nl

Andrea Bolioli
CELI, Turin
abolioli@celi.it

Malvina Nissim
University of Groningen
University of Bologna
m.nissim@rug.nl

Viviana Patti
University of Turin
patti@di.unito.it

Paolo Rosso
Universitat Politècnica de València
prosso@dsic.upv.es

Abstract

English. The SENTIment POLarity Classification Task (SENTIPOLC), a new shared task in the Evalita evaluation campaign, focused on sentiment classification at the message level on Italian tweets. It included three subtasks: *subjectivity classification*, *polarity classification*, and *irony detection*. SENTIPOLC was the most participated Evalita task with a total of 35 submitted runs from 11 different teams. We present the datasets and the evaluation methodology, and discuss results and participating systems.

Italiano. *Descriviamo modalità e risultati della campagna di valutazione di sistemi di sentiment analysis (SENTIment POLarity Classification Task), proposta per la prima volta a "Evalita–2014: Evaluation of NLP and Speech Tools for Italian". In SENTIPOLC è stata valutata la capacità dei sistemi di riconoscere il* sentiment *espresso nei messaggi Twitter in lingua italiana. Sono stati proposti tre sotto-task:* subjectivity classification, polarity classification *e un sotto-task pilota di* irony detection. *La campagna ha suscitato molto interesse e ricevuto un totale di* 35 *run inviati da* 11 *gruppi di partecipanti.*

1 Introduction

The huge amount of information streaming from online social networking and micro-blogging platforms such as Twitter, is increasingly attracting the attention of researchers and practitioners. The fact that the over 30 teams participated in the Semeval 2013 shared task on Sentiment Analysis in English tweets (Nakov et al., 2013) is indicative in itself.

Several frameworks for detecting sentiments and opinions in social media have been developed for different application purposes, and Sentiment Analysis (SA) is recognized as a crucial tool in social media monitoring platforms providing business services. Extracting sentiments expressed in tweets has been used for several purposes: to monitor political sentiment (Tumasjan et al., 2011), to extract critical information during times of mass emergency (Verma et al., 2011), to detect moods and happiness in a given geographical area from geotagged tweets (Mitchell et al., 2013), and in several social media monitoring services.

Overall, the linguistic analysis of social media has become a relevant topic of research, naturally relying on resources such as sentiment annotated datasets, sentiment lexica, and the like. However, the availability of resources for languages other than English is usually rather scarce, and this holds for Italian as well (Basile and Nissim, 2013; Bosco et al., 2013). The organisation of the SENTIPOLC shared task, articulated in three sub-tasks, was thus aimed at providing reliably annotated data as well as promoting the development of systems towards a better understanding and processing of how sentiment is conveyed in tweets.

2 Task description

The main goal of SENTIPOLC is sentiment analysis at the message level on Italian tweets. We devised three sub-tasks, with increasing complexity.

Task 1: Subjectivity Classification: *a system must decide whether a given message is subjective or objective.*

This is a standard task on recognising whether a 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.*

Sentiments expressed in tweets are typically categorized as positive, negative or neutral, but a message can contain parts expressing both positive and negative sentiment (mixed sentiment). Differently from most SA tasks, chiefly the Semeval 2013 task, in our data positive and negative polarities are *not* mutually exclusive. This means that a tweet can 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 3.2 provides further explanation and examples.

Task 3 (Pilot): 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 et al., 2014), and platforms monitoring the sentiment in Twitter messages experienced the phenomenon of wrong polarity classification in ironic messages (Bosco et al., 2013). Indeed, the presence of ironic devices in a text can work as an unexpected "polarity reverser" (one says something "good" to mean something "bad"), thus undermining systems' accuracy. In order to investigate this issue, our dataset includes ironic messages, and we devised a pilot subtask concerning irony detection.

The three tasks are meant to be completely independent. For example, a team could take part in the polarity classification task, which only applies to subjective tweets, without tackling Task 1. For each task, each team could submit two runs:

- **constrained**: 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;
- **unconstrained**: using additional data for training, as more sentiment annotated tweets.

Participants willing to submit an unconstrained run for a given task were required to also submit a constrained run for the same task.

3 Development and Test Data

3.1 Corpora Description

The data that we are using for this shared task is a collection of tweets derived from two existing corpora, namely SENTI-TUT (Bosco et al., 2013) and TWITA (Basile and Nissim, 2013). Both corpora have been revised according to the new annotation guidelines specifically devised for this task (see Section 3.3 for details).

There are two main components of the data: a *generic* and a *political* collection. The latter has been extracted exploiting specific keywords and hashtags marking political topics, while the former is composed of random tweets on any topic. Each tweet is thus also marked with a "topic" tag.

A tweet is represented as a sequence of comma-separated fields, namely the Twitter id, the subjectivity field, the positive polarity field, the negative polarity field, the irony field, and the topic field. Apart from the id, which is a string of numeric characters, the value of all the other fields can be either "0" or "1". For the four classes to annotate, 0 and 1 mean that the feature is absent/present, respectively. For the topic field, 0 means "generic" and 1 means "political".

3.2 Manual annotation

The fields with manually annotated values are: `subj`, `pos`, `neg`, `iro`. While these classes could be in principle independent of each other, the following constraints hold in 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), 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 flavour, 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.

Table 1 summarises the combinations allowed in our annotation scheme. Information regarding manual annotation and the possible combinations was made available to the participants when the development set was released.

The SENTI-TUT section of the dataset was previously annotated for polarity and irony[2]. The tags

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

[2] For the annotation process and inter-annotator agreement for the TW–NEWS and TW–FELICITTA portions of

Table 1: Combinations of values allowed by our annotation scheme

`subj`	`pos`	`neg`	`iro`	**description**
0	0	0	0	an objective tweet example: *l'articolo di Roberto Ciccarelli dal manifesto di oggi* `http://fb.me/1BQVy5WAk`
1	0	0	0	a subjective tweet with neutral polarity and no irony example: *Primo passaggio alla #strabrollo ma secondo me non era un iscritto*
1	1	0	0	a subjective tweet with positive polarity and no irony example: *splendida foto di Fabrizio, pluri cliccata nei siti internazionali di Photo Natura* `http://t.co/GWoZqbxAuS`
1	0	1	0	a subjective tweet with negative polarity and no irony example: *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	a subjective tweet with positive and negative polarity (mixed polarity) and no irony example: *Dati negativi da Confindustria che spera nel nuovo governo Monti. Castiglione: "Avanti con le riforme"* `http://t.co/kIKnbFY7`
1	1	0	1	a subjective tweet with positive polarity, and an ironic twist example: *Letta: sicuramente non farò parte del governo Monti . e siamo un passo avanti. #finecorsa*
1	0	1	1	a subjective tweet with negative polarity, and an ironic twist example: *Botta di ottimismo a #lInfedele: Governo Monti, o la va o la spacca.*

POS, NEG, MIXED and NONE[3] in Senti–TUT were automatically mapped in the following values for the SENTIPOLC's `subj`, `pos`, `neg`, and `iro` annotation fields: POS ⇒ 1100; NEG ⇒ 1010; MIXED ⇒ 1110; NONE ⇒ 0000. However, the original Senti–TUT annotation scheme did only partially match the one proposed for this task, in particular regarding the ironic tweets, which were annotated just as HUM in SENTI–TUT, without polarity. Thus, for each tweet tagged as HUM (ca. 800 tweets), two annotators independently added the polarity dimension. The inter-annotator agreement at this stage was $\kappa = 0.259$. In a second round, a third annotator attempted to solve the disagreements (ca. 33%). Tweets where all three annotators had a different opinion (ca. 10%) were discussed jointly for the final label assignment. Note that all the HUM cases that showed no or mixed polarity were considered simply humorous rather than ironic, and marked as 1000 or 1110, respectively.

The TWITA section of the dataset had to be completely re-annotated, as irony annotation was missing, and the three labels adopted in the original data (positive, negative, and neutral, where neutral stood both for objective tweets and subjective tweets with mixed polarity, see (Basile and Nissim, 2013)), were not directly transferrable to the new scheme. The annotation was performed by four experts in three rounds. Round one saw two annotators independently mark each tweet. Inter-annotator agreement was measured at $\kappa = .482$ for Task 1, $\kappa = 0.678$ for positive labels and $\kappa = 0.638$ for negative labels in Task 2, and at $\kappa = 0.353$ for Task 3. In round two, a third annotator made a decision on the disagreements from round one, and in round three a fourth annotator had to decide on those cases where disagreements were left by the previous two rounds. Tweets where all four annotators had a different opinion amounted to just nine cases, and were discussed jointly for the final label assignment.

Finally, to ensure homogenous annotation over the whole dataset, annotators of one subset checked the annotation of the other. No divergences in the guidelines' interpretation surfaced.

3.3 Distribution and data format

Participants were provided with a development set (SentiDevSet henceforth), consisting of 4,513 tweets encoded as described in 3.2. The dataset is the same for all three subtasks.

Due to Twitter's privacy policy, tweets cannot be distributed directly, so participants were also provided with a web interface based on the use of RESTful Web API technology, through which they could download the tweet's text on the fly for all the ids provided.[4]

However, some tweets for which ids were distributed, might be not available anymore at download time for various reasons: Twitter users can

SENTI–TUT see (Bosco et al., 2013; Bosco et al., 2014).

[3] Four annotators collectively reconsidered the set of tweets tagged by NONE in order to distinguish the few cases of subjective, neutral, not-ironic tweets (1000). The original Senti–TUT scheme did not allow such finer distinction.

[4] `http://www.di.unito.it/~tutreeb/sentipolc-evalita14/tweet.html`.

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 were asked to equip their runs with the information about the number of tweets actually retrieved from SentiDevSet.

The format of the dataset provided by the Web interface is as follows:

"id","subj","pos","neg","iro","top","text"

where the field `text` is to be filled using the procedure available on the website mentioned above. 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 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) only contains values for the `idtwitter` and `top` fields. In other words, the development data contains the first six columns annotated, while the test data contains values only in the first (id) and last (topic) columns. In both cases, the `idtwitter` allows to fetch the Twitter message. The distribution of combinations in both datasets is given in Table 2.

Table 2: Distribution of labels in gold standard

combination	SentiDevSet	SentiTestSet
0 0 0 0	1276 (28%)	501 (26%)
1 0 0 0	270 (6%)	111 (6%)
1 0 1 0	1182 (26%)	546 (28%)
1 0 1 1	493 (11%)	209 (11%)
1 1 0 0	895 (20%)	425 (22%)
1 1 0 1	71 (2%)	27 (1%)
1 1 1 0	326 (7%)	116 (6%)
total	4513 (100%)	1935 (100%)

4 Evaluation

4.1 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, 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: 10 (positive polarity), 01 (negative polarity), 11 (mixed polarity), 00 (no polarity). Accordingly, we evaluate positive polarity and negative polarity independently by computing precision, recall and F-score for both classes (0 and 1):

$$precision^{pos}_{class} = \frac{\#correct^{pos}\_class}{\#assigned^{pos}\_class}$$

$$precision^{neg}_{class} = \frac{\#correct^{neg}\_class}{\#assigned^{neg}\_class}$$

$$recall^{pos}_{class} = \frac{\#correct^{pos}\_class}{\#total^{pos}\_class}$$

$$recall^{neg}_{class} = \frac{\#correct^{neg}\_class}{\#total^{neg}\_class}$$

$$F^{pos}_{class} = 2\frac{precision^{pos}_{class} recall^{pos}_{class}}{precision^{pos}_{class} + recall^{pos}_{class}}$$

$$F^{neg}_{class} = 2\frac{precision^{neg}_{class} recall^{neg}_{class}}{precision^{neg}_{class} + recall^{neg}_{class}}$$

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

$$F^{pos} = (F^{pos}_0 + F^{pos}_1)/2$$

$$F^{neg} = (F^{neg}_0 + F^{neg}_1)/2$$

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

$$F = (F^{pos} + F^{neg})/2$$

4.3 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`):

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

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

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

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 Participants and Results

A total of 11 teams from four different countries participated in at least one of the three tasks of SENTIPOLC. Table 3 provides an overview of the teams, their affiliation, and the number of tasks they took part in, with how many runs in total.

Almost all teams participated to both subjectivity and polarity classification subtasks. Most of the submissions were constrained: 9 out of 12 for subjectivity classification; 11 out of 14 for polarity classification; 7 out of 9 for irony detection. In particular, three teams (uniba2930,UNITOR,IRADABE) participated with both a constrained and an unconstrained run on the subtasks of interest. Unconstrained systems did not show to improve performance, but actually decreased it, with the exception of UNITOR's systems, whose unconstrained runs performed better than the constrained ones.

Because of the downloading procedure which we had to implement to comply to Twitter's policies (described in Sec. 3.3), not all teams necessarily tested their systems on the same set of tweets. Differences turned out to be minimal, but to ensure evaluation was performed over an identical dataset for all, we evaluated all participating systems on the union of their classified tweets, which amounted to 1734 (1930-196) [5].

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. Detailed scores for all classes and all tasks are available in the Appendix.

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 12 total submissions from 9 teams. The highest F-score was achieved by uniba2930 at 0.7140 (constrained run). All participating systems show an improvement over the baseline.

Table 4: Task 1: F-scores for constrained (F(C)) and unconstrained runs (F(U)).

rank	team	F(C)	F(U)
1	uniba2930	0.7140	0.6892
2	UNITOR	0.6871	0.6897
3	IRADABE	0.6706	0.6464
4	UPFtaln	0.6497	–
5	ficlit+cs@unibo	0.5972	–
6	mind	0.5901	–
7	SVMSLU	0.5825	–
8	fbkshelldkm	0.5593	–
9	itagetaruns	0.5224	–
10	baseline	0.4005	–

5.2 Task2: polarity classification

Table 5 shows results for the polarity classification task, which with 14 submissions from 11 teams was the most popular subtask. Again, the highest F-score was achieved by uniba2930 at 0.6771 (constrained). Also in this case, all participating systems show an improvement over the baseline.[6]

Table 5: Task 2: F-scores for constrained (F(C)) and unconstrained runs (F(U)).

rank	team	F(C)	F(U)
1	uniba2930	0.6771	0.6638
2	IRADABE	0.6347	0.6108
3	CoLingLab	0.6312	–
4	UNITOR	0.6299	0.6546
5	UPFtaln	0.6049	–
6	SVMSLU	0.6026	–
7	ficlit+cs@unibo	0.5980	–
8	fbkshelldkm	0.5626	–
9	mind	0.5342	–
10	itagetaruns	0.5181	–
11	Itanlp-wafi*	0.5086	–
12	baseline	0.3718	–
	*amended run	0.6637	–

5.3 Task3: irony detection

Table 6 shows results for the irony detection task, which attracted 9 submissions from 7 teams. The highest F-score was achieved by UNITOR at 0.5959 (unconstrained run) and 0.5759 (constrained run). While all participating systems show an improvement over the baseline, this time some systems score very close to it, highlighting the complexity of the task.

[5] It turned out that five of the 1935 tweets in SentiTestSet were duplicates.

[6] After the task deadline, the Itanlp-wafi team reported about an error of the conversion script from their internal format to the official one. They submitted, then, the correct run. Official ranking was not revised, but the evaluation of the correct run is shown in the table (marked by star symbol).

Table 3: Teams participating to SENTIPOLC

team	institution	country	tasks	runs
CoLingLab	CoLing Lab – University of Pisa	IT	T2	1
IRADABE	U Politecnica de Valencia / U Paris 13	ES/FR	T1,T2,T3	6
SVMSLU	Minsk State Linguistic University	BY	T1,T2,T3	3
UNITOR	University of Roma Tor Vergata	IT	T1,T2,T3	6
UPFtaln	TALN – Universitat Pompeu Fabra	ES	T1,T2,T3	3
fbkshelldkm	Fondazione Bruno Kessler (FBK-IRST)	IT	T1,T2,T3	3
ficlit+cs@unibo	FICLIT-University of Bologna	IT	T1,T2	2
italianlp-wafi	ItaliaNLP Lab – ILC (CNR)	IT	T2	1
itgetaruns	Ca' Foscari University – Venice	IT	T1,T2,T3	3
mind	University of Milano-Bicocca	IT	T1,T2,T3	3
uniba2930	CS – University of Bari	IT	T1,T2	4

Table 6: Task 3: F-scores for constrained (F(C)) and unconstrained runs (F(U)).

rank	team	F(C)	F(U)
1	UNITOR	0.5759	0.5959
2	IRADABE	0.5415	0.5513
3	SVMSLU	0.5394	–
4	itagetaruns	0.4929	–
5	mind	0.4771	–
6	fbkshelldkm	0.4707	–
7	UPFtaln	0.4687	–
8	baseline	0.4441	–

6 Discussion and Conclusions

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

Most participants restricted themselves to the provided data and submitted constrained systems. Only three teams submitted uconstrained 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 necessary 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.

As noticed also in the context of similar evaluation campaigns for the English language (Nakov et al., 2013; Rosenthal et al., 2014), most systems used supervised learning (uniba2930, mind, IRADABE, UNITOR, UPFtaln, SVMSLU, itanlp-wafi, CoLingLab, fbkshelldkm). The most popular algorithm was SVM, 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 were used, including word-based, syntactic and semantic (mostly lexicon-based) features. The best team in Task1 and Task2, uniba2930, specifically mentions that in leave-one-out experiments, (distributional) semantic features 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 Task3, 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.

Almost all participants relied on various sentiment lexicons. At least six teams (uniba2930, UPFtaln, fbkshelldkm, ficlit+cs@unibo, UNITOR, IRADABE) used information from Senti-

WordNet (Esuli et al., 2010), either using the already existing Sentix (Basile and Nissim, 2013) or otherwise. 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 were used for pre-processing, such as tokenisers, POS-taggers, and parsers.

The majority of systems participating in more than one subtask adopted classification strategies including some form of interdependency among the tasks, with different directions of dependency.

Overall, through a first comparative analysis of the systems' behaviour which we can only briefly summarise here due to space constraints, we can make some observations related to aspects specific to the SENTIPOLC tasks. First, ironic expressions do appear to play the role of polarity reversers, undermining the accuracy of sentiment classifiers. Second, recognising mixed sentiment (tweets tagged as 1110) was hard for our participants, even harder than recognising neutral subjectivity (tweets tagged as 1000). Further and deeper investigations will be matter of future work.

To conclude, the fact that SENTIPOLC was the most popular Evalita 2014 task is indicative of the great interest of the NLP community on sentiment analysis in social media, also in Italy.

Acknowledgments

We would like to thank Manuela Sanguinetti, Cristina Bosco, and Marco Del Tredici for their help in annotating the dataset, and Sergio Rabellino (ICT staff, Dipartimento di Informatica, Turin) for his precious technical support. The last author gratefully acknowledges the support of EC WIQ-EI IRSES (Grant No. 269180) and MICINN DIANA-Applications (TIN2012-38603-C02-01).

References

V. Basile and M. Nissim. 2013. Sentiment analysis on Italian tweets. In *Proc. of WASSA 2013*, pages 100–107, NAACL 2013, Atlanta, Georgia.

C. Bosco, V. Patti, and A. 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.

C. Bosco, L. Allisio, V. Mussa, V. Patti, G. Ruffo, M. Sanguinetti, and E. Sulis. 2014. Detecting happiness in Italian tweets: Towards an evaluation dataset for sentiment analysis in Felicittà. In B. Schuller et al., editors, *Proc. of ESSSLOD 2014*, pages 56–63, LREC 2014, Reykjavik, Iceland.

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

D. Davidov, O. Tsur, and A. Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proc. of CoNLL '10*, pages 107–116, Stroudsburg, PA, USA.

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

R. González-Ibáñez, S. Muresan, and N. Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proc. ACL-HLT'11 - Short Papers - Volume 2*, pages 581–586, Stroudsburg, PA, USA.

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

L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth. 2013. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5), 05.

P. Nakov, S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in Twitter. In *Proc. of SemEval 2013*, pages 312–320.

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

A. Reyes, P. Rosso, and T. Veale. 2013. A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, 47(1):239–268.

A. Reyes and P. Rosso. 2014. On the Difficulty of Automatically Detecting Irony: Beyond a Simple Case of Negation. *Knowledge and Information Systems*, 40(3):595–614.

S. Rosenthal, A. Ritter, P. Nakov, and V. Stoyanov. 2014. Semeval-2014 Task 9: Sentiment analysis in Twitter. In *Proc. of SemEval 2014*, pages 73–80, Dublin, Ireland.

A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. 2011. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proc. of ICWSM-11*, pages 178–185, Barcelona, Spain.

S. Verma, S. Vieweg, W. Corvey, L. Palen, J. H. Martin, M. Palmer, A. Schram, and K. M. Anderson. 2011. Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency. In *Proc. of the 5th International AAAI Conference on Weblogs and Social Media*, 385–392.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Appendix: Detailed results per class for all tasks

Results of task 1

run	rank	Combined F-score	Prec. (0)	Rec. (0)	F-score (0)	Prec. (1)	Rec. (1)	F-score (1)	team
Constrained	1	0.7140	0.6976	0.5271	0.6005	0.8498	0.8064	0.8275	uniba2930
	2	0.6871	0.5768	0.5872	0.5819	0.8582	0.7358	0.7923	UNITOR
	3	0.6706	0.6247	0.4669	0.5344	0.8284	0.7862	0.8067	IRADABE
	4	0.6497	0.6565	0.3868	0.4868	0.8099	0.8155	0.8127	UPFtaln
	5	0.5972	0.4512	0.4449	0.4480	0.8029	0.6974	0.7464	ficlit+cs@unibo
	6	0.5901	0.4115	0.6473	0.5031	0.8484	0.5632	0.6770	mind
	7	0.5825	0.4363	0.4048	0.4200	0.7917	0.7037	0.7451	SVMSLU
	8	0.5593	0.3791	0.5311	0.4424	0.8050	0.5828	0.6761	fbkshelldkm
	9	0.5224	0.3479	0.3026	0.3237	0.7571	0.6883	0.7211	itagetaruns
	10	0.4005	0.0000	0.0000	0.0000	0.7308	0.8861	0.8010	baseline
Unconstrained	1	0.6897	0.6062	0.5491	0.5762	0.8496	0.7617	0.8032	UNITOR
	2	0.6892	0.6937	0.4629	0.5553	0.8317	0.8148	0.8232	uniba2930
	3	0.6464	0.4729	0.7335	0.5750	0.8955	0.5989	0.7178	IRADABE

Results of task 2

			Positive polarity							Negative polarity							
run	**rank**	**Combined F-score**	**Prec. (0)**	**Rec. (0)**	**F-score (0)**	**Prec. (1)**	**Rec. (1)**	**F-score (1)**	**F-score**	**Prec. (0)**	**Rec. (0)**	**F-score (0)**	**Prec. (1)**	**Rec. (1)**	**F-score (1)**	**F-score**	**team**
Constrained	1	0.6771	0.8102	0.8364	0.8231	0.7195	0.4162	0.5274	0.6752	0.7474	0.6890	0.7170	0.6882	0.5995	0.6408	0.6789	uniba2930
	2	0.6347	0.7782	0.8547	0.8147	0.7265	0.2998	0.4245	0.6196	0.7067	0.7107	0.7086	0.6822	0.5213	0.5910	0.6498	IRADABE
	3	0.6312	0.7976	0.7806	0.7890	0.5810	0.4109	0.4814	0.6352	0.6923	0.6701	0.6810	0.6384	0.5201	0.5732	0.6271	CoLingLab
	4	0.6299	0.7949	0.7704	0.7824	0.5604	0.4092	0.4730	0.6277	0.7225	0.6013	0.6564	0.6138	0.6018	0.6078	0.6321	UNITOR
	5	0.6049	0.7782	0.8004	0.7892	0.5766	0.3386	0.4267	0.6079	0.6804	0.6079	0.6421	0.5909	0.5351	0.5616	0.6019	UPFtaln
	6	0.6026	0.7943	0.7337	0.7628	0.5126	0.4303	0.4679	0.6153	0.6627	0.6239	0.6427	0.5856	0.4960	0.5371	0.5899	SVMSLU
	7	0.5980	0.8223	0.5943	0.6899	0.4373	0.5785	0.4981	0.5940	0.6546	0.7663	0.7060	0.6876	0.3901	0.4978	0.6019	ficlit+cs@unibo
	8	0.5626	0.7511	0.8525	0.7986	0.6277	0.2081	0.3126	0.5556	0.6573	0.5495	0.5986	0.5472	0.5339	0.5405	0.5695	fbkshelldkm
	9	0.5342	0.7403	0.7528	0.7465	0.4097	0.2522	0.3122	0.5293	0.6141	0.6089	0.6115	0.5300	0.4166	0.4665	0.5390	mind
	10	0.5181	0.7297	0.8158	0.7703	0.4313	0.1605	0.2339	0.5021	0.6097	0.7700	0.6805	0.6203	0.2819	0.3877	0.5341	itagetaruns
	11	0.5086	0.8106	0.4365	0.5675	0.3636	0.6420	0.4643	0.5159	0.7722	0.2620	0.3913	0.4989	0.7894	0.6114	0.5013	Itanlp-wafi*
	12	0.3718	0.7101	0.9039	0.7954	0.0000	0.0000	0.0000	0.3977	0.5573	0.9114	0.6917	0.0000	0.0000	0.0000	0.3459	baseline
		0.6637	0.8144	0.8048	0.8096	0.6521	0.4462	0.5298	0.6697	0.7287	0.6682	0.6971	0.6614	0.5800	0.6180	0.6576	*amended run
Unconstrained	1	0.6638	0.8189	0.7696	0.7935	0.5969	0.4780	0.5309	0.6622	0.7400	0.6654	0.7007	0.6658	0.5984	0.6303	0.6655	uniba2930
	2	0.6546	0.8212	0.7748	0.7973	0.6080	0.4815	0.5374	0.6673	0.7378	0.5994	0.6615	0.6208	0.6237	0.6223	0.6419	UNITOR
	3	0.6108	0.8204	0.6266	0.7105	0.4565	0.5556	0.5012	0.6058	0.6822	0.6635	0.6727	0.6266	0.5040	0.5587	0.6157	IRADABE

Results of task 3

run	rank	Combined F-score	Prec. (0)	Rec. (0)	F-score (0)	Prec. (1)	Rec. (1)	F-score (1)	team
Constrained	1	0.5759	0.9312	0.6956	0.7963	0.2675	0.5294	0.3554	UNITOR
	2	0.5415	0.8967	0.7849	0.8371	0.2400	0.2521	0.2459	IRADABE
	3	0.5394	0.8990	0.7630	0.8254	0.2274	0.2857	0.2533	SVMSLU
	4	0.4929	0.8829	0.7754	0.8257	0.1566	0.1639	0.1602	itagetaruns
	5	0.4771	0.8933	0.6235	0.7344	0.1570	0.3655	0.2197	mind
	6	0.4707	0.8766	0.7931	0.8328	0.1176	0.1008	0.1086	fbkshelldkm
	7	0.4687	0.8795	0.8889	0.8842	0.2800	0.0294	0.0532	UPFtaln
	8	0.4441	0.8772	0.8995	0.8882	0.0000	0.0000	0.0000	baseline
Unconstrained	1	0.5959	0.9208	0.7630	0.8345	0.3063	0.4286	0.3573	UNITOR
	2	0.5513	0.9139	0.7086	0.7983	0.2387	0.4202	0.3044	IRADABE