Overview of the EVALITA 2018 Task on Irony Detection in Italian Tweets (IronITA)

Alessandra Teresa Cignarella
Simona Frenda
Dipartimento di Informatica
Università degli Studi di Torino, Italy
PRHLT Research Center
Universitat Politècnica de València, Spain
{cigna,frenda}@di.unito.it

Valerio Basile, Cristina Bosco
Viviana Patti
Dipartimento di Informatica
Università degli Studi di Torino, Italy
{basile,bosco,patti}@di.unito.it

Paolo Rosso
PRHLT Research Center
Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Abstract

English. IronITA is a new shared task in the EVALITA 2018 evaluation campaign, focused on the automatic classification of irony in Italian texts from Twitter. It includes two tasks: 1) irony detection and 2) detection of different types of irony, with a special focus on sarcasm identification. We received 17 submissions for the first task and 7 submissions for the second task from 7 teams.

Italiano. *IronITA è un nuovo esercizio di valutazione della campagna di valutazione EVALITA 2018, specificamente dedicato alla classificazione automatica dell'ironia presente in testi estratti da Twitter. Comprende due task: 1) riconoscimento dell'ironia e 2) riconoscimento di diversi tipi di ironia, con particolare attenzione all'identificazione del sarcasmo. Abbiamo ricevuto 17 sottomissioni per il primo task e 7 per il secondo, da parte di 7 gruppi partecipanti.*

1 Introduction

Irony is a figurative language device that conveys the opposite of literal meaning, profiling intentionally a secondary or extended meaning. Users on the web usually tend to use irony like a creative device to express their thoughts in short-texts like tweets, reviews, posts or commentaries. But irony, as well as other figurative language devices, for example metaphors, is very difficult to deal with automatically. For its traits of recalling another meaning or obfuscating the real communicative intention, it hinders correct sentiment analysis of texts and, therefore, correct opinion mining. 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.

Considering the majority of state-of-the-art studies in computational linguistics, *irony* is often used as an umbrella-term which includes satire, sarcasm and parody due to fuzzy boundaries among them (Marchetti et al., 2007). However, some linguistic studies focused on *sarcasm*, a particular type of verbal irony defined in Gibbs (2000) as "a sharp or cutting ironic expression with the intent to convey scorn or insult". Other scholars concentrated on cognitive aspects related on how such figurative expressions are processed in the brain, focusing on key aspects influencing processing (see for instance the "defaultness" hypothesis presented in Giora et al. (2018)).

The importance to detect irony and sarcasm is also very relevant for reaching better predictions in Sentiment Analysis, for instance, what are the real opinion and orientation of users about a specific subject (product, service, topic, issue, person, organization, or event).

IronITA is organized in continuity with previous shared tasks of the past years within the context of the EVALITA evaluation campaign (see for instance the irony detection subtask proposed at SENTIPOLC in the 2014 and 2016 editions (Basile et al., 2014; Barbieri et al., 2016)). It is also inspired by the recent experience within the SemEval2018-Task3 *Irony detection in English tweets* (Van Hee et al., 2018). The shared task we propose for Italian is specifically dedicated to

irony detection taking into account both the classical binary classification task (irony vs not irony), and a related subtask, which gives to participants the possibility to reason on different types of irony. Differently from SemEval2018-Task3, we indeed ask the participants to distinguish sarcasm as a specific type of irony. This is motivated by the growing interest for detecting sarcasm, which is characterized by sharp tones and aggressive intention (Gibbs, 2000; Joshi et al., 2017; Sulis et al., 2016) often present in interesting domains such as politics and hate speech (Sanguinetti et al., 2018).

2 Task Description

The task consists in automatically annotating messages from Twitter for irony and sarcasm. It is organized in a main task (Task A) centered on irony, and a second task (Task B) centered on sarcasm, whose results will be separately evaluated. Participation was allowed to both the tasks (Task A and Task B) or to Task A only.

Task A: Irony Detection. Task A consists in a two-class (or binary) classification where systems have to predict whether a tweet is ironic or not.

Task B: Different types of irony with special focus on sarcasm identification. Sarcasm has been recognized in Bowes and Katz (2011) with a specific target to attack (Attardo, 2007; Dynel, 2014), more offensive and delivered with a cutting tone (rarely ambiguous). According to Lee and Katz (1998) hearers perceive aggressiveness as the feature that distinguishes sarcasm. Provided a definition of sarcasm as a specific type of irony, Task B consists in a multi-class classification where systems have to predict one out of the three following labels: **i) sarcasm**, **ii) irony not categorized as sarcasm** (i.e. other kinds of verbal irony or descriptions of situational irony which do not show the characteristics of sarcasm), and **iii) not-irony**.

The proposed tasks encourage the investigation of this linguistic devices. Moreover, providing a dataset from social media (Twitter), we focus on texts especially hard to be dealt with, because of their shortness and because they will be analyzed out of the context where they were generated.

The participants are allowed to submit either "constrained" or "unconstrained" runs (or both, within the submission limits). The constrained runs have to be produced by systems whose only training data is the dataset provided by the task organizers. On the other hand, the participant teams are encouraged to train their systems on additional annotated data and submit the resulting unconstrained runs.

We implemented two straightforward baseline systems for the task. *baseline-mfc* (Most Frequent Class) assigns to each instance the majority class of the respective task, namely `not-ironic` for task A and `not-sarcastic` for task B. *baseline-random* assigns uniformly random values to the instances. Note that for task A, a class is assigned randomly to every instance, while for task B the classes are assigned randomly only to eligible tweets who are marked `ironic`.

3 Training and Test Data

3.1 Composition of the datasets

The data released for the shared task come from different source datasets, namely: Hate Speech Corpus (HSC) (Sanguinetti et al., 2018) and the TWITTIRÒ corpus (Cignarella et al., 2018), composed of tweets from LaBuonaScuola corpus (TW-BS) (Stranisci et al., 2016), Sentipolc corpus (TW-SENTIPOLC), Spinoza corpus (TW-SPINO) (Barbieri et al., 2016).

In the test data we have the same sources, and in addition some tweets from the TWITA collection, that were annotated by the organizers of the SENTIPOLC 2016 shared task, but were not exploited during the 2016 campaign (Barbieri et al., 2016).

3.2 Annotation of the datasets

The annotation process involved four Italian native speakers and focused only on the finer-grained annotation of sarcasm in the ironic tweets, since the presence of irony was already annotated in the source datasets. It began by splitting in two halves the dataset and assigning the annotation task for each portion to a different couple of annotators. In the following step, the final inter-annotator agreement (IAA) has been calculated on all the dataset. Then, in order to achieve an agreement on a larger portion of data, the effort of the annotators has been focused only on the detected cases of disagreement. In particular, the couple previously involved in the annotation of the first half of the corpus produced a new annotation for the tweets in disagreement of the second portion of the dataset, while the couple involved in the annotation of the second half of the corpus did the same on the first

	TRAINING SET				TEST SET				
	IRONIC	NOT-IRO	SARC	NOT-SARC	IRONIC	NOT-IRO	SARC	NOT-SARC	TOTAL
TW-BS	467	646	173	294	111	161	51	60	
TW-SPINO	342	0	126	216	73	0	32	41	2,886
TW-SENTIPOLC	461	625	143	318	0	0	0	0	
HSC	753	683	471	282	185	119	106	79	1,740
TWITA	0	0	0	0	67	156	28	39	223
TOTAL	**3,977**				**872**				**4,849**

Table 1: Distribution of tweets according to the topic

portion of the dataset. After that, the cases where the disagreement persists have been discarded as too ambiguous to be classified (131 tweets).

The final IAA calculated with Fleiss' kappa is $\kappa = 0.56$ for the tweets belonging to the TWITTIRÒ corpus and $\kappa = 0.52$ for the data from the HSC corpus and it is considered *moderate*[1] and satisfying for the purpose of the shared task.

In this process the annotators relied on a specific definition of "sarcasm", and followed detailed guidelines[2]. In particular we defined **sarcasm** as *a kind of sharp, explicit and sometimes aggressive irony, aimed at hitting a specific target to hurt or criticize without excluding the possibility of having fun* (Du Marsais et al., 1981; Gibbs, 2000). The factors we have taken into account for the annotation are, the presence of:

1. a clear **target**,
2. an obvious **intention** to hurt or criticize,
3. **negativity** (weak or strong).

We have also tried to differentiate our concept of "sarcasm" from that of "satire", often present in tweets. For us, satire aims to ridicule the target as well as criticize it. Differently from sarcasm, satire is solely focused on a more negative type of criticism and moved by a personal and angry emotional charge.

A single training set has been provided for both tasks A and B, which includes 3,977 tweets. Following, a single test set has been distributed for both tasks A and B, which includes 872 tweets, hence creating an $82\% - 18\%$ balance between training and test data. Table 1 shows the distribution of ironic and sarcastic tweets among the different source/topic datasets cited in Section 3.1.

Additionally the IronITA datasets overlap with the data released for *HaSpeeDe*, the task of Hate Speech Detection (Bosco et al., 2018). In the training set we count 781 overlapping tweets, while in the test set we count an overlap of just 96 tweets.

3.3 Data Release

The data were released in the following format[3]:

```
idtwitter text irony sarcasm topic
```

where `idtwitter` is the Twitter ID of the message, `text` is the content of the message, `irony` is 1 or 0 (respectively for ironic and not ironic tweets), `sarcasm` is 1 or 0 (respectively for sarcastic and not sarcastic tweets), and `topic` refers to the source corpus from where the tweet has been extracted.

The training set includes for each tweet the annotation for the `irony` and `sarcasm` fields, according to the format explained above. Instead, the test set only containes values for the `idtwitter`, `text` and `topic` fields.

4 Evaluation Measures

Task A: Irony detection. Systems have been evaluated against the gold standard test set on their assignment of a 0 or 1 value to the `irony` field. We measured the precision, recall and F1-score of the prediction for both the `ironic` and `not-ironic` classes:

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

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

$$F1_{class} = 2\frac{precision_{class} recall_{class}}{precision_{class} + recall_{class}}$$

The overall F1-score is the average of the F1-scores for the `ironic` and `not-ironic` classes (i.e. macro F1-score).

[1] According to the parameters proposed by Fleiss (1971).

[2] For more details on this regard, please refer to the guidelines: `https://github.com/AleT-Cig/IronITA-2018/blob/master/Definition%20of%20Sarcasm.pdf`

[3] Link to the datasets: `http://www.di.unito.it/~tutreeb/ironita-evalita18/data.html`

topic	irony	sarcasm	**text**
TWITTIRÒ	0	0	*@SteGiannini @sdisponibile Semmai l'anno DELLA buona scuola. De la, in italiano, non esiste*
TWITTIRÒ	1	1	*#labuonascuola Fornitura illimitata di rotoli di carta igienica e poi, piano piano, tutti gli altri aspetti meno importanti.*
HSC	1	0	*Di fronte a queste forme di terrorismo siamo tutti sulla stessa barca. A parte Briatore. Briatore ha la sua.*
HSC	1	1	*Anche oggi sono in arrivo 2000migranti dalla Libia avanti in italia ce posto per tutti vero @laraboldrini ? Li puoi accogliere a casa tua*

Table 2: Examples for each combinations

Task B: Different types of irony. Systems have been evaluated against the gold standard test set on their assignment of a 0 or 1 value to the `sarcasm` field, assuming that the `irony` field is also provided as part of the results.

We have measured the precision, recall and F1-score for each of the three classes:

- not-ironic
  `irony` = 0, `sarcasm` = 0
- ironic-not-sarcastic
  `irony` = 1, `sarcasm` = 0
- sarcastic
  `irony` = 1, `sarcasm` = 1

The evaluation metric is the macro F1-score computed over the three classes. Note that for the purpose of the evaluation of task B, the following combination is always considered wrong:

- `irony` = 0, `sarcasm` = 1

Our scheme imposes that a tweet can be annotated as sarcastic only if it is also annotated as ironic, which correspond to interpreting sarcasm as a specific type of irony, as reported in Table 2.

5 Participants and Results

A total of 7 teams, both from academia and industry sector participated to at least one of the two tasks of IronITA. Table 3 provides an overview of the teams, their affiliation, and the tasks they took part in.

Four teams participated to both tasks A and B. Teams were allowed to submit up to four runs (2 constrained and 2 unconstrained) in case they implemented different systems. Furthermore, each team had to submit at least a constrained run. Participants have been invited to submit multiple runs to experiment with different models and architectures. However, they have been discouraged from submitting slight variations of the same model. Overall we have 17 runs for Task A and 7 runs for Task B.

5.1 Task A: Irony Detection

Table 4 shows the results for the irony detection task, which attracted 17 total submissions from 7 different teams. The best scores are achieved by the ItaliaNLP team (Cimino et al., 2018) that, with a constrained run, obtained the best score for both the `ironic` and `not-ironic` class, thus obtaining the highest averaged F1-score of 0.731.

Among the unconstrained systems, the best F1-score for the `not-ironic` class is achieved by the X2Check team (Di Rosa and Durante, 2018) with F = 0.708, and the best F1-score for the `ironic` class is obtained by the UNITOR team (Santilli et al., 2018) with F = 0.733.

All participating systems show an improvement over the baselines, with the exception of the only unsupervised system (venses-itgetarun, see details in Section 6).

team name	**id**	**F1-score**		
		not-iro	**iro**	**macro**
ItaliaNLP	1	**0.707**	**0.754**	**0.731**
ItaliaNLP	2	0.693	0.733	0.713
UNIBA	1	0.689	0.730	0.710
UNIBA	2	0.689	0.730	0.710
X2Check	1	**0.708**	0.700	**0.704**
UNITOR	1	0.662	0.739	0.700
UNITOR	2	0.668	**0.733**	0.700
X2Check	2	0.700	0.689	0.695
Aspie96	1	0.668	0.722	0.695
X2Check	2	0.679	0.708	0.693
X2Check	1	0.674	0.693	0.683
UO_IRO	2	0.603	0.700	0.651
UO_IRO	1	0.626	0.665	0.646
UO_IRO	2	0.579	0.678	0.629
UO_IRO	1	0.652	0.577	0.614
baseline-random		0.503	0.506	0.505
venses-itgetarun	1	0.651	0.289	0.470
venses-itgetarun	2	0.645	0.195	0.420
baseline-mfc		0.668	0.000	0.334

Table 4: Results Task A. Unconstrained runs are marked by grey background.

5.2 Task B: Different types of irony

Table 5 shows the results for the different types of irony task, which attracted 7 total submis-

team name	institution	tasks
ItaliaNLP	ItaliaNLP group ILC-CNR	A,B
UNIBA	University of Bari	A
X2Check	App2Check srl	A
UNITOR	University of Roma "Tor Vergata"	A,B
Aspie96	University of Torino	A,B
UO_IRO	CERPAMID, Santiago de Cuba / University of Informatics Sciences, Havana	A
venses-itgetarun	Ca' Foscari University of Venice	A,B

Table 3: Participants

sions from 4 different teams. The best scores are achieved by the UNITOR team that with an unconstrained run obtained the highest macro F1-score of 0.520.

Among the constrained systems, the best F1-score for the `not-ironic` class is achieved by the ItaliaNLP team with F1-score = 0.707, and the best F1-score for the `ironic` class is obtained by the Aspie96 team (Giudice, 2018) with F1-score = 0.438. The best score for the `sarcastic` class is obtained by a constrained run of the UNITOR team with F1-score = 0.459. The best performing UNITOR team is also the only team that participated to Task B with an unconstrained run.

team name	id	F1-score			
		not-iro	iro	sarc	macro
UNITOR	2	**0.668**	**0.447**	**0.446**	**0.520**
UNITOR	1	0.662	0.432	**0.459**	**0.518**
ItaliaNLP	1	**0.707**	0.432	0.409	0.516
ItaliaNLP	2	0.693	0.423	0.392	0.503
Aspie96	1	0.668	**0.438**	0.289	0.465
baseline-random		0.503	0.266	0.242	0.337
venses-itgetarun	1	0.431	0.260	0.018	0.236
baseline-mfc		0.668	0.000	0.000	0.223
venses-itgetarun	2	0.413	0.183	0.000	0.199

Table 5: Results Task B. Unconstrained runs are marked by grey background.

All participating systems show an improvement over the baselines, with the exception of the only unsupervised system (venses-itgetarun, see details in Section 6).

6 Discussion

We compare the participating systems according to the following main dimensions: classification framework (approaches, algorithms, features), text representation strategy, use of additional annotated data for training, external resources (e.g. sentiment lexica, NLP tools, etc.), and interdependency between the two tasks. This discussion is based on the information contained in the reports submitted by the participants (we received 6 reports out of 7 participating teams) and on the answers to a questionnaire sent by the organizers to the participants.

System architecture. Most submitted runs to IronITA are produced by supervised machine learning systems. In fact, all but one systems are supervised, although the nature and complexity of their architectures varies significantly. UNIBA (Basile and Semeraro, 2018) and UNITOR use Support Vector Machine (SVM) classifiers, with different parameter settings. UNITOR, in particular, employs a multiple kernel-based approach to create two SVM classifiers that work on the two tasks. X2Check uses several models based on Multinomial Naive Bayes and SVM in a voting ensemble. Three systems implemented deep learning neural networks for the classification of irony and sarcasm. Sequence-learning networks were a popular choice, in the form of Bidirectional Long Short-term Memory Networks (used by ItaliaNLP and UO_IRO (Ortega-Bueno and Medina Pagola, 2018)) and Gated Recurrent Units (Aspie96). The venses-itgetarun team proposed the only unsupervised system submitted to IronITA. The system is based on an extension of the ITGETARUN rule-based fully symbolic semantic parser (Delmonte, 2014). The performance of the venses-itgetarun system is penalized mainly by its low recall (see the detailed results on the task website).

Features. In addition to explore a broad spectrum of supervised and unsupervised architectures, the submitted systems leverage different kinds of linguistic and semantic information extracted from the tweets. Word n-grams of varying size are used by ItaliaNLP, UNIBA, and X2Check. Word embeddings were used as features by three systems, namely ItaliaNLP (built with word2vec on a concatenation of ItWaC[4] and a custom tweet corpus), UNITOR (built with

[4] `https://www.sketchengine.eu/itwac-italian-corpus/`

word2vec on a custom Twitter corpus) and UNIBA (built with Random Indexing (Sahlgren, 2005)) on a subset of TWITA (Basile et al., 2018). Affective lexicons were also employed to extract polarity-related features from the words in the tweets, by UNIBA, ItaliaNLP and UNITOR and UO_IRO (see the "Lexical Resources" section for details on the lexica). UNIBA and UO_IRO also computed sentiment variation and contrast in order to extract the ironic content from the text. Features derived from sentiment analysis are also employed by the unsupervised system venses-itgetarun. Aspie96 performs its classification based on the single characters of the tweet. Finally, a great number of other features is employed by the systems, including stylistic and structural features (UO_IRO), special tokens and emoticons (X2Check). See the details in the EVALITA proceedings (Caselli et al., 2018).

Lexical Resources. Several systems employed affective resources, mainly as a tool to compute the sentiment polarity of words and each tweet. ItaliaNLP used two affective lexica generated automatically by means of distant supervision and automatic translation. UNIBA used an automatic translation of SentiWordNet (Esuli and Sebastiani, 2006). UNITOR used the Distributed Polarity Lexicon by Castellucci et al. (2016). UO_IRO used the affective lexicon derived from the OpeNER project (Russo et al., 2016) and a polarity lexicon of emojis by Kralj Novak et al. (2015). venses-itgetarun used several lexica, including some specifically built for ITGETARUNS and a translation of SentiWordNet (Esuli and Sebastiani, 2006).

Additional training data. Three teams took the opportunity to send unconstrained runs along with constrained runs. X2Check included in the unconstrained training set a balanced version of the SENTIPOLC 2016 dataset, Italian tweets annotated with irony (Barbieri et al., 2016). UNITOR used for their unconstrained runs a dataset of 6,000 tweets obtained by distant supervision (searching for the hashtag #ironia — #irony). UO_IRO employed tweets annotated with fine-grained irony from TWITTIRÒ (Cignarella et al., 2018).

The team ItaliaNLP did not send unconstrained runs, although they used the information about polarity of Italian tweets from the SENTIPOLC 2016 dataset (Barbieri et al., 2016) and the data annotated for hate speech from the HaSpeeDe task at EVALITA 2018 (Bosco et al., 2018). We do not consider their runs unconstrained, because the phenomena annotated in the data they employed are different from irony.

Interdependency of tasks. Since the tasks A and B are inherently linked (a tweet can be sarcastic only if it is also ironic), some of the participating teams leveraged this information in their classification systems. ItaliaNLP employed a Multi-task learning approach, thus solving the two tasks simultaneously. UNITOR adopted a cascade architecture where only tweets that were classified as ironic were passed through to the sarcasm classifier. In the system by venses-itgetarun, the decision on whether to assign a tweet to `sarcasm` or `irony` is based on the contemporary presence of features common to the two tasks.

7 Concluding remarks

Differently from the previous sub-tasks on irony detection in Italian language proposed as part of the previous SENTIPOLC shared tasks, having Sentiment Analysis as reference framework, the IronITA tasks specifically focus on the irony and sarcasm identification.

Comparing the results for irony detection obtained within the SENTIPOLC sub-task (the best performing system in the 2016 edition reached $F = 0.5412$ and in 2014 $F = 0.575$) with the ones obtained in IronITA, it is worth to notice that a dedicated task on irony detection leaded to a remarkable improvement of the scores, with the highest value here being $F = 0.731$.

Surprisingly, scores for Italian are in line with those obtained at SemEval2018-Task3 on irony detection in English tweets, even if a lower amount of linguistic resources is available for Italian than for English, especially in term of affective lexica, a type of resource that is frequently exploited in this kind of task. Actually, some teams used resources provided by the Italian NLP community also in the framework of previous EVALITA's edition (e.g. additional information from annotated corpora as SENTIPOLC, HaSpeeDe and POST-WITA).

The good results obtained in this edition can be read also as a confirmation that linguistic resources for Italian language are increasing in quantity and quality, and they are helpful also for a very challenging task as irony detection.

Another interesting factor in this edition is the use of the innovative deep learning techniques, mirroring the growing interest in deep learning by the NLP community at large. Indeed, the best performing system is based on a deep learning approach revealing its usefulness also for irony detection. The high performance of deep learning methods is an indication that irony and sarcasm are phenomena involving more complex features than n-grams and lexical polarity.

The number of participants in task B was lower. Even though we wanted to encourage the investigation in the identification of sarcasm, we are aware that addressing the finer-grained task to discriminate between irony and sarcasm is still really difficult.

In hindsight, the organization of such a shared task, specifically dedicated to irony detection in Italian tweets, and also focused on diverse types of irony has been a hazard. It was intended to foster research teams in the exploitation of lexical and affective resources in Italian, developed in our NLP community and to encourage the investigation especially on data about politics and immigration.

Our proposal for this shared task arose from the intuition that a better recognition of figurative language like irony in social media data could also lead to a better resolution of other Sentiment Analysis tasks such as Hate Speech Detection (Bosco et al., 2018), Stance Detection (Mohammad et al., 2017), and Misogyny Detection (Fersini et al., 2018). IronITA wanted to be a first try-out and a first stimulus in this challenging field.

Acknowledgments

V. Basile, C. Bosco and V. Patti were partially supported by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media-IhatePrejudice*, S1618_L2_BOSC_01). The work of S.Frenda and P. Rosso was partially funded by the Spanish research project SomEMBED TIN2015-71147-C2-1-P (MINECO/FEDER).

References

Salvatore Attardo. 2007. Irony as relevant inappropriateness. In H. Colston and R. Gibbs, editors, *Irony in language and thought: A cognitive science reader*, pages 135–172. Lawrence Erlbaum.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the Evalita 2016 sentiment polarity classification task. In *Proceedings of 3rd Italian Conference on Computational Linguistics (CLiC-it 2016) & 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, Naples, Italy. CEUR.org.

Pierpaolo Basile and Giovanni Semeraro. 2018. UNIBA - Integrating distributional semantics features in a supervised approach for detecting irony in Italian tweets. In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 sentiment polarity classification task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, Pisa, Italy. Pisa University Press.

Valerio Basile, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term Social Media Data Collection at the University of Turin. In *Proceedings of the 5th Italian Conference on Computational Linguistics (CLiC-it 2018)*, Turin, Italy. CEUR.org.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the Evalita 2018 Hate Speech Detection Task. In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Andrea Bowes and Albert Katz. 2011. When sarcasm stings. *Discourse Processes: A Multidisciplinary Journal*, 48(4):215–236.

Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. EVALITA 2018: Overview of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In *Proceedings of 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2016. A language independent method for generating large scale polarity lexicons. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. ELRA.

Alessandra Teresa Cignarella, Cristina Bosco, Viviana Patti, and Mirko Lai. 2018. Application and Analysis of a Multi-layered Scheme for Irony on the Italian Twitter Corpus TWITTIRÒ. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. ELRA.

Andrea Cimino, Lorenzo De Mattei, and Felice Dell'Orletta. 2018. Multi-task Learning in Deep Neural Networks at EVALITA 2018. In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Rodolfo Delmonte. 2014. A linguistic rule-based system for pragmatic text processing. In *Proceedings of Fourth International Workshop EVALITA 2014*, Pisa. Edizioni PLUS, Pisa University Press.

Emanuele Di Rosa and Alberto Durante. 2018. Irony detection in tweets: X2Check at Ironita 2018. In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

César Chesneau Du Marsais, Jean Paulhan, and Claude Mouchard. 1981. *Traité des tropes*. Le Nouveau Commerce.

Marta Dynel. 2014. Linguistic approaches to (non) humorous irony. *Humor - International Journal of Humor Research*, 27(6):537–550.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genova, Italy.

Elisabetta Fersini, Maria Anzovino, and Paolo Rosso. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval. In *Proceedings of 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*. CEUR-WS.org.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*.

Raymond W. Gibbs. 2000. Irony in talk among friends. *Metaphor and symbol*, 15(1-2):5–27.

Rachel Giora, Adi Cholev, Ofer Fein, and Orna Peleg. 2018. On the superiority of defaultness: Hemispheric perspectives of processing negative and affirmative sarcasm. *Metaphor and Symbol*, 33(3):163–174.

Valentino Giudice. 2018. Aspie96 at IronITA (EVALITA 2018): Irony Detection in Italian Tweets with Character-Level Convolutional RNN. In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Aditya Joshi, Pushpak Bhattacharyya, and Mark James Carman. 2017. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50(5):73:1–73:22.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PLOS ONE*, 10(12):1–22, 12.

Christopher J. Lee and Albert N. Katz. 1998. The differential role of ridicule in sarcasm and irony. *Metaphor and Symbol*, 13(1):1–15.

A. Marchetti, D. Massaro, and A. Valle. 2007. *Non dicevo sul serio. Riflessioni su ironia e psicologia*. Collana di psicologia. Franco Angeli.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):26.

Reynier Ortega-Bueno and José E. Medina Pagola. 2018. UO_IRO: Linguistic informed deep-learning model for irony detection. In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Irene Russo, Francesca Frontini, and Valeria Quochi. 2016. OpeNER sentiment lexicon italian - LMF. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa.

Magnus Sahlgren. 2005. An introduction to random indexing. In *In Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Andrea Santilli, Danilo Croce, and Roberto Basili. 2018. A Kernel-based Approach for Irony and Sarcasm Detection in Italian. In *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Marco Stranisci, Cristina Bosco, Delia Irazú Hernández Farías, and Viviana Patti. 2016. Annotating Sentiment and Irony in the Online Italian Political Debate on #labuonascuola. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. ELRA.

Emilio Sulis, D. Irazú Hernández Farías, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo. 2016. Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*, 108:132 – 143. New Avenues in Knowledge Bases for Natural Language Processing.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*.

Appendix: Detailed results per class for all tasks

ranking	team name	run type	run id	precision (non-ironic)	recall (non-ironic)	F1-score (non-ironic)	precision (ironic)	recall (ironic)	F1-score (ironic)	average F1-score
1	ItaliaNLP	c	1	0.785	0.643	0.707	0.696	0.823	0.754	0.731
2	ItaliaNLP	c	2	0.751	0.643	0.693	0.687	0.786	0.733	0.713
3	UNIBA	c	1	0.748	0.638	0.689	0.683	0.784	0.730	0.710
4	UNIBA	c	2	0.748	0.638	0.689	0.683	0.784	0.730	0.710
5	X2Check	u	1	0.700	0.716	0.708	0.708	0.692	0.700	0.704
6	UNITOR	c	1	0.778	0.577	0.662	0.662	0.834	0.739	0.700
7	UNITOR	u	2	0.764	0.593	0.668	0.666	0.816	0.733	0.700
8	X2Check	u	2	0.690	0.712	0.700	0.701	0.678	0.689	0.695
9	Aspie96	c	1	0.742	0.606	0.668	0.666	0.789	0.722	0.695
10	X2Check	c	2	0.716	0.645	0.679	0.676	0.743	0.708	0.693
11	X2Check	c	1	0.697	0.652	0.674	0.672	0.715	0.693	0.683
12	UO_IRO	u	2	0.722	0.517	0.603	0.623	0.800	0.700	0.651
13	UO_IRO	u	1	0.667	0.590	0.626	0.631	0.703	0.665	0.646
14	UO_IRO	c	2	0.687	0.501	0.579	0.606	0.770	0.678	0.629
15	UO_IRO	c	1	0.600	0.714	0.652	0.645	0.522	0.577	0.614
16	*baseline-random*	*c*	*1*	*0.506*	*0.501*	*0.503*	*0.503*	*0.508*	*0.506*	*0.505*
17	venses-itgetarun	c	1	0.520	0.872	0.651	0.597	0.191	0.289	0.470
18	venses-itgetarun	c	2	0.505	0.892	0.645	0.525	0.120	0.195	0.420
19	*baseline-mfc*	*c*	*1*	*0.501*	*1.000*	*0.668*	*0.000*	*0.000*	*0.000*	*0.334*

Detailed results of Task A (Irony Detection)

ranking	team name	run type	run id	precision (non-ironic)	recall (non-ironic)	F1-score (non-ironic)	precision (ironic)	recall (ironic)	F1-score (ironic)	precision (sarcastic)	recall (sarcastic)	F1-score (sarcastic)	average F1-score
1	UNITOR	u	2	0.764	0.593	0.668	0.362	0.584	0.447	0.492	0.407	0.446	0.520
2	UNITOR	c	1	0.778	0.577	0.662	0.355	0.553	0.432	0.469	0.449	0.459	0.518
3	ItaliaNLP	c	1	0.785	0.643	0.707	0.343	0.584	0.432	0.518	0.338	0.409	0.516
4	ItaliaNLP	c	2	0.751	0.643	0.693	0.340	0.562	0.423	0.507	0.319	0.392	0.503
5	Aspie96	c	1	0.742	0.606	0.668	0.353	0.575	0.438	0.342	0.250	0.289	0.465
6	*baseline-random*	*c*	*1*	*0.506*	*0.501*	*0.503*	*0.267*	*0.265*	*0.266*	*0.239*	*0.245*	*0.242*	*0.337*
7	venses-itgetarun	c	1	0.606	0.334	0.431	0.341	0.210	0.260	0.500	0.009	0.018	0.236
8	*baseline-mfc*	*c*	*1*	*0.501*	*1.000*	*0.668*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.223*
9	venses-itgetarun	c	2	0.559	0.327	0.413	0.296	0.132	0.183	0.000	0.000	0.000	0.199

Detailed results of Task B (Sarcasm Detection)