UINAUIL: A Unified Benchmark for Italian Natural Language Understanding

UINAUIL: A Unified Benchmark for Italian Natural Language Understanding

Valerio Basile*, Livio Bioglio*, Alessio Bosca**, Cristina Bosco*, and Viviana Patti*

*University of Turin, Italy, email: name.surname@unito.it
*H-FARM, email: alessio.bosca@h-farm.com

Abstract

This paper introduces the Unified Interactive Natural Understanding of the Italian Language (UINAUIL), a benchmark of six tasks for Italian Natural Language Understanding. We present a description of the tasks and software library that collects the data from the European Language Grid, harmonizes the data format, and exposes functionalities to facilitates data manipulation and the evaluation of custom models. We also present the results of tests conducted with available Italian and multilingual language models on UINAUIL, providing an updated picture of the current state of the art in Italian NLU.

Video: https://www.youtube.com/watch?v=rZWKl9cPTbk

1 Introduction

Large Language Models (LLM) have revolutionized the field of Natural Language Processing. In the span of a few years, the common practice for most NLP tasks shifted from building ad-hoc models trained on task-specific data to fine-tuning general-purpose language models trained in a self-supervised fashion. The focus of the evaluation practices shifted accordingly, from measuring the impact of different features and neural architectures on the prediction performance, to assessing the predictive power of LLMs applied to a variety of NLP tasks. This has been possible, at least in part, due to the standardization proposed in Devlin et al. (2019), where four general task formats are claimed to represent the structure of most NLP tasks: text classification, sentence pair classification, sequence labeling, and question answering.

In this scenario, benchmarks have emerged that collect a number of tasks falling into the four mentioned categories, with the purpose of evaluating LLMs in a fair and reproducible environment. Perhaps the best known of such benchmarks is GLUE (Wang et al., 2018, Language Understanding Evaluation), a set of nine sentence classification and sentence pair classification Natural Language Understanding tasks. SuperGLUE (Wang et al., 2019) was presented not long after GLUE with the goal of proposing a harder set of NLU tasks, given the high performance reached by LLMs on GLUE not long after its release.

While English is covered by benchmarks such as GLUE and SuperGLUE, the situation differs sensibly when we turn to other languages, with only a few NLU benchmarks available for non-English languages (Shavrina et al., 2020; Kakwani et al., 2020; Xu et al., 2020; Wilie et al., 2020; Adesam et al., 2020). These are useful resources for the international NLP community, and a great example of language equality (Rehm et al., 2021). However, these benchmarks are mostly static collection of datasets with no additional tool to facilitate data gathering and management, evaluation, and automation in general (except for leaderboards, which are usually maintained by the respective authors).

In this paper, we present UINAUIL (Unified Interactive Natural Understanding of the Italian Language), an integrated benchmark for Italian NLU, with three main goals:

- G1 Filling the gap in Italian NLU evaluation in the era of LLMs by proposing one integrated benchmark as opposed to a myriad of individual shared task datasets.
- G2 Raising the bar on the automation level of NLU benchmarks, in order to create more accessible and user-friendly benchmarks.
- G3 Set the example via a use case to encourage scholars and NLP practitioners to publish modern, integrated benchmarks for underrepresented languages.

2 A Unified Benchmark for Italian NLP

UINAUIL is a set of six tasks originally proposed as shared tasks for evaluation of Italian NLP. In this

section, we describe the background of the tasks and how they are integrated into UINAUIL.

2.1 EVALITA

In order to select a set of NLU tasks to include into UINAUIL, we analysed the archives of EVALITA. Started in 2007, the "Evaluation campaign for Language Technology in Italian" has been organized every two or three years, with the latest edition currently ongoing in 2023[1]. EVALITA provides a common framework of shared tasks, where participating systems are evaluated on a wide variety of NLP tasks. The number of tasks of EVALITA has grown over time, from 5 tasks in the first edition in 2007, to the 14 tasks of the 2020 edition (Passaro et al., 2020). At the same time, the nature of the proposed tasks has evolved, progressively including a larger variety of exercises oriented to semantics and pragmatics, but without neglecting more classical tasks like part-of-speech tagging and parsing. Since the 2016 edition, EVALITA registered an increased focus on social media data, especially Twitter, and the use of shared data across tasks (Basile et al., 2017).

2.2 EVALITA4ELG

The European Language Grid (ELG) is an European project[2] whose aim is to establish a platform and marketplace for the European industrial and academic research community around language Technologies (Rehm et al., 2020). ELG is an evolution of META-NET[3] and its sistem projects T4ME, CESAR, METANET4U, and META-NORD. According to the META-NET Strategic Research Agenda for Multilingual Europe 2020 (Rehm and Uszkoreit, 2013), ELG will represent the "European Service Platform for Language Technologies". At the time of this writing, ELG counts 7,200 corpora, 3,707 tools and services, plus a large number of other language resources such as lexicons, models, and grammars,[4] for both EU official and EU candidate languages, as well as a number of non-EU languages, such as languages spoken by EU immigrants or languages of political and trade partners. The platform has an interactive web user interface and APIs. Crucially for the benchmark presented in this paper, a Python library is maintained by ELG, which greatly facilitates the programmatic acces to its resources.

In 2020, the project EVALITA4ELG "Italian EVALITA Benchmark Linguistic Resources, NLP Services and Tools for the ELG platform"[5] started, with the aim of collecting all the language resources developed during the several past editions of EVALITA, and integrate them into the ELG. The project succeeded to collect, harmonize and upload 43 corpora and 1 lexical/conceptual resource from all editions of EVALITA from 2007 to 2020 (Basile et al., 2022), among which are found the ones we selected for UINAUIL, described in the next section.

2.3 Tasks

For the first version of UINAUIL, we selected six tasks from EVALITA. We aimed at selecting a representative sample of tasks in terms of their level of language analysis and target phenomenon. Moreover, we selected tasks with different formats, and proposed at different editions of EVALITA. Table 1 summarized the six tasks of UINAUIL, described in detail in the rest of this section.

2.3.1 Textual Entailment

In the textual entailment task of EVALITA 2009 (Bos et al., 2009), participants are asked to submit systems that classify ordered pairs of sentences according to the logical relation holding between them. In particular, a text (T), with respect to an hypothesis (H), can be labeled as either ENTAILED or NOT ENTAILED.

2.3.2 EVENTI

EVENTI, from EVALITA 2014 (Tommaso et al., 2014) is a shared task on Temporal Processing for Italian. The task is built around ITA-TimeBank (Caselli et al., 2011), a large manually annotated dataset of events and temporal expressions. The dataset follows the TimeML tagging standard, where events and time expressions are labeled as spans of single or multiple tokens, and they may be associated with attributes. The shared task is articulated into four subtasks related to the prediction of the extent of events and timex, their classification, and the relations insisting between event/event or event/timex pairs.

In UINAUIL, we include the subtask B, which involves the tagging of events and their classification

[1] http://www.evalita.it/
[2] https://cordis.europa.eu/project/id/825627
[3] http://www.meta-net.eu/
[4] https://live.european-language-grid.eu/
[5] http://evalita4elg.di.unito.it/

Acronym	Full name	Task type	Size (training/test)
Textual Entailment	Textual Entailment	Sentence pair classification	400/400
EVENTI	Event detection & classification	Sequence labeling	5,889/917
FactA	Factuality classification	Sequence labeling	2,723/1,816
SENTIPOLC	Sentiment Polarity Classification	Sentence classification	7,410/2,000
IronITA	Irony Detection	Sentence classification	3,777/872
HaSpeeDe	Hate Speech Detection	Sentence classification	6,839/1,263

Table 1: Summary of the tasks included in UINAUIL.

according to one of the following classes: ASPECTUAL (phase or aspect in the description of another event); I_ACTION (intensional action); I_STATE (event that denotes stative situations which introduce another event); PERCEPTION (event involving the physical perception of another event); REPORTING (action of declaring something, narrating an event, informing about an event); STATE (circumstance in which something obtains or holds true); OCCURRENCE (other type of event describing situations that happen or occur in the world).

The task is a sequence labeling problem, therefore the classes are associated at the token level, prefixed with a B (beginning of a labelled span) or a I (inside a labelled span), while O (outside) denotes the tokens that do not belong to any event.

2.3.3 FactA

The Event Factuality Annotation (FactA) task was part of EVALITA 2016 (Minard et al., 2016). In this task, the participant systems are challenged to profile the factuality of events in the text by means of three attributes, namely: certainty, time, and polarity. In UINAUIL, we included the first subtask of FactA, that is, the labeling of certainty, whereas spans of tokens from the input text associated with an event are labeled with one of three classes: CERTAIN (the source is certain about the mentioned event); NON_CERTAIN (the source is not certain about the mentioned event); UNDERSPECIFIED (the certainty about the mentioned event is not specified). Like EVENTI, FactA is a sequence labeling task and therefore the annotation is at the token level following the BIO standard.

2.3.4 SENTIPOLC

The SENTIment POLarity Classification (SENTIPOLC) shared task was proposed for the first time at EVALITA 2014 (Basile et al., 2014) and then re-run in 2016 (Basile et al., 2014). SENTIPOLC is divided into three subtasks, two of which are binary classification tasks where systems are challenged to predict subjectivity and irony in Italian tweets. The other task is polarity prediction, where systems have to predict two independent binary labels for positivity and negativity, with all four possible combinations of values allowed. The polarity prediction task is included in UINAUIL, with a slight change of format: instead of two independent binary labels, the task in UINAUIL is cast as a four-value multiclass classification task with labels POSITIVE, NEGATIVE, NEUTRAL, and MIXED.

2.3.5 IronITA

The shared task on Irony Detection in Italian Tweets (Cignarella et al., 2018, IronITA) is a shared task focused on the automatic detection of irony in Italian tweets, from EVALITA 2018. The task comprises two subtasks with differing by their level of granularity. The first subtask is a binary classification of tweets into ironic vs. non-ironic. The second task adds the level of sarcasm to the classification, conditioned on the presence of irony in the tweets. In UINAUIL, we included the first task, as a sentence-level binary classification task with the labels IRONIC and NOT IRONIC.

2.3.6 HaSpeeDe

Hate Speech Detection (HaSpeeDe) from EVALITA is a classification task from EVALITA 2020 (Sanguinetti et al., 2020), updated re-run of the same task from EVALITA 2018 (Bosco et al., 2018). The task invites participants to classify social media data from Twitter and Facebook as hateful, aggressive, and offensive. The complete shared task comprises binary classification (HATE vs. NOT HATE), a cross-domain subtask, stereotype detection and the identification of nominal utterances linked to hateful content. In UINAUIL, we included the training and test data for the main classification task only.

3 The UINAUIL library

In addition to defining the benchmark, we created a software library to access the data and functionalities of UINAUIL. We developed a Python module for downloading the task data and metadata, represented in a structurally consistent way. Furthermore, the library provides an implementation of evaluation metrics. The UINAUIL package is available on pip / PyPI, and can be installed with the command:

```
$ pip install uinauil
```

Listing 1: How to install the UINAUIL package.

The package depends only on two non-standard packages: `elg`, the library maintained by the ELG project for accessing to the resources of the European platform [6] (see Section 2.2), and `scikit-learn`, a well known library for Machine Learning and Data Analysis [7]. Otherwise, since the UINAUIL library is fully contained into a single file, developers can directly download the source file and save it in the working folder.

Once installed, the library can be used to download the resources of the tasks described in Section 2.3, and to evaluate the predictions of a personal model through standard performance metrics selected by us for each task. The data are contained in Python standard structures (lists and dictionaries) and are divided into training and test sets, according to the original split for each task represented in the ELG repository. The list of available tasks is stored into a proper attribute:

```
>>> import uinauil as ul
>>> ul.tasks

{
  'haspeede': {
      'id': 7498,
      'task': 'classification'
    },
  'textualentailment': {
    'id': 8121,
    'task': 'pairs'
  },
  'eventi': {
    'id': 7376,
    'task': 'sequence'
  },
  'sentipolc': {
    'id': 7479,
    'task': 'classification'
  },
  'facta': {
    'id': 8045,
    'task': 'sequence'
  },
  'ironita': {
    'id': 7372,
    'task': 'classification'}
  }
```

Listing 2: List of available tasks.

The `tasks` variable contains information in a dictionary format, where each key is the name of a task (used to acces the task data in UINAUIL), while the value contains its identifier on the ELG platform and the type of task. An example of usage of the library for a learning and evaluation pipeline is the following:

```
1  import uinauil as ul
2
3  # load a task, for example 'facta'
4  task = ul.Task('facta')
5
6  # get training and test set of the task
7  train = task.data.training_set  # train
8  test = task.data.test_set       # test
9
10 # train the model on the training set
       and make prediction on test set
11 ...
12 pred = <make predictions on test set>
13
14 # evaluate model on standard metrics
15 scores = task.evaluate(pred)
16 print(scores)
```

Listing 3: Quickstart for UINAUIL package.

Line 1 imports the UINAUIL library, while Line 4 downloads the resources of a task, in the example the FactA task described in Section 2.3.3. The authentication on ELG is handled by the *elg* library and is equipped with a caching mechanism in order to minimize the requests for logins on the platform. The UINAUIL library also checks whether the data was previously downloaded before connecting to ELG. Lines 7 and 8 store the training and test sets in local variables. These are represented as lists of instances for classification tasks, or lists of lists of tokens for sequence labeling tasks. In turn, instances and tokens are dictionaries pairing text and labels. At this point of the code example, a model can be trained on the training data, or the labels can be predicted otherwise, depending on the implemented approach — the library is agnostic to specific classification models. On Line 15 the predictions are used to evaluate the model with the `evaluate` method of UINAUIL, that calculates several standard performance metrics chosen specifically for each task as follows:

- For sequence labeling tasks, the performance

[6] https://pypi.org/project/elg/
[7] scikit-learn.org/

is evaluated only with accuracy, calculated as the ratio of hits over all the tokens.

- For all the remaining tasks, the performance metrics are accuracy on all classes, then precision, recall and F1 for each single class and their macro average.

In addition to these core functionalities, the UINAUIL library contains several metadata that helps programmers to understand the resources of each task, including the list of key names of features and target, a brief description of the meaning of each feature, the list of possible values of the target and their meaning, and others. The complete list of variables and methods of UINAUIL library is available on the Github repository of the project [8]. Also present on the repository are several examples of use of the library on several common Machine Learning models in form of Python notebooks, and the complete leaderboards by task.

4 Evaluation

In order to test the library, and to offer the community a first set of results on Italian NLU tasks, we conducted a series of experiments with the aim of setting a baseline for all the tasks of the benchmark. The experiments consist in fine-tuning pre-trained language models for Italian (plus a multilingual one) on the training data of each task, and testing their prediction against the corresponding test data, computing the appropriate evaluation metrics.

4.1 Experimental setting

We implemented this series of experiments with *simpletransformers*[9], a Python library that facilitates LLM fine-tuning and prediction. Simpletransformers automatically downloads pre-trained models from the Huggingface repository[10], and provides functions for training and classification. We built scripts that collect data through the UINAUIL library, fine-tune LLMs, produce the predictions with simpletransformers, and finally use UINAUIL again to compute the relevant evaluation metrics. We kept the hyperparameter optimization at a minimum, on purpose, since the goal of these experiments is not that of achieving a high performance, as much as producing a fair (while still high) baseline, and a comparison between models across tasks. All models are fine-tuned for exactly 2 epochs, with a fixed learning rate of 10^{-4}. All experimental results are averages of five runs.

4.2 Models

Here we briefly describe the LLMs used in the baseline experiments. The string in brackets is the identifier of the model in Huggingface.

- ALBERTO (m-polignano-uniba/bert_uncased_L-12_H-768_A-12_italian_alb3rt0) is the first LLM that has been proposed for the Italian language (Polignano et al., 2019). This model is based on BERT and it is trained on a collection of 200 million posts from Twitter from TWITA (Basile et al., 2018).

- ITALIAN BERT (dbmdz/bert-base-italian-uncased, dbmdz/bert-base-italian-xxl-uncased) is a LLM maintained by the MDZ Digital Library at Bavarian State, based on ELECTRA (Clark et al., 2020) and trained on a Wikipedia dump, the OPUS corpora collection (Tiedemann and Nygaard, 2004), and the Italian part of the OSCAR corpus (Abadji et al., 2021) for a total of about 13 million tokens. The model comes in two variants, the regular one and a larger one (XXL).

- MULTILINGUAL BERT (bert-base-multilingual-uncased) is one of the first models released together with the BERT architecture itself (Devlin et al., 2019). It is trained on text in 102 languages from Wikipedia with a masked language model goal. Although it has been surpassed in performance for many NLP tasks, Multilingual BERT has been widely adopted, also because pre-trained language models for languages other than English are often unavailable or smaller than their English counterparts.

4.3 Results

Table 2 shows the results of the baseline systems on the UINAUIL tasks. Focusing on the sequence labeling tasks EVENTI and FactA, we notice how the model does not make substantial difference for the latter (factuality classification), while there is a 0.02 point difference in performance for the former (event classification). The larger model (Italian BERT XXL) is the one obtaining the best

[8] https://github.com/valeriobasile/uinauil
[9] https://simpletransformers.ai/
[10] https://huggingface.co/models

Model	Textual Entailment				SENTIPOLC				EVENTI
	P	R	F1	Acc.	P	R	F1	Acc.	Acc.
ITALIAN BERT	.441	.497	.404	.538	.741	.721	.716	.646	.916
ITALIAN BERT XXL	.391	.495	.379	.541	.764	.741	**.740**	.675	**.936**
ALBERTO	.427	.500	.391	.529	.727	.688	.691	.621	.925
MULTILINGUAL BERT	.445	.524	**.430**	.544	.660	.653	.645	.559	.925

Model	IronITA				HaSpeeDe				FactA
	P	R	F1	Acc.	P	R	F1	Acc.	Acc.
ITALIAN BERT	.737	.736	.735	.736	.786	.785	.785	.785	.907
ITALIAN BERT XXL	.769	.765	**.764**	.765	.792	.791	**.791**	.791	.908
ALBERTO	.744	.743	.742	.742	.744	.742	.741	.741	**.909**
MULTILINGUAL BERT	.710	.709	.709	.709	.743	.740	.739	.739	**.909**

Table 2: Baseline results on all task of UINAUIL project: Textual Entailment, SENTIPOLC, IronITA and HaSpeeDe in terms of macro-averaged precision (P), recall (R), and F1-score (F1), and accuracy; EVENTI and FactA in terms of token-level accuracy.

performance on EVENTI, as well as on all the sentence classification tasks SENTIPOLC, IronITA, and HaSpeeDe.

Interestingly, for Textual Entailment, Multilingual BERT is the best model. This is also the only sentence pair classification task of the benchmark, indicating how the pre-training strategy of LLMs (e.g., stronger emphasis on single text vs. sentence pair) has an impact on its performance on different tasks.

In absolute terms, the performances of all baselines on Textual Entailment are quite poor, with an accuracy slightly higher than 0.5 and a very low F1 score, around 0.4. This shows how even if this task has been published over a decade ago, there is still ample room for improvement.

The performance on the baselines on the classification tasks are all in line with the reported state of the art, validating the standardization proposed with our benchmark.

5 Conclusions

In this paper we presented UINAUIL (Unified Interactive Natural Understanding of the Italian Language), an integrated benchmark for Italian NLU. Its purposes are manifold: to fill the gap in Italian NLU evaluation by proposing one integrated benchmark; to create more accessible and user-friendly benchmarks for Italian NLU; to encourage scholars to publish modern, integrated benchmarks for under-represented languages.

UINAUIL is implemented in Python library, publicly available via pip/PyPI, that permits to easily download resources in Italian Language for six different NLU tasks, that can be used by programmers and researchers to train and evaluate their NLP models. UINAUIL is built with automation as principle, with the main goal of minimizing the overhead for a user who wants to evaluate a NLU model for Italian. In effect, only a few lines of code are sufficient before and after the main logic of a model, in order to retrieve the data and evaluate the model.

In this paper, we presented in details each task included into UINAUIL and the main features of the Python library, including a sample quickstart code for its most common functionalities. Furthermore, we evaluated the performances of several common NLP models for Italian Language on each task of UINAUIL. The results show that current models represent a high-performing baseline, especially for sentence classification tasks, while there is still room for improvement for Italian NLU, as shown by the performance on the textual entailment task. However, it should be noted that the goal of this paper is mainly to present the UINAUIL benchmark. A thorough analysis of the results of all available models, while out of our present scope, is a natural next development of this work.

As further developments, we plan to add other tasks to the project, accordingly to the future developments in the field of NLP for Italian Language. We also plan to implement the leaderboard as a service in ELG, besides the Github repository, so that users can submit their results autonomously, leveraging the provided authentication.

6 Ethical and legal statement

This work uses data that have been previously reviewed and published. As such, we find no particular ethical issue to be discussed beyond what is already discussed by the original articles presenting the datasets. One particular dataset, however, contains sensitive data: the HaSpeeDe shared task data made of tweets annotated for hate speech. While the user mentions in these tweets were anonymized by the authors of the dataset to protect the mentioned people's privacy, the texts still contain explicit and implicit expressions of hatred that may result hurtful to some readers.

The download of the datasets is managed through the European Language Grid. As part of the procedure, the user is informed about the terms and conditions of each individual dataset and must accept the licence before downloading the data. Furthermore, the ELG platform tracks the data exchange in order to comply with the European General Data Protection Regulation (GDPR).

7 Limitations

In this paper, in addition to a benchmark for Italian NLU and a Python library implementing it, we presented the results of pre-trained language models fine-tuned for the six tasks in the benchmark. The models we selected are widely used for the Italian language, but they are not the only available ones. Moreover, the scenario moves fast, with newer and larger language models being published regularly. The evaluation conducted in this paper, therefore, can only be a partial snapshot of the current panorama, while the UINAUIL library stands as an easy tool to evaluate new models as they are published with minimal overhead.

References

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event)*, pages 1 – 9, Mannheim. Leibniz-Institut für Deutsche Sprache.

Yvonne Adesam, Aleksandrs Berdicevskis, and Felix Morger. 2020. Swedishglue – towards a swedish test set for evaluating natural language understanding models. Technical report, University of Gothenburg.

Pierpaolo Basile, Viviana Patti, Francesco Cutugno, Malvina Nissim, and Rachele Sprugnoli. 2017. Evalita goes social: Tasks, data, and community at the 2016 edition. *Italian Journal of Computational Linguistics*, 3-1.

Valerio Basile, Andrea Bolioli, Viviana Patti, Paolo Rosso, and Malvina Nissim. 2014. Overview of the evalita 2014 sentiment polarity classification task. *Overview of the Evalita 2014 SENTIment POLarity Classification Task*, pages 50–57.

Valerio Basile, Cristina Bosco, Michael Fell, Viviana Patti, and Rossella Varvara. 2022. Italian NLP for everyone: Resources and models from EVALITA to the European language grid. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 174–180, Marseille, France. European Language Resources Association.

Valerio Basile, Mirko Lai, and Manuela Sanguinetti. 2018. Long-term social media data collection at the university of turin. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*, pages 1–6. CEUR-WS.org.

Johan Bos, Fabio Massimo Zanzotto, and Marco Pennacchiotti. 2009. Textual entailment at evalita 2009. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, volume 9, Reggio Emilia, Italy.

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 Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263, pages 1–9, Torino. CEUR Workshop Proceedings (CEUR-WS. org).

Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. 2011. Annotating events, temporal expressions and relations in Italian: the it-timeml experience for the ita-TimeBank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 143–151, Portland, Oregon, USA. Association for Computational Linguistics.

Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2018. Overview of the Evalita 2018 task on Irony Detection in Italian Tweets (IRONITA). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, volume 2263, pages 1–9, Torino. CEUR Workshop Proceedings (CEUR-WS. org).

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, pages 1 – 18.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Anne-Lyse Minard, Manuela Speranza, Tommaso Caselli, and Fondazione Bruno Kessler. 2016. The EVALITA 2016 event factuality annotation task (FactA). In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, volume 1749, Napoli. CEUR Workshop Proceedings (CEUR-WS. org).

Lucia C Passaro, Maria Di Maro, Valerio Basile, and Danilo Croce. 2020. Lessons learned from evalita 2020 and thirteen years of evaluation of italian language technology. *IJCoL. Italian Journal of Computational Linguistics*, 6(6-2):79–102.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*, pages 1–6. CEUR-WS.org.

Georg Rehm, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajič, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiljevs, Orians Anvari, Andis Lagzdiņš, Jūlija Meļņika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, José Manuel Gómez-Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch. 2020. European language grid: An overview. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3366–3380, Marseille, France. European Language Resources Association.

Georg Rehm, Federico Gaspari, German Rigau, Maria Giagkou, Stelios Piperidis, Annika Grützner-Zahn, Natalia Resende, Jan Hajic, and Andy Way. 2021. The european language equality project: Enabling digital language equality for all european languages by 2030. *The Role of National Language Institutions in the Digital Age*, page 17.

Georg Rehm and Hans Uszkoreit. 2013. *META-NET strategic research agenda for multilingual Europe 2020*. Springer.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. HaSpeeDe 2@EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, pages 1–9, Online. CEUR Workshop Proceedings (CEUR-WS. org).

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 1183–1186, Lisbon, Portugal. European Language Resources Association (ELRA).

Caselli Tommaso, R Sprugnoli, Speranza Manuela, and Monachini Monica. 2014. EVENTI EValuation of Events and Temporal INformation at Evalita 2014. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, volume 2, pages 27–34. Pisa University Press.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages

353–355, Brussels, Belgium. Association for Computational Linguistics.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.