Is EVALITA Done? On the Impact of Prompting on the Italian NLP Evaluation Campaign.

Valerio Basile[1]

[1]University of Turin, C.so Svizzera 185, 10147, Italy

Abstract

Prompt-based learning is a recent paradigm in NLP that leverages large pre-trained language models to perform a variety of tasks. With this technique, it is possible to build classifiers that do not need training data (zero-shot). In this paper, we assess the status of prompt-based learning applied to several text classification tasks in the Italian language. The results indicate that the performance gap towards current supervised methods is still relevant. However, the difference in performance between pre-trained models and the characteristic of the prompt-based classifier of operating in a zero-shot fashion open a discussion regarding the next generation of evaluation campaigns for NLP.

Keywords

Prompt-based learning, Text Classification, Benchmarking

1. Introduction

Shared tasks and evaluation campaigns are a pillar of the research in Natural Language Processing. The constant effort by the community to organize, maintain, and update shared tasks allows researchers to test their models and algorithms in systematic ways, compare the performance fairly, and apply them to new languages and domains. An important byproduct of the organization of a shared task is typically novel data, which gets distributed across the research community.

Perhaps the best known, long-running evaluation campaign in the field of Natural Language Processing is SemEval[1]. Originating in 1998, this initiative was at first called SensEval and focused on semantic-related tasks. Over the years, the campaign evolved to include a large variety of shared tasks in NLP. Some evaluation campaigns are focused on specific tasks or research areas, such as PAN[2] for digital text forensics and stylometry. Alternatively, shared tasks are sometimes organized in a standalone fashion, or linked to an event such a workshop like Threat, Aggression and Cyberbullying (TRAC)[3]. Finally, several research communities gravitating around a specific geographic area or interested in a specific language organize their

NL4AI 2022: Sixth Workshop on Natural Language for Artificial Intelligence, November 30, 2022, Udine, Italy [1]

valerio.basile@unito.it (V. Basile)
0000-0001-8110-6832 (V. Basile)

[1] https://semeval.github.io/

[2] http://pan.webis.de/

[3] https://sites.google.com/view/trac2022/

own NLP evaluation campaigns, such as GermEval[4] for German or IberLEF (previously known as IberEval)[5] for Spanish and other Iberian languages.

EVALITA is the "periodic evaluation campaign of Natural Language Processing (NLP) and speech tools for the Italian language"[6]. Started in 2007, EVALITA was held seven times in 2007, 2009, 2011, 2014, 2016, 2018, and 2020, and its eighth edition is scheduled for 2023. The retrospective article by Passaro et al.[2] describes a healthy community, reflected by a growing number of shared tasks proposed at each edition, culminating with the 14 tasks at EVALITA 2020 [3]. At the same time, more interestingly for this paper, the number of *classification* tasks has consistently grown over the years. This phenomenon has become apparent in the 2018 edition EVALITA [4], where a single system was submitted to four different tasks (ABSITA [5] GxG [6], HaSpeeDe [7], and IronITA [8]) and ranked first in most of the individual subtasks [9]. This system was able to achieve very high results on all the tasks by leveraging multi-task learning. While this advancement was rightly praised, it also spurred the didscussion about the format of the shared tasks organized at EVALITA, i.e., if many tasks follow the same format (text classification), then the evaluation campaign may be shifting its focus towards learning models, with less regard for the underlying language phenomena.

The latest edition of EVALITA in 2020 confirmed this trend, with at least four "pure" text classification tasks (AMI [10], SARDISTANCE [11], HaSpeeDe 2 [12], and TAG-it [13]) and a few more where classification is partially involved important role (DANKMEMES [14] and ATE_ABSITA [15]).

In this paper, we revisit a number of tasks from the past editions of EVALITA in the light of the newest technologies available for NLP. We focus on classification tasks (Section 4), although in principle the experiment could be extended to other forms of inference over textual data. In particular, we consider the recently proposed paradigm of prompt-based learning (Section 2), which makes use of large pre-trained language models (Section 3) to perform classification in a zero-shot fashion. With the right combiniation of parameters, prompt-based zero-shot classifiers often performs surprisingly well, therefore raising important questions about the future of the evaluation in NLP:

> R1: Is supervised learning becoming obsolete in NLP, along with the need for training data?

If pre-trained language models can provide acceptable predictions without training data, in particular superior to those of classical, pre-neural machine learning models, then perhaps the baseline methods typically associated with shared tasks should be rethought.

> R2: Should zero-shot methods become the new baseline for NLP tasks?

The rest of this paper presents an experiment where a number of language models are used in combination with prompt-based learning and tested against benchmarks provided by EVALITA, in order to answer these questions.

[4] https://germeval.github.io/

[5] https://sites.google.com/view/iberlef2022

[6] https://www.evalita.it

2. Methodology

Prompt-based learning [16] is a recent paradigm which gained enormous traction in the NLP community, applied, among other tasks, to zero-shot classification. In a nutshell, prompt-based classification makes use of large pre-trained language models to map labels to handcrafted or automatically derived natural language expressions. The plausibility of the instance to classify, augmented with the prompt, determines the label without the need for further training or fine-tuning. Prompting for NLP is an active area of research. Solutions have been proposed for automatically inducing prompts [17], to improve the learning process, e.g. with calibration [18], and to adapt the method to few-shot learning [19].

In this paper, we propose an experiment of classification with prompts and pre-trained models with purposely simplistic characteristics. For each binary classification task, we create exactly two verbalizations, one for each label. The template for the verbalizations is fixed and it belongs to one of two types, namely text classification and author profiling. Furthermmore, the templates provide exactly one slot which is filled with exactly one word. Table 1 illustrates the verbalizations associated to each label in our experiments. The verbalizations are manually crafted, without any effort to optimize them or tuning any parameters.

Label	Template	Positive filler	Negative filler
irony	Questa frase è [MASK] *(EN) This sentence is* [MASK]	ironica *(EN) ironic*	normale *(EN) normal*
hate		offensiva *(EN) offensive*	normale *(EN) normal*
subjective		soggettiva *(EN) subjective*	oggettiva *(EN) objective*
positive		positiva *(EN) positive*	normale *(EN) normal*
negative		negativa *(EN) negative*	normale *(EN) normal*
misogyny		misogina *(EN) misogynous*	normale *(EN) normal*
aggressiveness		aggressiva *(EN) aggressive*	normale *(EN) normal*
man/woman	*(EN) L'autore di questa frase è* [MASK] The author of this sentence is [MASK]	uomo *(EN) man*	donna *(EN) woman*

Table 1
Verbalizations associated with binary labels.

The experiment is implemented with OpenPrompt [20], a Python library that streamlines the process of creating templates and verbalizers, up to the prediction of labels on textual data.[7]

[7] https://github.com/thunlp/OpenPrompt

3. Models

The classification power of prompt-based learning is only as good as the pre-trained model that serves as the basis for the classification algorithm. In this section, we briefly describe the three models used in the experiments presented in this paper. The models are based on Bidirectional Encoder Representations from Transformers [21, BERT], a popular and high-performing language model based on transformers [22].

Two of the models used in this paper are monolingual and have been created specifically to encode the properties of the Italian language. The third model is multilingual, i.e., trained on text from multiple languages

3.1. AlBERTo

The first neural language model that has been proposed for the Italian language is called AlBERTo [23]. AlBERTo is based on BERT and trained on a collection of 200 million posts from Twitter from the corpus TWITA [24]. The hyperparameter setting of AlBERTo mimics the first base model for English, with 12 hidden layers, 768-dimensional embeddings, and 12 attention heads, for a total of 110 million parameters. AlBERTo is available from the Huggingface model repository[8] with the identifier:
`m-polignano-uniba/bert_uncased_L-12_H-768_A-12_italian_alb3rt0`

3.2. MDZ Italian BERT

The MDZ Digital Library team at the Bavarian State Library published a set of BERT and ELECTRA [25] models trained on a Wikipedia dump, the OPUS corpora collection [26], and the Italian part of the OSCAR corpus [27] for a total of about 13 million tokens. The architecture of the network is for the most part the same as AlBERTo: 12 hidden layers, 768-dimensional embeddings, and 12 attention heads. The Italian BERT model used for the experiments in this paper is available from the Huggingface model repository with the identifier:
`dbmdz/bert-base-italian-xxl-uncased`

3.3. Multilingual BERT

The multilingual BERT, in its cased and uncased variants, is one of the first models released together with the BERT architecture itself [21]. 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. The Multilingual BERT model used for the experiments in this paper is available from the Huggingface model repository with the identifier:
`bert-base-multilingual-uncased`

[8] https://huggingface.co/models

4. Tasks

Six shared tasks have been selected from the past three editions of EVALITA, one from EVALITA 2016 [28], four from EVALITA 2018 [4], and one from EVALITA 2020 [3]. All tasks are classification tasks, and more specifically *binary* classification tasks, i.e., where the label to predict for each textual instance can have one of two possible values. Table 2 summarizes the tasks selected for the experiments presented in this paper and statistics on their size and label distribution.

Task	Label	Pos. labels	Neg. labels	Total
IronITA	irony	435	437	872
HaSpeeDe (TW)	hate	324	676	1000
HaSpeeDe (FB)	hate	677	323	1000
HaSpeeDe 2	hate	622	641	1263
AMI	misogyny	500	500	1000
	aggressiveness	176	824	1000
SENTIPOLC	subjective	1305	695	2000
	positive	352	1648	2000
	negative	770	1230	2000
	irony	235	1765	2000
GxG (CH)	man/woman	100	100	200
GxG (DI)	man/woman	37	37	74
GxG (JO)	man/woman	100	100	200
GxG (TW)	man/woman	3000	3000	6000
GxG (YT)	man/woman	2200	2200	4400

Table 2
The six EVALITA shared tasks used as benchmarks in this paper and the distribution of the labels in their test sets.

For all the shared tasks, we downloaded the test set textual data and labels from the European Language Grid[9] (ELG) [29]. The ELG is a recently proposed platform for Language Technology in Europe funded by the Horizon 2020 scheme. The main goal of ELG is to create an open and shared linguistic benchmark for Italian on a large set of representative tasks. The EVALITA4ELG project [30][10] integrated a large number of datasets and other resources, including pre-trained models and systems, from all editions of EVALITA to date into the ELG. It is therefore sufficient to register an account on the platform and the data can be accessed programmatically with the official ELG Python library.

4.1. IronITA

The EVALITA 2018 Task on Irony Detection in Italian Tweets [8, IronITA] is a shared task focused on the automatic detection of irony in Italian tweets. The shared task is articulated in two subtasks with increasing 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,

[9] https://live.european-language-grid.eu/
[10] https://live.european-language-grid.eu/meta-forum-2022/project-expo/evalita4elg

conditioned on the presence of irony in the tweets. For the experiments of this task, we only consider the first subtask.

4.2. HaSpeeDe and HaSpeeDe 2

Hate Speech Detection (HaSpeeDe) is a classification task that was run twice, at EVALITA 2018 [7] and 2020 [12], with similar scheme but updating the dataset from one edition to the other. The task focuses on the classification of hateful, aggressive, and offensive content in social media data from Twitter and Facebook. The first edition of HaSpeeDe features a binary classification task (hate vs. not hate) and a cross-domain subtask. In this paper, we used the test set of the first two subtasks, i.e., binary classification of hate on Twitter (TW) and Facebook (FB). HaSpeeDe 2 proposed a couple of additional subtasks, namely stereotype detection and the identification of nominal utterances linked to hateful content. For the purpose of this paper, we only used the data and labels from the main subtask of HaSpeeDe 2.

4.3. AMI

The Automatic Misogyny Identification shared task at EVALITA 2020 [10] proposes a benchmark for the classification of misogynistic and aggressive content towards women in Italian tweets. The main task is a double binary classification where systems are called to label tweets with two independent labels: misogynous vs not misogynous and aggressive vs. not aggressive. Furthermore, the second subtask of AMI introduces a synthetic dataset to measure the fairness of misogyny classification models. In this paper, we only used the binary classification data from the first subtask of AMI (misogyny and aggressiveness).

4.4. SENTIPOLC

The Sentiment Polarity Classification task (SENTIPOLC) was organized at EVALITA 2014 [31] and 2016 [32], with the second edition including the data used for the prevvious one plus a new test set. The task is focused on sentiment analysis on Italian tweets, with three classification tasks: subjectivity, polarity, and irony. The main task, classification of polarity is cast as a double binary classification task, where systems must produce two independent labels for positive and negative sentiment found in the text. In this way, the SENTIPOLC annotation scheme is able to encode poositive and negative sentiment, as well as neutral (both the positive and negative labels are absent) and mixed sentiment (both the positive and negative labels are present). For the experiments in this paper, we use the test sets of all the four binary classification tasks of SENTIPOLC 2016.

4.5. GxG

The Cross-Genre Gender Prediction task [6, GxG] was organized at EVALITA 2018. The shared task falls in the area of author profiling, in particularly asking participant systems to predict whether the author of a short text is a man or a woman. The texts come from five different sources: Twitter (TW), YouTube (YT), children writings (CH), newspapers (JO, for journalism), and personal diaries (DI). GxG places an emphasis on cross-dataset prediction, where a model is trained on a set of data from one domain (or source, in this case) and predictions are made on data from a different one. For this paper, we use the five sets independently, since no training is involved in our experiment. In this binary classification task, there is no natural negative and positive label, therefore we impose the arbitrary mapping man=negative label; woman=positive label.

5. Results

In this section, we present the results of the experiment of prompt-based classification on EVALITA tasks. The results are presented separately for each task, because evaluation metrics may vary from one task to another — accuracy, F1-score of the positive class, and macro-averaged F1-score are used. Moreover, we present the results, in Tables 4–7, along with the baseline(s) and best systems according to the reports of the individual tasks.

Task	System	Score
IronITA task A	Prompt-based$_{AlBERTo}$	.419
	Prompt-based$_{ItalianBERT}$	.469
	Prompt-based$_{MultilingualBERT}$	.573
	Baseline (most frequent class)	.334
	Baseline (random)	.505
	Best system (ItaliaNLP)	.731

Table 3
Results on IronITA (irony detection) in terms of macro-averaged F1-score.

Task	System	Score
HaSpeeDe-FB	Prompt-based$_{AlBERTo}$	.534
	Prompt-based$_{ItalianBERT}$	.613
	Prompt-based$_{MultilingualBERT}$	.505
	Baseline (most frequent class)	.244
	Best system (ItaliaNLP)	.828
HaSpeeDe-TW	Prompt-based$_{AlBERTo}$	.625
	Prompt-based$_{ItalianBERT}$	.590
	Prompt-based$_{MultilingualBERT}$	.507
	Baseline (most frequent class)	.403
	Best system (ItaliaNLP)	.799
HaSpeeDe 2 task A	Prompt-based$_{AlBERTo}$	.526
	Prompt-based$_{ItalianBERT}$	.583
	Prompt-based$_{MultilingualBERT}$	.537
	Baseline (most frequent class)	.336
	Baseline (Support Vector Machine)	.721
	Best system (TheNorth)	.808

Table 4
Results on the two editions of HaSpeeDe (hate speech detection) in terms of macro-averaged F1-score.

Task	System	Score
AMI task A	Prompt-based$_{AlBERTo}$	.573
	Prompt-based$_{ItalianBERT}$	.509
	Prompt-based$_{MultilingualBERT}$	.422
	Baseline (most frequent class)	.665
	Best system (jigsaw)	.741

Table 5
Results on AMI (misogyny identification) in terms of average between the macro-averaged F1-score of the two classes *misogyny* and *aggressiveness*.

Task	System	Score
SENTIPOLC task 1: subjectivity	Prompt-based$_{AlBERTo}$	.374
	Prompt-based$_{ItalianBERT}$	.501
	Prompt-based$_{MultilingualBERT}$	.443
	Baseline (most frequent class)	.394
	Best system (Unitor)	.744
SENTIPOLC task 2: polarity	Prompt-based$_{AlBERTo}$	.470
	Prompt-based$_{ItalianBERT}$	.498
	Prompt-based$_{MultilingualBERT}$	.476
	Baseline (most frequent class)	.416
	Best system (UniPI)	.663
SENTIPOLC task 3: irony	Prompt-based$_{AlBERTo}$	.374
	Prompt-based$_{ItalianBERT}$	.400
	Prompt-based$_{MultilingualBERT}$	.412
	Baseline (most frequent class)	.468
	Best system (tweet2check)	.541

Table 6
Results on SENTIPOLC (sentiment analysis) in terms of macro-averaged F1-score for task 1 and 3, and average of the macro-averages F1-scores of the two classes *positive* and *negative* for task 2.

The results of this experiment show that prompt-based classification (at least, this simplified version of it) usually beats trivial baselines, but otherwise underperforms with respect to supervised models on benchmarks for the Italian language. This is expected, since the method is fully zero-shot. The results on GxG, the only task related to author profiling, are closer to the best performing systems of the shared task, indicating an expressive power of the language models beyond the standing meaning of the text. Interestingly, the results vary widely between pre-trained language models, with none of the three models being clearly superior to the others across tasks.

6. Discussion and Conclusion

The *Betteridge's law of headlines*[11] states that "any headline that ends in a question mark can be answered by the word no". This paper is no exception: the answer to the question **Is EVALITA**

[11] https://web.archive.org/web/20090226202006/http://www.technovia.co.uk/2009/02/techcrunch-irresponsible-journalism.html

Task	System	Score
GxG CH	Prompt-based$_{AlBERTo}$	.550
	Prompt-based$_{ItalianBERT}$	.570
	Prompt-based$_{MultilingualBERT}$	.595
	Best system (ItaliaNLP)	.640
GxG DI	Prompt-based$_{AlBERTo}$	.581
	Prompt-based$_{ItalianBERT}$	.554
	Prompt-based$_{MultilingualBERT}$	.527
	Best system (ItaliaNLP)	.676
GxG JO	Prompt-based$_{AlBERTo}$	.560
	Prompt-based$_{ItalianBERT}$	.565
	Prompt-based$_{MultilingualBERT}$	.545
	Best system (UniOR)	.585
GxG TW	Prompt-based$_{AlBERTo}$	.542
	Prompt-based$_{ItalianBERT}$	.577
	Prompt-based$_{MultilingualBERT}$	.529
	Best system (ItaliaNLP)	.595
GxG YY	Prompt-based$_{AlBERTo}$	.510
	Prompt-based$_{ItalianBERT}$	.536
	Prompt-based$_{MultilingualBERT}$	.483
	Best system (ItaliaNLP)	.555

Table 7
Results on GxG (gender prediction) in terms of accuracy.

done? is certainly **no**. The prompt-based systems presented in this papers are far from the classification performance of their supervised counterparts on the EVALITA benchmarks. This result is in stark contrast to results reported on English benchmarks[12]. Moreover, the performance of the two Italian models and the multilingual model tested in this paper are unstable, with some models apparently more fit to certain tasks than others, raising the question whether the subpar performance is due to the method or the underlying language-specific pre-trained models. However, the results of the prompt-based models could be undermined by the lack of optimization of verbalizers and templates. There is certainly space for improvement, which was not the main focus of this paper, including an analysis of the disagreement between verbalizers, and of the actual output of the prompt-based models.

It is worth noting that this new technology allows us to create zero-shot classifiers for rather abstract language classification problems. Recent literature indicates that often few training instances (few-shot learning) are sufficient to increase the performance of prompt-based classifiers greatly [33]. Considering that the experiments in this paper make use only of the most basic elements of prompt-based classification, this paradigm should be regarded as a new frontier, not only for the advancement of text classification methodology, but also for its evaluation. Supervised learning in NLP is perhaps not on its way to obsolescence (R1), but the growing literature on zero-shot classification indicates at least that there is a new player on the field. Would it make sense to organize a shared task as part of an evaluation campaign like EVALITA where training data is not provided at all (R2)? The first results presented in this

[12] https://github.com/thunlp/OpenPrompt/tree/main/results/

paper seem to indicate that this is the case, paving the way for evaluation campaigns focused on zero-shot learning for NLP.

References

[1] D. Nozza, L. Passaro, M. Polignano, Preface to the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: D. Nozza, L. C. Passaro, M. Polignano (Eds.), Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022) co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2022), November 30, 2022, CEUR-WS.org, 2022.

[2] L. C. Passaro, M. Di Maro, V. Basile, D. Croce, Lessons learned from evalita 2020 and thirteen years of evaluation of italian language technology, IJCoL. Italian Journal of Computational Linguistics 6 (2020) 79–102.

[3] V. Basile, D. Croce, M. D. Maro, L. C. Passaro, EVALITA 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian, in: V. Basile, D. Croce, M. D. Maro, L. C. Passaro (Eds.), Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020, volume 2765 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 1–7. URL: http://ceur-ws.org/Vol-2765/overview.pdf.

[4] T. Caselli, N. Novielli, V. Patti, P. Rosso, Evalita 2018: Overview on the 6th evaluation campaign of natural language processing and speech tools for italian, in: T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), 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), Turin, Italy, December 12-13, 2018, volume 2263 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 1–6. URL: http://ceur-ws.org/Vol-2263/paper001.pdf.

[5] P. Basile, D. Croce, V. Basile, M. Polignano, Overview of the evalita 2018 aspect-based sentiment analysis task (absita), in: T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), 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, CEUR Workshop Proceedings (CEUR-WS. org), Torino, 2018, pp. 10–16.

[6] F. Dell'Orletta, M. Nissim, Overview of the Evalita 2018 cross-genre gender prediction (GxG) task, in: T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), 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, CEUR Workshop Proceedings (CEUR-WS. org), Torino, 2018, pp. 1–9.

[7] C. Bosco, F. Dell'Orletta, F. Poletto, M. Sanguinetti, M. Tesconi, Overview of the EVALITA 2018 hate speech detection task, in: T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), 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, CEUR Workshop Proceedings (CEUR-WS. org), Torino, 2018, pp. 1–9.

[8] A. T. Cignarella, S. Frenda, V. Basile, C. Bosco, V. Patti, P. Rosso, Overview of the Evalita 2018 task on Irony Detection in Italian Tweets (IRONITA), in: T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), 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, CEUR Workshop Proceedings (CEUR-WS. org), Torino, 2018, pp. 1–9.

[9] A. Cimino, L. D. Mattei, F. Dell'Orletta, Multi-task learning in deep neural networks at EVALITA 2018, in: T. Caselli, N. Novielli, V. Patti, P. Rosso (Eds.), 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), Turin, Italy, December 12-13, 2018, volume 2263 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 1–10. URL: http://ceur-ws.org/Vol-2263/paper013.pdf.

[10] E. Fersini, D. Nozza, P. Rosso, AMI@EVALITA2020: Automatic misogyny identification, in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), CEUR Workshop Proceedings (CEUR-WS. org), Online, 2020, pp. 1–8.

[11] A. T. Cignarella, M. Lai, C. Bosco, V. Patti, P. Rosso, SardiStance@EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets, in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), CEUR Workshop Proceedings (CEUR-WS. org), Online, 2020, pp. 1–10.

[12] M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, HaSpeeDe 2@EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task, in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), CEUR Workshop Proceedings (CEUR-WS. org), Online, 2020, pp. 1–9.

[13] A. Cimino, F. Dell'Orletta, M. Nissim, TAG-it@EVALITA2020: Overview of the topic, age, and gender prediction task for italian, in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), CEUR Workshop Proceedings (CEUR-WS. org), Online, 2020, pp. 1–9.

[14] M. Miliani, G. Giorgi, I. Rama, G. Anselmi, G. E. Lebani, DANKMEMES@EVALITA2020: The memeing of life: memes, multimodality and politics, in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), CEUR Workshop Proceedings (CEUR-WS. org), Online, 2020, pp. 1–9.

[15] L. De Mattei, G. De Martino, A. Iovine, A. Miaschi, M. Polignano, G. Rambelli, ATE_ABSITA@EVALITA2020: Overview of the aspect term extraction and aspect-based sentiment analysis task, in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), CEUR Workshop Proceedings (CEUR-WS. org), Online, 2020, pp.

1–8.

[16] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys (CSUR) (2022).

[17] G. Cui, S. Hu, N. Ding, L. Huang, Z. Liu, Prototypical verbalizer for prompt-based few-shot tuning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 7014–7024. URL: https://aclanthology.org/2022.acl-long.483. doi:10.18653/v1/2022.acl-long.483.

[18] Z. Zhao, E. Wallace, S. Feng, D. Klein, S. Singh, Calibrate before use: Improving few-shot performance of language models, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 12697–12706. URL: https://proceedings.mlr.press/v139/zhao21c.html.

[19] T. Le Scao, A. Rush, How many data points is a prompt worth?, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2627–2636. URL: https://aclanthology.org/2021.naacl-main.208. doi:10.18653/v1/2021.naacl-main.208.

[20] N. Ding, S. Hu, W. Zhao, Y. Chen, Z. Liu, H. Zheng, M. Sun, OpenPrompt: An open-source framework for prompt-learning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 105–113. URL: https://aclanthology.org/2022.acl-demo.10. doi:10.18653/v1/2022.acl-demo.10.

[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.

[23] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, Alberto: Italian BERT language understanding model for NLP challenging tasks based on tweets, in: R. Bernardi, R. Navigli, G. Semeraro (Eds.), Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019, volume 2481 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 1–6. URL: http://ceur-ws.org/Vol-2481/paper57.pdf.

[24] V. Basile, M. Lai, M. Sanguinetti, Long-term social media data collection at the university of turin, in: E. Cabrio, A. Mazzei, F. Tamburini (Eds.), Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018, volume 2253 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 1–6. URL: http://ceur-ws.org/Vol-2253/paper48.pdf.

[25] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as

discriminators rather than generators, in: International Conference on Learning Representations, 2020, pp. 1 – 18. URL: https://openreview.net/forum?id=r1xMH1BtvB.

[26] J. Tiedemann, L. Nygaard, 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), European Language Resources Association (ELRA), Lisbon, Portugal, 2004, pp. 1183–1186. URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/320.pdf.

[27] J. Abadji, P. J. O. Suárez, L. Romary, B. Sagot, Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus, in: H. Lüngen, M. Kupietz, P. Bański, A. Barbaresi, S. Clematide, I. Pisetta (Eds.), Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event), Leibniz-Institut für Deutsche Sprache, Mannheim, 2021, pp. 1 – 9. URL: https://nbn-resolving.org/urn:nbn:de:bsz:mh39-104688. doi:10.14618/ids-pub-10468.

[28] P. Basile, F. Cutugno, M. Nissim, V. Patti, R. Sprugnoli, et al., Evalita 2016: Overview of the 5th evaluation campaign of natural language processing and speech tools for italian, in: 3rd Italian Conference on Computational Linguistics, CLiC-it 2016 and 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, EVALITA 2016, volume 1749, CEUR-WS, 2016, pp. 1–4.

[29] G. Rehm, M. Berger, E. Elsholz, S. Hegele, F. Kintzel, K. Marheinecke, S. Piperidis, M. Deligiannis, D. Galanis, K. Gkirtzou, P. Labropoulou, K. Bontcheva, D. Jones, I. Roberts, J. Hajič, J. Hamrlová, L. Kačena, K. Choukri, V. Arranz, A. Vasiļjevs, O. Anvari, A. Lagzdiņš, J. Meļņika, G. Backfried, E. Dikici, M. Janosik, K. Prinz, C. Prinz, S. Stampler, D. Thomas-Aniola, J. M. Gómez-Pérez, A. Garcia Silva, C. Berrío, U. Germann, S. Renals, O. Klejch, European language grid: An overview, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 3366–3380. URL: https://aclanthology.org/2020.lrec-1.413.

[30] V. Basile, C. Bosco, M. Fell, V. Patti, R. Varvara, Italian NLP for everyone: Resources and models from EVALITA to the European language grid, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 174–180. URL: https://aclanthology.org/2022.lrec-1.19.

[31] V. Basile, A. Bolioli, V. Patti, P. Rosso, M. Nissim, Overview of the evalita 2014 sentiment polarity classification task, Overview of the Evalita 2014 SENTIment POLarity Classification Task (2014) 50–57.

[32] F. Barbieri, V. Basile, D. Croce, M. Nissim, N. Novielli, V. Patti, Overview of the evalita 2016 sentiment polarity classification task, in: P. Basile, A. Corazza, F. Cutugno, S. Montemagni, M. Nissim, V. Patti, G. Semeraro, R. Sprugnoli (Eds.), 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), Napoli, Italy, December 5-7, 2016, volume 1749 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016, pp. 1–11. URL: http://ceur-ws.org/Vol-1749/paper_026.pdf.

[33] T. Schick, H. Schütze, Exploiting cloze-questions for few-shot text classification and natural language inference, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 255–269. URL: https://aclanthology.org/2021.eacl-main.20. doi:10.18653/v1/2021.eacl-main.20.