EVALITA 2023: Overview of the 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian

Mirko Lai[1], Stefano Menini[2], Marco Polignano[3], Valentina Russo[4], Rachele Sprugnoli[5] and Giulia Venturi[6]

[1]*University of Turin*
[2]*Fondazione Bruno Kessler*
[3]*University of Bari "Aldo Moro"*
[4]*Logogramma s.r.l.*
[5]*University of Parma*
[6]*Institute for Computational Linguistics "A. Zampolli" (CNR-ILC), Pisa*

Abstract

EVALITA provides a shared framework for evaluating and comparing different Nautural Language Processing (NLP) and speech systems across various tasks suggested and organized by the Italian research community. These tasks represent scientific challenges and allow testing of methods, resources, and systems on shared benchmarks related to linguistic open issues and real-world applications, including considering multilingual and/or multi-modal perspectives. The EVALITA 2023 edition consisted of 13 different tasks grouped into four research areas: Affect, Authorship Analysis, Computational Ethics, and New Challenges in Long-standing Tasks. The participation saw 42 groups from 12 different countries, indicating an increasing international interest, partly due to the proposal of multilingual tasks. The final workshop showcases the results obtained and highlights the growing interest in using deep learning techniques based on Large Language Models as a new trend. Overall, EVALITA serves as a valuable platform for Italian and international researchers to explore NLP-related challenges, develop solutions, and foster discussions within the community.

Keywords

NLP, Evaluation, Italian, Speech, Tools

1. Introduction

The Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA) is the biennial initiative aimed at promoting the development of language and speech technologies for the Italian language. EVALITA is promoted by the Italian Association of Computational Linguistics (AILC)[1] and it is endorsed by the Italian Association for Artificial Intelligence (AIxIA)[2] and the Italian Association for Speech Sciences (AISV)[3].

EVALITA provides a shared framework where different systems and approaches can be scientifically evaluated and compared with each other with respect to a large variety of tasks, suggested and organized by the Italian research community. The proposed tasks represent scientific challenges where methods, resources, and systems can be tested against shared benchmarks representing linguistic open issues or real-world applications, possibly in a multilingual and/or multi-modal perspective. The collected datasets provide big opportunities for scientists to explore old and new problems concerning NLP in Italian as well as to develop solutions and discuss NLP-related issues within the community. Some tasks are traditionally present in the evaluation campaign, while others are completely new.

This paper introduces the tasks proposed at EVALITA 2023 and provides an overview of the participants and systems whose descriptions and obtained results are reported in these Proceedings. The EVALITA 2023 final workshop, held in Parma on September 7-8th, counts 13 different tasks. In particular, the selected tasks are grouped into four research areas (tracks) according to their objective and characteristics, namely: (i) *Affect*; (ii) *Authorship Analysis*; (iii) *Computational Ethics*; (iv) *New Challenges in Long-standing Tasks*.

This edition was participated by 42 groups whose members have affiliations in 12 different countries. The high number of tasks is in line with a clear trend towards an increasing volume of proposed tasks at EVALITA. In fact, we have witnessed a significant progression from the 5 tasks organized in the first EVALITA campaign in 2007 to a peak of 14 tasks in the latest 2020 edition. Although EVALITA is generally promoted and targeted to the Italian research community, this edition saw increas-

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
mirko.lai@unito.it (M. Lai); menini@fbk.eu (S. Menini); marco.polignano@uniba.it (M. Polignano); vrusso@logogramma.com (V. Russo); rachele.sprugnoli@unipr.it (R. Sprugnoli); giulia.venturi@ilc.cnr.it (G. Venturi)

[1]http://www.ai-lc.it
[2]http://www.aixia.it
[3]http://www.aisv.it

ing international participation, partly due to our strong encouragement for the submission of multilingual tasks. This confirms a general trend of internationalization for the campaign, which reached its maximum this year, as discussed further.

This overview is organized as follows: in Section 2 a brief description of the tasks belonging to the various areas is reported. Section 3 discusses the participation in the workshop referred to several aspects, from the research area to the affiliation of authors. Section 4 describes the criteria used to assign the best system across tasks award, made by an ad-hoc committee starting from the suggestions of task organizers and reviewers. Finally, section 5 points out both the obtained results and the future of the workshop.

2. EVALITA 2023 Tracks and Tasks

In the 2023 edition of EVALITA, 13 different tasks were proposed, peer-reviewed, and accepted. Data were produced by the task organizers and made available to the participants. For the future availability of this data, we are going to release them on GitHub[4], in accordance with the terms and conditions of the respective data sources. Such a repository will also reference alternative repositories managed by the task organizers. The tasks of EVALITA 2023 are grouped according to the following tracks corresponding to four broad research areas:

Affect

EMit - Categorical Emotion Detection in Italian Social Media [1]. It aims to provide the first evaluation framework for emotion detection in Italian texts at EVALITA, following the categorical approach and offering novel annotated data. It presents two subtasks: *i)* Subtask A, which consists of an emotion detection challenge, and *ii)* Subtask B, which introduces a novel problem of target detection of the expressed emotion.

EmotivITA - Dimensional and Multi-dimensional Emotion Analysis [2]. The first shared task for Italian that follows the dimensional approach in emotion analysis. It introduces a new Italian dataset annotated with the Valence, Arousal, and Dominance dimensions and has two subtasks: *i)* Dimensional emotion regression and *ii)* Multi-dimensional emotion regression.

Authorship Analysis

PoliticIT - Political Ideology Detection in Italian Texts [3]. It aims to extract politicians' ideology information from a set of tweets in Italian framed as a binary and a multiclass classification. The task is designed to be privacy-preserving and accompanied by a subtask targeting the identification of self-assigned gender as a demographic trait.

GeoLingIt - Geolocation of Linguistic Variation in Italy [4]. The first shared task on the geolocation from social media posts comprising content in language varieties other than standard Italian (i.e., regional Italian, and languages and dialects of Italy). It is articulated into two subtasks: *i)* coarse-grained geolocation, aiming at predicting the region in which the variety expressed in the post is spoken, and *ii)* fine-grained geolocation, aiming at predicting its exact coordinates.

LangLearn - Language Learning Development [5]. The first shared task on automatic language development assessment aimed at developing and evaluating systems to predict the evolution of the written language abilities of learners across several time intervals. It was conceived to be multilingual, relying on written productions of Italian and Spanish learners, and representative of L1 and L2 learning scenarios.

Computational Ethics

HaSpeeDe 3 - Political and Religious Hate Speech Detection [6]. The third edition of a shared task on the detection of hateful content in Italian tweets. Differently from the two previous editions (organized within EVALITA 2018 and 2020), it explores hate speech in strong polarised debates, concerning politics and religion. Participants are asked to predict hate speech in both in- and out-domain settings, using either only the provided textual content of the tweet or any kind of external data.

HODI - Homotransphobia Detection in Italian [7]. The first shared task for the automatic detection of homotransphobia in Italian. The challenge is organized into two subtasks: *i)* Subtask A focuses on the binary textual classification of homotransphobic tweets, *ii)* Subtask B is concerned with the identification of rationales for explainability in the form of textual spans of text.

MULTI-Fake-DetectiVE - MULTImodal Fake News Detection and VErification [8]. The first task on fake news detection in Italian that explores multimodality and wants to address the problem from two perspectives, represented by the two subtasks: *i)* sub-task 1 aimed to evaluate the effectiveness of multimodal fake news detection systems,

[4] https://github.com/evalita2023/datasets

ii) sub-task 2, which consists in gaining insights into the interplay between text and images. Both perspectives were framed as classification problems.

ACTI – Automatic Conspiracy Theory Identification [9]. The first shared task based exclusively on comments published on conspiratorial channels of telegram. It is articulated into two subtasks: *i)* Conspiratorial Content Classification consisting in identifying conspiratorial content and *ii)* Conspiratorial Category Classification about specific conspiracy theory classification.

New Challenges in Long-Standing Tasks

NERMuD - Named-Entities Recognition on Multi-Domain Documents [10]. It consists in extracting and classifying persons, organizations, and locations from documents in various domains. It is articulated into two subtasks: *i)* Domain-agnostic classification, where participants are required to identify and classify entities from different types of texts, i.e., news, fiction, and political speeches, using a single model, and *ii)* Domain-specific classification, where a different model can be used for each text type.

CLinkaRT - Linking a Lab Result to its Test Event in the Clinical Domain [11]. It is a relation extraction task based on clinical cases taken from the E3C corpus, i.e., Italian written documents reporting statements of clinical practice. The task consists in identifying test results and measurements and linking them to the textual mentions of the laboratory tests and measurements from which they were obtained.

WiC-ITA - Word-in-Context task for Italian [12]. The first shared task at EVALITA on determining if a word occurring in two different sentences has the same meaning or not. It has been modeled as both a binary classification and a ranking problem.

DisCoTEX - Assessing DIScourse COherence in Italian TEXts [13]. The first shared task focused on modeling discourse coherence for Italian real-word texts. It was organized into two independent tasks: a more traditional one, aimed at evaluating whether models are able to distinguish well-organized documents from corrupted ones, and a less explored one, which assesses the models' performance on texts evaluated for coherence by human raters.

3. Participation

EVALITA 2023 attracted the interest of a large number of researchers from academia and industry, for a total of 42 single teams composed of about 109 individuals participating in one or more of the 13 proposed tasks. After the evaluation period, **51** system descriptions were submitted (reported in these proceedings), i.e., a **12**% percentage decrease with respect to the previous EVALITA edition [14].

Moreover, task organizers allowed participants to submit more than one system result (called runs), for a total of 246 submitted runs. Table 1 shows the different tracks and tasks along with the number of participating teams and submitted runs. The data reported in the table is based on information provided by the task organizers at the end of the evaluation process. Such data represents an overestimation with respect to the systems described in the proceedings. The trends are similar, but there are differences due to groups participating in more than one task and groups that have not produced a system report.

Unlike previous EVALITA editions, the organizers were not discouraged from distinguishing the submissions between unconstrained and constrained runs[5]. In fact, some of them introduced subtasks based on external resources used for training, while others required both a constrained and an unconstrained run. Alternatively, they allowed participants the freedom to utilize external resources or augment the distributed datasets. This decision was motivated by the expectation that most participants would employ pre-trained Neural Language Models. Thus, the organizers wanted to assess the participants' creativity in adopting strategies beyond solely relying on these models.

Participation was quite imbalanced across different tracks and tasks, as reported in Figure 1: each rectangle represents a task whose size reflects the number of participants, while the color indicates the corresponding track.

In line with the past edition of EVALITA [14], the development of systems dedicated to identifying unethical behaviors or malicious intentions in texts, spanning various aspects of human society, remains a topic of significant interest to the community. In fact, as evidenced by the high participation, the shared tasks grouped under the "Computational Ethics" track obtained the most attention. However, for the first time, this year the second most participated track was the "Authorship Analysis" one, which is focused on analyzing text writing styles to capture diverse author characteristics. This is a quite new result since the same typology of track had a relatively

[5] A system is considered *constrained* when using the provided training data only; on the contrary, it is considered *unconstrained* when using additional material to augment the training dataset or to acquire additional resources.

Track	Task	Teams	Runs
Affect	EMit	4	8
	EmotivITA	2	5
Authorship Analysis	PoliticIT	7	8
	GeoLingIt	6	35
	LangLearn	5	18
Computational Ethics	HaSpeeDe 3	6	29
	HODI	8	22
	MULTI-Fake-DetectiVE	4	6
	ACTI	8	81
New Challenges in Long-standing Tasks	NERMuD	1	2
	CLinkaRT	3	6
	WiC-ITA	4	9
	DisCoTEX	3	19

Table 1
Number of participating teams and number of runs organized by track and task. The data reported is an overestimation with respect to the systems described in the proceedings (e.g. teams participating in more than a task are counted according to the number of tasks they participated in).

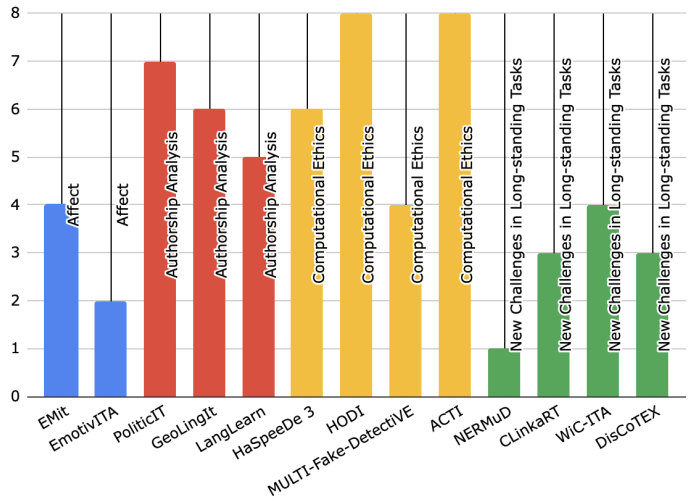

Figure 1: Number of participating teams organized by track (color) and task. The blue color is adopted for the track "*Affect*", the red color for "*Authorship Analysis*", yellow for "*Computational Ethics*", and green for "*New Challanges in Long-standing Tasks*".

low number of participants during the 2020 campaign. This shows the interest of the NLP community towards new and potentially more challenging areas of natural language understanding. It is worth noting that this year for the first time we introduced a new track solely dedicated to evaluating systems of emotions detection from two diverse perspectives (the "Affect" track). Additionally, we decided to keep the "New Challenges in Long-standing Tasks" track. Even if this track was among the least participated, the rationale behind this choice was to offer benchmarks for more conventional NLP tasks to the latest generation of Large Language Models. These models served as the foundation for the majority of the approaches devised by the participants, as illustrated in Section 5

It is worth noting that we also received a considerable number of tasks presented for the first time at EVALITA. Besides the two tasks centered around modeling different aspects of affect, namely *EMit* and *EmotivITA*, among them we can find *GeoLingIt*, *LangLearn*, *HODI*, and *ACTI*, which introduced novel problems. Interestingly, two of these newly introduced tasks received the highest num-

ber of submissions, showing the interest of the community in taking on new challenges.

In contrast to the 2020 edition, which saw a total of over 180 task organizers or participants, EVALITA 2023 experienced reduced participation. However, it is worth noting that the authorship of the 172 proceedings authors, including both participants and task organizers, reflects a greater diversity in terms of their origins, spanning 15 different countries. Notably, 70% of these contributors come from Italy, while the remaining 30% come from Institutions and companies abroad. The group of the 63 task organizers have affiliations in 6 countries (79% from Italy while 21% from Institutions and companies abroad). In summary, a noticeable increase was observed in the number of task organizers, particularly those affiliated with institutions abroad. In fact, the proportion of organizers with foreign affiliations more than doubled with respect to the previous edition, rising from 10% to 21% of the total organizers. This indicates a growing international interest in EVALITA. Notably, 6 out of the 13 tasks were organized by authors with mixed affiliations, combining both Italian and foreign institutions. This statistic aligns with one of the innovations we introduced this year. Indeed, during the call for tasks period, we encouraged the proposal of multilingual tasks, where participants were provided with datasets in both Italian and other languages. Up until now, only two tasks, namely *LangLearn* and *WiC-ITA*, provided participants with datasets in Italian and Spanish, and English, respectively. Although only a small number of organizers embraced this suggestion, we see it as a promising first step towards achieving a more international profile for EVALITA in the future.

As a last remark, we would like to notice that this year we had four teams that participated in multiple tasks. Among them, one team employed the same approach for two tasks (*HODI* and *HaSpeeDe 3*), while two other teams utilized distinct methods for two tasks each (*LangLearn* and *WiC-ITA*, and *LangLearn* and *DisCoTEX*). Particularly innovative was the approach taken by a single team, which submitted results for all 13 tasks, employing variations of the same model. In Section 5, we discuss how this feat was accomplished through the utilization of instruction-based models fine-tuned on all the EVALITA 2023 datasets using task-specific prompts.

4. Best System Across Tasks Award

In line with the previous edition, we confirmed the award to the best system across-task. The award was introduced with the aim of fostering student participation in the evaluation campaign and in the workshop.

A committee of 5 members (Felice Dell'Orletta, Bernardo Magnini, Azzurra Mancini, Stefano Menini, Viviana Patti) was asked to choose the best system across tasks. Three of the five members come from academia while two of them are from industry. The composition of the committee is balanced with respect to the level of seniority as well as to their academic background (computer science-oriented vs. humanities-oriented). In order to select a short list of candidates, the task organizers were invited to propose one candidate system participating in their tasks (not necessarily top-ranking). The committee was provided with the list of candidate systems and the criteria for eligibility, based on:

- *novelty* with respect to the state of the art;
- *originality*, in terms of identification of new linguistic resources, identification of linguistically motivated features, and implementation of a theoretical framework grounded in linguistics;
- *critical insight*, paving the way to future challenges (deep error analysis, discussion on the limits of the proposed system, discussion of the inherent challenges of the task);
- *technical soundness* and *methodological rigor*.

We collected **5** system nominations from the organizers of **9** tasks from across all tracks. The candidate systems are authored by 13 authors, among whom 6 are Ph.D. students. The award recipient(s) will be announced during the final EVALITA workshop, during the plenary session.

5. Final Remarks

The widespread adoption of Large Language Models (LLMs) was evident in the EVALITA 2023 challenge. LLMs, such as GPT-3 and its variants, have revolutionized the NLP landscape due to their ability to learn from large amounts of data and generate contextually relevant responses. These models have shown remarkable performance across various NLP tasks, and their usage was prominent in this edition of EVALITA. The confirmation of the massive use of LLMs underscores their effectiveness and potential in advancing NLP technology.

Traditional supervised learning approaches heavily rely on annotated data, which can be expensive and time-consuming to obtain. In response to this challenge, many participants in EVALITA 2023 proposed a semi-supervised approach using the prompting technique. The prompting technique involves providing the model with a few example inputs or a prompt to guide its response generation. This method allows leveraging limited labeled data while utilizing the model's language understanding capabilities to generalize to unseen instances. The adoption of the prompting technique showcases the interest in exploring more efficient and resourceful ways to tackle NLP tasks.

A noteworthy development in EVALITA 2023 was a team that participated in all tasks using the same approach, facilitated by prompt-based LLMs fine-tuning. While this approach showed promise, it also highlighted an essential observation: the performance of LLMs varies significantly across different NLP tasks. While LLMs are powerful models, they may not excel uniformly in all linguistic challenges. This underscores the need to understand the strengths and limitations of LLMs and to fine-tune them specifically for each task to achieve optimal results.

Another important outcome of the EVALITA 2023 challenge was the substantial increase in participation from groups outside Italy, making it one of the most attended editions by international teams. The rising international interest can be attributed to the growing significance of NLP and speech technologies on a global scale. The encouragement for multilingual tasks and the availability of shared datasets might have attracted researchers from different countries to participate actively. This trend signifies the growing impact and international recognition of the EVALITA initiative, facilitating collaboration and knowledge exchange among NLP communities worldwide.

To sum up, EVALITA 2023 outcomes demonstrate the dominance of LLMs in NLP, the exploration of semi-supervised approaches, the significance of task-specific fine-tuning, and the increasing internationalization of the initiative. These outcomes contribute to advancing the field of NLP, encouraging further research, and fostering a diverse and collaborative NLP community.

Acknowledgments

We would like to thank our sponsors: Talia[6], Almawave[7], Aptus.AI[8] and Logogramma[9]. Our gratitude goes also to the University of Parma for hosting the event. In addition, we sincerely thank the Best System award committee for providing their expertise and experience. Moreover, we acknowledge the AILC Board members for their trust and support. We warmly thank our invited speaker Julio Gonzalo, for having shared his knowledge and insights with his talk. Last but not least, we would like to thank all the task organizers and participants who made this edition special with their enthusiasm and creativity.

[6] https://talia.cloud/
[7] https://www.almawave.com/it/
[8] https://www.aptus.ai/
[9] https://www.logogramma.com/

References

[1] C. Alzetta, D. Brunato, F. Dell'Orletta, A. Miaschi, K. Sagae, C. H. Sánchez-Gutiérrez, G. Venturi, LangLearn at EVALITA 2023: Overview of the Language Learning Development Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[2] G. Gafà, F. Cutugno, M. Venuti, EmotivITA at EVALITA 2023: Overview of the Dimensional and Multidimensional Emotion Analysis Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[3] D. Russo, S. M. Jiménez Zafra, J. A. García-Díaz, T. Caselli, M. Guerini, L. A. Ureña López, R. Valencia-García, PoliticIT at EVALITA 2023: Overview of the Political Ideology Detection in Italian Texts Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[4] A. Ramponi, C. Casula, GeoLingIt at EVALITA 2023: Overview of the Geolocation of Linguistic Variation in Italy Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[5] O. Araque, S. Frenda, R. Sprugnoli, D. Nozza, V. Patti, EMit at EVALITA 2023: Overview of the Categorical Emotion Detection in Italian Social Media Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[6] M. Lai, F. Celli, A. Ramponi, S. Tonelli, C. Bosco, V. Patti, HaSpeeDe3 at EVALITA 2023: Overview of the Political and Religious Hate Speech Detection task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[7] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the first Shared Task on Homotransphobia Detection in Italian, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[8] A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L. C. Passaro, M. Sabbatini, MULTI-Fake-DetectiVE at EVALITA 2023: Overview of the MULTImodal Fake News Detection and VErification Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[9] G. Russo, N. Stoehr, M. Horta Ribeiro, ACTI at EVALITA 2023: Overview of the Conspiracy Theory Identification Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[10] A. Palmero Aprosio, T. Paccosi, NERMuD at

EVALITA 2023: Overview of the Named-Entities Recognition on Multi-Domain Documents Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[11] A. Begoña, G. Karunakaran, A. Lavelli, B. Magnini, M. Speranza, R. Zanoli, ClinkaRT at EVALITA 2023: Overview of the Task on Linking a Lab Result to its Test Event in the Clinical Domain, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[12] P. Cassotti, L. Siciliani, L. C. Passaro, M. Gatto, P. Basile, WiC-ITA at EVALITA 2023: Overview of the EVALITA2023 Word-in-Context for ITAlian Task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[13] D. Brunato, D. Colla, F. Dell'Orletta, I. Dini, D. P. Radicioni, A. A. Ravelli, DisCoTEX at EVALITA 2023: Overview of the assessing DIScourse COherence in Italian TEXts task, in: M. Lai, alii (Eds.), Proceedings of EVALITA 2023, CEUR.org, September 7th-8th 2023, Parma, 2023.

[14] V. Basile, D. Croce, M. Di Maro, L. C. Passaro, EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), CEUR.org, Online, 2020.