Evaluating Text-To-Text Framework for Topic and Style Classification of Italian texts

Michele Papucci[1,2], Chiara De Nigris[1], Alessio Miaschi[3] and Felice Dell'Orletta[2,3]

[1]*Università di Pisa, Pisa*
[2]*TALIA S.r.l., Pisa*
[3]*Istituto di Linguistica Computazionale "A. Zampolli" (ILC-CNR), ItaliaNLP Lab, www.italianlp.it, Pisa*

Abstract

In this paper, we propose an extensive evaluation of the first text-to-text Italian Neural Language Model (NLM), IT5 [1], on a classification scenario. In particular, we test the performance of IT5 on several tasks involving both the classification of the topic and the style of a set of Italian posts. We assess the model in two different configurations, single- and multi-task classification, and we compare it with a more traditional NLM based on the Transformer architecture (i.e. BERT). Moreover, we test its performance in a few-shot learning scenario. We also perform a qualitative investigation on the impact of label representations in modeling the classification of the IT5 model. Results show that IT5 could achieve good results, although generally lower than the BERT model. Nevertheless, we observe a significant performance improvement of the Text-to-text model in a multi-task classification scenario. Finally, we found that altering the representation of the labels mainly impacts the classification of the topic.

Keywords

transformers, text-to-text, t5, bert, topic classification, style classification

1. Introduction and Motivation

Over the past few years, the text-to-text paradigm has become one of the most widely adopted approach in the development of state-of-the-art Neural Language Models (NLMs) [3, 4, 5]. The basic idea of this paradigm, inspired by previous unifying frameworks for NLP tasks [6, 7, 8], is to consider each task as a text-to-text task, i.e. getting text as input data and producing new text as output.

This unifying framework has proven to be a particularly effective transfer learning method, often outperforming previous models, e.g. BERT [9], in data-poor settings. Nevertheless, few works proposed systematic evaluations of such models in different classification scenarios and in comparison with more traditional NLMs. Among these, [3] showed that T5 achieves comparable, if not better performance, with previous state-of-the-art models on the most popular NLP benchmarks, e.g. GLUE [10] and SQuAD [11]. [12], instead, demonstrated that T5

NL4AI 2022: Sixth Workshop on Natural Language for Artificial Intelligence, November 30, 2022, Udine, Italy [2]
m.papucci@studenti.unipi.it (M. Papucci); c.denigris@studenti.unipi.it (C. Nigris); alessio.miaschi@ilc.cnr.it (A. Miaschi); felice.dellorletta@ilc.cnr.it (F. Dell'Orletta)
https://github.com/michelepapucci (M. Papucci); https://alemiaschi.github.io/ (A. Miaschi); http://www.italianlp.it/people/felice-dellorletta/ (F. Dell'Orletta)
0000-0003-4251-7254 (M. Papucci); 0000-0002-0736-5411 (A. Miaschi); 0000-0003-3454-9387 (F. Dell'Orletta)

Attribute	Description	Values
Age	Age of the writer	Five ranges: 0-19, 20-29, 30-39, 40-49 and 50-100
Gender	Gender of the writer	M, F
Topic	Topic of the post	Eleven possible categories: ANIME, AUTO-MOTO, BIKES, CELEBRITIES, ENTERTAINMENT, NATURE, MEDICINE-AESTHETIC, METAL-DETECTING, SMOKE, SPORTS, TECHNOLOGY

Table 1
TAG-it labels description.

outperforms BERT in a document ranking task, especially in a data-poor setting with limited training data. Inspecting the performance of 6 different NLMs on a sentiment analysis task, [13] found that T5 is the second best performing model, next only to XLNet [14].

Whereas, focusing on languages other than English, [1] compared the performance of their IT5 with other multilingual and Italian models, showing e.g. that IT5 base outperforms BERT on SQuAD-IT [15], the extractive question answering task for the Italian language. Similar results have been obtained by [16] while measuring the performance of their Brazilian Portuguese T5 model (PTT5) against the ones obtained with BERTimbau, a BERT model pre-trained on the brWaC corpus [17]. Comparing the models on three different evaluation tasks for the Portuguese language (i.e. semantic similarity and entailment prediction [18] and NER [19]), they showed that PTT5 achieves competitive performance with BERTimbau, although the latter obtained slightly better results.

Building on these previous studies, in this work we propose an evaluation of the first text-to-text Transformer model developed for the Italian language, IT5 [1], on several classification tasks. More specifically, we performed our experiments on two different classification scenarios, single-task and multi-task, and we compared the performance of IT5 against those obtained with an Italian version of BERT. Furthermore, in order to verify the ability of the model in a data-poor setting, we also tested its performance in a few-shot learning scenario. Finally, following the approach devised by [20], we performed a more in-depth analysis to test the impact of label representations in modeling the classification of the IT5 model.

The remainder of the paper is organized as follows: in Sec. 2 and 3 we introduce the dataset and the models used in our experiments, in Sec. 4 we describe the experimental setting, in Sec. 5 and 6 we discuss the obtained results and in Sec. 7 we conclude the paper.

Contributions: In this paper we: i) proposed an extensive evaluation of IT5 performance on three different classification tasks based on Italian sentences; ii) we tested the performance of the model in different scenarios (single- and multi-task classification) and we compared them with those obtained with another Transformer especially suited for classification tasks; iii) we studied the behavior of the model in a data-poor setting by measuring its performance in few-shot learning scenario; iv) we verified the impact of label modification on IT5' performance.

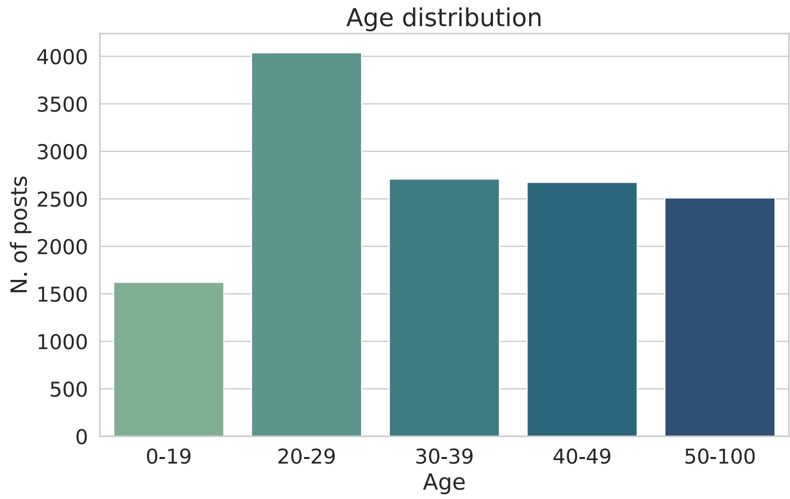

Figure 1: Age distribution.

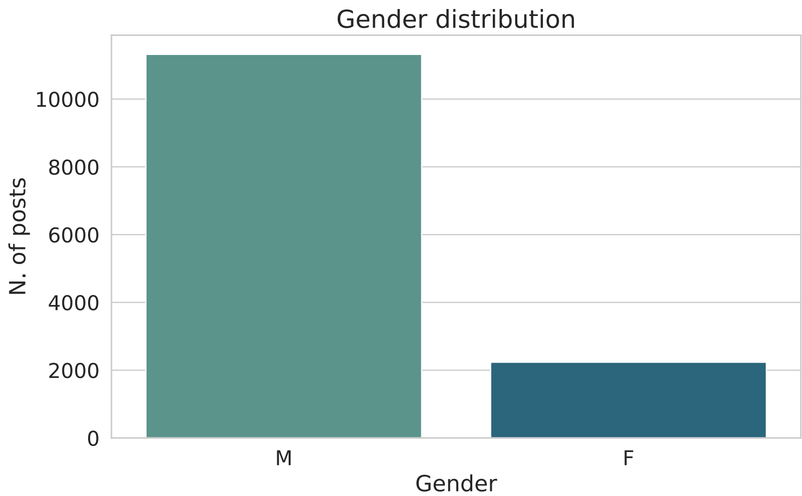

Figure 2: Gender distribution.

2. Data

In order to perform our experiments, we relied on posts extracted from TAG-IT [21], the profiling shared task presented at EVALITA 2020 [22]. The dataset, based on the corpus defined in [23], consists of more than 10,000 posts written in Italian and collected from different blogs. Each post is labeled with three different labels: age and gender of the writer and topic. The details and the statistics about the dataset are reported in Table 1 and Figures 1, 2 and 3.

As it can be noticed from the Figures, the *Age* variable presents a quite balanced distribution

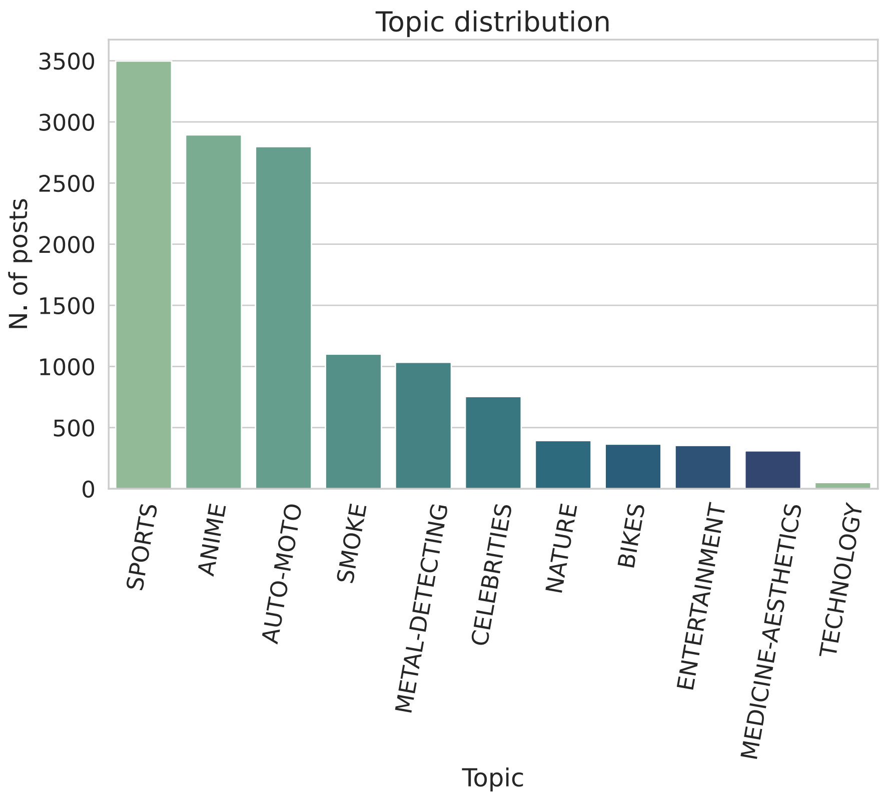

Figure 3: Topic distribution.

among the five classes, especially for the three intervals between *30* and *100*. For what concerns the *Gender* attribute, we can observe that the majority of posts were written by male users, thus determining a strongly unbalanced distribution of the two classes. The last variable, *Topic*, presents 11 labels, with 3 of them (ANIME, SPORTS and AUTO-MOTO) having more than 2,500 posts each.

In order to have enough data to fine-tune our pre-trained models, we decided to modify the original task as defined in [21]. Instead of predicting the three labels of a given collection of texts (multiple posts), we fine-tuned our models to predict age, gender and topic from each single post. Moreover, since a fair amount of sentences were quite short, we decided to remove those shorter than 10 tokens. At the end of this process, we obtained a dataset consisting of 13553 posts as training set and 5055 posts as test set.

3. Models

In what follows, we discuss more in detail the characteristics of the models used in our experiments.

IT5 We used the T5 base version pre-trained on the Italian language [1][1]. In particular, the model was trained on the Italian sentences extracted from a cleaned version of the mC4 corpus [24], a multilingual version of the C4 corpus including 107 languages. As discussed in [3], in order to compare different architectures (e.g. T5 and BERT), it would be ideal to analyze models with meaningful similarities, e.g. having a similar number of parameters or amount of computation to process an input-output sequence. Since T5 with n layers has approximately the same number of parameters as a BERT with $2n$ layers but also the same amount of computational cost of an n-layers BERT, in order to achieve the fairest comparison of the two Transformers, we decided to use the base version of IT5 (220M parameters).

BERT In order to compare the performance of IT5 with that of another Transformer model generically used in classification scenarios, we relied on a pre-trained Italian BERT. Specifically, we used the base cased BERT (12 layers, 110M parameters) developed by the MDZ Digital Library Team, available trough the Huggingface's *Transformers* library [25][2]. The model was trained using Wikipedia and the OPUS corpus [26].

4. Experimental Setting

As we already introduced in Sec. 1, we performed our experiments on two different classification scenarios: i) single-task and ii) multi-task classification. For what concerns the single-task scenario, we both fine-tuned BERT and IT5 three times in order to create three different single-task sequence classification models, one for each variable. To perform fine-tuning with the BERT model, we converted the three target variables into numeric labels. On the other hand, the target variables were verbalized empirically as follows for the IT5 model:

- *Gender*: values have been transformed in *uomo* and *donna*;
- *Topic*: values have been translated in Italian, written in lowercase and truncated into a single word (e.g. *MEDICINE-AESTHETIC* into *medicina*), thus resulting in the following list: *anime*, *automobilismo*, *bici*, *sport*, *natura*, *metalli*, *medicina*, *celebrità*, *fumo*, *intrattenimento*, *tecnologia*;
- *Age*: values have been left unchanged.

Moreover, following the *Fixed-prompt LM tuning* approach (see [27] for an overview), we added a prefix to each input when fine-tuning the IT5 model. This approach implies providing a textual template that is then applied to every training and test example. Fixed-prompt LM tuning has been already successfully explored for text classification, allowing more efficient learning [28, 29, 30]. In our experiments, we tested three different prefixes, one for each classification task: *"Classifica argomento"*, *"Classifica età"* and *"Classifica genere"*.

Concerning instead the multi-task classification, each sentence has been presented three times during the training phase of the two models, each one with the appropriate label and, in the case of IT5, with the appropriate prefix.

[1] https://huggingface.co/gsarti/it5-base
[2] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

Model	Topic		Age		Gender	
	Macro	**Weighted**	**Macro**	**Weighted**	**Macro**	**Weighted**
Dummy (S)	0.09	0.17	0.20	0.22	0.50	0.68
Dummy (MF)	0.04	0.10	0.09	0.14	0.44	0.69
BERT Random	0.14	0.34	0.26	0.27	0.56	0.74
IT5 Random	0.14	0.34	0.20	0.26	0.36	0.74
BERT	0.50	0.64	0.32	0.33	0.76	0.84
IT5	0.19	0.41	0.16	0.22	0.31	0.70
Multi-task						
MT BERT	0.56	0.67	0.32	0.33	0.75	0.84
MT IT5	0.31	0.52	0.16	0.23	0.33	0.71

Table 2
Macro and Weighted average F-Score for all the models and according to the tree classification variables. Results obtained with the multi-task models are also reported (*MT BERT/IT5*).

Few-Shot Learning In order to evaluate the performance of IT5 also in a context with little data available, we decided to carry out our classification experiments in a few-shot learning scenario. Specifically, we divided the original dataset into 5 equal subsets (1/5 = 2,710, 2/5 = 5,420, 3/5 = 8,130, 4/5 = 10,840, 5/5 = 13,554) and then we monitored the performance trend of both IT5 and BERT at increasing intervals of data samples: 0/5, 1/5, 2/5, 3/5, 4/5 and 5/5 of the TAG-IT dataset.

4.1. Baseline and Evaluation

We relied on two different typology of models as baseline. The first one is based on two dummy classifiers: i) *most frequent classifier* (Dummy (MF)), which always predict the most frequent label for each input sequence and ii) *stratified dummy classifier* (Dummy(S)), that generates predictions by respecting the class distribution of the training data. Moreover, in order to assess the impact of the pre-training phase of the two Transformer models, we also used a BERT Italian (*BERT Random*) and an IT5 model (*IT5 Random*) with randomly initialized weights.

We used F-Score (macro and weighted) as evaluation metric for all the experiments.

5. Results

Single-task Classification results are reported in Table 2. As we can observe, Transformer models outperformed the dummy baselines in almost all the classification tasks. The only exception concerns the performance of IT5 on the *Age* prediction task, for which the stratified dummy classifier obtained the same scores. It should be considered that the *Age* classification task appears to be the most complex task, regardless of the model taken into account. In fact, the best performing model (BERT) obtained only .11 points more than the baseline. The complexity in predicting the age ranges could be due to the fact that the task requires more sophisticated information rather than the simple identification of textual clues. On the other hand, the classifiers that achieved best results are those trained to predict the gender and the topic of each post. This result is in line with [21], where the authors suggested that textual clues seem to be more indicative of these dimensions than age. Moreover, the higher scores obtained for the gender classification task could also be indicative of the fact that, differently from the other two, gender prediction was cast as a binary task.

When we look at the performance obtained by the randomly initialized BERT and IT5, we note that the latter achieved results close to those of the pre-trained models. Indeed in some cases, e.g. *IT5* on the *Age* and *Gender* prediction tasks, the Random model gets better results. This seems to suggest that the pre-training phase of IT5 did not allow the model to encode enough useful information in order to improve its performance on the selected tasks. On the other hand, the pre-training phase had a strong impact on BERT performance, since the pre-trained model outperformed the Random one in all classification tasks.

If we focus instead on the differences between the two models, we can clearly notice that BERT performed best in all three configurations. In particular, IT5 achieved fairly reasonable results in comparison with BERT for simpler tasks, such as *Gender* and *Topic* classification. For what concerns the *Age* prediction task instead, we observed a performance drop, with a difference in terms of weighted F-Score of .17 points. A possible explanation for this behavior could be due to the fact that, differently from BERT, T5 has to produce the label by generating open text, thus making the prediction more complex from a computational point of view. In this regard, it is important to notice that for our experiments we relied on the base version of IT5, which, despite being bigger in terms of parameters than BERT base, is still quite smaller than the best-performing model (T5-11B) presented in [3]. Moreover, it should be pointed out that in some cases IT5 generated labels that did not belong to those defined in Sec. 4, but which actually turned out to be more accurate than the original ones. This is the case, for instance, of a few posts labelled with *fumo* (en. *smoke*) that were predicted instead by IT5 with the label *tabacco* (en. *tobacco*). We will inspect more in detail this behavior in Sec. 6. We also found that sometimes IT5 was not able to generate meaningful labels, but rather produced only punctuation marks or single letters. Nevertheless, we only identified a few isolated cases of them (less than 5 for what concerns *Topic* classification), which had no real impact on the overall performance of the model. We would like to also point out that the *IT5 Random* model does not generate unexpected labels like the pre-trained one does. This could be another motivation for its better performance in the two cases of *Age* and *Gender* classification.

Multi-task Observing the results obtained in the multi-task setting, we notice a significant increase in the performance of IT5. In fact, while BERT achieved a consistent boost only in the *Topic* prediction scenario, T5 performances improve significantly in all classification tasks, with an average improvement of around .06 points more (in terms of weighted F-Score) than during single-task classification. This is particularly evident with regard to *Topic* and *Age* classification, while the scores obtained for the *Gender* prediction task remain roughly the same. This result could suggest that, besides having more data for the fine-tuning phase, the IT5 model particularly benefits from learning multiple tasks at a time, thus improving its generalization abilities.

Few-Shot Learning Figures 4, 5 and 6 report the results obtained with the few-shot learning classification scenario. As we can see, the trend is quite different between the two models. In

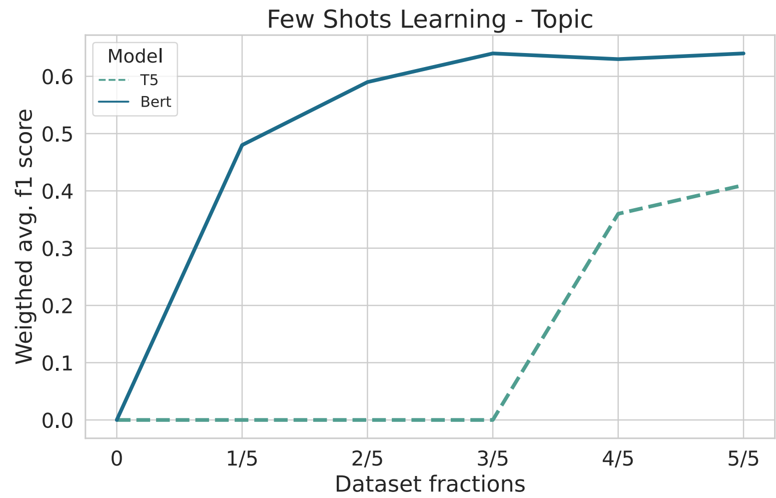

Figure 4: Few-Shot Learning results for *Topic* classification.

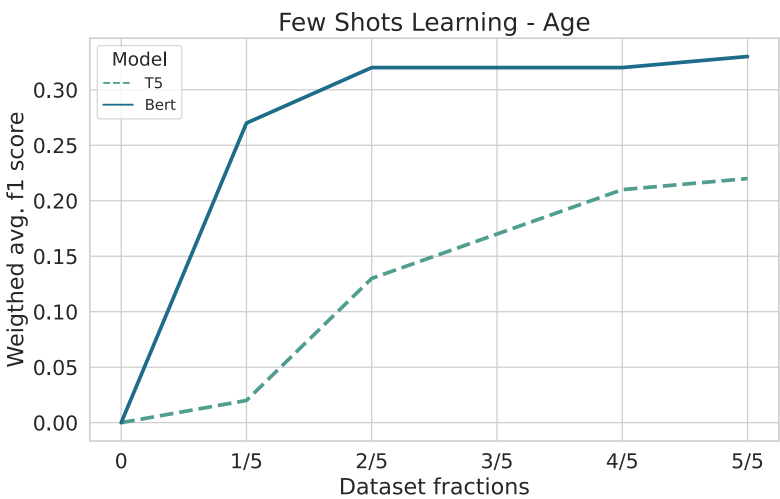

Figure 5: Few-Shot Learning results for *Age* classification.

fact, while BERT performance shows a fairly regular increase across the 5 fractions of the dataset, in the IT5 model we observe a quite constant improvement only for the *Age* prediction task. Interestingly, for what concerns *Topic* and *Gender* classification, IT5 makes correct predictions only after being exposed to 4/5 of the entire dataset. This behavior appears to be in line with what we already noticed during multi-task classification, namely that having more data available for the fine-tuning phase allows the model to perform better, and consequently, to obtain results closer to those of BERT. This seems to be further suggested by the fact that, unlike IT5, BERT obtains strong performance already from the early portions of the datasets but then it tends to

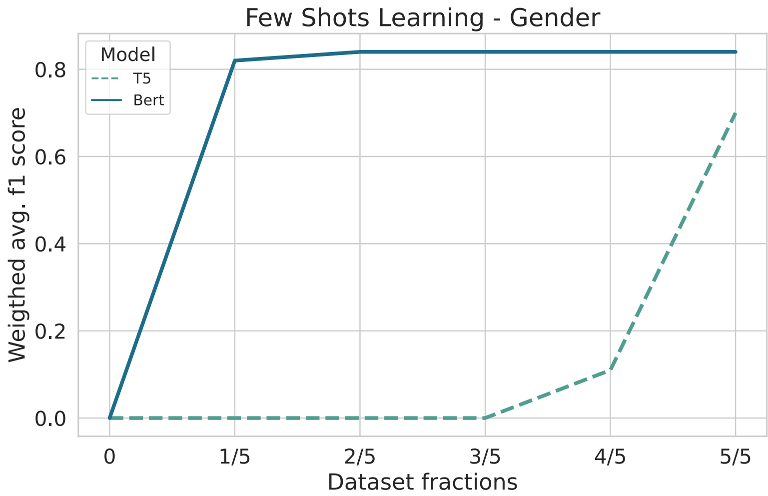

Figure 6: Few-Shot Learning results for *Gender* classification.

Sentence	**Predicted Label**	**Correct Label**
Che bell'acqua e che bei vitellini! Grande Pres.!	animali	celebrità
Perchè non l'alcool alimentare essendo neutro? E costa pure meno	alcol	fumo
terza miscela svizzera champagne eccellente! non vedo l'ora di tornare da two lions per altre miscele	bevande	fumo

Table 3
Examples of IT5 predictions.

remain quite stable, showing an improvement of only a few points in the remaining portions. This is especially the case of the *Gender* prediction task, where the accuracy of the BERT model in predicting the correct labels is roughly the same (.84 in terms of weighted F-Score) even after seeing 2/5 of the original dataset. Nevertheless, in the case of zero-shot learning, both models are unable to correctly classify the posts occurring in the test set of the three datasets.

6. Label Analysis

As described in [3], one of the issues of the Text-to-text framework applied in a classification scenario is that the model could outputs text on a task that does not correspond to any of the possible labels. However, as we already observed in the previous section, in some cases it seems that IT5 was able to generate more appropriate labels that those originally defined for the task, thus suggesting generalization abilities. For instance, as we can observe from the examples in Table 3, the labels predicted for the three input posts are not among those expected for the *Topic* prediction task. Nevertheless, by looking at the posts, the labels predicted by T5 might be

Model	Topic		Age		Gender	
	Macro	**Weighted**	**Macro**	**Weighted**	**Macro**	**Weighted**
IT5	0.19	0.41	0.16	0.22	0.31	0.70
IT5 shuffled	0.07	0.17	0.11	0.17	0.29	0.69

Table 4
Macro and Weighted F-Scores for the classification tasks obtained with IT5 using correct and shuffled labels (*IT5 shuffled*).

Labels	Macro	Weighted
m/f	0.32	0.70
uomo/donna	0.31	0.70

Table 5
Macro and Weighted F-Score on the *Gender* prediction task using *m/f* and *uomo/donna* as target variables.

considered more appropriate predictions.

Inspired by such behavior, we decided to further investigate the generalization abilities of the IT5 model by measuring the impact of different labels on model performance. More specifically, we decided to produce a shuffled version of each dataset by randomly replacing the labels with each other. Results are reported in Table 4. As we can see, the most significant variations in model performance concern the *Topic* and *Age* classification tasks. In particular, we can observe a drastic performance drop for what concerns *Topic*, with a difference between the predictions on correct and shuffled labels of more than .24 points in terms of Weighted F-Score. Moreover, it is interesting to note that the scores obtained with the shuffled labels are also lower than those obtained by the randomly initialized IT5 (0.17 vs. 0.34). This result seems to suggest that the IT5 model is indeed able to learn some specific lexical correlations between the encoding of the input tokens and of the labels during the fine-tuning phase and that these correlations are no longer observable after the shuffling process. This is also corroborated by the fact that, when presented with shuffled data, the model stopped generating new and more specific labels for the input sequences.

If we look instead at the results obtained with the *Gender* dataset, we can notice that shuffling the labels does not have a significant effect on the performance of the model. This is a clear evidence that, unlike *Topic*, the *Gender* prediction task does not present a direct lexical connection between the input sequence and the label. As a result, the model tends to memorize the information available in the fine-tuning data rather than derive generalities exploiting the knowledge learned during the pre-training phase.

Finally, inspired by the work of [20], we conducted further analysis on the effect of the strings used to represent labels on model performance. In particular, we decided to replace the labels used for the *Gender* prediction task (i.e. *uomo* and *donna*) with the original tags defined in the TAG-IT dataset, i.e. *m* and *f*. As shown in Table 5, modifying the label representation did not affect the performance of IT5, which obtained basically the same results in both configurations. This seems to confirm once again that for tasks that do not show an explicit relationship between input samples and labels, the choice of the label largely does not affect model performance.

7. Conclusions

In this paper, we proposed an extensive evaluation of the first Italian text-to-text model, IT5, on different classification tasks based on Italian sentences. Specifically, we chose to exploit the TAG-it dataset in order to measure the performance of the model in different classification scenarios.

First, we evaluated IT5 in a high data setting, assessing its performance during single- and multi-task classification and comparing them with the ones obtained by fine-tuning an Italian version of BERT. Results showed that IT5 is able to achieve quite good results, especially in *Topic* and *Gender* classification, and that its performance increases significantly when fine-tuned in a multi-task manner. Nevertheless, we found that BERT outperformed IT5 in all classification tasks.

Next, we tested the model in a poor data setting by measuring its performance in a few-shot learning scenario. Once again, IT5 achieved lower scores with respect to BERT, which obtained satisfactory results even in a context with very few data available (e.g. 1/5 of the entire dataset). A possible explanation of these results could be that given the high complexity of predicting the correct label by generating open text, it may be necessary to employ bigger text-to-text models to outperform models that are explicitly designed for solving classification tasks. Regardless of the classification scenario, we noticed that, especially for the *Topic* prediction task, IT5 occasionally generated labels that were not among those defined in the TAG-it dataset and that such labels often proved to be more indicative of the topic than the original ones. This result suggested that the model is indeed able to identify lexical clues indicative of the topic although in some cases it does not associate them with the labels that were originally defined for the task.

Finally, we investigated the impact of modifying the classification labels on IT5 performance. In particular, by shuffling at random the values of the original labels, we found that the model achieved generally lower scores and this is especially true for the classification of the topic. Nevertheless, experimenting with the *Gender* prediction task, we found that the choice of label representation does not affect significantly the model performance.

References

[1] G. Sarti, M. Nissim, It5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[2] 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.

[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer., J. Mach. Learn. Res. 21 (2020) 1–67.

[4] V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler,

T. Le Scao, A. Raja, et al., Multitask prompted training enables zero-shot task generalization, in: The Tenth International Conference on Learning Representations, 2022.
[5] V. Aribandi, Y. Tay, T. Schuster, J. Rao, H. S. Zheng, S. V. Mehta, H. Zhuang, V. Q. Tran, D. Bahri, J. Ni, et al., Ext5: Towards extreme multi-task scaling for transfer learning, in: International Conference on Learning Representations, 2021.
[6] B. McCann, N. S. Keskar, C. Xiong, R. Socher, The natural language decathlon: Multitask learning as question answering, arXiv preprint arXiv:1806.08730 (2018).
[7] N. S. Keskar, B. McCann, C. Xiong, R. Socher, Unifying question answering, text classification, and regression via span extraction, arXiv preprint arXiv:1904.09286 (2019).
[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners (????).
[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[10] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. R. Bowman, Glue: A multi-task benchmark and analysis platform for natural language understanding, in: 7th International Conference on Learning Representations, ICLR 2019, 2019.
[11] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2383–2392. URL: https://aclanthology.org/D16-1264. doi:10.18653/v1/D16-1264.
[12] R. Nogueira, Z. Jiang, R. Pradeep, J. Lin, Document ranking with a pretrained sequence-to-sequence model, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 708–718. URL: https://aclanthology.org/2020.findings-emnlp.63. doi:10.18653/v1/2020.findings-emnlp.63.
[13] K. Pipalia, R. Bhadja, M. Shukla, Comparative analysis of different transformer based architectures used in sentiment analysis, in: 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), IEEE, 2020, pp. 411–415.
[14] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems 32 (2019).
[15] D. Croce, A. Zelenanska, R. Basili, Neural learning for question answering in italian, in: International Conference of the Italian Association for Artificial Intelligence, Springer, 2018, pp. 389–402.
[16] D. Carmo, M. Piau, I. Campiotti, R. Nogueira, R. Lotufo, Ptt5: Pretraining and validating the t5 model on brazilian portuguese data, arXiv preprint arXiv:2008.09144 (2020).
[17] J. A. Wagner Filho, R. Wilkens, M. Idiart, A. Villavicencio, The brWaC corpus: A new open resource for Brazilian Portuguese, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1686.
[18] L. Real, E. Fonseca, H. Gonçalo Oliveira, The assin 2 shared task: a quick overview, in: International Conference on Computational Processing of the Portuguese Language, Springer, 2020, pp. 406–412.
[19] D. Santos, N. Seco, N. Cardoso, R. Vilela, Harem: An advanced ner evaluation contest for

portuguese, in: quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006), 2006.

[20] X. Chen, J. Xu, A. Wang, Label representations in modeling classification as text generation, 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: Student Research Workshop, 2020, pp. 160–164.

[21] Cimino, Dell'Orletta, Nissim, Tag-it – topic, age and gender prediction, EVALITA (2020).

[22] V. Basile, M. Di Maro, D. Croce, L. Passaro, Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian, in: 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2020, volume 2765, CEUR-ws, 2020.

[23] A. Maslennikova, P. Labruna, A. Cimino, F. Dell'Orletta, Quanti anni hai? age identification for italian., in: CLiC-it, 2019.

[24] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, 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. 483–498. URL: https://aclanthology.org/2021.naacl-main.41. doi:10.18653/v1/2021.naacl-main.41.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[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. URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/320.pdf.

[27] 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, arXiv preprint arXiv:2107.13586 (2021).

[28] T. Schick, H. Schütze, Few-shot text generation with pattern-exploiting training, arXiv preprint arXiv:2012.11926 (2020).

[29] 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.

[30] T. Gao, A. Fisch, D. Chen, Making pre-trained language models better few-shot learners, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics

and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 3816–3830. URL: https://aclanthology.org/2021.acl-long.295. doi:10.18653/v1/2021.acl-long.295.