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

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

[1] Istituto di Linguistica Computazionale "Antonio Zampolli", ItaliaNLP Lab, Pisa
[2] TALIA S.R.L., Pisa
[3] Università di Pisa

Premise: The text-to-text paradigm has become one of the most widely adopted approach in the development of state-of-the-art NLMs, offering stunning performances even in data-poor settings for a variety of NLP tasks.

Premise: The text-to-text paradigm has become one of the most widely adopted approach in the development of state-of-the-art NLMs, offering stunning performances even in data-poor settings for a variety of NLP tasks.

Open Issue: Are text-to-text NLMs really suited for any kind of task? How do they fare in tasks that are usually tackled without the use of generative NLMs?

Premise: The text-to-text paradigm has become one of the most widely adopted approach in the development of state-of-the-art NLMs, offering stunning performances even in data-poor settings for a variety of NLP tasks.

Open Issue: Are text-to-text NLMs really suited for any kind of task? How do they fare in tasks that are usually tackled without the use of generative NLMs?

Overall Results: No correlation was found between the sets of representation (ranked by *similarity* with the class they represent) and the model weighted F-score.

Three different classification tasks: Gender (2 classes), Topic (11 classes) and Age classification (5 classes).

Data distribution: The Age distribution is skewed towards the 20-29 range but overall is balanced, Gender is heavily skewed towards the Male class and Topic presents both very high frequency classes (Sports, Anime and Auto-moto) and very low frequency classes (Technology, Medicine-Aesthetics).

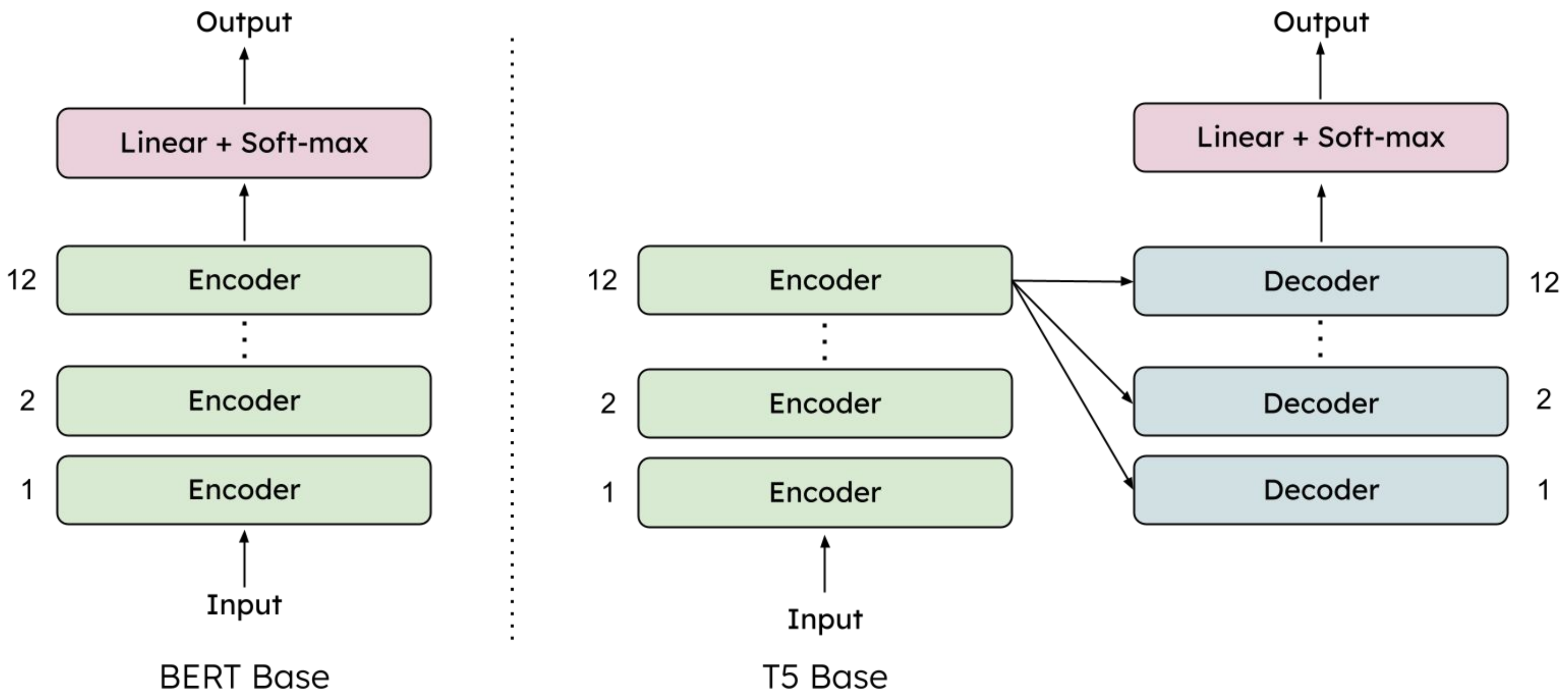
The classification objective has been changed from profiling an author given a collection of posts, to predict one of the three classes, or all three, given a single post.

Dataset dimension: 13.553 posts formed the training dataset and 5055 the test dataset.

IT5: Is a T5 (Raffel et al., 2020) pre-trained for the Italian language. The model is trained on the Italian sentences extracted from a cleaned version of the mC4 corpus (Xue et al., 2021), a multilingual version of the C4 corpus including 107 languages.

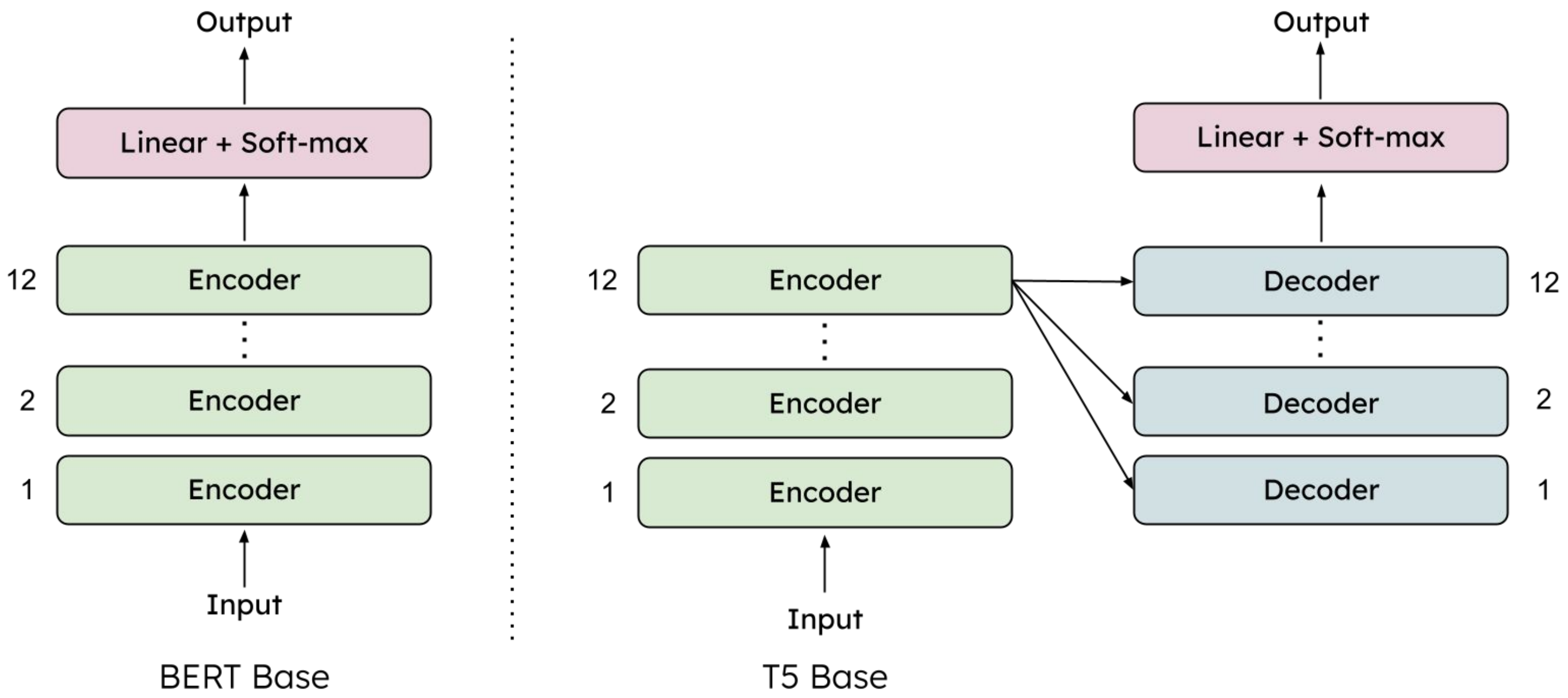
IT5: Is a T5 (Raffel et al., 2020) pre-trained for the Italian language. The model is trained on the Italian sentences extracted from a cleaned version of the mC4 corpus (Xue et al., 2021), a multilingual version of the C4 corpus including 107 languages.

BERT (Devlin et al., 2018): We used the cased BERT pre-trained for the Italian language using Wikipedia and the OPUS corpus (Tiedemann et al., 2004) by the MDZ Digital Library Team.



Model Selection

talia



We decided to use **BERT Base** (110M parameters) and **IT5 Base** (220M parameters).

Model Selection

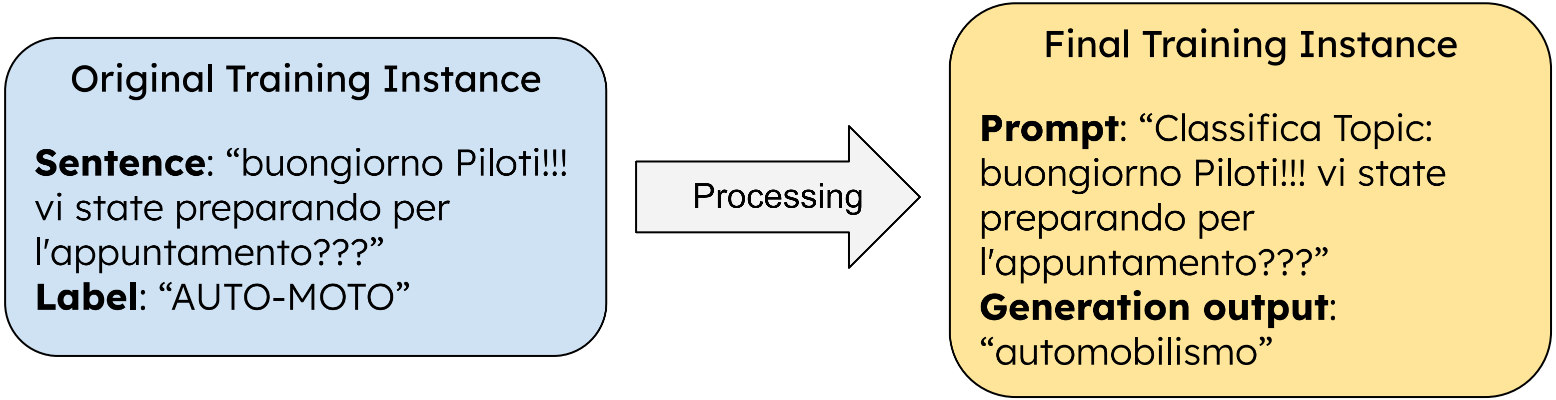
Class labels have been verbalized into single words. Ex: for Topic classification, *MEDICINE-AESTETHICS* → *medicina*

Class labels have been verbalized into single words. Ex: for Topic classification, *MEDICINE-AESTETHICS* $\longrightarrow$ *medicina*

A task prefix was added to each sentence to form the prompt. Topic: *“Classifica argomento”*, Age: *“Classifica età”*, Gender: *“Classifica genere”*

Class labels have been verbalized into single words. Ex: for Topic classification, *MEDICINE-AESTETHICS* → *medicina*

A task prefix was added to each sentence to form the prompt. Topic: *“Classifica argomento”*, Age: *“Classifica età”*, Gender: *“Classifica genere”*



IT5 Pre-processing

Single Task: We fine-tuned three BERT models and three T5 models, one for each task (Gender, Topic and Age classification).

Single Task: We fine-tuned three BERT models and three T5 models, one for each task (Gender, Topic and Age classification).

Multi-task: Each sentence has been presented three times for the fine-tuning of both multitask models, each time with the appropriate label and, for T5, the correct task prefix.

Single Task: We fine-tuned three BERT models and three T5 models, one for each task (Gender, Topic and Age classification).

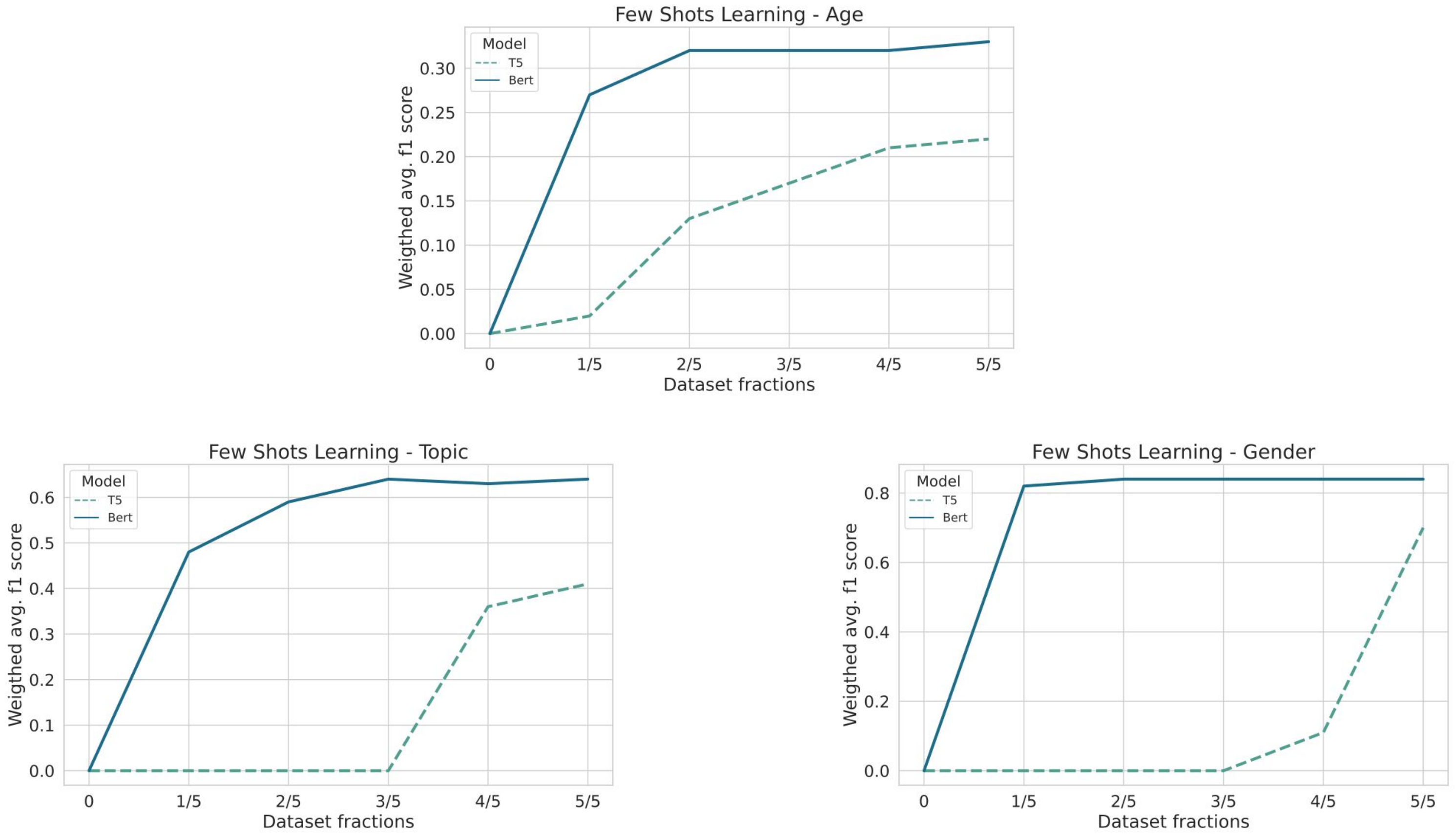
Multi-task: Each sentence has been presented three times for the fine-tuning of both multitask models, each time with the appropriate label and, for T5, the correct task prefix.

Few-shot: We evaluated the performance of the single tasks models using increasing intervals of data samples (1/5, 2/5, 3/5, 4/5 of the training dataset).

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

Macro and Weighted average F-Score for all models and for all classification tasks. In **bold** the highest result for each task.

Single- and Multi-task results



Few-shot results

talia

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	alcool	fumo
terza miscela svizzera champagne eccellente! non vedo l’ora di tornare da two lions per altre miscele	bevande	fumo

IT5 making *wrong* but meaningful predictions.

IT5 wrong predictions examples

To test the impact of *lexical connections* between the prompts and the labels, we repeated the single-task experiments with a shuffled dataset:

To test the impact of *lexical connections* between the prompts and the labels, we repeated the single-task experiments with a shuffled dataset:

- **Topic**: the labels have been shuffled randomly;

To test the impact of *lexical connections* between the prompts and the labels, we repeated the single-task experiments with a shuffled dataset:

- **Topic**: the labels have been shuffled randomly;
- **Gender**: the labels *uomo* and *donna* have been swapped;

To test the impact of *lexical connections* between the prompts and the labels, we repeated the single-task experiments with a shuffled dataset:

- **Topic**: the labels have been shuffled randomly;
- **Gender**: the labels *uomo* and *donna* have been swapped;
- **Age**: the labels have been mixed trying to maximize the ordinal distance between the original and the shuffled label;

To test the impact of *lexical connections* between the prompts and the labels, we repeated the single-task experiments with a shuffled dataset:

- **Topic**: the labels have been shuffled randomly;
- **Gender**: the labels *uomo* and *donna* have been swapped;
- **Age**: the labels have been mixed trying to maximize the ordinal distance between the original and the shuffled label;

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

Macro and Weighted F-Score for the three classification tasks done with IT5 trained on the original and on the shuffled dataset (*IT5 shuffled*).

We evaluated the performance of IT5 in both **single- and multi-task classification** scenarios: comparing it to a BERT that use the same amount of computation, the latter performed better.

We evaluated the performance of IT5 in both **single- and multi-task classification** scenarios: comparing it to a BERT that use the same amount of computation, the latter performed better.

We tested the model performance in a **few-shot learning** scenario. Again, BERT performed better, requiring less data than T5 to achieve satisfactory results.

We evaluated the performance of IT5 in both **single- and multi-task classification** scenarios: comparing it to a BERT that use the same amount of computation, the latter performed better.

We tested the model performance in a **few-shot learning** scenario. Again, BERT performed better, requiring less data than T5 to achieve satisfactory results.

We tested the importance of **label representation** for the three tasks and found out that for tasks with an explicit lexical connection between the prompt and the label, the choice of representation for the label have a strong impact on performances.

Thank you for your attention!

Contacts

Mail: michele.papucci97@gmail.com
Twitter: @mpapucci_
Linkedin: linkedin.com/in/michelepapucci

talia