Evaluating Pre-Trained Transformers on Italian Administrative Texts

Serena Auriemma[1], Martina Miliani[1,2], Alessandro Bondielli[3], Lucia C. Passaro[3] and Alessandro Lenci[1]

[1]*Department of Philology, Literature, and Linguistics, University of Pisa*
[2]*University for Foreigners of Siena*
[3]*Department of Computer Science, University of Pisa*

Abstract

In recent years, Transformer-based models have been widely used in NLP for various downstream tasks and in different domains. However, a language model explicitly built for the Italian administrative language is still lacking. Therefore, in this paper, we decided to compare the performance of five different Transformer models, pre-trained on general purpose texts, on two main tasks in the Italian administrative domain: Name Entity Recognition and multi-label document classification on Public Administration (PA) documents. We evaluate the performance of each model on both tasks to identify the best model in this particular domain. We also discuss the effect of model size and pre-training data on the performances on domain data. Our evaluation identifies UmBERTo as the best-performing model, with an accuracy of 0.71, an F1 score of 0.89 for multi-label document classification, and an F1 score of 0.87 for NER-PA.

Keywords

Natural Language Processing, Evaluation of Neural Language Models, Domain Language, Public Administration

1. Introduction

Today, language technologies are indispensable for facilitating interaction between citizens and the Public Administration (PA). Natural Language Processing tools represent a practical resource for managing the vast data in the PA domain. They can be leveraged in many ways to extract the implicit information in administrative texts and transform it into structured data that are more easily accessible, manageable, shareable, and secure.

In recent years, pre-trained language models are gradually emerging as a new paradigm in Natural Language Processing. The main advantage of these models is that they are already trained on self-supervised tasks and they can be fine-tuned for a variety of downstream tasks with a relatively small amount of data and training iterations. In addition, some of these

AIxPA 2022: 1st Workshop on AI for Public Administration, December 2nd, 2022, Udine, IT
serena.auriemma@phd.unipi.it (S. Auriemma); m.miliani@studenti.unistrasi.it (M. Miliani); alessandro.bondielli@unipi.it (A. Bondielli); lucia.passaro@unipi.it (L. C. Passaro); alessandro.lenci@unipi.it (A. Lenci)
https://colinglab.humnet.unipi.it/people/miliani/ (M. Miliani); https://scholar.google.com/citations?user=zcXQk6YAAAAJ&hl=en (A. Bondielli); https://luciacpassaro.github.io/ (L. C. Passaro); https://people.unipi.it/alessandro_lenci/ (A. Lenci)
0000-0003-1124-9955 (M. Miliani); 0000-0001-5790-4308 (A. Lenci)

models, such as BERT [1] and RoBERTa [2], have achieved significant improvements in several NLP tasks, and Transformers [3] lie at the core of many recent architectures [4]. Pre-trained language models are advantageous in contexts where labelled data are limited while large amounts of unlabelled data are available. This is the case for underrepresented languages or specific domains. Indeed, several attempts have also been made to adapt these models to specific domains via additional pre-training on domain data, which improved their performance[5]. For example, BERT, one of the most widely used Transformer models, has several variants in different domains: MedBERT [6] and BioBERT [7] are further pre-trained on clinical and biomedical data, respectively, to solve medical-related tasks, such as biomedical NER or disease prediction; SciBERT [8] is fine-tuned on a suite of tasks including sequence tagging, sentence classification and dependency parsing on data from a variety of scientific domains; Legal BERT [9] is pre-trained on English legislative documents, contracts and trial records for legal text classification and sequence tagging. Nevertheless, a language model specifically built for the Italian administrative domain is still lacking. Thus, we sought to explore the use of generic models pre-trained on general text data for a specialized domain, such as PA. Domain-agnostic pre-trained models are often used as a baseline to which compare the performance of domain-adapted models[10, 7]; alternatively, their performance is usually compared with that of non-Transformer-based models, such as Conditional Random Field (CRF), Convolutional Neural Network(CNN), Long Short Term Memory (LSTM) [11, 12]. Very few approaches evaluate the performance of different Transformer models on domain data, before choosing the one to adapt, and most of them concern the medical domain. For instance, Polignano et al.[13] created a hybrid NER model for analyzing textual medical diagnoses in Spanish, based on a configuration of different Transformer-based models and a CRF, to detect the best performing one for the task. For the same purpose, Khan et al. [14] carried out a comparison of nine Transformer-based models on a classification task of mentioning health twitter in English. Hence, we decided to compare the performance of different Transformer-based models on two main NLP downstream tasks, which we consider particularly congenial to the needs of PA: Named Entity Recognition and Document Classification.

Named Entity Recognition (NER) is the task of identifying and classifying entities mentioned in unstructured text into predefined categories, e.g., proper names of people, places, organizations, dates, etc. [15]. A NER for PA, such as the one proposed by [16], includes other classes that are particularly relevant to the administrative domain, such as the references to other administrative documents, the legislative references, and organizations related explicitly to PA. Identifying relevant entities in a document can facilitate intelligent access and search within documents by municipalities.

Similarly, the document classification task associates each document with a label related to its content. It has extensive applications, including topic labeling, sentiment classification, and spam detection [17]. Since PA handles documents treating different aspects of the municipality, such as *Environment*, *Urban Planning*, and *Public Education*, automatically classifying documents according to their content can smooth the information retrieval process and simplify the management and protocol of PA's document repository. Transformer-based models have been shown to handle these two tasks with remarkable results [18, 19, 20, 21]. Moreover, various off-the-shelf models are available in Italian or have multilingual versions. Thus, we decided to compare their performances in these two tasks in the domain of Italian administrative

documents without any additional domain pre-training. Given the inherent differences of the administrative language style from the standard Italian on which these models were pre-trained, it is undoubtedly interesting to compare their performances on these two tasks to assess the capabilities of such models in the PA scenario.

Generally, Transformer-based models vary in size, number of parameters, pre-training objective, and pre-training data. Hence, it is a challenging decision to select one of these models for a specific downstream task. Therefore, we decided to assess their performance in the administrative domain to identify which model best suits each of these tasks and for this particular domain.

The main contributions of this work are:

- a comparative evaluation of five Transformer-based models on two main downstream tasks, namely NER and multi-label document classification in the administrative domain;
- the creation of a new Italian dataset for multi-label document classification in the administrative domain, which we called ATTO;
- a comparison between Transformer-based models and a PA-specialized machine learning model on NER.

The rest of this paper is organized as follows. Section 2 provides a brief overview of the Transformer models we considered for comparison. Section 3 presents the experimental details. We describe the datasets, the parameter configuration used for fine-tuning the models on the NER-PA and multi-label document classification tasks, the baselines, and the metrics used for evaluation. Section 4 reports on the results and discusses them. Finally, Section 5 draws some conclusions and highlights future works.

2. Transformer Models

In this work we focus on Transformer-based models pre-trained on the Italian language or available in a multilingual version, to identify which model represents the best choice for handling administrative data, even in the absence of adaptive domain pre-training. We compare the performance on NER-PA and PA multi-label document classification of 5 Transformer-based language models: BERT-base-Ita,[1] UmBERTo,[2] Multilingual BERT,[3] XLM-RoBERTa,[4] and GePpeTto.[5] .

BERT [1] is a bidirectional Transformer model pre-trained on a multi-task objective (Masked Language Modeling and Next Sentence Prediction) over vast amounts of text and can be fine-tuned for various NLP downstream tasks. The Italian version of BERT, henceforth BERT-Ita, was pre-trained on a Wikipedia dump and various texts from the OPUS corpora collection[6] for a total corpus size of 13 GB of texts.

UmBERTo is a RoBERTa-based Language Model trained on a subset of the OSCAR corpus[7]

[1] https://huggingface.co/dbmdz/bert-base-italian-uncased
[2] https://huggingface.co/Musixmatch/umberto-wikipedia-uncased-v1
[3] https://huggingface.co/bert-base-multilingual-uncased
[4] https://huggingface.co/xlm-roberta-base
[5] https://huggingface.co/LorenzoDeMattei/GePpeTto
[6] https://opus.nlpl.eu
[7] https://oscar-corpus.com

Table 1
Transformer-based language models characteristics.

Model Name	Based on	Model size	Corpus size	Multilingual
BERT-Ita	BERT-base	109M	13GB	No
mBERT	BERT-base	177M	-	Yes
UmBERTo	RoBERTa-base	110M	70GB	No
XLM-RoB.	RoBERTa-base	278M	2.5TB	Yes
GePpeTto	GPT-2	108M	13GB	No

containing about 70 GB of plain text. Compared to BERT, RoBERTa was trained longer, over more data, and with larger training batch size. In addition, the training process was carried out with a dynamic masking of tokens in place of the Next Sentence Prediction task [2]. This resulted in an improved performance of the model on various GLUE [22] benchmark results. UmBERTo is an expanded version of RoBERTa that contains two new features: *Sentence Piece Model* and *Whole Word Masking*. Sentence Piece Model is a language independent tokenizer that generates sub-word units specific to the chosen vocabulary size and corpus language. With Whole Word Masking (WWM), if at least one of the tokens created by Sentence Piece Tokenizer is chosen as a mask, the mask is applied to the entire word. Thus, only whole words are masked, not sub-words.

The multilingual version of BERT (mBERT) [1] achieves language understanding by training the MLM task with shared vocabulary and weights on Wikipedia texts from the 104 top languages. Each training sample is a monolingual document, and there are no explicitly designed cross-linguistic objectives or cross-linguistic data. Despite this, mBERT is successful in cross-linguistic generalisation [23].

Another multilingual encoder is XLM-RoBERTa. This model was pre-trained on a significantly larger amount of data, 2.5 TB of clean Common Crawl data in 100 different languages. Like mBERT, XLM-RoBERTa pre-training task is solely monolingual MLM. It achieves state-of-the-art results on several multilingual benchmarks, including XNLI, MLQA, and NER, outperforming mBERT [24].

GePpeTto [25] is the first autoregressive language model for Italian. It is built using the GPT-2 architecture [26]. The latter is a scale-up of GPT [27], with 10x parameters and 10x training data. GPT-2 is a unidirectional generative Transformer model trained on next token prediction given all of the previous words within some text. Its Italian version, GePpeTto was trained on a collection of Wikipedia Text and the ItWac corpus [28], amounting to almost 13GB.

The Transformer-based models we discussed differ with regard to pre-training objective, size, and number of parameters. Moreover, they are also pre-trained on corpora of different sizes, as highlighted in Table 1.

3. Experimental settings

We aimed to compare the performance of five Transformer-based models on two different tasks related to the administrative domain. To this end, we fine-tuned these five models on

two main downstream tasks in the administrative domain: NER-PA and multi-label document classification. More specifically, we chose the `base-cased` version of all the models and fine-tuned each model on two datasets, one for each task. We describe the used datasets in Section 3.1. We compared their performance in terms of precision, recall, F1 score, and accuracy while analyzing the model pre-training and architecture effect on each task, as described in Section 3.5.

3.1. Datasets

In order to fine-tune the models on the NER task for the administrative domain, we selected a corpus containing documents from the Italian Public Administration, i.e. the PA Corpus. The corpus is annotated with domain entity labels. As for the multi-label document classification task, we proposed a new dataset, named the ATTO corpus, specifically for this purpose. Unfortunately, the raw annotated dataset used in this paper cannot be released due to sensitive information being present in the data. However, trained models (both for NER and document classification) are available via huggingface[8].

3.1.1. PA Corpus.

It was first proposed in [16] for building a Named Entity Recognizer, named INFORMed PA, specifically designed for the administrative domain. INFORMed PA extended the traditional NER classes [29] (i.e. Person, Locations, Organizations, and Miscellaneous Entities) with other classes representing the administrative domain. As for the traditional classes, the dataset includes the class LOC, which is used for marking both geo-political entities (e.g., "Comune di Pisa") and locations (e.g., "via S. Maria 36, Pisa"); the class PER, which is used for identifying a physical subjects; the class ORG, used for marking organizations such as Companies. As for the PA-specific classes, it includes the following additional classes: ORG_PA, LAW, and ACT. ORG_PA is used for labelling organizations specifically related to PA (e.g., the Municipality Departments). LAW denotes instead legislative references. Finally, ACT marks references to other administrative documents. In particular, ACT is further divided in several sub-classes. Specifically, the annotation distinguishes their type (ACT_T), number (ACT_N), date (ACT_D), functional tokens (ACT_X) and unparsable tokens (ACT_U). For example, the ACT *Delibera di Giunta Comunale numero 53 del 23/10/2016* is annotated as follows: *Delibera di Giunta comunale* (ACT_T) *numero* (ACT_X) *53* (ACT_N) *del* (ACT_X) *23/10/2016* (ACT_D), while the act *DD/67/2012* is annotated as ACT_U.

The corpus contains 460 documents taken from the *Albo Pretorio Nazionale* for a total of 724,623 tokens. The first 100 documents of the corpus were annotated using the aforementioned Name Entities (NE) by two annotators, including one domain expert. Then, a CRF model was trained using these documents and it was used to automatically annotate new documents. Finally, two other annotators manually revised the entire output. The distribution of the different NEs in the corpus, as well as examples, is listed in Table 2.

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

Table 2
The distribution of Named Entities in the PA Corpus.

Tag	Freq.	Example
LAW	5217	*art. 183 comma 7 del D.Lgs. n. 267/2000*
LOC	4498	*Comune di Pisa*
PER	3706	*Mario Rossi*
ORG	3594	*Consip Spa*
ACT	2240	*Determina n. 4 del 12/02/2011*
ORG_PA	2074	*Sezione Anagrafe*

3.1.2. ATTO.

The ATTO (**A**dministrative **T**exts labeled by **TO**pic) corpus is a dataset that includes administrative texts that are labeled with one or more topics (e.g., *Environment, Construction, Urban Planning, Social, etc.*) pertaining to PA. The dataset was built upon a domain ontology created in the context of the project SEMPLICE.[9] The ontology contains approximately 2,700 administrative domain terms that domain experts divided into 13 classes, each corresponding to a different sub-sector of the PA (e.g. *Environment, Construction, Urban Planning, Social, etc.*). We collected up to about 1,000 documents for each of the 13 classes of topics. The collected data are a subset of a dataset including documents from several Italian municipalities annotated and indexed by topic. The complete list of considered topics, as well as their distribution, are shown in Table 3.

The corpus was built by querying the whole PA document collection by topic. As each document was associated with more than one topic in the original dataset, we only selected unique documents (i.e., a document retrieved through two or more different topics was added only once to the ATTO dataset). We further filtered the documents, to obtain high-quality data. Specifically, we removed documents containing OCR-level errors, which were identified by means of shallow matching rules. Finally, we removed documents with a length higher than 600 words. The final version of ATTO includes 11,019 documents.

3.2. Named Entity Recognition in the PA domain

In order to evaluate the performances of the chosen Transformer models on Named Entity Recognition for PA (NER-PA), we fine-tuned each model on the PA corpus (see Sec. 3.1.1). First, the data are pre-processed to unify cross-sentence entities, merging sentences with a common entity. Then, we split the dataset into 70% training, 20% validation, and 10% test. In addition to the test set, we evaluated the models on a different test set of 25 documents from 25 various municipalities following the original approach proposed in [16]. The goal of this further analysis is to test the performance of the models on different templates and ways of referring to entities employed by the various municipalities. As for the actual training, we fine-tuned each model for five epochs with a learning rate of 2e-5. Due to differences in model size and limited hardware availability, we chose to vary the training batch size from model to model to best fit our available memory, to obtain the best possible results for each model. The training

[9] SEMantic instruments for PubLIc administrators and CitizEns : www.semplicepa.it

Table 3
The distribution of the topic labels in the ATTO corpus.

Topic	Number of Documents
Environment	999
Demographics	716
Advocacy	1433
Tenders and Contracts	3928
Trade and Business	210
Culture, tourism and sport	381
Constructions	1358
Personnel	1625
Public Education	951
Information services	1613
Financial services	5380
Social	1419
Urban Planning	1567
Total	*11019*

was performed on a desktop computer equipped with an NVIDIA TITAN RTX graphic card. BERT-Ita, mBERT and XLM-Roberta were fine-tuned with a batch size of 16, while GePpeTto and UmBERTo with a batch size of 8 and 4, respectively. All models belong to the Huggingface Transformers library.[10]

3.3. Multi-label document classification

As for the multi-label document classification on the ATTO dataset (see Sec. 3.1.2), the training process was straightforward. We chose to perform 5-fold cross-validation, in order to ensure the reliability of the results in absence of a test set. We first performed a preliminary evaluation in order to find the best hyperparameters. By considering the per-epoch training and validation loss, we concluded that all the models could be trained for 10 epochs before starting to overfit. Thus, we fine-tuned our models for 10 epochs using a batch size of 16, 2e-5 as learning rate, and a maximum sequence length of 512. The final results for each model are obtained by averaging the performances on each training fold.

3.4. Baselines

For what concerns NER, we used as baseline the results obtained on the same datasets by Informed PA [16], a PA-specialized machine learning model based on the Stanford NER, using a Conditional Random Field (CRF) as a learning algorithm.

As for the baseline of the multi-label document classification task, we implemented a Bidirectional LSTM (bi-LSTM) model. We constructed the model with one bi-LSTM layer with 256 neurons and a single dense output layer with 13 neurons since we have 13 labels in the dataset. We chose the Sigmoid as the activation function and the binary-cross entropy as the

[10]https://huggingface.co/docs/Transformers/index

Table 4
Results for 5-fold cross validation on multi-label document classification.

Model Name	F1-score	Accuracy
BERT-Ita	0.881	0.692
mBERT	0.857	0.647
XLM-RoBERTa	0.883	0.692
UmBERTo	**0.892**	**0.708**
GePpeTto	0.870	0.671
Bi-LSTM	0.386	0.545

loss function. We also pre-processed the documents to feed the bi-LSTM removing line breaks, typycal of administrative acts layout. We fed the neural network with 100-dimensional vector representations of the first 250 tokens of each document obtained using Word2Vec [30]. We performed 5-fold cross-validation on the ATTO corpus, as for all the other Transformer-based models, and we trained the bi-LSTM for ten epochs with a batch size of 128. The final results of the model are obtained by averaging the result of each training fold.

3.5. Evaluation Metrics

As the two considered tasks are fairly standard, we evaluated the performances on common metrics for classification. We used micro and macro averages for precision, recall, and F1-score for the NER-PA task. For the multi-label document classification, accuracy, precision, recall and micro average F1-score were considered. Note that the micro average F1-score on multi-label classification corresponds to the harmonic mean between precision and recall, while accuracy refers to the number of data points for which the predicted set of labels is identical to the true one over all the dataset.

4. Results and Discussion

Table 4 summarizes the results of the models on the multi-label document classification task. Differences in performances are clearly minimal across all Transformer-based models, which significantly outperform the baseline with regards to F1-score. The best performing model was UmBERTo with an F1-score of 0.89 and an accuracy of 0.71. BERT-Ita and XLM-RoBERTa both achieved an accuracy of 0.69 and an F1-score of 0.88, followed closely by GePpeTto which achieved 0.67 accuracy and 0.87 F1-score. The worst performing model among Transformers was mBERT, whose accuracy and F1-score were 0.65 and 0.86, respectively. Ultimately, the bi-LSTM model achieved an accuracy of 0.55 and an F1-score of 0.39, both far below the average score obtained by Transformers.

These findings indicate that the most suited model for a multi-label classification task on administrative documents is UmBERTo. This is probably due to the fact that, with respect to the other monolingual Italian models (i.e., BERT-Ita and GePpeTto, UmBERTo), it has a bigger size and was pre-trained on more data. While, comparing Bert-Ita and GePpeTto, despite the models having approximately the same number of parameters and the same amount of data used

for pre-training, it appears that BERT-Ita represents a more suitable architecture to perform a multi-label classification task than the generative model GePpeTto.

Regarding the multilingual models, our results confirm what has been reported in the literature for the general domain, i.e., the monolingual models perform better than their multilingual counterparts [31]. Indeed, mBERT was the model with the lowest scores in both accuracy and F1-score, and, although XLM-RoBERTa is almost three times as large as BERT-Ita in terms of parameters, their performance was identical.

As highlighted in Table 5, which shows the precision, recall, and F1-score obtained by the models for each topic class, the ranking of the models is quite consistent across all classes: the best-performing model, UmBERTo, obtained the highest F1-score in most cases, followed by Bert-Ita and XLM-Roberta, GePpeTto, mBERT, and lastly the bi-LSTM model. Only two classes constitute an exception: *Trade and Business* and *Culture, tourism, and sport*. In these cases, the best-performing models were GePpeTto and XLM-Roberta, although the scores are still lower when compared to the results of the other classes. That is probably related to the number of documents labeled with these classes in the dataset, as they have fewer instances than all others (see Table 3). Regarding the baseline, we can observe that the bi-LSTM performs poorly on almost the whole dataset. Only in the *Financial Service* class, the most represented in the corpus, the bi-LSTM achieves an F1-score of 0.88, comparable to that of the Transformer models.

Table 6 provides the results of each model for the NER-PA task in terms of precision, recall, and F1-score. These results refer to the scores obtained on the PA Corpus test set. Table 7 reports the results obtained by the models on the additional dataset of 25 administrative documents from 25 different municipalities.

As for the multi-label document classification, the best performing model was UmBERTo, with a micro average precision of 0.86, a recall of 0.89, and F1-score of 0.87. XLM-RoBERTa obtained the same recall score as UmBERTo, but with lower precision and F1-score, respectively of 0.83 and 0.86. BERT-Ita and mBERT achieved an equal precision of 0.83 and F1-score of 0.85. BERT-Ita performed better than mBERT in terms of recall, with 0.88 compared to 0.87. Again, the model that performed worse was GePpeTto with a micro average precision of 0.69, recall of 0.80, and F1-score of 0.74, deviating from the best model by about 0.13 F1-score points.

On the 25 document dataset, the best-performing Transformer models was XLM-RoBERTa, scoring 0.8, 0.88, and 0.84 on micro average precision, recall and F1-score, respectively. The second best performing model was UmBERTo, followed by BERT-Ita and mBERT. GePpeTto was the worst performing model with 0.63 precision, 0.78 recall, and an F1-score of 0.7.

This performance comparison highlights that for the token classification task, the amount of monolingual data used for pre-training appears to play a crucial role, together with the model's size. We observed that the best performing model on the PA corpus was UmBERTo, which is pre-trained on a much larger sample of Italian texts with respect to all the other models. On the 25 document sample, which includes a higher variability in terms of how Named Entities are mentioned in the text, the larger model in terms of parameters, namely XLM-RoBERTa, has a slight advantage over UmBERTo. Moreover, results show that a rather small generative model, in comparison with current state-of-the-art ones such as GPT-3 [32], is the least suited for fine-tuning on a token classification task.

If we look at the detailed scores for each class on the 25 documents dataset, we can also observe that the classes for which the Transformer models struggle the most are ORG_PA

Table 5
Performance comparison of the models on the ATTO corpus

Doc Label	Metric	BERT-Ita	mBERT	XLM-RoB	UmBERTo	GePpeTto	Bi-LSTM
Environment	P	0.853	0.798	0.861	0.865	0.844	0.400
	R	0.782	0.734	0.810	0.839	0.759	0.043
	F1	0.815	0.763	0.834	**0.852**	0.798	0.074
Advocacy	P	0.919	0.929	0.922	0.949	0.904	0.849
	R	0.868	0.833	0.851	0.866	0.869	0.402
	F1	0.893	0.878	0.885	**0.905**	0.886	0.539
Tenders and Contracts	P	0.876	0.859	0.869	0.886	0.845	0.823
	R	0.880	0.874	0.898	0.914	0.890	0.530
	F1	0.878	0.867	0.883	**0.900**	0.867	0.638
Trade and Business	P	0.932	0.948	0.948	0.971	0.803	0.550
	R	0.378	0.283	0.260	0.157	0.550	0.072
	F1	0.532	0.432	0.403	0.268	**0.652**	0.125
Culture, tourism and sport	P	0.781	0.760	0.780	0.767	0.743	0.200
	R	0.551	0.520	0.567	0.557	0.585	0.003
	F1	0.645	0.615	**0.655**	0.645	0.654	0.005
Demographics	P	0.937	0.910	0.932	0.939	0.918	0.874
	R	0.909	0.868	0.889	0.930	0.876	0.234
	F1	0.923	0.888	0.909	**0.934**	0.896	0.321
Constructions	P	0.862	0.861	0.863	0.878	0.842	0.790
	R	0.821	0.800	0.856	0.871	0.816	0.314
	F1	0.840	0.829	0.859	**0.874**	0.829	0.444
Personnel	P	0.881	0.858	0.890	0.902	0.880	0.859
	R	0.860	0.787	0.866	0.866	0.844	0.239
	F1	0.870	0.821	0.878	**0.883**	0.861	0.363
Public Education	P	0.861	0.839	0.855	0.856	0.842	0.861
	R	0.837	0.789	0.847	0.843	0.802	0.183
	F1	0.849	0.812	**0.851**	0.849	0.821	0.239
Information services	P	0.874	0.867	0.866	0.878	0.885	0.922
	R	0.822	0.806	0.844	0.844	0.799	0.433
	F1	0.847	0.835	0.855	**0.860**	0.840	0.587
Financial services	P	0.951	0.940	0.949	0.954	0.950	0.908
	R	0.965	0.953	0.966	0.968	0.954	0.858
	F1	0.958	0.946	0.958	**0.961**	0.952	0.876
Social	P	0.858	0.761	0.837	0.855	0.820	0.373
	R	0.791	0.732	0.811	0.809	0.790	0.053
	F1	0.823	0.746	0.824	**0.831**	0.804	0.085
Urban Planning	P	0.904	0.887	0.903	0.916	0.898	0.890
	R	0.832	0.808	0.815	0.836	0.810	0.538
	F1	0.867	0.845	0.856	**0.874**	0.851	0.669

Table 6
Performance comparison of the Transformer model and INFORMed PA on the PA corpus.

Model	Metric	ACT	LAW	LOC	ORG	ORG_{PA}	PER	MacAvg	MicAvg
BERT-Ita	P	0.893	0.831	0.764	0.763	0.784	0.888	0.849	0.832
	R	0.909	0.878	0.826	0.818	0.823	0.880	0.877	0.877
	F1	0.898	0.854	0.794	0.790	0.803	0.884	0.861	0.854
mBERT	P	0.918	0.810	0.769	0.727	0.774	0.889	0.855	0.829
	R	0.881	0.867	0.822	0.815	0.809	0.899	0.862	0.867
	F1	0.895	0.838	0.795	0.769	0.791	0.894	0.856	0.848
XLM-RoB.	P	0.891	0.821	0.787	0.755	0.754	0.908	0.848	0.834
	R	0.945	0.881	0.833	0.820	0.857	0.897	0.901	0.890
	F1	0.916	0.850	0.809	0.786	0.802	0.902	0.873	0.861
UmBERTo	P	0.916	0.846	0.808	0.795	0.785	0.908	0.872	0.858
	R	0.942	0.877	0.841	0.838	0.828	0.900	0.899	0.890
	F1	**0.928**	**0.861**	**0.824**	**0.816**	**0.806**	**0.904**	**0.885**	**0.873**
GePpeTto	P	0.833	0.641	0.640	0.574	0.579	0.776	0.738	0.694
	R	0.851	0.761	0.733	0.678	0.773	0.817	0.802	0.800
	F1	0.824	0.696	0.683	0.622	0.662	0.796	0.758	0.744
INFORMed PA	P	0.788	0.827	0.702	0.709	0.616	0.837	0.746	-
	R	0.891	0.842	0.740	0.689	0.777	0.878	0.803	-
	F1	0.836	0.834	0.720	0.698	0.686	0.857	0.772	-

and ACT_U (see Tab. 7 and 8). As for ORG_PA, this may be due to the fact that the names of the organizations are more closely related to the individual Public Administration entity (e.g., *Settore LL.PP.* - Public Work Sector; *Ufficio politiche abitative* - Housing Policies Office). Given that this dataset includes documents from 25 different municipalities, it is understandable that even the best-performing models, namely UmBERTo and XLM-RoBERTa, may encounter difficulty handling such variability in the nomenclature of PA organizations. On the other hand, UmBERTo and XLM-RoBERTa seem to handle well the class ACT_U, which labeled the unparsable token, i.e., the reference to the PA act expressed as a unique string that can vary in terms of templates, such as *n.12/2013; 67/96; 57/2008.* On the contrary, BERT-Ita, mBERT, and GePpeTto performed poorly on this class.

We also computed the macro average of precision, recall, and F1-score for each model and dataset to make our data comparable with the results obtained by INFORMed PA [16]. Recall that INFORMed PA is a NER for the Italian PA based on the Stanford NER using a Conditional Random Field (CRF) as learning algorithm. The results of INFORMed PA on the PA Corpus are shown in Table 6, while the results on the additional dataset of 25 documents are reported in Table 7. By comparing the performance of our models with INFORMed PA on the PA Corpus, we observed that it performed almost on par with GePpeTto, the worst performing Transformer-based model, on all metrics. It obtained a precision of 0.74, 0.80 for recall, and an F1-score of 0.77. In this regard, the best-performing Transformer model, UmBERTo, outperformed INFORMed PA by a wide margin. Specifically, it scored 0.13 points higher in accuracy, 0.10 points higher in

Table 7
Performance comparison of the Transformer model and INFORMed PA on 25 documents dataset.

Model	Measure	ACT	LAW	LOC	ORG	ORG_{PA}	PER	MicAvg	MacAvg
BERTIta	P	0.876	0.804	0.692	0.557	0.598	0.907	0.790	0.794
	R	0.788	0.927	0.789	0.775	0.688	0.914	0.857	0.803
	F1	0.811	0.861	0.738	0.648	0.640	0.910	0.822	0.785
mBERT	P	0.861	0.828	0.676	0.551	0.504	0.880	0.780	0.774
	R	0.757	0.949	0.759	0.773	0.656	0.930	0.851	0.785
	F1	0.780	0.884	0.715	0.643	0.570	0.904	0.814	0.762
XLM-RoB.	P	0.876	0.807	0.720	0.563	0.579	0.909	0.796	0.796
	R	0.902	0.932	0.768	0.790	0.753	0.943	0.880	0.870
	F1	0.889	0.865	**0.743**	0.657	0.654	**0.926**	0.836	0.829
UmBERTo	P	0.877	0.836	0.665	0.579	0.538	0.911	0.796	0.792
	R	0.906	0.936	0.770	0.760	0.677	0.918	0.870	0.859
	F1	0.890	0.883	0.714	0.657	0.600	0.915	0.831	0.822
GePpeTto	P	0.597	0.628	0.532	0.447	0.386	0.737	0.630	0.572
	R	0.713	0.816	0.648	0.702	0.602	0.812	0.780	0.715
	F1	0.646	0.710	0.584	0.547	0.471	0.773	0.697	0.631
INFORMed PA	P	0.975	0.949	0.799	0.802	0.871	0.914	0.914	0.885
	R	0.848	0.962	0.691	0.769	0.796	0.869	0.836	0.822
	F1	**0.907**	**0.955**	0.741	**0.785**	**0.832**	0.891	**0.873**	**0.852**

recall, and 0.12 points higher in F1-score. All the others models outperformed INFORMed PA of at least 0.11 points in precision, 0.6 in recall, and 0.9 in F1-score. Surprisingly, the results of INFORMed PA on the 25 documents dataset exceed those of the Transformer-based models in terms of both precision and F1 score. We speculate that the CRF model is better able in dealing with classes whose instance show a high degree of linguistic regularity. In particular, the model seems to benefit from linguistic and shallow features such as word shape, n-grams, PoS-tags, and the presence of complex terms. By performing a class-wise comparison between models, we can observe that Transformer-based ones are more prone to errors in extracting Organization classes (both ORG and ORG_PA), LAW and ACTS while remaining effective on the others, which are arguably less dependent on the domain. In this case, in fact, their prediction capabilities are closer to InformedPA. These results show that language models based on the Transformer encoder architecture can guarantee higher performance when adequately trained on domain data than traditional tools such as INFORMed PA. Conversely, when there is some variability between the data on which the models are evaluated and the data used during the fine-tuning, the performance of Transformers tends to suffer with respect to that of a CRF model.

This means that to guarantee high performance on domain-specific linguistic data, such as administrative texts, and in domain-specific tasks, such as the recognition of Public Administration entities, the domain adaptation of the model via additional pre-training on domain-specific data may prove to be necessary. In this way, the model should be better able to derive an adequate representation of domain-specific terms and cope with the high degree of variability

Table 8
Precision, Recall and F1-score achieved by the models in the sub-sections of the ACT_PA class on the 25 documents dataset

Model	Measure	ACT_D	ACT_N	ACT_T	ACT_U	ACT_X
BERTIta	P	0.936	0.969	0.817	0.800	0.859
	R	0.903	0.939	0.884	0.320	0.890
	F1	**0.919**	0.954	0.849	*0.457*	0.874
mBERT	P	0.928	0.968	0.814	0.714	0.881
	R	0.912	0.929	0.868	0.200	0.877
	F1	**0.920**	0.948	0.840	*0.312*	0.879
XLM-RoB.	P	0.929	0.948	0.807	0.808	0.886
	R	0.920	0.929	0.901	0.840	0.922
	F1	**0.924**	0.939	0.852	0.824	0.904
UmBERTo	P	0.937	0.948	0.837	0.759	0.905
	R	0.897	0.902	0.911	0.880	0.938
	F1	**0.916**	0.925	0.873	0.815	0.921
GePpeTto	P	0.884	0.782	0.502	0.000	0.819
	R	0.922	0.951	0.823	0.000	0.871
	F1	0.903	0.858	0.624	*0.000*	0.844

in how entities and acts are mentioned in bureaucracy documents. Thus, one of our future goals will be to make a domain adaptation of the model that performed best in this comparison and to extend further the suite of tasks more closely related to the PA domain on which the model will be assessed.

5. Conclusion

In this paper, we aimed to confront the performance of various Transformer-based models on administrative texts. Our goal was to evaluate the performance of generic pre-trained models on administrative data and to identify the most suitable model for this type of task in this particular domain. To this end, we considered a multi-label document classification task and a Named Entity Recognition task on Public Administration texts.

Among the different kinds and sizes of Transformer models, UmBERTo was shown to be the best option to handle both text and token classification tasks. It achieved the highest accuracy of 0.71 and F1-score of 0.89 on the multi-label classification task and the highest micro average precision, recall, and F1-score on the NER-PA task of 0.86, 0.89, and 0.87, respectively.

While on multi-label document classification, no clear differences emerged among Transformer-based models, which outperformed the bi-LSMT model by a wide margin, we have observed a much more significant variance in performances for the NER-PA task. This is especially true if we consider encoder-only models such as UmBERTo and BERT and encoder-decoder ones such as GePpeTto. Indeed, the latter performed worst in the NER-PA, both in comparison with other Transformer-based models and with INFORMed PA [16].

Furthermore, we have observed that in the case of high variability in the data, document structures, and the way in which entities are expressed, the best performances are obtained with the domain-adapted CRF model INFORMed PA. These findings lead us to surmise that as a future direction, it may be crucial to adapt Transformer-based language models to this domain via additional pre-training steps in order to improve their performances.

In addition, we also aim at extending the kind of tasks and the number of datasets related to the Italian administrative domain in order to have more data available for the training and the evaluation of the models. We also plan to anonymize these data to share them.

Acknowledgments

This research has been funded by the Project "ABI2LE (Ability to Learning)", funded by Regione Toscana (POR Fesr 2014-2020). The project partners are University of Pisa, CoLing Lab and the companies 01Semplice s.r.l. (coordinator), IPKOM s.r.l., and Insurance OnLine S.p.A.

References

[1] 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.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[4] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang, Pre-trained models for natural language processing: A survey, Science China Technological Sciences 63 (2020) 1872–1897.

[5] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8342–8360. URL: https://aclanthology.org/2020.acl-main.740. doi:10.18653/v1/2020.acl-main.740.

[6] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, NPJ digital medicine 4 (2021) 1–13.

[7] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[8] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).
[9] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Legal-bert: The muppets straight out of law school, arXiv preprint arXiv:2010.02559 (2020).
[10] P. Grouchy, S. Jain, M. Liu, K. Wang, M. Tian, N. Arora, H. Ngai, F. K. Khattak, E. Dolatabadi, S. A. Koçak, An experimental evaluation of transformer-based language models in the biomedical domain, arXiv preprint arXiv:2012.15419 (2020).
[11] R. Joshi, A. Gupta, Performance comparison of simple transformer and res-cnn-bilstm for cyberbullying classification, arXiv preprint arXiv:2206.02206 (2022).
[12] C. Lothritz, K. Allix, L. Veiber, J. Klein, T. F. D. A. Bissyande, Evaluating pretrained transformer-based models on the task of fine-grained named entity recognition, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 3750–3760.
[13] M. Polignano, M. de Gemmis, G. Semeraro, Comparing transformer-based ner approaches for analysing textual medical diagnoses., in: CLEF (Working Notes), 2021, pp. 818–833.
[14] P. I. Khan, I. Razzak, A. Dengel, S. Ahmed, Performance comparison of transformer-based models on twitter health mention classification, IEEE Transactions on Computational Social Systems (2022) 1–10. doi:10.1109/TCSS.2022.3143768.
[15] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE Transactions on Knowledge and Data Engineering 34 (2020) 50–70. doi:10.1109/TKDE.2020.2981314.
[16] L. C. Passaro, A. Lenci, A. Gabbolini, Informed pa: A ner for the italian public administration domain, in: Fourth Italian Conference on Computational Linguistics CLiC-it, 2017, pp. 246–251.
[17] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 1480–1489. URL: https://aclanthology.org/N16-1174. doi:10.18653/v1/N16-1174.
[18] C. Lothritz, K. Allix, L. Veiber, J. Klein, T. F. D. A. Bissyande, Evaluating pretrained transformer-based models on the task of fine-grained named entity recognition, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 3750–3760.
[19] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel, N. Dehak, Hierarchical transformers for long document classification, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2019, pp. 838–844.
[20] Z. Shaheen, G. Wohlgenannt, E. Filtz, Large scale legal text classification using transformer models, arXiv preprint arXiv:2010.12871 (2020).
[21] A. Adhikari, A. Ram, R. Tang, J. Lin, Docbert: Bert for document classification, arXiv preprint arXiv:1904.08398 (2019).
[22] 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, arXiv preprint arXiv:1804.07461 (2018).
[23] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual BERT?, in: Pro-

ceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4996–5001. URL: https://aclanthology.org/P19-1493. doi:10.18653/v1/P19-1493.

[24] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).

[25] L. De Mattei, M. Cafagna, F. Dell'Orletta, M. Nissim, M. Guerini, Geppetto carves italian into a language model, arXiv preprint arXiv:2004.14253 (2020).

[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[27] A. Radford, T. Salimans, Gpt: Improving language understanding by generative pre-training. arxiv (2018).

[28] M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, The wacky wide web: a collection of very large linguistically processed web-crawled corpora, Language resources and evaluation 43 (2009) 209–226.

[29] E. F. Sang, F. De Meulder, Introduction to the conll-2003 shared task: Language-independent named entity recognition, arXiv preprint cs/0306050 (2003).

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems 26 (2013).

[31] 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: 6th Italian Conference on Computational Linguistics, CLiC-it 2019, volume 2481, CEUR, 2019, pp. 1–6.

[32] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.