# AdminSet and AdminBERT: a Dataset and a Pre-trained Language Model to Explore the Unstructured Maze of French Administrative Documents

**Thomas Sebbag[1,2], Solen Quiniou[1], Nicolas Stucky[1], Emmanuel Morin[1],**

[1] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France,
[2] Explore, Carquefou, France,
**Correspondence:** thomas.sebbag@univ-nantes.fr, solen.quiniou@univ-nantes.fr,
nicolas.stucky@etu.univ-nantes.fr, emmanuel.morin@univ-nantes.fr

## Abstract

In recent years, Pre-trained Language Models (PLMs) have been widely used to analyze various documents, playing a crucial role in Natural Language Processing (NLP). However, administrative texts have rarely been used in information extraction tasks, even though this resource is available as open data in many countries. Most of these texts contain many specific domain terms. Moreover, especially in France, they are unstructured because many administrations produce them without a standardized framework. Due to this fact, current language models do not process these documents correctly. In this paper, we propose AdminBERT, the first French pre-trained language model for the administrative domain. Since interesting information in such texts corresponds to named entities and the relations between them, we compare this PLM with general domain language models, fine-tuned on the Named Entity Recognition (NER) task applied to administrative texts, as well as to a Large Language Model (LLM) and to a language model with an architecture different from the BERT one. We show that taking advantage of a PLM for French administrative data increases the performance in the administrative and general domains, on these texts. We also release AdminBERT as well as AdminSet, the pre-training corpus of administrative texts in French and the subset AdminSet-NER, the first NER dataset consisting exclusively of administrative texts in French.

## 1 Introduction

Most public administrations worldwide use the same mechanism to purchase goods and services and to build or renovate public works: public procurement (Potin et al., 2023), also known as requests for proposals. France has added a transparency mechanism to the process. A public administration must publish, in open access, the content of the discussions and decisions taken on the selected service provider for the management of its territories. This rule induces the production of a large amount of textual data from very heterogeneous sources without a standard framework. Such volume and diversity make these documents time and resource-consuming for humans to analyze, yet they contain important information. In particular, the relationships between public administrations and economic actors are interesting to extract in order to better understand the economic network of a given geographical area.

NLP techniques are a means to extract these relationships. To do so, the first step would be to identify economic actors (persons, organizations) and then to localize their activities (localizations). French administrative texts, in addition to their lack of standardized structure, contain very formal syntax and expressions about the domain concerned. This makes them complex for general language models to process. Fine-tuned models for the administrative domain would be a solution for processing these documents but, to our knowledge, there are no such specialized language models that can be used for French administrative documents. In order to build a French specialized language model for the administrative domain, and to fine-tune it for the NER task, French administrative documents as well as documents labeled with the entities of interest are needed, but no such documents are currently available.

In this paper, we propose a French dataset of administrative documents (and a labeled subset of it) as well as pre-trained language models. Our contributions can be summarized as follows:

1. AdminSet[1], an open source dataset crawled from several French online administrative sources produced between 2020 and 2022;

[1] https://huggingface.co/datasets/taln-ls2n/Adminset



*Proceedings of the 31st International Conference on Computational Linguistics*, pages 392–406
January 19–24, 2025. 

2. AdminSet-NER[2], a labeled dataset dedicated to the NER task consisting of documents produced in 2023;

3. AdminBERT[3][4], the first PLM in French, for the administrative domain, based on the RoBERTa architecture;

4. A comparison between general language models, and language models fine-tuned to administrative texts.

## 2 Related Work

### 2.1 Resources

In terms of resources, since NER is a historical task, we can find a considerable number of datasets, such as CoNLL03[5], WikiNER[6] or MultiNERD[7], to name some of them. As mentioned above, we didn't find a NER dataset for French administrative data, but several studies have been carried out over the years, into administrative data using different approaches to administrative data. One notable example is the task of automatic translation.

An early project in this area is the Dutch Parallel Corpus, proposed by De Clercq and Perez (2010), which aims to align Dutch, English, and French resources. At the same time, we can also find MultiUN, a parallel corpus based on the official documents of the United Nations Organization (Eisele and Chen, 2010). More recently, some parallel corpus dedicated to administrative data have appeared in NLP literature, such as a corpus in Croatian and Italian, which is based only on administrative texts (Brkic Bakaric and Lalli Pacelat, 2019). Besides NLP, we found research on administrative texts in Information Retrieval. Falkner et al. (2019) has presented an Optical Character Recognition (OCR) and classification model dedicated to English Request For Proposal.

Rusinol et al. (2013) proposed a framework to extract information from French administrative invoices, or in a more recent approach, Sharma Kafle et al. (2023) works on semantic context combined with layout information for classification extraction on French administrative invoices. Koptelov et al. (2023) also worked on extracting rules from urban planning documents, texts directly produced by local French local administrations. Potin et al. (2023) published the first dataset consisting only of French requests for proposals dedicated to the monitoring of public policies. Unfortunately, this dataset did not meet our needs compared to the data we can access. The author strongly criticized the original data source (Tenders Electronic Daily [8]), and the variety of document types was too limited compared to ours, as described in Section 3.1.

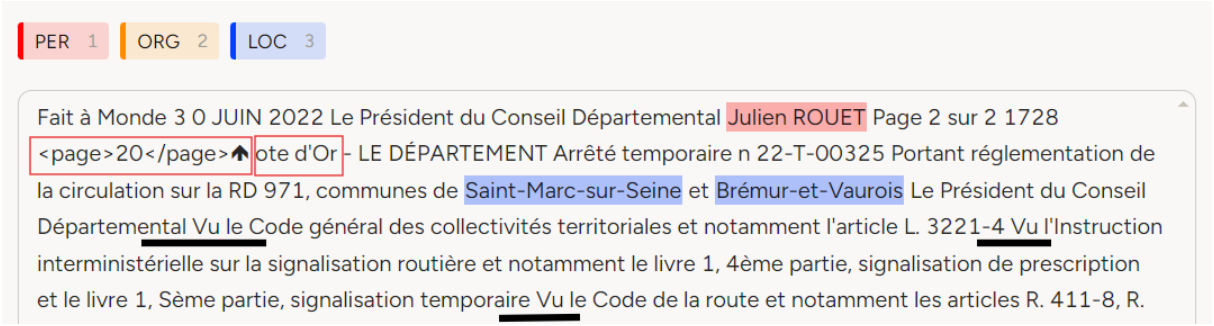
### 2.2 Named Entity Recognition (NER) Models

NER is a well-established task in the field of NLP, and it has numerous applications. In French, we can mention Flair (Akbik et al., 2018), which proposes a NER model in several languages, including French. More recently, Bourdois, Loïck (2024) (NERmemBERT), proposed an improvement of the original CamemBERT-Base dedicated to NER. Zaratiana et al. (2024) have proposed GLiNER, an alternative to the BERT architecture for NER that can be applied to arbitrary data and entities.

From an administrative data perspective it seems at date that no PLM or fine-tuned model has been dedicated to it, at this time. We found work on German, Heinisch and Lušicky (2020) using NER, though they didn't specify their approach, to provide the Austrian portal language with administrative texts and entities such as names of municipalities, names of politicians, common first names and last names of people, etc. On the English resources, Desmet et al. (2020) used American administrative texts to solve text classification and NER on biomedical entities for the public health sector using a Bi-LSTM CRF network.

Regarding French administrative data, we did not found any work focusing on the NER task. Among works in French, the closest work used financial corpora data to perform relation extraction through the NER task with the toolkit Spacy[9]. Jabbari et al. (2020) cover a wide range of a large spectrum of financial entities such as organization names, persons, roles, assets, currencies, locations, or even legal sanctions. Also, Bizrel is a corpus of French press documents proposed in Khaldi et al. (2020) focusing on standard entities (person, orga-

[2] https://huggingface.co/datasets/taln-ls2n/Adminset-NER
[3] https://huggingface.co/taln-ls2n/AdminBERT-16GB
[4] https://huggingface.co/taln-ls2n/AdminBERT-4GB
[5] https://huggingface.co/datasets/conll2003
[6] https://github.com/kata-ai/wikiner
[7] https://github.com/Babelscape/multinerd
[8] https://ted.europa.eu/en/
[9] https://spacy.io/api/entityrecognizer/




**English translation:** DONE at Monde, 30 JUNE 2022 The President of the Departmental Council Julien ROUET Page 2 of 2 <page>20</page> ote d'Or - THE DEPARTMENT Temporary order No. 22-T-00325 Concerning the regulation of traffic on RD 971, in the municipalities of Saint-Marc-sur-Seine and Brémur-et-Vaurois The President of the Departmental Council Considering the General Code of Local Authorities, and in particular Article L. 3221-4 Considering the Interministerial Instruction on Road Signage, and in particular Book 1, 4th part, Regulatory Signage, and Book 1, 5th part, Temporary Signage Considering the Highway Code, and in particular Articles R. 411-8, R.

Figure 1: Example of an administrative text after using an OCR. The noise due to the OCR is framed in red, the lack of punctuation between the sentences is underlined in black the entities mentioned in this text are highlighted in light red and light blue

nization, localization) dedicated to territorial competitive intelligence analysis, also using Spacy.

In another category, working on LLMs, Faysse et al. (2024) used, among other sources, administrative data provided by the French government to build his train set. However, these texts are exclusively from parliamentary debates, which can be considered as being closer to legal data than to purely administrative data.

## 3 Introduction to French Administrative Documents

### 3.1 Administrative Data

In order to provide a comprehensive overview of the various French administrative documents, we have divided them into three categories (more details are given in Table 7, in the Appendix).

**Local and Regional Councils** They contain the transcripts of debates and decisions on the provision of services to the community, at a local or regional level. Debate documents are the longest in this category because they contain the detailed discussions of the councils, while decisions are summaries of those discussions and focus on the resulting actions.

**Applications for Public Procurement** They are the preliminary step of the first category, when an administration decides to make a request for proposal on the economic market and describes its needs. In this category, we find documents related to the provision of a service, with the details of what is expected to be selected by the public administration. There are also documents dedicated to consultation regulations, which contain similar information on the provision of services combined with the current rule compliance. Finally, the largest documents in this category are the technical and administrative reports. They can be up to 90 pages long and contain a lot of domain specifications or legal reminders.

**Legal decrees** They represent the legal decisions taken on the demand of citizens or for mandatory regulation, regarding a punctual event. Those documents could also be very long, up to 50 pages.

### 3.2 Challenges of Administrative Documents

We face several challenges related to the diversity of sources for these documents. The fact that France does not have regulatory measures to standardize the format of these documents has led to the creation of semi-structured documents, like the one shown in Figure 1.

In addition, most of these documents are PDF documents (an example is given in the appendix A.3) and OCR tools are required to extract their text. This can add some noise to the extracted text, for example if the documents contain headers for mandatory references (for example, ID numbers, date of issue, place of issue of these documents, or page references). In addition, erroneous handwritten human annotations can also cause ty-

pographical errors, such as missing punctuation marks that results in missing line breaks.

Since these documents were written by local elected officials, some political or administrative knowledge is assumed to be implicitly known by all the participants, making them more difficult to understand for non-expert readers. Although some aspects are considered challenging, it is also a wonderful opportunity to compare existing baselines on these aspects and to understand the limitations of "real world data."

## 4 Creation of French Administrative Datasets and PLMs

### 4.1 AdminSet: a French Corpus of Administrative Documents

First, we present AdminSet, the first open corpus crawled on French administrative data from different text sources produced by French public administrations. It consists of more than 50 million text fragments, representing 16 GB of OCR texts, crawled from texts published between January 2020 and December 2022 by French administrations of all sizes. In order to cover the administrative domain as exhaustively as possible, we have collected a wide range of administrative contexts within numerous topics of requests, for services or debate issues, from all the categories given in Table 1. We can see that reports are the most common type of documents. The legal decrees contained in the collections of administrative acts are also a consequent part of this dataset.

| Category | #Docs | #Words |
|---|---|---|
| Special Administrative Condition Reports | 15,5 M | 810,4 M |
| Deliberations | 10 M | 702,5 M |
| Collections of Administrative Acts | 6,5 M | 551,8 M |
| Specific Technical Specification Reports | 6 M | 305,7 M |
| Official Reports | 5,2 M | 296,5 M |
| Rules for Consultation | 4,9 M | 257,3 M |
| Public Procurement | 712 K | 42,2 M |
| Prefectorial Regulations | 743 K | 57 M |
| Decisions | 528 K | 37,7 M |
| **Total** | **50 M** | **2,8 B** |

Table 1: Document sources of AdminSet

To collect administrative data, we relied on internal tools, using multiple scrappers to collect text from numerous administrative websites, through open access regulations. We then extracted the text of the PDF documents using OCR tools. As administrative documents are long documents, they exceed the limit of 512 tokens (which is BERT maximum size input). To create text fragments of up to 512 words, we used punctuation marks to separate the initial text according to these marks. If no punctuation mark were present, we used Langchain [10], to create text fragments. Langchain has a module called `tiktoken` with a chunk size of up to 450 tokens. We only cleaned up the special characters in the text, but we didn't change the case of the characters (especially the randomly appearing uppercase characters) or remove the noise from the OCR that was caused by page headers. In fact, our goal is to obtain models that can be applied to noisy data, in order to improve the performance on new data from the administrative domain.

### 4.2 AdminSet-NER: an Annotated Corpus of French Administrative Documents

Since there is currently little research on the administrative domain, we have created the first dataset dedicated to NER, called AdminSet-NER. These texts were extracted from French public administration websites between November and December 2023 and then manually annotated. The result is 814 annotated texts. We considered only three types of documents among our main categories: official reports, public decisions, and legal decrees. Indeed, based on our knowledge of these data, we focused on these three categories as they contain, on average, the highest number of named entities per document, making them ideal for NER.

| | Train | Validation | Test | Total |
|---|---|---|---|---|
| #Documents | 583 | 146 | 85 | 814 |
| #Words | 36 082 | 10 113 | 946 | 46 195 |
| Vocabulary size | 7 523 | 3 140 | 491 | 5 070 |
| #ORG mentions | 1 869 | 377 | 75 | 2 246 |
| #PER mentions | 1 311 | 545 | 261 | 1 856 |
| #LOC mentions | 620 | 261 | 24 | 883 |

Table 2: Statistics on AdminSet-NER various subsets

From Table 2, we observe that the majority of entities mentioned in these texts are `ORG` mentions (2 246 in total), followed by `PER` mentions (1 856 in total) and `LOC` mentions as the minority (883 in total). The training set consists of 7 523 unique tokens. We notice a slight underrepresentation of the label `LOC`, even though this is in line with the actual frequency of this type of entity in administrative texts. The test set was built separately from the training and the validation datasets, using data published in November 2024. Since public procure-

[10] https://python.langchain.com

ment is seasonal, having a one-year gap between the training/validation and the test sets seems to limit the repetition of entities as much as possible. The test set thus consists of 85 text fragments with 491 unique tokens.

Four non-expert human annotators (students from the first year of our NLP master's program) labeled the data with an overall agreement rate of 86%, as computed by Label Studio[11], based on exact matches. The entity labels are the usual domain labels, following the IOB2 tagging scheme: `PER` (persons), `ORG` (organizations), and `LOC` (localizations). We used NERmemBERT-3-entities to pre-annotate the entities, and then used Label Studio to manually correct the pre-annotations.

### 4.3 AdminBERT: a PLM for French Administrative Documents

In order to build AdminBERT, a language model adapted to the administrative domain, we carried out a continuous pre-training strategy of a French monolingual language model on our corpus AdminSet, starting from the weights and configuration of the CamemBERT-Base model. We trained two PLMs, depending on the size of the training dataset: AdminBERT 4GB (on a randomly selected subset of AdminSet, representing 4GB) and AdminBERT 16GB (on the full version of AdminSet). We applied the Masked Language Modeling (MLM) task with whole-word masking, following the method of Martin et al. (2020). We ran 10 000 steps over 2.5 hours using 24 A100 GPUs, on the Jean Zay supercomputer, with a batch size of 96 per GPU, a learning rate of 1e-4, on three epochs (following the recommendations of Labrak et al. (2023) regarding the limited impact of a high number of epochs, on the quality of the training).

## 5 Experimental Setup

### 5.1 Baseline Models

We describe some existing pre-trained models used as baselines in our comparative study.

**CamemBERT-Base (Martin et al., 2020)** is a RoBERTa-based model pre-trained from scratch on the French subset of the OSCAR corpus (138 GB). This model is the basis of our administrative language model. A NER version of this model, called CamemBERT-NER, was trained on the French part of WikiNER and is available on Hugging Face[12].

**NERmemBERT-3-entities (Bourdois, Loïck, 2024)** is a BERT model backbone with CamemBERT-Base tokenizers and weights, fine-tuned to the French NER task on three entity labels (ORG, PER, and LOC). The model was trained on a dataset obtained by concatenating the French parts of several NER datasets, including the French part of WikiNER (Nothman et al., 2013), Wikiann (Rahimi et al., 2019), MultiNERD (Tedeschi and Navigli, 2022), MultiCoNER v2 (Malmasi et al., 2022) and an industrial dataset called Pii-masking-200k proposed by Ai4Privacy. This model is our main baseline for comparing our results, as it is the state-of-the-art model for the French NER task.

**Wikineural-NER (Tedeschi et al., 2021)** is a variant of the multilingual BERT-based neural model of Mueller et al. (2020), where words are represented using the mean of their subword representations, instead of their own representations. This model is combined with a Bi-LSTM and a CRF model for a fine-tuning on the Wikineural dataset on 9 languages jointly.

**GLiNER (Zaratiana et al., 2024)** is a bidirectional transformer encoder, a variant of BERT with a span representation layer added on the same level as the feed-forward network layer. The model was trained by instruction on the Pile-NER [13] dataset derived from the Pile Corpus (Gao et al., 2020), resulting in a very flexible model capable of identifying any entity.

**Mixtral 7x8B (Jiang et al., 2024)** is a Sparse Mixture of Experts (SMoE) Large Language Model (LLM), using the same decoder-only transformer architecture as Mistral 7B (Jiang et al., 2023). The main difference is the 8 feed-forward blocks or *experts* added to the architecture, which theoretically gives access to a 47B parameter model but uses only 13B active parameters during inference.

### 5.2 Fine-tuned Models

We used the training and the validation set of AdminSet-NER to fine-tune both our PLMs (AdminBERT-NER 16GB and AdminBERT-NER

[11] Available at https://labelstud.io/

[12] https://huggingface.co/Jean-Baptiste/camembert-ner

[13] https://huggingface.co/datasets/Universal-NER/Pile-NER-type

4GB) as well as the following baseline models: CamemBERT-Base, NERmemBERT-Base-3-entities, NERmemBERT-Large-3-entities, and Wikineural-NER (in the remainder of the paper, the names corresponding fine-tuned models are followed by "FT"). Each fine-tuning was performed during 5 runs, using the same seeds for all the models, as well as a learning rate of 1e-4, a batch size of 3, a weight decay of 0.1 during 10 epochs with an early stopping patience of 3.

### 5.3 Evaluation Protocol

For the LLM evaluation, we had to prompt Mixtral 7x8B in one shot, giving it our three labels (PER, LOC, ORG) to extract, context information to recognize the labels, an example of the desired output in the Conll scheme, and the text to extract (see details in Appendices A.4 and A.5). Once the prediction was generated, we manually aligned its output to our reference, in order to compute the evaluation scores. Finally, GLiNER was given the original full texts from the test set and the three labels to extract. We then created a Conll file from its predictions and computed the scores from it.

### 5.4 Evaluation Metrics

We adopted the standard NER evaluation methodology, calculating the precision (P), recall (R) and F1-score (F1), based on the exact match between predicted and actual entities. To evaluate the models, we used the macro-average measures provided by the scikit-learn tool (Kramer, 2016). To evaluate each model, we performed five runs and averaged the results on the test set.

## 6 Experiments and Results

We evaluated the performance of the general language models as well as of the language models fine-tuned on the NER task, for the administrative domain, using the test set of AdminSet-NER.

### 6.1 Results with General Language Models

We first consider language models trained on general web data and evaluate them on the AdminSet-NERtest set, as shown in Table 3.

If we compare the results obtained by the French BERT-based language models, we observe very low performances, with the best model achieving an F1-score of 5.67%. This shows that the named entities of the administrative domain may be relatively different from those of the general domain. Wikineural-NER gives a slight improvement with an F1-score of 12.54%. The subword representation used in this model could be useful to better identify the named entities of the administrative domain.

| Model | P | R | F1 |
|---|---|---|---|
| CamemBERT-NER | 10.29 | 11.06 | 4.17 |
| NERmemBERT-Large-3-entities | 3.31 | 19.73 | 5.58 |
| NERmemBERT-Base-3-entities | 9.32 | 10.79 | 5.67 |
| Wikineural-NER | 16.32 | 20.16 | 12.54 |
| Mixtral 7x8B | 43.61 | 33.95 | 35.53 |
| GLiNER | **75.00** | **42.00** | **53.92** |

Table 3: Results of the general language models on the test set of AdminSet-NER (results are given in %)

Mixtral 7x8B showed improved performance, achieving an F1-score of 35.53%. Its mixed results may be due to a tendency to overlabel named entities (see Section 6.5 for an analysis of its errors). GLiNER achieved the best results, with an F1-score of 53.92%, showing a better ability to identify French named entities, although it was trained only on English data and did not see any administrative data during training. It will be interesting to see if a future version of this model would be adapted to French and to evaluate the performance of such a model.

### 6.2 Results with Fine-tuned Language Models

We then consider language models fine-tuned on the training and validation sets of AdminSet-NER (as described in Section 5.2) and compare them to our fine-tuned AdminBERT-NER PLMs, as shown in Table 4.

| Model | P | R | F1 |
|---|---|---|---|
| Wikineural-NER FT | 77.49 | 75.40 | 75.70 |
| NERmemBERT-Large FT | 77.43 | 78.38 | 77.13 |
| CamemBERT FT | 77.62 | 79.59 | 77.26 |
| NERmemBERT-Base FT | 77.99 | 79.59 | 78.34 |
| AdminBERT-NER 4GB | 78.47 | 80.35 | 79.26 |
| AdminBERT-NER 16GB | **78.79** | **82.07** | **80.11** |

Table 4: Results of the fine-tuned language models on the test set of AdminSet-NER (results are given in %)

We observe that, even with a small labeled training set used for the fine-tuning, all the fine-tuned models perform significantly better on our administrative test set. Wikineural-NER achieves the lowest results, although it performs better than the original model without the fine-tuning. We hypothesize that it reaches the limit of what a multilingual model can do when applied to a specialized domain

such as the administrative domain. Then, the general French language model (CamemBERT-Base), as well as the two French language models already specialized for the NER task (both forms of the NERmemBERT models), benefit even more from the fine-tuning with an F1-score multiplied by 25.

In terms of F1-score, our AdminBERT-NER 16GB model has the best score with 80.11%. The lighter version comes in second with an F1-score of 79.26%. Looking at the detailed results on the entity labels (given in Appendix A.2), we observe the same behavior for AdminBERT-NER 16GB as for any of our fine-tuned models: PER mentions are very well recognized (98.80% and 99.64% for words at the beginning of PER mentions and words within PER mentions, respectively), while ORG mentions are less well recognized (79.34% and 81.99% for words at the beginning of ORG mentions and words within ORG mentions, respectively). LOC mentions are the most difficult to detect for all the models (58.41% and 53.45% for words at the beginning of LOC mentions and words within LOC mentions, respectively): one reason is that they are the less numerous labels in our training dataset. Another reason is the ambiguity of administration names, which can be either a localization or an organization, the type chosen depending on the context of the sentences.

The continuous pre-training strategy allowed us to improve the performance by almost 2 points compared to the best general language model. Adaptation to this type of vocabulary and noisy structure is one way to achieve better results. However, there are still several steps to be taken before achieving the same results for NER as the state-of-the-art models on general data. Overall, the challenge of overcoming the noise that alters the model's ability to correctly predict entities is the first step towards progress.

### 6.3 Results on Binary NER Classification

To deepen our understanding of the strengths and weaknesses of AdminBERT-NER 16GB, we analyze a binary NER classification, as shown in Table 5. Here, we just want to see if a word is considered to be part of a named entity or not, regardless of its entity label.

In this experiment, we wanted to see if our model is able to reach the state of the art models, in terms of NER, regardless of the labels of the entities. To do so, we trained a binary classifier for entity recognition, with two classes: the NE class corresponds to words detected as part of a named entity, while the Other class corresponds to words outside of named entities. The results are promising, as we obtained an F1-score of 92.05% on the NE class. This is good enough for the next step of our project, which is to detect relations between the identified named entities.

| | P | R | F1 |
|---|---|---|---|
| NE class | 92.32 | 91.79 | 92.05 |
| Other class | 99.07 | 99.13 | 99.10 |
| Macro Avg | 95.69 | 95.46 | 95.57 |

Table 5: Results with AdminBERT-NER 16GB with 1 entity label to detect (results are given in %)

### 6.4 Results on the WikiNER Dataset

To evaluate the robustness of AdminBERT 16GB in a general domain, we fine-tuned it on the French part of the WikiNER dataset[14]. The results are shown in Table 6.

| Model | F1 |
|---|---|
| AdminBERT 16GB FT WikiNER | 92.00 |
| NERmemBERT-Large-3-entities | 99.00 |

Table 6: Results of AdminBERT 16GB FT WikiNER on the WikiNER dataset compared to NERmemBERT-Large-3-entities (results are given in %)

The model achieves an F1-score of 92.00%, which is very satisfactory for a language model trained on a specific domain, although it is still far behind the state-of-the-art language model on this dataset, which achieves 99.00%. This experiment shows that a language model adapted to the administrative domain can be used on general data, while the other way does not work. In the future, including administrative data from scratch during the training of French language models could improve performance in different domains.

### 6.5 Discussions: Error Analysis

During the fine-tuning of our models, we observed some common difficulties across all models.

First, at the end of our second annotation campaign, we noticed that all models had trouble recognizing certain patterns such as "Commune de" (EN: *Municipality of*) or "Ville de" (EN: *City of*) as part of an entity, which resulted in mislabeling

[14] https://huggingface.co/datasets/Jean-Baptiste/wikiner_fr

the next token in LOC when it was an ORG. To reduce the problem, we added more texts with this pattern as more content-rich paragraphs, which had a positive effect when we evaluated the final training set. Despite these improvements, we still need more data to help the model stop mislabeling ORG and LOC, as we ended up with a training dataset of 756 text fragments.

Second, we also looked at the CamemBERT-Base and Wikineural-NER tokenizers. Digging into our data, we realized that most entities representing an administration were not well tokenized, creating too many subtokens to be recognized as an entity. For example, the municipality of Ollioules is tokenized like this: `-Commune`, `-D' O`, `l`, `lio`, `ules`. We consider for a while to improve the tokenizer. However, to affect the tokenizer, we would have to train a language model from scratch, which is much more time-consuming and resource-intensive than continuing pre-training, even if it doesn't affect the tokenizer. Labrak et al. (2024) showed for the medical domain that despite significant improvement in token processing during tokenization, they did not observe fundamental differences on downstream tasks such as NER. Despite the low tokenization performance, the transformer-based model manages to overcome it. Nevertheless, it seems very interesting to investigate how to do the strongest tokenization or add metadata to improve such language models.

On a more model-centric analysis, Mixtral 7x8B was over-labeling entities. For example, people functions were a recurring mistake of labeling as `PER`. For example, in " The Mayor of Deauville", which does not mention any names, Mixtral 7x8Bhas labeled the word Mayor as `PER` instead of `O`. Dates were a source of error as they were regularly labeled as `ORG`. Demonstrated a recurring problem with entity boundaries, many tokens were labeled "I-" without a token labeled "B-" in front of them. For example, the mention of "Communauté d'Agglomération Les Sorgues du Comtat" (EN: *Les Sorgues du Comtat Agglomeration Community*) was labeled with: `Communauté B-ORG`, `d'Agglomération I-ORG`, `Les O`, `Sorgues O`, `du O`, `Comtat I-ORG`. This mislabeling could be caused by the way we created the prompt to get predictions in BIO format to make it comparable to our study. The LLM was able to understand the administrative context even when it overlabeled, showing that this generative model is not the best option for this task.

## 7 Conclusion

In this work, we have proposed the first adaptation to the administrative domain of a Transformer-based language model based on the RoBERTa architecture for French and its version trained for NER. An evaluation study of these specific models was carried out on an aggregated collection of public administrative data. Our study proposes a first approach of Named Entity Recognition tailored to French administrative documents. Our model, AdminBERT, is a first step in the exploration of the French administrative domain applied to Territorial Competitive Intelligence. In addition, we have shown that further pre-training can improve the performance of the models on the NER task, but the AdminSet-NER needs further refinement. The impact of tokenization needs to be further investigated, or disambiguation techniques, such as the addition of metadata, need to be tried to improve performance on this task. AdminSet and AdminSet-NER as well as the AdminBERT pre-trained models have been released[15].

AdminBERT has considerable potential for dealing with unstructured data, such as French administrative data. It will be interesting to see if it can overcome certain limitations that general unstructured data models face when dealing with different tasks. Looking ahead, applying new tasks to the domain will open up new possibilities for future evaluation. We are also excited to build on our NER foundation to deepen the extraction of relations in the French administrative domain. We strongly believe that it is possible to extract interactions between the public administration and the private sector or associations, from the data produced by these public administrations. The non-uniform structure of public documents is also an excellent challenge for the information retrieval domain, inspiring us to find innovative ways to adapt the extraction of valuable information from them. All these possible tasks will improve our understanding of the domain and contribute to the development of graph knowledge networks for territorial competitive intelligence analysis.

## 8 Limitations

It is important to discuss some of the limitations of our experiments. First, for evaluation purposes on "real world data", we intentionally limited the preprocessing of our administrative texts to make them

[15] https://huggingface.co/taln-ls2n

noisy in order to better reflect the reality of administrative documents. This approach may introduce typographical errors, as a large proportion of these documents were written by humans during administrative meetings. Additionally, our data may lack diversity as we mainly chose data from 2020 to 2022, which may not fully represent the language diversity used in this document, especially during the COVID-19 period with fewer local economic activities.

The robustness of the model to noise should be experimented with to observe the impact of noise during the training. It will be interesting to integrate different amounts and types of noise into the training set to quantify its impact on the model. In addition, we did not use any specific process to select our documents or to avoid repetition of the discussion topics. These aspects will be addressed in future work.

## Acknowledgments

We would like to thank the anonymous reviewers for their thoughtful and insightful comments, which we found very helpful in improving the paper. We would also like to thank the undergraduate students who helped us with the three annotation campaigns.

This work was performed using HPC resources from GENCI-IDRIS (Grant 2024-AD011015204). This work was financially supported by the company Explore.

## References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Bourdois, Loïck. 2024. NERmembert-large-3entities .

Marija Brkic Bakaric and Ivana Lalli Pacelat. 2019. Parallel Corpus of Croatian-Italian Administrative Texts. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 11–18, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Orphée De Clercq and Maribel Montero Perez. 2010. Data Collection and IPR in Multilingual Parallel Corpora. Dutch Parallel Corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Bart Desmet, Julia Porcino, Ayah Zirikly, Denis Newman-Griffis, Guy Divita, and Elizabeth Rasch. 2020. Development of Natural Language Processing Tools to Support Determination of Federal Disability Benefits in the U.S. In *Proceedings of the 1st Workshop on Language Technologies for Government and Public Administration (LT4Gov)*, pages 1–6, Marseille, France. European Language Resources Association.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Andreas Falkner, Cristina Palomares, Xavier Franch, Gottfried Schenner, Pablo Aznar, and Alexander Schoerghuber. 2019. Identifying Requirements in Requests for Proposal: A Research Preview. In *Requirements Engineering: Foundation for Software Quality*, pages 176–182, Cham. Springer International Publishing.

Manuel Faysse, Patrick Fernandes, Nuno M. Guerreiro, Ant'onio Loison, Duarte M. Alves, Caio Corro, Nicolas Boizard, Joao Alves, Ricardo Rei, Pedro Henrique Martins, Antoni Bigata Casademunt, François Yvon, André Martins, Gautier Viaud, C'eline Hudelot, and Pierre Colombo. 2024. Croissantllm: A truly bilingual french-english language model. *ArXiv*, abs/2402.00786.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *ArXiv*.

Barbara Heinisch and Vesna Lušicky. 2020. The Austrian Language Resource Portal for the Use and Provision of Language Resources in a Language Variety by Public Administration – a Showcase for Collaboration between Public Administration and a University. In *Proceedings of the 1st Workshop on Language Technologies for Government and Public Administration (LT4Gov)*, pages 28–31, Marseille, France. European Language Resources Association.

Ali Jabbari, Olivier Sauvage, Hamada Zeine, and Hamza Chergui. 2020. A French Corpus and Annotation Schema for Named Entity Recognition and Relation Extraction of Financial News. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2293–2299, Marseille, France. European Language Resources Association.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. *arXiv preprint*.

Hadjer Khaldi, Amine Abdaoui, Farah Benamara, Grégoire Sigel, and Nathalie Aussenac-Gilles. 2020. Classification de relations pour l'intelligence économique et concurrentielle. In *27ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2020)*, volume 2 of *Traitement Automatique des Langues Naturelles*, pages 27–39, Nancy, France. ATALA : Association pour le traitement automatique des langues. Backup Publisher: ATALA (Association pour le Traitement Automatique des Langues).

Maksim Koptelov, Margaux Holveck, Bruno Cremilleux, Justine Reynaud, Mathieu Roche, and Maguelonne Teisseire. 2023. Towards a (Semi-)Automatic Urban Planning Rule Identification in the French Language. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

Oliver Kramer. 2016. Scikit-Learn. In Oliver Kramer, editor, *Machine Learning for Evolution Strategies*, pages 45–53. Springer International Publishing, Cham.

Yanis Labrak, Adrien Bazoge, Béatrice Daille, Mickael Rouvier, and Richard Dufour. 2024. How Important Is Tokenization in French Medical Masked Language Models? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8223–8234, Torino, Italia. ELRA and ICCL.

Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. DrBERT: A robust pre-trained model in French for biomedical and clinical domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16207–16221, Toronto, Canada. Association for Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

David Mueller, Nicholas Andrews, and Mark Dredze. 2020. Sources of Transfer in Multilingual Named Entity Recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, 194:151–175.

Lucas Potin, Vincent Labatut, Pierre-Henri Morand, and Christine Largeron. 2023. FOPPA: An Open Database of French Public Procurement Award Notices From 2010–2020. *Scientific Data*, 10(1). ArXiv:2305.18317 [cs].

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Marcal Rusinol, Tayeb Benkhelfallah, and Vincent Poulain dAndecy. 2013. Field Extraction from Administrative Documents by Incremental Structural Templates. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1100–1104, Washington, DC, USA. IEEE.

Dipendra Sharma Kafle, Eliott Thomas, Mickael Coustaty, Aurélie Joseph, Antoine Doucet, and Vincent Poulain d'Andecy. 2023. Subgraph-Induced Extraction Technique for Information (SETI) from Administrative Documents. In Mickael Coustaty and Alicia Fornés, editors, *Document Analysis and Recognition – ICDAR 2023 Workshops*, volume 14194, pages 108–122. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined Neural and Knowledge-based Silver Data Creation for Multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Tedeschi and Roberto Navigli. 2022. MultiNERD: A Multilingual, Multi-Genre and Fine-Grained Dataset for Named Entity Recognition (and Disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model

for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

# A Appendix

## A.1 Details on the administrative documents selected in AdminSet

| | |
|---|---|
| **Documents extracted from councils** | |
| Official reports | It will contain the transcript of the discussion append in the meeting, it is the longest documents we can find for reporting what happens during the meeting. |
| Deliberations | It sets out the deliberations following the debate that took place during a public administration council. |
| Decisions | It is a decision taken by the administrative authority following discussions by a council. |
| **Applications for public offers** | |
| Public Procurement | This is a notice published in a legal gazette by a contracting authority or entity to inform businesses of the award of one or more contracts. |
| Special Administrative Conditions Report | This is a contractual document that sets out the administrative clauses specific to the public contract. |
| Specific Technical Specifications Report | It details the subject of the contract and the nature and scope of the purchaser's requirements. |
| Rules for Consultation | It is a complementary document of Public Procurement |
| **Legal decrees** | |
| Collection of Administrative Acts | Contains regulatory acts (decrees, decisions, etc.) and deliberations of a regulatory administration taken by the local authority's. |

Table 7: Details on the administrative document types from AdminSet

## A.2 Detailed results, per category, on the NER evaluation, for the three best performing models

| | **AdminBERT-NER 16GB** | | | **AdminBERT-NER 4GB** | | | **NERmemBERT-Base FT** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| B-LOC | 54.10 | 63.46 | 58.41 | 53.85 | 53.85 | 53.85 | 54.55 | 57.69 | 56.07 |
| B-ORG | 80.67 | 78.05 | 79.34 | 73.39 | 73.98 | 73.68 | 77.39 | 72.36 | 74.79 |
| B-PER | 97.64 | 100.00 | 98.80 | 96.88 | 100.00 | 98.41 | 96.83 | 98.39 | 97.60 |
| I-LOC | 50.00 | 57.41 | 53.45 | 49.02 | 46.30 | 47.62 | 38.54 | 62.96 | 47.89 |
| I-ORG | 79.00 | 85.22 | 81.99 | 80.75 | 84.73 | 82.69 | 78.87 | 75.37 | 77.08 |
| I-PER | 99.28 | 100.00 | 99.64 | 98.57 | 100.00 | 99.28 | 97.87 | 100.00 | 98.92 |
| O | 99.47 | 98.93 | 99.20 | 99.29 | 99.06 | 99.18 | 99.39 | 98.98 | 99.19 |

Table 8: Detailed results with AdminBERT-NER 16GB, AdminBERT-NER 4GB and NERmemBERT-Base FT, on the test set of AdminSet-NER (results are given in %). B-X indicates the first word of a X entity whereas I-X indicates a word inside a X entity (replace X by LOC, ORG or PER) and O indicates a word not belonging to an entity of the three considered types (the word is thus considered to be outside the considered entity labels)

## A.3 Example of the first page of a decree issued by the municipality of Grasse

Grasse

**Arrêté temporaire n°22-AT-0225**
**Portant réglementation du stationnement et de la circulation**

**CHEMIN DES BASSES RIBES**

Le Maire de la ville de Grasse,
**VU** le Code général des collectivités territoriales et notamment les articles L. 2213-1 à L. 2213-6
**VU** le Code de la route et notamment les articles R. 411-8, R. 411-21-1, R. 413-1 et R. 417-10
**VU** l'Instruction interministérielle sur la signalisation routière et notamment le livre 1, 4ème partie, signalisation de prescription
**VU** l'arrêté municipal portant délégation de signature en date du 6 juin 2020
**VU** la demande en date du 04/08/2022 émise par ETS RUSSO THIERRY demeurant 2879 route de Grasse 06530 SAINT CEZAIRE représentée par Monsieur Thierry RUSSO aux fins d'obtenir un arrêté de réglementation du stationnement et de la circulation
**CONSIDÉRANT** que la réalisation de travaux (Entretien des voies et de l'espace public / Entretien du réseau routier (élagage, fauchage, curage de fossé) rend nécessaire d'arrêter la réglementation appropriée du stationnement et de la circulation, afin d'assurer la sécurité des usagers, le 05/09/2022 CHEMIN DES BASSES RIBES

### ARRÊTE

**Article 1**
Le 05/09/2022, les prescriptions suivantes s'appliquent du 298 au 307 CHEMIN DES BASSES RIBES :

- La circulation est alternée par K10 de 9h à 16h ;
- Le dépassement des véhicules, autres que les deux-roues, est interdit, de 9h à 16h ;
- Le stationnement des véhicules est interdit de 9h à 16h. Par dérogation, cette disposition ne s'applique pas aux véhicules de l'entreprise exécutant les travaux, véhicules de police et véhicules de secours. Le non-respect des dispositions prévues aux alinéas précédents est considéré comme gênant au sens de l'article R. 417-10 du code de la route et passible de mise en fourrière immédiate ;
- La vitesse maximale autorisée des véhicules est fixée à 30 km/h de 9h à 16h ;

**Article 2**
La signalisation réglementaire conforme aux dispositions de l'Instruction Interministérielle sur la signalisation routière sera mise en place par le demandeur, ETS RUSSO THIERRY.

**Article 3**
Le Maire de la ville de Grasse est chargé de l'exécution du présent arrêté qui sera publié et affiché conformément à la réglementation en vigueur.

### A.4 Prompt used to evaluate Mixtral 7x8B

```
Tu es un modèle de reconnaissance d'entité nommées, tu dois extraire trois classes
PER, ORG, LOC

Contexte :
PER correspond à des Prénom et Nom représentant une personne
ORG correspond à des noms d'entreprises, d'associations, d'administration publique
LOC correspond à des localisations généralistes, nom de ville, région, département
ou zone géographique.

En sortie le texte doit être dans le format BIO comme suit :
    * Je - O
    * suis - O
    * Thomas - B-PER
    * Martin - I-PER
.

Voici le texte :

Après s'être assuré que le receveur a repris dans ses écritures le montant de chacun
des soldes figurant au bilan de l'exercice 2020, celui de tous les titres de recette
émis et celui des mandats de paiement ordonnancés, qu'il a procédé à toutes les
opérations d'ordre qu'il lui a été prescrit de passer dans ses écritures, et qu'ainsi
la balance de sortie peut être arrêtée comme suit :
22 <page>22</page>
Résultat à la clôture de l'exercice 2020 Part affectée à l'investissement 2021 Résultat
de l'exercice 2021 Résultats de clôture de 2021
Investissement -1 486 788,59 0,00 32 860,07 -1 453 928,52
Exploitation 1 161 042,46 0,00 78 235,77 1 239 278,23
TOTAL -325 746,130,00 111 095,84 -214 650,29
Mme Françoise BRISSON :
Je vais vous présenter le budget d'assainissement de Machecoul, puis celui de
Saint-Même pour la dernière année, puisque nous avons délibéré le 4 novembre pour
clôturer le budget de Saint-Même et le rattacher à celui de Machecoul, pour devenir
le budget d'assainissement de Machecoul-Saint-Même, à partir du 1er janvier 2022.
```

### A.5 English Translation of the Prompt used to evaluate Mixtral 7x8B

```
You are a named entity recognition model, you need to extract three classes PER, ORG,
LOC

Context:
PER corresponds to first and last names representing a person
ORG corresponds to the names of companies, associations and public authorities
LOC corresponds to general locations, the name of a town, region, department or
geographical area.

The output text must be in BIO format, as follows:
    * I - O
    * I am - O
    * Thomas - B-PER
    * Martin - I-PER
```

.

Here is the text:

After ensuring that the Receiver has entered in his accounts the amount of each of the balances appearing in the balance sheet for the financial year 2020, that of all the revenue orders issued and that of the payment orders authorized, that he has carried out all the operations of order that he has been prescribed to enter in his accounts, and that the outgoing balance can thus be established as follows:
22 <page>22</page>
Year-end result 2020 Portion allocated to investment 2021 Year-end result 2021
Investment -1 486 788.59 0.00 32 860.07 -1 453 928.52
Operating 1 161 042.46 0.00 78 235.77 1 239 278.23
TOTAL -325,746.13 0.00 111,095.84 -214,650.29
Mme Françoise BRISSON:
I am going to present the Machecoul sanitation budget, then the Saint-Même budget for the last year, as we voted on 4 November to close the Saint-Même budget and attach it to the Machecoul budget, to become the Machecoul-Saint-Même sanitation budget, from 1 January 2022.