

1 Presentation of the task

This work is part of the Socface project, which aims to collect and process handwritten nominal lists from the censuses between 1836 and 1936 using automatic handwriting recognition. Its aim is to leverage this archival material to produce a database of all individuals who lived in France between 1836 and 1936, which will be used to analyze social changes over a 100-year period.

Within the scope of this particular work, we focus on information typing. Indeed, an automatic handwriting recognition system has been used to transcribe individual information from these archive documents, and the task at hand involves typing each recognized piece of information such as name, first name, age, year of birth, nationality, etc. Thus, we are tasked with developing a system to tag all the information of an individual, evaluating this system, and analyzing the conditions for its application to the entire corpus.

The rest of this work is organized as follows: first, we will provide a presentation and analysis of the data at our disposal; then, we will analyze the strengths and weaknesses of the models suitable for this task; finally, we will present the protocol of our experiment along with the results obtained and formulate recommendations.

2 Presentation and description of the data

The database we have contains information on 25 448 individuals who were recorded between 1836 and 1936. These data were extracted from census records using an automatic handwriting recognition system. For each person, we have at least one of the following pieces of information: last name, first name, profession, family relationship (link), date of birth, nationality, place of residence (lob), age, marital status, education level, employer, maiden name, and census observations. The table below summarizes the number of data entries for each of these variables, the percentage is given relative to the total of 25 448 individuals.

Variable	surname	surname household	first name	age	civil status	link
number of entries	19 159 (75.3%)	5640 (22.2%)	24 931 (98.0%)	16 436 (64.6%)	10 705 (42.0%)	20 728 (81.2%)
Variable	occupation	birth date	nationality	lob	employer	
number of entries	14 760 (58.0%)	7344 (28.9%)	13 314 (52.3%)	9 236 (36.3%)	2 851 (11.2%)	

Table 1: Number of entries for each variables

It should be noted that the information regarding maiden name and education level is not provided. In the following section, we will proceed with the detailed analysis of each variable under study with the only purpose to have a deep understanding of data base before designing our system.

2.1 Descriptive Statistics

2.1.1 Nationality, occupation and employer

As presented in Figure 3 in annex, out of the 13 314 individuals whose nationality is known, an overwhelming majority of 99% are French. Additionally, there are 73 Polish, 27 Belgian, and 23 Spanish individuals, while other nationalities are marginally represented. Regarding occupation data, among the 14 760 individuals for whom this information is available, 19% are unemployed. There is a significant presence of farmers, accounting for 17% , followed by categories such as laborers, housewives, and seamstresses suggesting that during this period, economic activities

were primarily concentrated in the primary sector. Finally, it is worth noting that 12% of the collected information indicates "néant".

Regarding employer information, out of the 2 851 individuals concerned, 31% are self-employed , while the gouvernement employs only 1% of individuals, ranking behind Madic (see Figure 3 and word clouds (a) and (b) in Figure 4 in the Annex).

2.1.2 Age and others variables

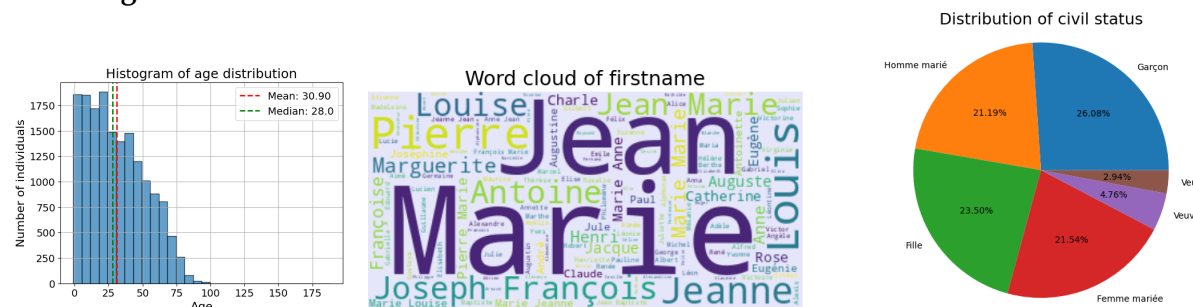


Figure 1: Age and civil status distributions

Since the age column was replaced by year of birth in 1906, we limited our analysis to age data from 1836-1906 censuses. For post-1906 censuses, calculating ages would require the exact census years, which we lack. With 16 436 age data points versus 7 344 year of birth ones, focusing on age provides ample insight into French demographics for our study's scope.

Assuming that the database provided to us is a representative sample of the population during this period, the histogram in Figure 1 and the box-plot in Figure 5 in annex show that the population was very young, with a need for births, and with half of the individuals having an age equal to or less than 28 years. Additionally, the first quartile suggests that 25% of individuals were aged 13 years or younger, while the third quartile indicates that 75% of individuals were aged 46 years or younger.

The pie chart in Figure 1 illustrates the distribution of civil statuses according to the censuses from 1836 to 1881, before these data were replaced by relationship to the head of the household. Word cloud in Figure 1 reveals that the most common first names are Marie, Jean, Pierre, Louis, François, and Jeanne, thus reflecting the French identity. For places of residence (lob), as show by the word cloud in figure 5 in Annex, Orléans and Querrien appear slightly more than the others, without significantly standing out. For the other variables, no notable information has been highlighted, as shown in Figure 6 in annex.

Note: To perform these descriptive statistics, we conducted various preprocessing steps, such as converting all ages into years, standardizing the names of different nationalities, etc. However, to ensure that our model is trained directly on data from the automatic translation of handwritten documents, we do not apply these transformations during the training of our models.

2.2 Estimating the target data size and identification of other usable external data

As mentioned earlier, the database we have contains information on 25 448 individuals. However, as mentioned on the website [1] of the Socface projet, the overall project encompasses 700 million individual records concerning individuals who resided in France between 1836 and 1936.

Regarding the exploitation of other external databases, we can approach the National Commission for Data Protection (CNIL) to obtain authorization and access to similar types of data on individuals who have lived or are living in France. This data could be useful for pre-training our models. Furthermore, since our work focuses on word classification, pre-trained LLMs on large textual corpora will be beneficial for this task. These corpora can be considered as external databases that have contributed to solving our task.

3 Models analysis

The main task involves **Named Entity Recognition** (NER) for information typing. Therefore, any models suitable for **token classification** can be used. The most well-known ones include:

BERT [3]: BERT is a language model based on Transformers that considers word context in both directions, allowing for more precise sentence understanding. Pre-trained on vast amounts of data, BERT has a strong grasp of language and can be easily fine-tuned for named entity recognition task. However, BERT is resource-intensive, with hundreds of millions of parameters, and the base BERT model was pre-trained on English corpora, making it less effective for tasks involving the French language. For instance, during tokenization, words may be excessively subdivided compared to a model trained on French corpora, such as Camembert (see Tableau 3 in Annex).

Camembert Model [5]: Camembert is a version of BERT models but is pre-trained on French corpora, making it better suited for our task. However, since we deal with non-contextual information, the word context understanding is not crucial. Consequently, the results from BERT and Camembert do not significantly differ. Like BERT, Camembert exists in various versions, each trained on different corpora with varying hyperparameters.

GPT Models [2]: GPT are another Transformer-based model initially designed for text generation. It can also be adapted for named entity recognition. Pre-trained on even larger and more diverse corpora, including multiple languages, GPT tends to perform better. However, these models are even more resource-intensive than BERT, with typically hundreds of billions of parameters. Note that the currently available open-source version is GPT-J, which is an alternative to GPT-3.

LSTM Models: While less resource-intensive, LSTMs are less common today due to the advancements of Transformer-based models. They often suffer from vanishing gradient issues during training and are generally harder to train.

Given the simplicity of our database and the fact that the information is in French, we chose to present the results of the base Camembert model, pre-trained on the OSCAR corpus (138 GB of text) with 110 million parameters. You can find the results of other models we fine-tuned in our GitHub report. Note that fine tuning all the BERT based model for NER task follow exactly the same step, we just have to change the name all the corresponding version.

4 Experimentation

4.1 Data Processing

As our task essentially involves named entity recognition, the first step is to format the data according to the required specifications. The pre-trained Camembert model takes input data in the format below, so we firstly preprocessed our dataset to fit this format.

```
{
  "tokens": ["Moullier", "Chometon", "Germaine", "Eurelie", "fleuriste", "femme", "1888", "Francais", "neant"],
  "ner_tags": ["B-surname", "I-surname", "B-firstname", "I-firstname", "B-occupation", "B-birth_date", "B-nationality", "0"]
}
```

The "B-" prefix indicates the beginning of an entity, while "I-" indicates an intermediate part of an entity that is not the beginning. In the example above, "Moullier" is the beginning of a surname, and "Chometon" is an intermediate part of a surname. Tokens labeled as "néant" are marked as "O" to signify that they are not part of the elements being targeted. Additionally, since there are no entries for "maiden_name" and "education_level", they are removed from the "ner_tags" list. Finally, any tokens labeled as 'Idem' are replaced with the appropriate information taken from the previous person in the list.

4.2 Split and distribution of entity types across splits

As depicted in Figure 7 in annex, we partitioned the dataset into training (80%), validation (10%), and test (10%) sets. Ensuring that the validation and test sets exhibit a similar distribution of data across different classes as the training set is crucial. We verified this by computing the distribution of tags across the various splits. The table below shows that the distribution of tags remains approximately consistent across the splits. Therefore, we expect the validation and test sets to offer a reliable assessment of the model's generalization capabilities.

Dataset	Surname	Firstname	Occupation	Age	Civil Status	Nationality
Train	15280	19898	11768	13067	8538	10646
Validation	1931	2510	1513	1673	1109	1304
Test	1948	2523	1479	1696	1058	1364

Dataset	Surname Household	Link	Birth Date	Lob	Employer	Observation
Train	4519	16544	5922	7423	2290	484
Validation	571	2082	703	916	290	56
Test	550	2102	719	897	271	62

Table 2: Dataset Distribution

4.3 Tokenization and label alignment

We also have to deal with the subword tokenization alignment problem that come with using Transformers. In fact, transformers use subword tokenization to split long words into their more common subwords, for example the surname "Desbroper" become "_Des", "bro", and "per". Common subwords would be reused across different words. This helps to keep the vocabulary size down, or conversely, helps to cover a larger input vocabulary using the same size of tokenizer vocabulary. However, our input sequence of "tokens" is space separated and our "ner_tags" are aligned with this space-separated token sequence. Since Transformers expect their input to be subword tokenized, the corresponding labels need to be aligned with the subword tokens as well. So for our " Desbroper " example, the corresponding labels will change from a single "B-surname" to " B-surname ", "IGN", and "IGN", i.e. the trailing subwords would be ignored.

4.4 Model and training

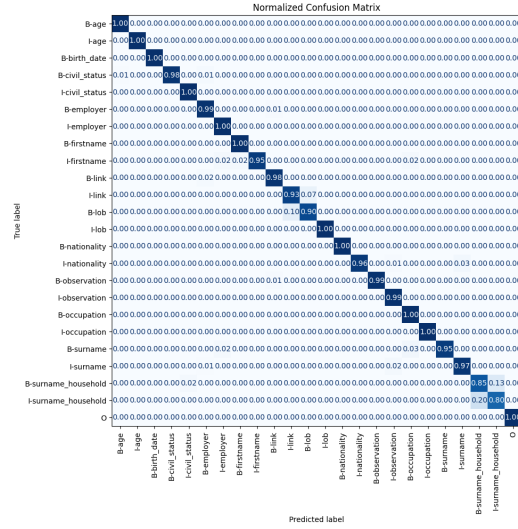
In the `AutoModelForTokenClassification.from_pretrained` call, we instantiate a `CamemBertForTokenClassification` object from HuggingFace[4]. This is a model with a CamemBERT encoder and a head consisting of a `torch.nn.Dropout` and a `torch.nn.Linear` layer for classifying the output of the pre-trained CamemBERT encoder into one of the entity types tags. We use the HuggingFace Trainer API to do the fine-tuning instead of native PyTorch. We fine-tuned our model on Google Colab on GPU T4 during 8 epochs with a batch size of 16 and a weight decay equals to 0.01.

4.5 Analysis of the results

We evaluate our trained model against the test split in two ways. Firstly, we compute the precision, recall, and f1-score for each entity type and produce the classification report. Secondly, for more graphical and easier to understand report of model performance, we compute the confusion matrix.

	precision	recall	f1-score	support
age	1.00	1.00	1.00	1631
birth_date	1.00	1.00	1.00	730
civil_status	1.00	1.00	1.00	1074
employer	0.99	0.97	0.98	302
firstname	1.00	1.00	1.00	2475
link	0.98	0.99	0.98	2042
lob	0.99	1.00	1.00	933
nationality	1.00	1.00	1.00	1327
observation	0.97	0.93	0.95	61
occupation	0.98	0.98	0.98	1495
surname	0.97	0.93	0.95	1889
surname_household	0.80	0.90	0.85	566
micro avg	0.98	0.98	0.98	14525
macro avg	0.97	0.97	0.97	14525
weighted avg	0.98	0.98	0.98	14525

(a) Classification report



(b) Confusion matrix

Figure 2: Results report

As evidenced by the classification report, the model demonstrates overall excellent performance, boasting precision, recall, and F1 scores surpassing 0.98 for most categories. However, it appears to slightly struggle with classifying "surname_household," with scores between 0.90 and 0.80. The confusion matrix also reveals a slightly challenges in distinguishing between the beginning of "surname_household" and an intermediate part that is not the beginning.

5 Conclusion

In the context of this project, we explored the task of information typing within the Socface project. Initially, we conducted a descriptive analysis of the available data. Despite similar performance among token classification models based on the BERT architecture, we ultimately chose the CamemBERT model due to its pre-training on French corpora. Although our initial dataset covered 25 448 individuals, our modeling experiments demonstrated that CamemBERT achieved excellent performance on the test set, with precision, recall, and F1 scores exceeding 0.98 for most categories. Given these results, this trained model can be directly used for information typing across 700 million individuals. We have thoroughly documented our code and each step of the experimentation process in our notebook. Additionally, increasing the size of the training data could further enhance the model's precision. Therefore, we recommend utilizing the CamemBERT model for information typing in Socface projet.

NB: In addition to Camembert, we fine-tuned other models based on the BERT architecture, and all these experiments can be found [in this GitHub repository](#)

References

- [1] <https://socface.site.ined.fr/en/the-project/methods/>.
- [2] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems*. 2020. URL: <https://proceedings.neurips.cc/paper/2020/hash/1457b4d2e9ac9e90f62ea9bc2a21e42b-Abstract.html>.
- [3] Jacob Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018). URL: <https://arxiv.org/abs/1810.04805>.
- [4] Hugging Face. *CamemBERT Documentation*. année. URL: https://huggingface.co/docs/transformers/en/model_doc/camembert.
- [5] Louis Martin et al. "CamemBERT: a tasty French language model". In: *arXiv preprint arXiv:1911.03894* (2019). URL: <https://arxiv.org/abs/1911.03894>.

6 Annex

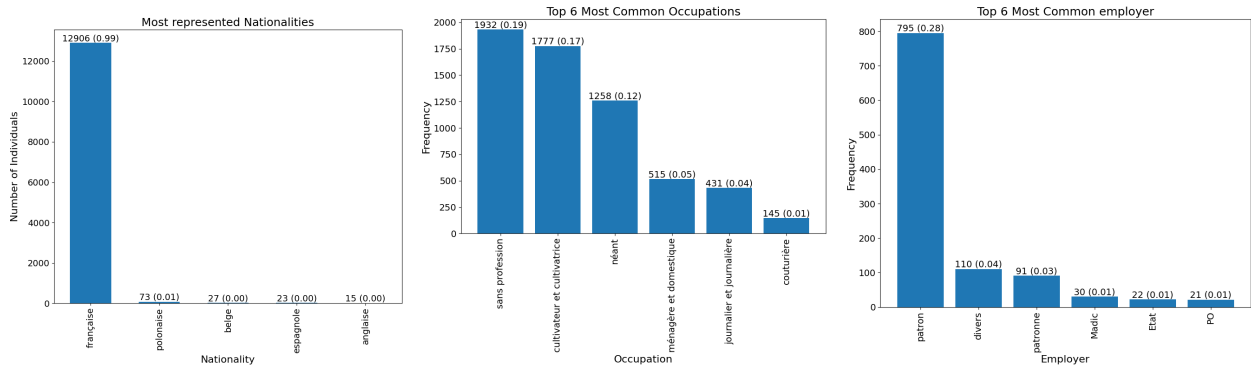
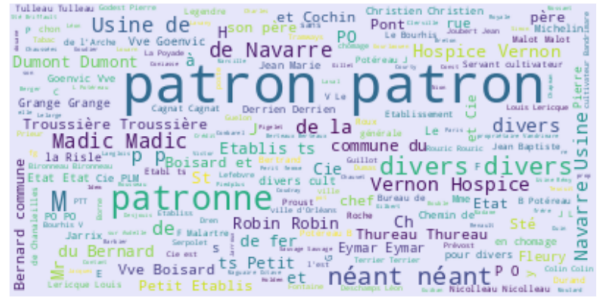


Figure 3: Repartition of nationality, occupation and employer

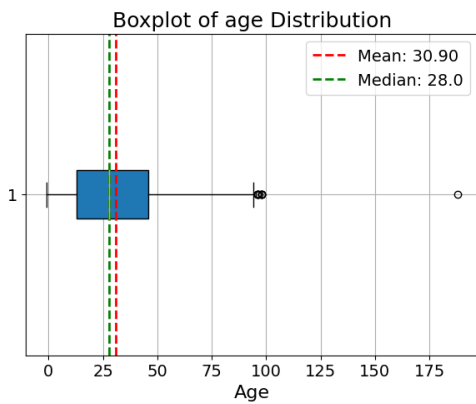


(a) Word cloud of occupations

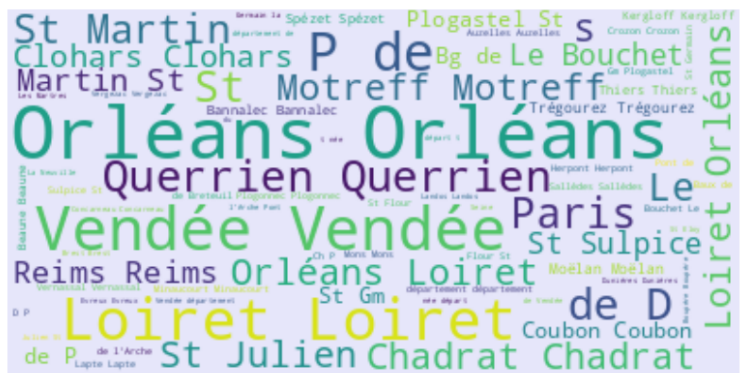


(b) Word cloud of employers

Figure 4: Word cloud of occupations and employers



(a) Age distribution



(b) Word cloud of lob

Figure 5: Age distribution and Word cloud of lob

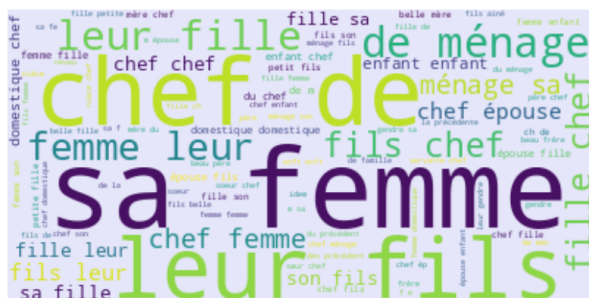





Figure 6: Word cloud of occupations and link

Subwords tokens from CamemBERT								
Tokens	<s>	_Cyril	le	_menuisier	_25	_Garçon	_français	</s>

Subwords tokens from BERT															
Tokens	[CLS]	Cyril	##le	menu	##isi	##er	25	G	##ar	##ç	##on	f	##ran	##çaise	[SEP]

```
Generating train split:  20326/0 [00:00<00:00, 91896.73 examples/s]
Generating validation split:  2563/0 [00:00<00:00, 34731.20 examples/s]
Generating test split:  2559/0 [00:00<00:00, 29789.85 examples/s]

DatasetDict({
  train: Dataset({
    features: ['tokens', 'ner_tags'],
    num_rows: 20326
  })
  validation: Dataset({
    features: ['tokens', 'ner_tags'],
    num_rows: 2563
  })
  test: Dataset({
    features: ['tokens', 'ner_tags'],
    num_rows: 2559
  })
})
```