Assessing Fine-Tuned NER Models with Limited Data in French: Automating Detection of New Technologies, Technological Domains, and Startup Names in Renewable Energy

Connor Maclean, Denis Cavallucci

machine learning & knowledge extraction

MDPI

Article

Assessing Fine-Tuned NER Models with Limited Data in French: Automating Detection of New Technologies, Technological Domains, and Startup Names in Renewable Energy

Connor MacLean * and Denis Cavallucci *

INSA Strasbourg, 67000 Strasbourg, France

* Correspondence: connor.mac_lean@insa-strasbourg.fr (C.M.); denis.cavallucci@insa-strasbourg.fr (D.C.)

Abstract: Achieving carbon neutrality by 2050 requires unprecedented technological, economic, and sociological changes. With time as a scarce resource, it is crucial to base decisions on relevant facts and information to avoid misdirection. This study aims to help decision makers quickly find relevant information related to companies and organizations in the renewable energy sector. In this study, we propose fine-tuning five RNN and transformer models trained for French on a new category, "TECH". This category is used to classify technological domains and new products. In addition, as the model is fine-tuned on news related to startups, we note an improvement in the detection of startup and company names in the "ORG" category. We further explore the capacities of the most effective model to accurately predict entities using a small amount of training data. We show the progression of the model from being trained on several hundred to several thousand annotations. This analysis allows us to demonstrate the potential of these models to extract insights without large corpora, allowing us to reduce the long process of annotating custom training data. This approach is used to automatically extract new company mentions as well as to extract technologies and technology domains that are currently being discussed in the news in order to better analyze industry trends. This approach further allows to group together mentions of specific energy domains with the companies that are actively developing new technologies in the field.

Keywords: natural language processing; named entity recognition; renewable energy; web-scraping

Citation: MacLean, C.; Cavallucci, D. Assessing Fine-Tuned NER Models with Limited Data in French: Automating Detection of New Technologies, Technological Domains, and Startup Names in Renewable Energy. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1953–1968. https://doi.org/10.3390/make6030096

Academic Editors: Elena Bellodi and Andreas Holzinger

Received: 29 May 2024
Revised: 16 August 2024
Accepted: 22 August 2024
Published: 27 August 2024

1. Introduction

In the context of transitioning to renewable energy, decision makers often find themselves overwhelmed by the increasing amount of noisy and unclear data on innovation. It is simply not possible to evaluate every single new company and product manually and in an efficient way. Non-automated approaches include manually analyzing data sources, searching for individual terms, and relying on systems such as Twitter/X hashtags. Manual annotation is a process that can be time-consuming, as we spent several hours manually annotating the 49 articles used for model training. Once models are trained, the annotation time is reduced from hours to seconds, while maintaining significant levels of accuracy. The goal of this paper is to develop an automated process using advancements in Natural Language Processing (NLP) to identify companies, technologies, and technological domains in order to quickly analyze trends in the industry. The goal of this project is to show the potential of automated analyses of news corpora using custom domain-specific Named Entity Recognition models trained on frugal datasets. Using recent open source tools such as [1–4] we are able to create a robust pipeline for data collection, data annotation, and model training. This approach allows for the creation of news corpora easily that can then be automatically analyzed by the trained model. This allows for insights to be quickly consulted without spending time poring over articles in order to find trending topics. The approach described in this article is applied specifically to the renewable energy

sector, but the approach is detailed in such a way as to be easily replicated and applied to other domains.

This study presents a comprehensive approach to enhancing Named Entity Recognition (NER) through the integration of advanced models such as fr_core_news_lg, Babelscape/Wikineural-Multilingual-Ner, CamemBERT, DistilCamemBERT, and Camembert NER. We address the challenges of entity disambiguation by proposing a robust methodology for data collection, preprocessing, annotation, and model training using spaCy [3] pipelines. Our research extends beyond traditional NER experimental frameworks by exploring the feasibility of training models on varying fractions (20%, 40%, 60%) of the dataset to evaluate the efficency of this approach when faced with frugal datasets. We further demonstrate practical applications in real-world scenarios such as identifying co-occurrences of organizations (ORG) and technologies (TECH) within articles.

1.1. Objective

The approach used involves finding mentions of startups and their new technologies in news media. This process is of course possible using indicators such as mentions using "#" on Twitter/X, but in order to develop a more robust system, we decided to use Named Entity Recognition (NER). This approach allows us to automatically find mentions of new companies, technological domains, and products directly from unstructured text on the web. Furthermore, fine-tuning a system using texts directly related to renewable energies will permit us to create a specialized system that can adapt to the various challenges presented by unstructured text and provide more accurate and nuanced insights into the evolving landscape of renewable energy startups and innovations.

This process is applied to news media, and allows for a corpus of specific articles to be quickly extracted and analyzed. We detail the web-scraping process and show that it is possible to create corpora using exact dates, publishing country, and language. The goal of this process is to quickly extract names of trending companies, technological domains, and potential products for decision makers to quickly orient themselves and their potential research and investments into certain sectors, products, or approaches.

1.2. Challenges

Several challenges to this project include fine-tuning models to recognize a new category as well as adapting the ORG category to include names of companies (startups) that it may never have seen in its original training phase. The hope with this fine-tuning approach is to evaluate the model's capacity to generalize to novel company names that may not have been specifically annotated in the model's original training data. As many companies, especially startups, were founded after the original training of these models, it is highly unlikely that all of the startups that are in the fine-tuning dataset will be contained in the original training data. This is to say, that although these models can effectively detect company names, it may be unlikely that they detect all novel company names. We will be evaluating their performance on these unseen companies and compare the models to their baseline performance on the category "ORG".

For this approach, we fine-tune several transformer and CNN models in order to add a category, "TECH". This category corresponds to a technological domain or a technological product. The specificity of this category is in the renewable energy domain. This means that while an entity "solar panel" will be correctly annotated by the fine-tuned models, an entity such as "iPhone" would not be annotated.

As most models have been previously trained on a large quantity of data in their training phases, the models used are able to adapt their weights using fine-tuning techniques. Fine-tuning is necessary when adding a new category as the model needs new data in order to learn and generalize. Fine-tuning is also necessary when examining domain-specific texts and adapting to the unique challenges presented by specific domains as well as new company names.

A further challenge faced here is disambiguation of entities between the "PERSON" and the "ORG" tags. We can see several entities tagged incorrectly, and this is due to the fact that some organizations do in fact use the name of a person for their organization. In this case, it is important to ensure a high quality of training data and to ensure a sufficient amount of examples in the new data for fine-tuning.

2. Similar Projects

Although we are seeking to effectively extract entities such as companies from news articles, effective systems have been created in the past. This work includes a paper on "Linked Enterprise Data for Fine Grained Named Entity Linking and Web Intelligence" [5]. This approach seeks to link named entities to company information using SQL databases and a Named Entity Recognition system that it names, "Recognyze". This system faced many challenges in preprocessing data for the NER system, such as disambiguation of casual company names, abbreviated names, and names which could be interpreted as people instead of companies.

In a previous paper, bidirectional LSTM models were used for health-domain named-entity recognition [6], using only a small, high-quality training dataset. Based on this work, it was shown that reliable NER pipelines were able to be created using fine-tuning techniques. With results showing an F1 score over 80% on all categories, we show that this approach is viable and presume that it can be improved by using recent advances in transformer [7] models.

For the use of NER in a similar project, an NER model was fine-tuned on several million domain-specific examples related to farming [8]. The paper shows that with several million annotations, the accuracy of NER models can achieve an F1 score over 97% on the task with sufficient data. The paper goes on to create a prototype of an automatic annotation framework that would allow for data to be quickly annotated in order to collect a sufficient amount of data for similar projects. Unfortunately, the tool is not currently accessible, but the work of this team demonstrates the effectiveness and robustness of developing domain-specific NER systems.

Other interesting work showing the effectiveness of NER on specific use cases includes the development of a custom NER model for pandemic outbreak surveillance using Twitter [9].

We have seen other business applications including resume ranking across multiple job domains using the same spaCy NER [3] model [10]. This approach detects quite a few different categories and aims to automatically extract information from resumes for recruiters to be able to analyze resumes without the need to manually scan each and every one. This approach achieves a significant performance in their initial tests and shows that automatically classifying information using NER is a viable approach and saves time for people and businesses.

A very complete project, although not concentrated on a commercial goal, shows the potential application for NER systems in the detection and classification of information. This approach, titled "Deep LearningBased Named Entity Recognition Models for Recipes" [11], shows the capacity of a model to adapt and generalize on new data. Their approach demonstrates that with a quality training dataset and clear, structured information to analyze, NER models can adapt well in a variety of scenarios. The work and projects performed on domain-specific NER models are impressive and show that this approach can be viable in the energy sector as well. For a new project, there are two important factors to consider. The first is the addition of new categories to NER models. The effectiveness of this approach depends on both the amount and quality of the annotated data. The second consideration is the model itself. Previous papers have used approaches based on statistical measures and ontologies behind their models such as "Linked Enterprise Data for Fine Grained Named Entity Linking and Web Intelligence" [5]. Recent advances, as shown by the previous articles using deep learning frameworks, show that similar, and even better performance, can be attained by using transformer models, RNN, and CNN models. These

approaches rely on capturing the context from the text itself in order to recognize both the entities themselves and the context in which they are found.

3. Methodology

3.1. Full Preprocessing and Training Pipeline

For the creation of the model, all of the steps discussed below are combined into a single pipeline. We start with the extraction and preprocessing of news articles so that we can be sure to target the maximum quantity of the relevant texts. Next, these texts are split into individual sentences and given to a first NER model. These annotated texts are then corrected, and the "TECH" category is added. Once the annotation phase is complete, the annotations are split into a training dataset (80%) and a validation dataset (20%).

Finally, the same training and validation data is fed to all five of our models for fine-tuning. The most efficient model is then used to annotate unseen texts and extract relevant information for decision makers. Below, the full pipeline is displayed to show the training process in Figure 1.

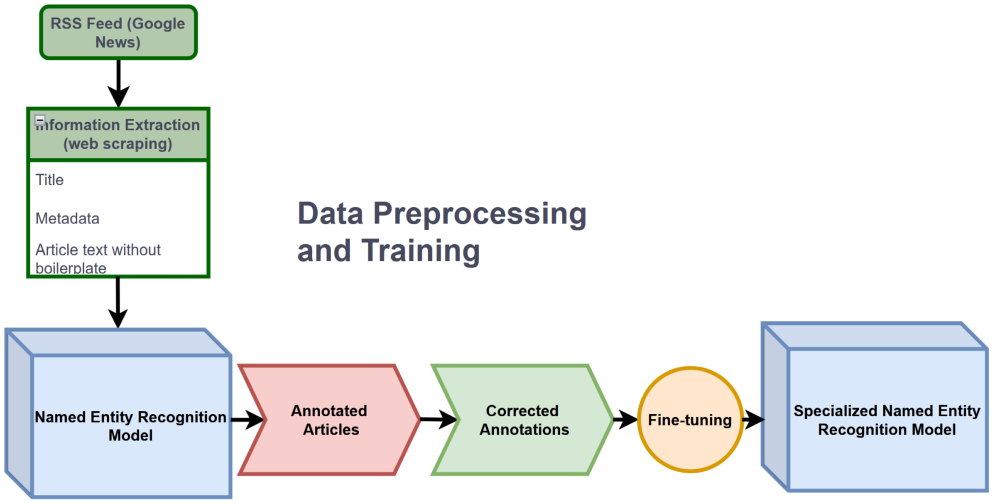

Figure 1. Pipeline.

3.2. Web-Scraping

For the construction of the training corpus, we decided to train the model on news articles that are as recent as possible. In order to do this, the library GNews [1] was used. This library allows for us to generate an RSS feed from a specific query in Google News. This library allows us to automatically filter articles by publication date, language, and country. Using this approach, it is possible to retroactively generate a corpus covering previous months or years of news on a specific topic. The limits of this approach are minimal, with the main downside being that the results only contain up to one hundred articles per day per request searched. This can also be viewed as a way to filter the resulting corpus; only the top one hundred results per day and per search term will be contained in the corpus. This approach allows for flexibility in corpus creation and allows for decision makers and analysts to very quickly extract relevant information from large quantities of unstructured text from news media.

The resulting RSS feed will contain up to one hundred articles for a specific date and for a specific search term. Using this RSS feed, we automatically classify the resulting articles by the request that was made and add fields for the date as well as the source of the media.

Once the RSS feed is created, the resulting article links are then opened using the BeautifulSoup [12] library. The text is then passed to the Justtext [13] library to remove

the maximum amount of boilerplate text. The titles of the articles, the date that they were scraped, the request, as well as the cleaned text are then stored in a CSV document.

The first version of the CSV document contains the text and metadata from 49 articles scraped using this methodology. As there is only a single annotator working on this project, the first training set is relatively small, containing 3260 annotated entities. The distribution of articles is shown in the Table 1.

Table 1. Categories of subjects used in training data.

Request	Translation	Number of Articles
énergies renouvelables	renewable energies	26
éolienne	wind turbines	17
hydroélectrique	hydroelectric	3
startup énergie	energy startup	3

All articles were written on either the 27 or the 28 of March 2024. While it is great to see that the category of "renewable energies" is well represented and composes 53% of the corpus, the "hydroelectric" and "energy startup" categories are underrepresented in the corpus. As the goal is to be able to find company names, technological domains, and products, this does not pose a problem in the case of fine-tuning the model, as we will see in the results.

3.3. spaCy NER

In order to assure the uniformity of the process of fine-tuning each model, we decided to use the spaCy [3] library. This library, one of the most widely used for NLP tasks, allows us to directly modify a simple configuration file in order to define the model, the hyperparameters, the vectors used, as well as the training and validation data used. Recently, the library allowed for the integration of the HuggingFace transformers library [2]. This new addition allows for us to simply test different models in order to fine-tune them and evaluate their performance on the task of Named Entity Recognition.

In the case of using spaCy's included models, the library automatically applies a processing pipeline to our text. This creates a "Doc" object inside of spaCy. This library allows us to directly access the component of the pipeline that will be used, NER in our case, and to modify it. For our fine-tuning, only the tokenizer and NER components will be used. This internal framework is show by Figure 2.

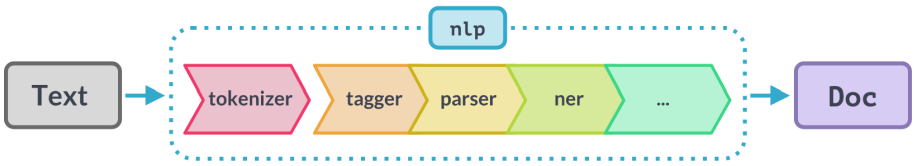

Figure 2. spaCy's language processing pipeline [3].

3.4. Preprocessing

After choosing our training corpus, the data were preprocessed in order to facilitate the the annotation and training processes. The process is as follows:

- All articles were split into individual sentences;
- All sentences were fed to spaCy, in order to create Doc objects;
- The tokenizer was applied to all texts;
- The default large French model "fr_core_news_lg" was used to preannotate texts with "ORG", "LOC", and "PER" labels.

This approach allows for us to split the data into chunks that are manageable both for the annotator and for the model.

3.5. Annotation

To begin the annotation stage, spaCy's [3] default large French model ("fr_core_news_lg") is used. This model, pre-trained on a corpus of French news, is used to predict LOC (location), MISC, ORG (organization/company), and PER (person) labels. All articles are split into sentences by spaCy's French tokenizer before being automatically annotated. This process saves time as the annotator(s) only need to correct annotations and add entities that were skipped by the model.

Once all articles are annotated, a script is run to convert the annotations into a format accepted by Doccano [4] in order to annotate using a graphical interface. During the manual annotation stage, previous annotations made by the model are corrected and new "TECH" (technological domain or product) annotations are manually added.

It is important to note that although we are the most interested in the "ORG" and "TECH" tags for the moment, all entities belonging to the other categories will also be annotated in order to avoid the "catastrophic forgetting problem" [14]. This issue leads to models losing the ability to detect the original categories as the models would generalize on the new, incorrectly annotated data. It is important to note that as this is a prototype designed to prove the validity of this approach, all annotations were exclusively made and verified by one annotator. Future iterations will include more annotators to ensure quality and inter-annotator agreement.

Annotation Guide

During the annotation stage, a small set of guidelines referenced in Table 2 was established to make sure that all entities are annotated in the same way.

Table 2. Annotation guidelines.

Entity Tag	Description
PERSON	proper noun: name of person, position; i.e., President of XXXX, first name + last name
LOC	location: city, country, place
ORG	proper noun: company, government, committee, etc.
TECH	technological field or product; i.e., solar energy, solar panels

As there is only one annotator, these guidelines were established so that the annotator can quickly consult the guide.

The annotation was performed following the BIO annotation scheme [15]. This annotation scheme attributes a "B" label for all tokens that are located at the beginning of an entity, an "I" label for all tokens that are located inside of the entity, and an "O" label for all other tokens. Below, the distribution for all B, I, and O tokens are shown for all three datasets. The tokens are listed in descending order, starting, of course with the "O" tokens.

To note, the novel "TECH" category is the most represented category in all three datasets, which was a conscious decision as we are highlighting the performance of these models to adapt to a novel category. A full overview of the labeling scheme is listed in Tables 3–5.

Table 3. Training dataset.

Entity Tag	Count
O	49,773
I-TECH	1488
B-TECH	1082
I-ORG	743

Table 3. *Cont.*

Entity Tag	Count
B-ORG	696
B-LOC	586
I-LOC	254
I-PER	202
B-PER	196
B-MISC	1
I-MISC	1

Table 4. Validation dataset.

Entity Tag	Count
O	12,524
I-TECH	411
B-TECH	271
I-ORG	208
B-ORG	182
B-LOC	161
I-LOC	84
I-PER	43
B-PER	40
B-MISC	1

Table 5. Test dataset.

Entity Tag	Count
O	9133
B-ORG	212
I-TECH	158
B-TECH	153
I-ORG	145
B-LOC	75
I-LOC	39
I-PER	30
B-PER	22

3.6. Training spaCy Pipelines

The internal process of fine-tuning a spaCy [3] model is detailed on their website. Using the default large French model "fr_core_news_lg", the model will parse the training data and learn the associations between the text and the entity labels that they are given. Next, in the validation phase, the model will use 20% of the training data in order to evaluate its progress. This process is iterative, with the model continually improving and correcting its weights. Once there is no significant change in performance, the model training will automatically finish. The best-performing model's weights and the last model's weights are then automatically saved.

The architecture used for most spaCy models is a Convolutional Neural Network (CNN) [16], which contrasts with the more recent transformer models [7] that are used in the comparison. An example of spaCy's CNN training process is dispayed in Figure 3. For this study, all transformer models will use the same iterative training process.

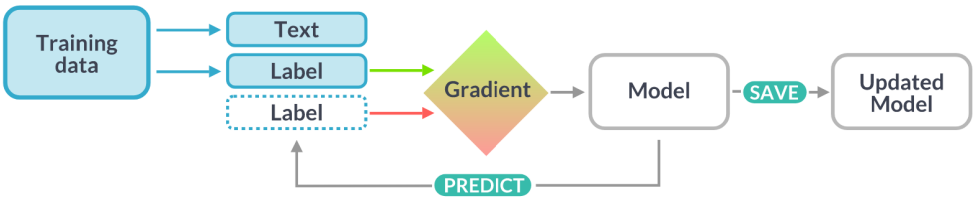

Figure 3. Training spaCy's included models [3].

3.7. Choice of Models

3.7.1. spaCy fr_core_news_lg

In the interest of having a complete test using the best tools available for the analysis of French, we decided on five separate models. As we are using the training framework provided by spaCy [3], we decided to use the default spaCy model, "fr_core_news_lg". This model differs from the others as it is a Convolutional Neural Network [16] and not a transformer [7] model like the four others. Nevertheless, this model is quicker to train, and has a performance that is close to that seen with the more advanced models.

3.7.2. Babelscape/Wikineural-Multilingual-Ner

For the second model, we decided to test the Wikineural-Multilingual-Ner model proposed by Babelscape [17]. This model is the most popular specialized Named Entity Recognition (NER) model that can be applied to French on the HuggingFace Hub [2].

3.7.3. CamemBERT

For an analysis of French transformer models on NER tasks, it is impossible to avoid discussing and using the CamemBERT model [18]. This model, based on BERT [19] is the first specialized model for French, and remains interesting to this day for its industrial and research applications.

The model has been trained on several downstream tasks, including NER, on which it improves the state of the art and asserts itself as a reference for French transformer models.

3.7.4. DistilCamemBERT

As with the DistilBERT model [20] that came before it, DistilCamemBERT [21] seeks to reduce the size of the CamemBERT model while also preserving its performance as much as possible.

As shown in the following Table 6 taken from the paper introducing DistilCamemBERT [21], we see a decrease in performance across almost all tasks, except for NER. The paper shows that there is a very slight increase in performance in DistilCamemBERT over the CamemBERT base model.

Table 6. Model accuracy of CamemBERT and DistilCamemBERT. All measurements are based on the F1 score.

Model	Sentiment (%)	NER (%)	NLI (%)	QA (%)
CamemBERT	95.74	88.93	81.68	79.57
DistilCamemBERT	97.57	89.12	77.48	62.65

3.7.5. Camembert NER

This model is one of the most popular NER models for French with over 2 million downloads on the HuggingFace Hub [2] as of April 2024. The model is a fine-tuned version of the CamemBERT [18] model discussed earlier. The CamemBERT model was fine-tuned on the wikiner-fr dataset [22] in order to further specialize the model on the task of Named Entity Recognition, for which CamemBERT was already trained [23].

4. Model Training

All models are trained using the same data and the same hyperparameters to ensure that no training is biased. The models are trained using 3260 annotations, split into 80% training and 20% validation sets. An additional 460 annotations were created and corrected from a random subset of articles. These annotations were used to evaluate all models so to ensure that they perform well when shown new data not used during training.

Hardware and Optimizer

Training was performed using a GeForce RTX 4070 Laptop GPU on a laptop with 16 GB of RAM. The Adam V1 optimizer was used. All models were trained using the default spaCy settings, which does not limit the number of epochs of training and only finishes training once no significant progress is being made from one epoch to the next.

5. Results and Discussion

The results, shown in Table 7 are below the baseline for most, but above the baseline performance in the case of CamemBERT. When considering the relatively small amount of training data that was used during fine-tuning, we can demonstrate the effectiveness of this approach and can assume that performance would increase with the addition of more annotated training data. Examining the results, we can see a clear difference between the models used. Performance varies between the categories, with the highest level of performance observed in the already trained PER category.

When comparing these models, we can observe that the one CNN model, spaCy fr_core_news_lg, has the most difficult adaptation to our data. In response to the performance of this model, the decision was made to examine the effectiveness of transformer models on this same task. As a result, this model is the only CNN represented in this paper. This model is the least capable of the five of generalizing to new data. As discussed, the main roadblocks in this project are the addition of a new category, "TECH", as well as expanding the "ORG" category to include names of new companies/startups that were not seen in the original training data before fine-tuning.

Table 7. Model F1 score evaluated on independent test data.

Model	F1 Score TECH	F1 Score ORG	F1 Score LOC	F1 Score PER
spaCy fr_core_news_lg	56.67%	54.35%	73.53%	92.68%
Babelscape	58.37%	71.81%	76.51%	89.36%
CamemBERT	**91.28%**	**89.98%**	**91.39%**	**100.00%**
DistilCamemBERT	66.67%	74.94%	63.75%	88.89%
CamemBERT NER	68.50%	85.10%	74.36%	95.45%

Here, we can see that the model that adapted the best to our new data is CamemBERT [18]. This model has an impressive performance and generalizes very well to the new TECH category as well as to the new additions to the ORG category. It is followed by CamemBERT NER [23], which is closely followed by DistilBERT [20]. It is important to note that the PER category is the least represented in the final test data and only contains 30 annotations.

6. Training with Limited Data

As data annotation is a resource-intensive task, it is crucial to assess the performance of Named Entity Recognition (NER) models on smaller, frugal datasets. To this end, we conducted experiments by training our best model on various fractions (20%, 40%, 60%) of the full training data. The 80% split is the standard amount of training data as the remaining 20% is used for the validation set. This approach helps demonstrate the feasibility of deploying NER models in scenarios where only limited annotated data are available, such as in specialized domains or emerging fields.

Our results, shown in Table 8, indicate that the models retain a significant level of performance even with reduced training data. Specifically, we observed that while there is a gradual decline in accuracy and F1 scores as the training dataset is reduced, the model trained on 40% of the data still retains moderately satisfactory performance compared to the model trained on and 80% of the data. The model trained on 60% of the data retains performance only a few percentage points below the model trained on 80% of our data. This suggests that our NER model is robust and can generalize well even with smaller datasets, making it a viable option for applications where data annotation resources are constrained.

Additionally, training with limited data highlights the importance of model efficiency and data quality. Our experiments show that careful selection and annotation of a smaller, high-quality dataset can often compensate for the lack of extensive annotated data. Future work could explore techniques such as data augmentation, active learning, and semi-supervised learning to further enhance the performance of NER models on frugal datasets, ensuring that they remain effective and reliable in various practical applications.

Table 8. Performance metrics for different training data proportions.

Training Data	Entity Type	Precision (%)	Recall (%)	F1 Score (%)
20%	TECH	71.63%	66.01%	68.71%
	ORG	68.78%	76.89%	72.61%
	LOC	73.17%	80.00%	76.43%
	PER	90.91%	90.91%	90.91%
40%	TECH	91.35%	62.09%	73.93%
	ORG	75.95%	84.91%	80.18%
	LOC	64.89%	81.33%	72.19%
	PER	95.65%	100.00%	97.78%
60%	TECH	88.71%	71.90%	79.42%
	ORG	81.55%	79.25%	80.38%
	LOC	63.22%	73.33%	67.90%
	PER	83.33%	90.91%	86.96%
80%	TECH	93.79%	88.89%	91.28%
	ORG	93.40%	86.79%	89.98%
	LOC	90.79%	92.00%	91.39%
	PER	100.00%	100.00%	100.00%

7. Practical Applications

Once the model is fine-tuned on our data, we can use it to automatically extract information from news articles related to renewable energies. To demonstrate its performance, we created a small corpus of the top 151 articles related to renewable energies and energy startups in France from Google News https://news.google.com (accessed on 5 April 2024). These articles, from Table 9 were scraped between 2 April and 4 April 2024.

Table 9. Articles per query: published between 2 April and 4 April 2024.

Query	English Translation	Count
énergies renouvelables	Renewable energies	138
startup énergie	Energy startups	13

These results confirm that our approach is possible and that the new category is able to be successfully integrated into the existing NER pipelines for French. With the best-performing model, the correct entities are extracted in roughly nine out of ten instances. This performance shows that a majority of technological domains, names and titles, company names, and locations can be automatically extracted and used to gain insight into industry trends.

To give an example of company and product names extracted, we see here that the company, Ombrea, is correctly annotated in Figure 4. We also have the "sliding shaders piloted by an algorithm" correctly annotated, which belongs to our new "TECH" category.

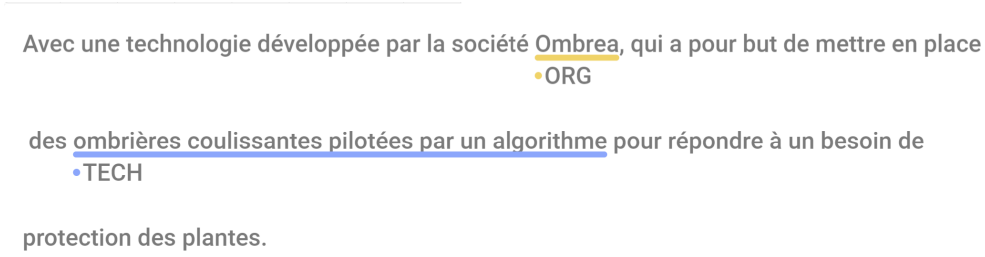

Figure 4. Correct annotations outside of training data predicted by the trained model.

To show the technological domains that are correctly annotated, we have the following example in Figure 5 that has been annotated by our newly fine-tuned model.

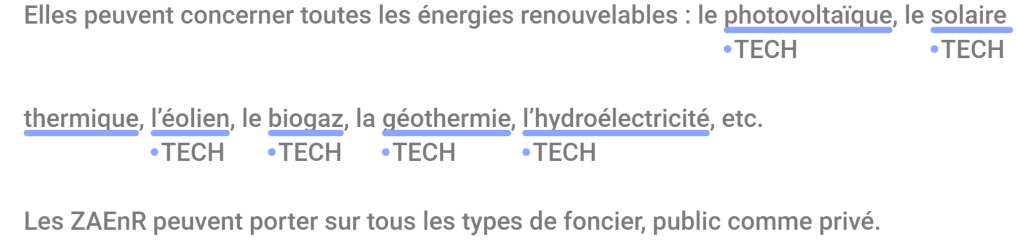

Figure 5. Energy domains correctly annotated by the model.

Here, we can see that all relevant technological domains related to renewable energy have been correctly found and labeled by the model. These domains are photovoltaic, solar, wind, biogas, geothermal, and hydroelectric energy.

These results allow us to extract information on trending companies, trending technological domains, and new products. A next step, below, is the extraction of company mentions in news articles based on a simple metric of the number of articles in which a given company is mentioned.

Below, in Table 10 are all the companies that appear in six articles or more in our small corpus from Table 9.

Table 10. All organizations mentioned in six or more articles.

Company	Number of Articles
TotalEnergies	17
Engie	17
EDF	8
Greenpeace	7
COP28	7
Enercoop	7
Ekwateur	6
Selectra	6
ilek	6
VertVolt	6

This approach allows for a quick analysis of mentions in news media, and as these data include the top 150 news articles in renewable energies and energy startups, they give a reliable image of which companies are currently being discussed.

When looking at the second category, and establishing a minimum of six articles for the item to be discussed in, we can also quickly extract the latest trending technological domains and products from the news, as shown in Table 11.

Table 11. All technological domains and products mentioned in six or more articles.

Technology	English Translation	Number of Articles
éolien	wind	33
solaire	solar	26
hydraulique	hydraulic	23
nucléaire	nuclear	20
panneaux solaires	solar panels	18
éoliennes	wind turbines	18
hydroélectricité	hydroelectricity	17
éolienne	wind turbine	16
photovoltaïque	photovoltaic	9
hydrogène	hydrogen	9
panneaux photovoltaïques	photovoltaic panels	9
biogaz	biogas	9
centrale solaire	solar power plant	8
hydroélectriques	hydroelectric	8
parcs éoliens	wind farms	7
hydroélectrique	hydroelectric	7
centrales nucléaires	nuclear power plants	7
géothermie	geothermal	7
l'énergie solaire	solar energy	6
batteries	batteries	6
énergie éolienne	wind energy	6

It is important to note that although we are interested in company names, the model is specialized in finding organization names. This means that even nonprofit organizations such as Greenpeace or conferences such as the COP28 are also detected by the model.

Once the list of companies is extracted, it is possible to access public databases on company financial information, websites, and contacts in order to manually evaluate companies that are recognized by the model.

Co-Occurrences of ORG and TECH in the Same Article

In order to have a better understanding of the technology and the companies that are discussed in a given corpus, we propose to analyze cases where an organization and a given technological domain are present in the same article. Using the same corpus given in Table 9, we count every article that contains ORG + TECH pairs. This count proposed in Figure 6 will add together the amount of articles that contain two given entities. When filtering, we find over a thousand pairs mentioned in at least two articles. For the sake of simplicity, we have condensed this list to include only the most frequent pairs of ORG + TECH. These lists allow us to immediately access the companies and domains that are discussed together, which often translates to the domain of activity for a given company.

To preprocess this data, companies were all converted to lowercase as some mentions had varied casing. All technological domains were converted to a base form in order to group plural and singular forms together.

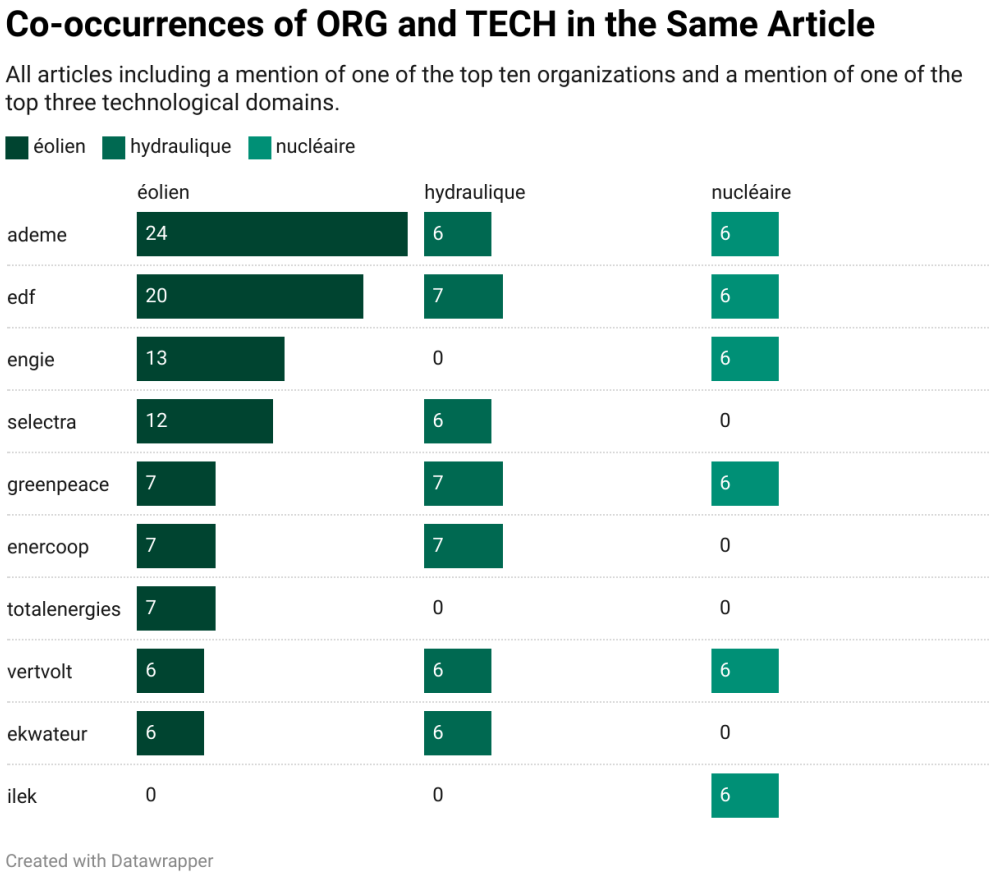

Figure 6. Co-occurrence of organizations and technological domains in the same article.

When filtering for the companies and technological domains that are mentioned the most frequently, we find certain companies and domains that are fairly well represented in our small news corpus. In order to show the information that was automatically extracted, we have ordered the top ten companies with the top three technological domains mentioned.

All data have been ordered by mentions of wind power (éolien); the other columns are hydraulic power and nuclear power.

In addition to showing the trending companies and their associated domains, this functionality has another use. This approach allows for the automatic classification of companies that are only mentioned once or twice. If they are present with a mention of a specific domain, we can associate the two in order to immediately classify previously unknown companies.

8. Conclusions

This approach allowed the creation of a Named Entity Recognition model specialized in detecting startup names as well as technological domains in the renewable energies sector. We have fine-tuned the model to recognize entities specifically related to this sector and have added a new category, "TECH". We see a high F1 score in all categories for the most optimal model, as well as performance above 65% on the new category in all CamemBERT models.

Using these models to extract information is a quick process that we have shown can have a high level of accuracy. With a small amount of code and a fine-tuned model, we can extract information from the web in a fraction of the time manual analyses would need. We have shown that it is possible to extract mentions of companies and new startups, as well as adding a new category. Statistically analyzing mentions of companies and technological domains allows for us to analyze current trends in the industry as well as to detect new technologies and companies. We have shown that regrouping and taking measures of co-occurrences of the two main categories allows for relevant information to be immediately extracted from a custom corpus.

With this new category, we have demonstrated that all that is needed is a few thousand high-quality, domain-specific examples. We can easily fine-tune a model and allow for it to generalize and find new entities in our data based on patterns that have been assimilated.

Using only a few thousand examples, it is possible to fine-tune a model capable of extracting entities correctly nine times out of ten. With this approach, spending only a few hours annotating data allows for an enormous amount of time saved for both analysts and decision makers.

Future Perspectives

As we have found that this approach of adding new, specific categories for technological domains and products is possible without an exorbitant amount of resources, it would be useful to continue this project. It is possible to use the newly developed model to annotate new texts, correct the annotations, then fine-tune the model to further improve its effectiveness.

With the addition of more annotated data, as well as a more diverse set of articles to annotate, the best-performing models can be fine-tuned in order to further increase performance and reliability. As a next step, we can also evaluate the effectiveness of training on restrained, frugal datasets in order to maximize performance.

As we have seen the successful creation of new categories with a small amount of high-quality data, it would be possible to continue expanding the capabilities of these models. A future possibility as well is to further discuss metrics that interest business use-cases to regroup and analyze the extracted information in useful ways. We can continue creating and fine-tuning new categories in line with information needed by business analysts.

Once the models are judged to be satisfactory for an industrial application, a database of trending companies, products, and domains can be established. Products can be linked to existing public patent information, information on the company, and scientific papers that the company has published. Together, these indicators and the information extracted can be a way to evaluate companies quickly in order to rate their performance and future viability.

Author Contributions: Conceptualization, C.M. and D.C.; methodology, C.M.; software, C.M.; validation, C.M. and D.C.; formal analysis, C.M.; investigation, C.M.; resources, C.M.; data curation, C.M.; writing—original draft preparation, C.M.; writing—review and editing, C.M. and D.C.; visualization, C.M.; supervision, D.C.; project administration, D.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was completed during a PhD contract funded by EDF-Discovery's contribution to the AIARD (Artificial Intelligence Aided Research and Development) industrial chair. The AIARD industrial chair is co-financed by the Grand EST region, the Eurométropole de Strasbourg and nine other companies, a list of which is available on the Chair's website (www.aiard.eu).

Data Availability Statement: All data used in this study has been generated using the tools and methods outlined in this article. Should any difficulties arise when attempting to recreate these experiments, all data used, including annotations made during the project, can be sent upon request.

Acknowledgments: This work was accomplished in the CSIP group at the ICube laboratory, INSA of Strasbourg during a PhD contract. We extend our appreciation to all colleagues for their support, collaboration, and ideas given during this project. Special appreciation is extended to the EDF (Électricité de France) Discovery Group for their invaluable collaboration throughout this project. Their support and insights have greatly enriched the outcomes of this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Abdullah, M. Gnews: Provide an API to Search for Articles on Google News and Returns a Usable JSON Response. Online Resource on GitHub. Available online: https://github.com/ranahaani/GNews (accessed on 5 April 2024).
2. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* **2019**, arXiv:1910.03771.
3. Honnibal, M.; Montani, I. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks, and Incremental Parsing. Published in 2017. Available online: https://spacy.io (accessed on 2 April 2024).
4. Nakayama, H.; Kubo, T.; Kamura, J.; Taniguchi, Y.; Liang, X. doccano: Text Annotation Tool for Human. 2018. Available online: https://github.com/doccano/doccano (accessed on 2 April 2024).
5. Weichselbraun, A.; Streiff, D.; Scharl, A. Linked Enterprise Data for Fine Grained Named Entity Linking and Web Intelligence. In Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS '14), Thessaloniki, Greece, 2–4 June 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 1–11, ISBN 978-1-4503-2538-7.
6. Unanue, I.J.; Borzeshi, E.Z.; Piccardi, M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *J. Biomed. Inform.* **2017**, *76*, 102–109. [CrossRef] [PubMed]
7. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017; Volume 30, pp. 5998–6008. Available online: https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf (accessed on 3 April 2024).
8. Kumar, M.; Chaturvedi, K.K.; Sharma, A.; Arora, A.; Farooqi, M.S.; Lal, S.B.; Lama, A.; Ranjan, R. An Algorithm for Automatic Text Annotation for Named Entity Recognition Using the spaCy Framework. Preprints 2023. Available online: https://typeset.io/pdf/an-algorithm-for-automatic-text-annotation-for-named-entity-3r6892x9.pdf (accessed on 3 April 2024).
9. Jayathilake, H.M. Custom NER Model for Pandemic Outbreak Surveillance Using Twitter. MSc Thesis, Robert Gordon University, Aberdeen, Scotland, 2021.
10. Satheesh, D.K.; Jahnavi, A.; Iswarya, L.; Ayesha, K.; Bhanusekhar, G.; Hanisha, K. Resume Ranking based on Job Description using SpaCy NER model. *Int. Res. J. Eng. Technol.* **2020**, *7*, 74–77.
11. Goel, M.; Agarwal, A.; Agrawal, S.; Kapuriya, J.; Konam, A.V.; Gupta, R.; Rastogi, S.; Niharika; Bagler, G. Deep Learning Based Named Entity Recognition Models for Recipes. Preprint. Available online: https://arxiv.org/abs/2402.17447 (accessed on 4 April 2024).
12. Richardson, L. Beautifulsoup4: Screen-Scraping Library. Available online: https://www.crummy.com/software/BeautifulSoup/bs4/ (accessed on 4 April 2024).
13. Pomikálek, J. jusText: Heuristic-Based Boilerplate Removal Tool. Available online: https://github.com/pomikalek/jusText (accessed on 4 April 2024).
14. Korbak, T.; Elsahar, H.; Kruszewski, G.; Dymetman, M. Controlling Conditional Language Models without Catastrophic Forgetting. *arXiv* **2022**, arXiv:2112.00791.
15. Ramshaw, L.A.; Marcus, M.P. Text Chunking Using Transformation-Based Learning. *arXiv* **1995**, arXiv:cmp-lg/9505040.
16. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]

17. Tedeschi, S.; Maiorca, V.; Campolungo, N.; Cecconi, F.; Navigli, R. WikiNEuRal: Combined Neural and Knowledge-based Silver Data Creation for Multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*; Moens, M.-F., Huang, X., Specia, L., Wen-tau Yih, S., Eds.; Association for Computational Linguistics: Punta Cana, Dominican Republic, 2021; pp. 2521–2533.
18. Martin, L.; Muller, B.; Suárez, P.J.O.; Dupont, Y.; Romary, L.; de La Clergerie, É.V.; Seddah, D.; Sagot, B. CamemBERT: A tasty French language model. *arXiv* **2019**, arXiv:1911.03894.
19. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
20. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
21. Delestre, C.; Amar, A. DistilCamemBERT: Une Distillation du Modèle Français CamemBERT. In CAp (Conférence sur l'Apprentissage Automatique), Vannes, France, July 2022. Available online: https://hal.archives-ouvertes.fr/hal-03674695 (accessed on 4 April 2024).
22. Nothman, J.; Ringland, N.; Radford, W.; Murphy, T.; Curran, J.R. Learning multilingual named entity recognition from Wikipedia. *Artif. Intell.* **2013**, *194*, 151–175. [CrossRef]
23. Polle, J.B. LSTM Model for Email Signature Detection. Medium, 24 September 2021. Available online: https://medium.com/@jean-baptiste.polle/lstm-model-for-email-signature-detection-8e990384fefa (accessed on 4 April 2024).