SUPPORT POOL
OF EXPERTS PROGRAMME

AI Possible Risks & Mitigations

Name Entity Recognition

by Isabel BARBERÁ

Table of Contents

Disclaimer by the Author: the examples and mentions of companies in this report are illustrative and do not imply that the author considers them the only or the best choice. The technology analysis presented in this report is based on the state of the art of the technology in August 2023.

Document submitted in September 2023

1. Background

Description of the task, main technologies used and references to some openly accessible examples.

Named Entity Recognition (NER) is an information extraction technique employed in natural language processing (NLP)[1]. NER is used to identify named entities such as names, organizations and locations within a document and classify them into predefined categories.

NER plays a vital role in NLP systems such as chatbots and search engines. Its application extends to diverse fields like healthcare, finance, human resources, customer support, higher education, and social media where NER can assist in extracting valuable information from different textual sources.

How does NER work?

Named Entity Recognition technology is based on three main methods: lexicon, rules, and machine learning[2]:

- Lexicon-based or dictionary-based approaches rely on a predefined list of terms from different sources such as pre-existing labeled datasets and online resources. In this approach, the input text is matched with entries in the lexicon to identify the named entities. This method might have trouble classifying new named entities and entities with ambiguous meaning or variations in spelling.
- Rule-based systems contain rules that are constructed either manually or automatically[3] and that are designed to detect entities based on specific patterns or criteria within the text.
- Supervised machine learning-based methods can automatically identify and classify named entities in new text by learning from annotated data. This method requires a significant amount[4] of annotated training data to estimate and fine-tune the parameters of the models. While earlier NER systems primarily relied on lexicon and hand-crafted rule-based approaches, modern techniques predominantly employ machine learning due to their ability to adapt and generalize well to various contexts and domains. Some NER systems combine multiple methods to enhance their performance and accuracy.[5]

Emerging NER systems that employ unsupervised machine learning Large Language Models like BERT[6], GPT-4, LlaMA and Mistral could offer an alternative approach that could help reduce the often time-consuming and costly process of annotating training data with labeled named entities. Though this is still a novel approach, it holds the potential to handle more complex tasks in comparison to traditional supervised methods.

[1] Natural language processing is an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyse large amounts of natural language data (Source: Wikipedia)

[2] Venkat N. Gudivada, in Handbook of Statistics, 2018

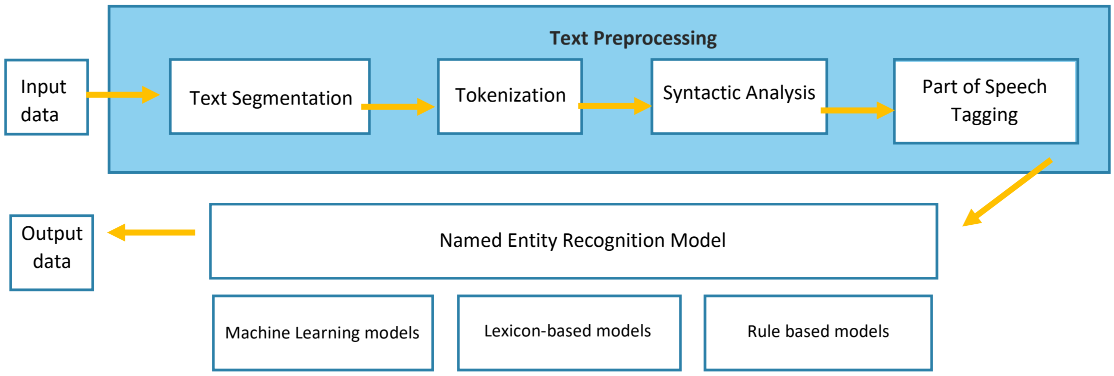
[3] The rules can be obtained using machine learning techniques to identify patterns and connections between words and their corresponding entity types.

[4] This differs depending on the model task and the dataset used. For instance, Twitter NER Corpus contains about 100,000 tokens, and OntoNotes 5.0 more than 1,5 million tokens. Source: https://huggingface.co/; https://www.defined.ai/blog/entity-recognition-datasets/.

[5] Keretna et al., 2014

[6] Devlin J, Chang M, Lee K, Toutanova K, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2019

The NER process typically involves the following steps:



The input data goes through a text preprocessing phase where the following subprocesses take place:

- Text Segmentation: in this initial stage, the textual data (input data) is divided into meaningful units, such as sentences or paragraphs. Text segmentation helps organize the text into manageable sections for further analysis.
- Tokenization: tokenization involves breaking down the segmented text into smaller units called tokens. Tokens are typically individual words, and they serve as the basic unit of analysis.
- Syntactic Analysis: it is often performed using techniques like parsing or dependency parsing[7] and it examines the grammatical structure and relationships between words in a sentence. Syntactic analysis helps determine how words are connected and their roles within the sentence, such as subject, verb or object.
- Part of Speech (POS) Tagging: POS tagging involves assigning grammatical tags to each token, indicating their respective part of speech (noun, verb, adjective, etc.).

After this categorization, further processing and analysis can take place depending on the specific requirements and goals of the application.

By segmenting and tokenizing the text, performing syntactic analysis, and applying POS tagging, NER systems can better understand the structure, relationships, and linguistic features of the text, what facilitates a more accurate identification and classification of named entities.

After the initial preprocessing steps are complete, the NER model is applied to the preprocessed text. During this step the NER model determines whether each token corresponds to a named entity and proceeds to categorize it. This results in an output containing identified named entities and their associated entity types.

Most NER models have the capacity to operate in numerous languages, however, their performance may vary due to factors such as the availability of training data and the complexity of the language. While English may have an advantage because there are more available resources in that language,

[7] Parsing is a core natural language processing technique that can be used to obtain the structure underlying sentences in human languages: Alonso, M.A; Gómez Rodríguez, C.; Vilares, J. "On the Use of Parsing for Named Entity Recognition". Appl. Sci. 2021

the effectiveness of NER models in other languages is continuously evolving. There are also multilingual models like BERT that have been trained in more than 104 different languages.

How are NER solutions available in the market?

NER systems are mostly offered as Software as a Service (SaaS) solution offering the possibility to use pre-trained models or custom models that customers can train with their own dataset.

Most vendors offer NER systems as a cloud solution via a system of APIs[8], what seems to be the preferred option for most customers because of their ease of integration and fast productivity. Though some providers of this technology offer a general system where models are already trained and shared by all the customers, there are also vendors that offer the possibility to have a custom model that the customer can train, finetune and delete when necessary[9].

Some vendors also offer the possibility for customers to host the models on-premises making the NER capabilities available in an own local environment. This can be a good alternative to comply with strict security and data governance requirements.

It is also possible to develop and implement your own NER solution in-house. There are different NER open-source tools available for this, such as:

- Stanford Named Entity Recognizer (SNER), a Java tool with pre-trained models for recognizing most people, places, businesses, and other entities.
- Natural Language Toolkit (NLTK), a Python library for natural language processing.
- SpaCy, a Python framework well known for its ease of use and implementation.

NER SaaS solution hosted in cloud	NER Third party solution hosted on premises	NER self-developed, hosted on premises
- Ready to use models that are trained by the vendor - Possibility to create your own self-trained models	Models trained by vendor or customer	Models trained by user

Data flow in a NER solution:

NER is commonly applied to digital text documents, such as web pages, social media posts, emails, or other types of digital publications. The text we aim to examine is referred to as the **Input Data**. The outcome obtained from the information extraction process after named entities have been recognized is known as the **Output Data.**

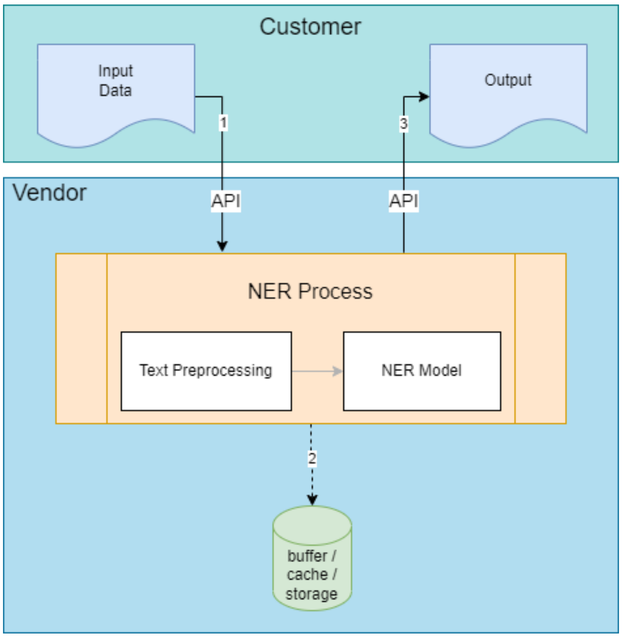
In the following examples we show three possible scenarios based on the different available service models:

1. A customer uses a NER third party solution hosted in the cloud[10]

[8] An Application Programming Interface is a way for two or more computer programs to communicate with each other (source: Wikipedia).
[9] https://learn.microsoft.com/en-us/azure/ai-services/language-service/custom-named-entity-recognition/overview
[10] Example of business solution offered by SAP with information about how to set up the API, the training data and the management of the models (see pages 22, 26, 28, 46 and 60): https://help.sap.com/doc/ca0dcc31534b42f69118b02d68237027/SHIP/en-US/2667f0700ab2436aaec4d9e7bddb2549.pdf

2. A customer uses a NER third party solution hosted on premises[11]
3. A user develops an own NER system

1. <u>Example of data flow diagram when using a third party NER system hosted in the cloud</u>



Step 1: The input data (source text) are transmitted digitally via an API from the customer's device(s) to the server(s) where the NER process will take place by the vendor in the cloud.
In most NER systems, all the steps part of the NER are integrated into a single pipeline that operates on the same server. However, the specific implementation could vary based on the architecture and design choices of the NER solution. It is important to understand where the different processes take place and if there are integrations with other parties to be able to assess possible data protection risks.

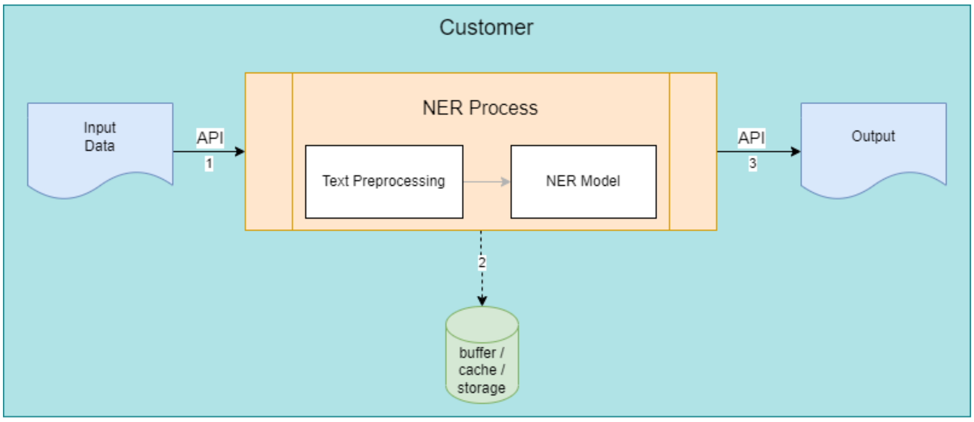
Step 2: The input and output data can be temporarily stored locally at the vendor's location in the cloud. The most common storage options are the following:
1. The data could be stored in a buffer only during the execution of the NER process. The vendor does not retain any data once it has sent the output to the customer.
2. The data could also be temporarily stored in cache[12] to be reused by other immediate processes. The data retention period is variable and depends on the cache memory capacity and configuration.
3. Another possible scenario though less common, is the storage of data in a persistent[13] storage layer such as a database or a cloud storage. This could be done for the analysis or processing of the data at a later stage.

[11] Example of NER on-premises third party solution: https://www.private-ai.com/ner
[12] Caching is usually implemented at a software level to reduce the computational overhead of reprocessing the same text or data and improve overall performance.
[13] Persistent storage means that the data remains intact even when the system is powered off or restarted.

The longer the data is stored in a vendor's system, the higher the risk of a data breach, unlawful repurpose or an infringement of the data storage limitation principle. In this specific case, number 1 (buffer) is the option with less risks since data is stored only during the process in memory. In option 2 (cache) though also with a low risk, data is usually stored for a longer period than in buffer and this can happen outside the process and even on a different location[14]. Option number 3 (storage location like a file or a database for instance) is the one with the highest risks since the storage can take place for a longer period.

Step 3: Once the NER process has finalized, the output data (recognized text elements, categories, their location in the source text and in some cases extra metadata) are sent back via an API to the customer.
In some cases, the input and/or the output data could be used by the vendor to retrain and fine-tune the NER model. Though this is usually done after informing the customers and obtaining their consent, it is important to verify it with the vendor.

2. <u>Example of data flow diagram when using third party NER systems hosted on premises</u>



Step 1: All data transfers and NER process take place internally at the customer's premises on their own servers within their data centers.

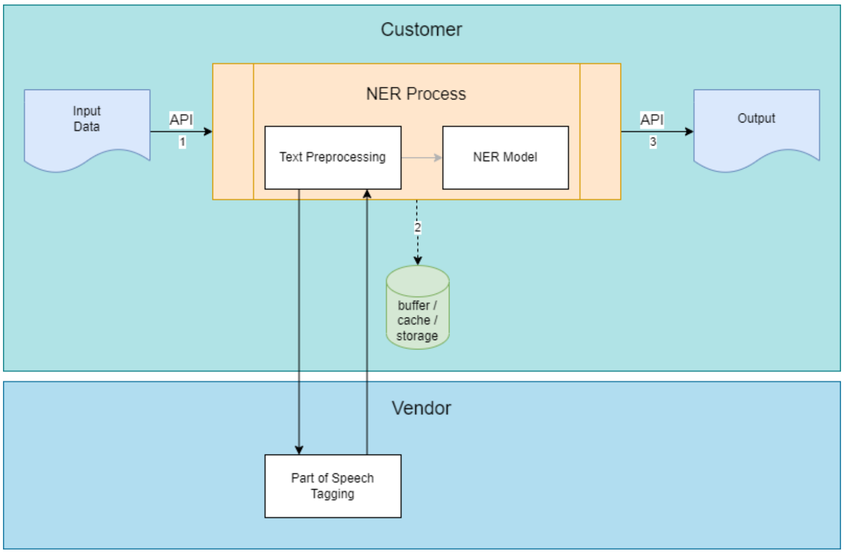
Step 2: The input and output data can be temporarily stored locally. The data could be stored in a buffer only during the execution of the NER process or could be temporarily stored in cache to be reused by other immediate processes.
The input and/or output data could also be stored in a storage location at the customer's premises. This could be done for analysis or processing of the data at a later stage or for auditing purposes.

Step 3: Once the NER process has finalized, the output data (recognized text elements, categories, their location in the source text and in some cases extra metadata) is produced.

[14] The location and method of caching can depend on the specific requirements of the NER system and the architecture of the application. For instance, caching could happen in a different location when using a distributed microservices architecture or a cloud-based caching.

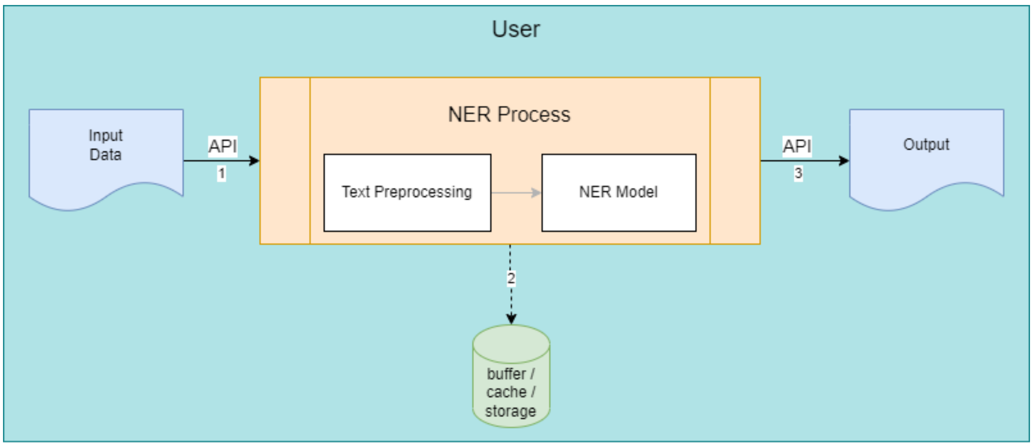
The input and/or output data could also be used to retrain and fine-tune the NER model that is also stored at the user's premises.

A self-hosted NER system from a third party provider can be set up in different ways depending on the architecture and design choices. The specific details can vary if the choice is a completely on premises set up or a hybrid one[15]in which some of the processes are still hosted at the vendor side. For instance, some of the steps in the text preprocessing phase could happen outside the customers premises (see image below).



It's important to review the documentation and architecture of the third party NER system to understand its data flows and whether there are data transfers to the vendor or other third-parties. Additionally, it is also important to assess the system's compatibility with the user's infrastructure, the security as well as any potential requited change in network or firewall configurations.

3. <u>Example of data flow diagram when using a self-developed NER system</u>



[15] Different steps from the preprocessing phase could take place in the cloud. This could be a decision due various reasons such as resource-intensive tasks, redundancy, expertise, etc.

The data flow in this scenario is similar to the one of example 2. In this use case, there are no data transfers to any NER vendor, and all the processes are executed on premises. All processes take place internally.

Performance measure[16] in NER systems: Precision, recall and F1 score

The most common metrics for assessing NER performance are precision, recall, and F1 score.

- *Precision*[17]: it assesses the performance of the model by measuring the ratio of correctly identified positive predictions (true positives) to the total number of positive predictions. It provides insight into the correctness of the predicted entities.
- *Recall*[18] or sensitivity: it evaluates the model's capability to identify the actual positive instances by measuring the ratio of correctly identified true positives to the total number of actual positive instances. It indicates how effectively the model captures the entities present in the data.
- *F1- score*[19]: it considers the harmonic mean of precision and recall, providing an overall assessment of the model's performance. It ranges from 0 to 1 where a higher F1 score (1) denotes a better quality classifier.

It's important to note that evaluation results can vary based on the quality of the training data, the complexity of the text, the presence of domain-specific or rare entities, and the specific criteria set for entity matching.

Limitations of the precision, recall and F1-score metrics:
In the field of NLP, there have been ongoing discussions regarding the adequacy of using the F1 score as a performance metric. When assessing NER models, it is typical to measure performance at the level of individual tokens and this simple schema ignores the possibility of partial matches or other scenarios. Since named entities can consist of multiple tokens, it could be more valuable to evaluate performance based on complete named entities rather than individual tokens.[20]

Accuracy[21] is another measure that calculates the ratio of corrected identified instances (both entities and non-entities) out of the total instances. Although frequently employed, accuracy can provide a misleading picture, particularly in situations involving imbalanced datasets where the count of non-entities considerably surpasses that of entities. That is one of the reasons why when measuring the performance of NER models, F1-score is often preferred over accuracy. NER uses often imbalanced data or unevenly distributed data. This means that there are usually more regular words (non-entities) than specific names, locations, or organizations (entities) in the text.

Currently state of the art (SOTA) performance reach for unsupervised model BERT[22] is 86% (F1-score) depending on the task assigned, while for supervised models we can find in the latest benchmark a

[16] https://paperswithcode.com/task/named-entity-recognition-ner/latest
[17] From the total of selected as true positive how many are really correct. Precision= True Positive / (True Positive + False Positive)
[18] From the total of correct answers how many where selected as true positive. Recall= True Positive / (True Positive + False Negative)
[19] F1-score = 2 * (Precision * Recall) / (Precision + Recall)
[20] https://github.com/MantisAI/nervaluate
[21] Accuracy = (True Positives + True Negatives) / Total Instances
[22] Turchin A, Masharsky S, Zitnik M "Comparison of BERT implementations for natural language processing of narrative medical documents", 2023

performance result of over 90% (F1-score) for different models[23]. For models using human-annotated datasets the performance is also above the 90%[24](F1-score).

<u>Issues that can affect the performance of a NER model:</u>

- Lexical ambiguities, spelling variations, uncommon named entities.
- The quality[25] of the input data being analyzed.
- Insufficient datasets to train the NER model.
- Contexts' ambiguity could difficult the correct identification and categorization of entities. This may be remedied by adding such examples to the training set, or by incorporating pattern matching algorithms[26] to the NER approach.
- Annotator bias can influence the model performance.[27]Though still under research, this could be the case if data is wrongly labeled or not enough entities are represented, which can lead to favoring certain results [28]

Common uses of NER technologies

NER techniques are currently being applied in a variety of use cases. Here are some examples:

- Legal text analysis[29]: to extract legal terms, case citations, and other relevant entities from legal documents, streamlining document review processes for lawyers and law firms.
- Research papers[30]: NER helps organize large amounts of research papers and scholarly articles in a well-structured manner.
- Human resources CV filtering[31]: NER expedites the hiring process by automatically filtering resumes and identifying candidates with specific skills.
- Healthcare[32]: NER tools extract critical information from lab reports and electronic health records, enabling faster data analysis and improved healthcare delivery.
- Classifying content for news providers[33]: NER helps news and publishing houses categorize articles by automatically identifying names, organizations, and places mentioned in them. This facilitates content organization and enables efficient content discovery.

---

[23] https://paperswithcode.com/sota/named-entity-recognition-ner-on-conll-2003; http://nlpprogress.com/english/named_entity_recognition.html
[24] In this paper the performance reached is 92.2%: Jackson M. Steinkamp, Wasif Bala, Abhinav Sharma, Jacob J. Kantrowitz, "Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes", 2020
[25] Low quality can be caused by a lack of annotation accuracy and consistency, data completeness, domain relevance, and if text is not free from errors, typos, and inconsistencies.
[26] These algorithms can be more sophisticated than simple regular expressions matching, being capable of finding contextual cues, relationships between words, and more complex patterns.
[27] Geva, Goldberg, Berant, 2019 "Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets";
[28] Shubhanshu Mishra, Sijun He, and Luca Belli. 2020. Assessing Demographic Bias in Named Entity Recognition. In Proceedings of the Bias in Automatic Knowledge Graph Construction - A Workshop at AKBC 2020, June 24, 2020, Virtual. ACM, New York, NY, USA, 12 pages.
[29] Example: https://www.lexisnexis.com/en-us/home.page
[30] Example: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7302801/
[31] Example: https://inda.ai/en/cv-parsing/
[32] Example: https://github.com/flairNLP/flair
[33] Example: https://metatext.io/blog/named-entity-recognition-for-news-articles

- In search engines[34]: to speed up the search process by associating relevant entities (tags) with each article.
- Content recommendation[35]: NER is used by news publishers to recommend similar articles based on extracted entities.
- Customer support[36]: NER can categorize customer feedback based on mentioned locations and products, allowing complaints to be assigned to relevant departments within the organization.
- Chatbots[37]: to identify relevant entities in user queries to generate appropriate responses.
- E-commerce[38]: NER extracts product names, brand names, and other relevant entities from customer reviews and descriptions, providing insights into customer preferences and improving product offerings.
- Content moderation[39]: NER, along with other NLP techniques, can be used to moderate harmful text content, recognizing terrorist propaganda, hate speech, harassment, and fake news.
- Finance[40]: NER extracts figures and names from documents such as loans and financial reports, and identifies names and companies mentioned in social media, assisting in analyzing profitability, credit risk, and monitoring market trends.

2. Data protection and privacy risk identification

Definition of the criteria to consider when identifying risks and their categorization

To help identify risks associated to the use of data extraction technologies like NER we can make use of a variety of risk factors.

Risk factors are conditions associated with a higher probability of undesirable outcomes. They can help to identify, assess, and prioritize potential risks. For instance, using health data and processing large volumes of data are risks factors with a high level of risk. Acknowledging them in your own use case, can help you identify related potential risks and their severity. In this case, an example of associated risk with a high severity could be 'a risk of violation of patients privacy due to a data breach'.

[34] Example: https://cloud.google.com/natural-language/docs/analyzing-entities; https://www.unleash.so/a/blog/how-nlp-makes-semantic-search-more-intuitive-and-accurate
[35] Example: https://www.recombee.com/?gclid=CjwKCAjw-vmkBhBMEiwAlrMeFxN7BdyJz7q8eOUm5i1tOcuMRK3Mmy7jKqMr-DmFF_YfzpZIBGa7KBoCheoQAvD_BwE; https://docs.developers.optimizely.com/recommendations/docs/content-recommendations
[36] Example: https://kanoki.org/2019/02/21/named-entity-recognition-how-to-automate-customer-support/
[37] Example: https://www.infobip.com/glossary/ner
[38] Example: https://unbxd.com/blog/ecommerce-internal-site-search-engine-importance
[39] Example: https://openai.com/blog/new-and-improved-content-moderation-tooling
[40] Example: https://github.com/hroptatyr/finner

The risk factors shown below are the result of analysing the contents of legal instruments such as the GDPR[41], the EUDPR[42], the EU Charter[43] and other applicable guidelines related to privacy and data protection[44].
The following risk factors can help us identify data protection and privacy high level risks in data extraction technologies like NER:

High level Risk / Important concerns	**Examples of applicability**
Sensitive & impactful purpose of the processing Using a NER system to decide on or prevent the exercise of fundamental rights of individuals, or about their access to a service, the execution or performance of a contract, or access to financial services is a concern, especially if these decisions will be automated without human intervention. Wrong decisions could have an adverse impact on individuals.	- When using NER in the health sector and the named entities are extracted from medical records and the output is used to take decisions related to the health and life of individuals - When NER is used in recruitment for filtering CV's - When NER is used in the context of court filings or social media for instance, and the output obtained is used to make decisions about the data subject. This could be the case in content moderation.
Processing sensitive data When the NER system is processing sensitive data such as: health data, special categories of data, personal data related to convictions and criminal offences, financial data, behavioural data, unique identifiers, location data, etc. This is a reason of concern since processing inappropriately this personal data could negatively impact individuals.	- When using NER for medical records or legal documents by courts, in banking sector, when used in customer support for the analysis of behavioural data, etc.
Large scale processing Processing high volumes of personal data is a reason of concern, especially if these personal data are sensitive. The higher the volume the bigger the impact in case of a data breach or any other situation that put the individuals at risk.	This could apply to most of NER use cases since data extraction technologies like NER are usually applied to large volumes of data.
Processing data of vulnerable individuals This is a concern because vulnerable individuals often require special protection. Processing their personal data without proper safeguards can lead to violations of their fundamental rights. Some examples of vulnerable individuals are children, elderly people, people with mental illness, disabled, patients, people at risk of social exclusion, asylum seekers, persons who access social services, employees, etc.	This could be the case when NER solutions are used in the health sector, at schools, social services organizations, government institutions, employers, etc.
Low data quality The low data quality of the input data and/or the training data is a concern bringing possible risks of inaccuracies in the generated output what could cause wrong identification of named entities and have other adverse impacts depending on the use case.	NER systems are not 100% accurate and there are not high quality datasets available for all domains and languages.
Insufficient security measures The lack of sufficient safeguards could be the cause of a data breach. Data could also be transferred to countries without an adequate level of protection.	- This could be the case if there are not sufficient safeguards implemented to protect the input data and the results of the processing. This could be applicable to any use case. - Data extraction technologies like NER are often offered as SaaS solutions which imply transfers to data processors or third parties.

[41] General Data Protection Regulation (2016/679)
[42] European Union Data Protection Regulation (Reg. 2018/1725)
[43] Charter of Fundamental Rights of the European Union (2012/C 326/02)
[44] Pag. 79, AEPD, "Risk Management and Impact Assessment in Processing of Personal Data", 2021

Presentation of examples of risks specific to NER

Technologies for data extraction like NER can present different types of privacy and data protection risks. The number and type of risks will depend on the use case, the context in which the technology is being applied as well as the different risks factors previously identified.
We are going to analyze different risks related to the procurement, development and use of this technology.

Data protection and privacy risks posed by the **procurement** of those types of AI systems:
NER solutions are frequently available as SaaS solution from third party providers. Due to the different type of configurations available and the required maintenance of the models used, the use of an external supplier is usually the preferred option for users of this technology.
Some third party NER systems, though rarely, can also be hosted on-premises.

Data Protection and Privacy Risks	**Risk description**	**GDPR Potential Impact**	**Examples**	**Risk applicable on service model provision**
Insufficient protection of personal data what eventually can be the cause of a data breach	Safeguards for the protection of personal data are not implemented or are insufficient	Infringement of Art. 32 Security of processing, Art. 5 (f) Integrity and confidentiality and Art. 9 Processing of special categories of personal data	NER systems that process text containing personal data could be not properly secured. This could be the case if for instance, transmission of data is not secure, data are not stored encrypted or with an adequate access control mechanism. It is important to verify that vendors have the necessary safeguards implemented.	• SaaS cloud • On-premises
Possible adverse impact on data subjects that could negatively impact fundamental rights	The output of the system could have an adverse impact on the individual if erroneous data are used for important decisions	Infringement of Art. 5 (d) Accuracy, Art. 5(a) Fairness, Art. 22 Automated individual decision-making, including profiling, Art. 25 Data protection by design and by default	A system providing output that is not accurate and does not provide with mechanisms to amend errors. Or when vendors claim their system offers certain performance, but this is not reproduced in real cases.	• SaaS cloud • On-premises
Lack of compliance with GDPR by not granting data subjects their right to data rectification and erasure	Data subjects' requests to access, rectify or to erase personal data cannot be completed	Infringement of Art. 16 and Art. 17: Right to rectification and right to erasure	A low-quality output could prevent a controller from finding all the data of a data subject in their data storage since the data cannot be matched properly. This could also be the case if there is not a possibility to search for the data subject's data in the output and to correct and delete data.	• SaaS cloud • On-premises
Unlawful repurpose of personal data	Personal data extracted is used for a different purpose	Infringement of Art. 5 (b) Purpose limitation, Art. 5(a) Lawfulness, fairness and transparency, Art. 29 Processing under the authority of the controller or processor	This could be the case if the supplier uses the input and/or output data for training the ML models without this being formally agreed on beforehand.	• SaaS cloud

Excessive storage of personal data	Input data and/or data extracted from images is being stored longer than necessary	Infringement of Art. 5 (e) Storage limitation	The system could be unnecessarily storing input data that is not directly relevant to the NER process. In some cases, the output could be stored by the vendor longer than necessary.	• SaaS cloud • On-premises
Unlawful transfer of personal data	Data are being processed in countries without an adequate level of protection	Infringement of Art. 44 General principle for transfers, Art. 45 Transfers on the basis of an adequacy decision, Art. 46 Transfers subject to appropriate safeguards	NER solutions could store and be processing the data in countries that do not offer enough safeguards.	• SaaS cloud

Data protection and privacy risks posed by the **development** of those types of AI systems:

The development of data extraction technologies can also face data protection and privacy risks. Risks could arise at different phases of the development life cycle, that is why it is important to implement an iterative process for the identification of this type of risks.

The development of a NER system typically involves training machine learning models on large datasets of annotated text. These datasets are manually labeled to identify and classify named entities within the text. The data used for training a NER system typically includes:

- Training Data: text documents or corpora that are annotated with labeled named entities. The annotations indicate the boundaries of the named entities (such as person names, organization names, locations, etc.) and their corresponding entity types (e.g. person, organization, location).
- Validation Data: a separate portion of the dataset is reserved for validation purposes during the model development process.
- Test Data: the other portion of the dataset that is used to evaluate the final performance of the trained NER system. The test data (as well as the training data) should be representative of the real-world scenarios and provide a fair assessment of the system's accuracy and reliability.

NER system developers often curate or collect their own datasets, which can include publicly available data, proprietary data, or datasets obtained through partnerships or collaborations. It is important to mention that training data can introduce certain risks in the development of NER systems. Here are a few key considerations:

- Biases in the annotation process can lead to imbalanced representation of entity types and introduce unfairness in entity recognition.
- Inaccurate or incomplete annotations can negatively impact the performance and generalization capabilities of the NER system.
- Inadequate coverage of entity types, languages, or domain-specific variations in the training data may limit the system's performance when encountering new or rare entities.
- Handling sensitive or private information in the training data requires proper precautions to ensure compliance with privacy regulations and ethical considerations.

The following table offers an overview of data protection and privacy risks that developers of NER systems should consider during the design and development phase. The idea behind this table is to make developers conscious of privacy by design choices that can help prevent risks:

Data Protection and Privacy Risks	Risk description	GDPR Potential Impact	Examples
Insufficient protection of personal data what eventually can be the cause of a data breach	Safeguards for the protection of personal data that is part of the training dataset are not implemented or are insufficient	Infringement of Art. 32 Security of processing, Art. 5 (f) Integrity and confidentiality and Art. 9 Processing of special categories of personal data	We could be using third party libraries, SDK[45]or applications for the development of the NER system, and we could be leaking data to these third parties. The system could be integrated with other systems internally and the transmission of input data could be insecure; data could be stored unencrypted and with inadequate access control mechanism. If using the cloud, this could be not configured according to security best practices.
Possible adverse impact on data subjects that could negatively impact fundamental rights	The output of the system could have an adverse impact on the individual if important decisions over individuals need to be taken based on incorrect data or if the data is not available.	Infringement of Art. 5 (d) Accuracy, Art. 5(a) Fairness, Art. 22 Automated individual decision-making, including profiling, Art. 25 Data protection by design and by default	We could have insufficient datasets to train the NER models, training data could be biased due to misrepresentation in the dataset of certain named entities or due to wrong annotations, preventing entities for being (correctly) identified and affecting the quality of the process.
Excessive storage of personal data	Input data and/or data extracted from the text are being stored longer than necessary	Infringement of Art. 5 (e) Storage limitation	This could be the case if training datasets containing personal data are stored for too long. But it could also be the case if the system is developed in a way where input and output data are automatically stored without offering the user the possibility for deletion.
Breach of the data minimization principle	Extensive processing of personal data for training the model	Infringement of Art. 5 (c) Data minimization	NER systems usually require large amounts of data to train the models.

<u>Data protection and privacy risks posed by the **use** of those types of AI systems:</u>

Users of NER technologies need to consider the risks related to their specific use cases and context. Making use of the risk factors or evaluation criteria can facilitate the identification of those risks. For instance, the criteria 'large-scale processing of personal data' can already trigger the identification of risky processing activities that could result in harm.

When using a NER solution, users have three different service model provisions available: SaaS solution from third party providers hosted in the cloud, third party solutions hosted on-premises and self-developed own solutions hosted on-premises.

[45] SDK stands for software development kit. SDK is a set of software-building tools for a specific platform.

Data Protection and Privacy Risks	**Risk description**	**GDPR Potential Impact**	**Examples**	**Risk applicable on service model provision**
Insufficient protection of personal data what eventually can be the cause of a data breach	Safeguards for the protection of personal data are not implement or are insufficient	Infringement of Art. 32 Security of processing, Art. 5 (f) Integrity and confidentiality, and Art. 9 Processing of special categories of personal data	NER systems that process text containing personal data could be not properly secured. This could be the case if transmission of input data is not secure, data are not stored encrypted or with an adequate access control mechanism. This is especially sensitive when we are processing special category of personal data like when using NER in medical records or legal documents containing criminal data.	• SaaS cloud • Third party on-premises • Self-developed
Possible adverse impact on data subjects that could negatively impact fundamental rights	The output of the system could have an adverse impact on the individual if erroneous data are used for important decisions.	Infringement of Art. 5 (d) Accuracy, Art. 5(a) Fairness, Art. 22 Automated individual decision-making, including profiling, Art. 25 Data protection by design and by default	Errors in the output could attribute incorrectly actions to an individual or group. This could have especially a big impact when using NER for the identification of sensitive words in medical records, legal documents and in content moderation systems.	• SaaS cloud • Third party on-premises • Self-developed
Possible adverse impact on data subjects and lack of compliance with GDPR requirement of providing human intervention for processing that can have a legal or important effect on the data subject	Data subjects are subjected to an automatic decision-making process without human intervention, and/or there is a processing of special categories of personal data	Infringement of Art. 22 Automated individual decision-making, including profiling, Art. 9 Processing of special categories of personal data	The output of a NER system could be used to make automatic decisions which produce legal effects or similarly significant effects on data subjects, this could be the case when using NER in the banking sector for identity verification, also in content moderation or when data are extracted from legal contracts or financial statements and is then analyzed automatically to identify non-compliant clauses, suspicious activities, or anomalies, triggering appropriate actions or alerts.	• SaaS cloud • Third party on-premises • Self-developed
Lack of compliance with GDPR by not granting data subjects their right to data rectification and erasure	Data subjects' requests to rectify or to erase personal data cannot be completed	Infringement of Art. 16 and Art. 17 Right to rectification and right to erasure	This could be the case If there is not a possibility to search for the data subject's data in the output and to correct and delete it.	• SaaS cloud • Third party on-premises • Self-developed
Excessive storage of personal data	Input data and/or output data are being stored longer than necessary	Infringement of Art. 5 (e) Storage limitation	In principle a NER system doesn't need to store the input data unless it is necessary for audit, verification, or archival purposes. The system should avoid unnecessary retention or storage of input data that is not directly relevant to the NER process. The storage of a NER system's output depends on the specific application and requirements. In some cases, the output could	• SaaS cloud • Third party on-premises • Self-developed

			be stored by the vendor longer than necessary. But this could also be the case on-premises by the user, if we are not applying data retention rules to our stored data.	
Breach of the data minimization principle	Extensive processing of personal data for training the model	Infringement of Art. 5 (c) Data minimisation	The amount of data required to train NER models can vary depending on the complexity of the task, the specific domain or language, and the desired level of performance.	• SaaS cloud • Third party on-premises • Self-developed
Unlawful transfer of personal data	Data are being processed in countries without an adequate level of protection	Infringement of Art. 44 , General principle for transfers, Art. 45 Transfers on the basis of an adequacy decision, Art.46 Transfers subject to appropriate safeguards	NER systems could store the data and be processing the data in countries that do not offer enough safeguards. This could also be the case with self-developed systems if we store the data in the cloud.	• SaaS cloud • Self-developed (if using Cloud)

3. Data protection and privacy risk assessment

Once risks have been identified, it is time to proceed with their classification. The actual risk level or risk classification will depend on the specific use case and context.

The GDPR outlines in Recital 90 the importance of establishing the context: "taking into account the nature, scope, context and purposes of the processing and the sources of the risk".
This is an important process when performing a privacy risk assessment to manage risks to the rights and freedoms of natural persons.
The following processes are[46]:
- assessing the likelihood and severity of the risks;
- treating the risks by mitigating the identified risks and in that way ensuring the protection of personal data and demonstrating compliance with the GDPR and EUDPR.

There are different risk management methodologies available to classify and assess risks. It is not the purpose of this document to define or establish a methodology to be used since this is a choice that should be left to each organization. But for the purpose of this document, we will use the international standards that have been previously referenced in the WP29[47] and the AEPD[48] Guidelines.

In general risk management terms, risk can be summarized in one equation:

Risk = Likelihood x Severity

[46] Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, Article 29 Data Protection Working Party, Last revision 2017
[47] ISO 31000:2009, Risk management — Principles and guidelines, International Organization for Standardization (ISO); ISO/IEC 29134, Information technology – Security techniques – Privacy impact assessment – Guidelines, International Organization for Standardization (ISO).
[48] ISO 31010:2019, Risk management — Risk Assessment Techniques, International Organization for Standardization (ISO)

This means that risk is the probability of an event occurring, multiplied by the potential impact or severity incurred by the event.

To assess the level of risk of the data protection and privacy risks identified when procuring, developing and using NER technologies, we first need to estimate the likelihood and severity of the identified risks happening.

3.1. Criteria to establish the likelihood of NER risks. How to assess likelihood.

To determine the likelihood of NER technologies we are using the following four level risk classification matrix:

Level of Consequence	**Likelihood Definition**
Very High	High likelihood of an event occurring
High	Substantial probability of an event occurring
Low	Low probability of an event occurring
Unlikely	There is no evidence of such a risk materializing in any case

Likelihood can only be determined based on specific risks and use cases. We will look later at a specific example to better understand how this process works.

3.2. Criteria to establish the severity of NER risks. How to assess severity.

To determine the severity of NER technologies we are using the four level risk classification matrix[49]:

Level of Severity	**Severity Definition**
Very Significant	It affects the exercise of fundamental rights and public freedoms, and its consequences are irreversible and/or the consequences are related to special categories of data or to criminal offences and are irreversible and/or it causes significant social harm, such as discrimination, and is irreversible and/or it affects particularly vulnerable data subjects, especially children, in an irreversible way and/or causes significant and irreversible moral or material losses.
Significant	The above cases when the effects are reversible and/or there is loss of control of the data subject over their personal data, where the extent of the data are high in relation to the categories of data or the number of subjects and/or identity theft of data subjects occurs or may occur and/or significant financial losses to data subjects may occur and/or loss of confidentiality of data subject or breach of the duty of confidentiality and/or there is a social detriment to data subjects or certain groups of data subjects
Limited	Very limited loss of control of some personal data and to specific data subjects, other than special category or irreversible criminal offences or convictions and/or negligible and irreversible financial losses and/or loss of confidentiality of data subject to professional secrecy but not special categories or infringement penalties
Very Limited	In the above case (limited), when all effects are reversible

The severity criteria are related to a loss of privacy that is experienced by the data subject but that may have further related consequences impacting other individuals and/or society.

[49] Pag. 77, AEPD, "Risk Management and Impact Assessment in Processing of Personal Data", 2021

3.3. Example of NER specific risk assessments

Use case: NER system for the analysis of legal documents

Scenario: We want to analyze legal documents from criminal cases to extract named entities such us name and location. The documents contain sensitive personal criminal information. We do not have the expertise to develop and host ourselves a NER system, so we are going to contract a third-party provider offering a SaaS solution in the cloud.

The following risk factors/ important concerns from section 2. could be applicable in our specific use case:

Risk factor	Use case applicability
Sensitive & impactful purpose of the processing	We are going to take decisions based on the output that could have an impact on rights or freedoms of individuals
Processing sensitive data	Personal data related to convictions and criminal offences
Large scale processing	The volume of data to be processed is high
Processing data of vulnerable individuals	Criminal records
Low data quality	We do not know if the dataset is of sufficient quality
Insufficient security measures	We might transfer personal data to states or organisations in other countries without an adequate level of protection. There could be a possibility of a data breach. Third party vendor solution has not been chosen yet; this must be taken into account when making a choice.

Based on the identified risks factors we are going to identify together with other stakeholders[50] the data protection and privacy risks that could arise with the NER implementation.

We are going to use as foundation the risks identified in section 2 for procurement, and we are going to assess what is the likelihood of the identified risks and assign to each risk one of the 4 likelihood classification levels from the matrix: Very high, High, Low, Unlikely.

Data Protection and Privacy Risks	Risk factor	Risk description	Likelihood	Reasoning
Insufficient protection of personal data what eventually can be the cause of a data breach	- Insufficient security measures - Processing sensitive data - Large scale processing - Processing data of vulnerable individuals	Safeguards for the protection of personal data are not implemented or are insufficient	Low	The third-party suppliers we have reviewed, have implemented security measures such as secure transmissions, strong access control measures, and data encryption at rest. We, as user/customer have also strong security processes implemented internally.
Possible adverse impact on data subjects that could negatively impact fundamental rights	- Low data quality - Sensitive & impactful purpose of the processing	The output of the system could have an adverse impact on the individual if	High	The probability of obtaining inaccurate data is present in NER systems. This probability is

[50] Meaningful involvement of different stakeholders during the risk assessment process: https://ecnl.org/sites/default/files/2023-03/Final%20Version%20FME%20with%20Copyright%20%282%29.pdf

	- Processing data of vulnerable individuals	erroneous data are used for important decisions.		higher if NER is combined with OCR[51]. Misidentification could have an important adverse impact on the data subject.
Lack of compliance with GDPR by not granting data subjects their right to data rectification and erasure	- Processing sensitive data - Processing data of vulnerable individuals - Sensitive & impactful purpose of the processing	Data subjects' requests to rectify or to erase personal data cannot be completed	Low	There is a very small possibility that not all data has been properly extracted and cannot be found during the search function. The system offers the possibility to delete data from the output since output is available in formats that are modifiable.
Unlawful repurpose of personal data	- Processing sensitive data - Processing data of vulnerable individuals - Sensitive & impactful purpose of the processing - Insufficient security measures	Personal data extracted is used for a different purpose	Low *This risk will be unlikely in an on-premises solution	Except in a case of unlawful processing by the third party provider, in principle is the probability of the data extraction results being used to retrain and fine-tune the NER model low. Most SaaS solutions delete the data or offer you the possibility to decide if you want to share the results for that purpose.
Excessive storage of personal data	- Processing sensitive data - Processing data of vulnerable individuals - Sensitive & impactful purpose of the processing - Insufficient security measures	Input data and/or data extracted from text are being stored longer than necessary	Low	For our use case we could consider for instance, vendors that offer a 24 hours automated deletion period what already reduces the likelihood of this risk. Vendors usually offer a 24 or max. 48 hours automated deletion period in SaaS solutions. Self-developed NER systems can be configured in a way that input and output can be immediately deleted or at a scheduled moment.
Unlawful transfer of personal data	- Insufficient security measures - Processing sensitive data - Processing data of vulnerable individuals	Data are being processed in countries without an adequate level of protection	Low	In our specific use case, we have decided to work with vendors that offer an adequate level of protection what already reduces the likelihood though not the impact.

After the likelihood assessment, we are going to assess what is the impact of the identified risks on the data subjects, individuals and society. Based on that impact/severity assessment, we will assign one of the 4 severity classification levels: Very significant, Significant, Limited, Very limited.

Data Protection and Privacy Risks	**Risk description**	**Likelihood**	**Severity**	**Reasoning**
Insufficient protection of personal data what eventually can be the cause of a data breach	Safeguards for the protection of personal data are not implemented or are insufficient	Low	Very significant	The documents contain very sensitive information, and a data breach could cause significant harm to the data subjects.
Possible adverse impact on data subjects that could negatively impact fundamental rights	The output of the system could have an adverse impact on the individual if erroneous data are used for important decisions	High	Very significant	An inaccurate output or not identifying all named entities on a document could have adverse consequences for instance in trails and criminal investigations.

[51] "An Analysis of the Performance of Named Entity Recognition over OCRed Documents", Hamdi, Jean-Caurant, Sidere, Coustaty, Doucet, 2022

Lack of compliance with GDPR by not granting data subjects their right to data rectification and erasure	Data subjects' requests to rectify or to erase personal data cannot be completed	Low	Significant	Not being able to rectify incorrect or not up to date information could have a significant impact on the data subject due to the nature of the data being processed.
Unlawful repurpose of personal data	Personal data extracted is used for a different purpose	Low *This risk will be unlikely in an on-premises solution	Very significant	This could have a big impact on the data subjects if for instance the vendor keeps a copy of the input and/or the output data and uses this afterwards for non-agreed purposes, especially due to the nature of the personal data contained in the documents.
Excessive storage of personal data	Input data and/or data extracted from images are being stored longer than necessary	Low	Significant	An unlimited or unlawful storage of personal data would worsen any data breach affecting stored data. And although data being properly protected while being stored would limit the harm caused to the data subject, it will still be an infringement of the GDPR.
Unlawful transfer of personal data	Data are being processed in countries without an adequate level of protection	Low	Significant	Transferring the data to a country that doesn't offer enough safeguards could bring significant risks to the data subjects.

4. Data protection and privacy risk treatment

Risk treatment criteria

i.e., mitigate, transfer, avoid or accept a risk.

The assessments of likelihood and severity will offer us the basis to obtain the risk level classification. Based on the four level classification used for likelihood and severity, we can use a matrix like the following to obtain the resulting final risk level classification: Very High, High, Medium, Low.

Likelihood	**Very High**	Medium	High	Very high	Very high
	High	Low	High	Very high	Very high
	Low	Low	Medium	High	Very high
	Unlikely	Low	Low	Medium	Very high
		Very limited	**Limited**	**Significant**	**Very Significant**
		Severity			

Based on this matrix we can classify the risks identified in our use case as follows:

Data Protection and Privacy Risks	**Risk description**	**Likelihood**	**Severity**	**Risk Level**
Insufficient protection of personal data what eventually can be the cause of a data breach	Safeguards for the protection of personal data are not implemented or are insufficient	Low	Very significant	Very High
Possible adverse impact on data subjects that could negatively impact fundamental rights	The output of the system could have an adverse impact on the individual if erroneous data are used for important decisions.	High	Very significant	Very High

Lack of compliance with GDPR by not granting data subjects their right to data rectification and erasure	Data subjects' requests to rectify or to erase personal data cannot be completed	Low	Significant	High
Unlawful repurpose of personal data	Personal data extracted is used for a different purpose	Low *This risk will be unlikely in an on-premises solution	Very significant	Very High
Excessive storage of personal data	Input data and/or data extracted from images are being stored longer than necessary	Low	Significant	High
Unlawful transfer of personal data	Data are being processed in countries without an adequate level of protection	Low	Significant	High

We have identified three risks with a very high level, and three with a high level. Best practices in risk management suggest that the mitigation of very high and high level risks should be prioritized.[52] The next step involves the implementation of a risk treatment plan.

Risk treatment involves developing options for mitigating the risks and preparing and implementing action plans. The appropriate treatment option should be chosen on a contextual basis and considering a feasibility analysis[53] like the following:

- Evaluate the type of risk and the available mitigation measures that can be implemented.
- Compare the potential benefits gained from implementing the mitigation against the costs and efforts involved.
- Assess the impact on the purpose that is being pursued by implementing the NER system.
- Evaluate what could be the reasonable expectations of individuals.
- Assess the impact mitigation measures could have on transparency and fairness of the processing.

An analysis of these criteria is essential to risk mitigation and risk management planning and helps in determining whether the risk mitigation is justifiable.

The most common risk treatment criteria are: **Mitigate, Transfer, Avoid** and **Accept**.
For each risk one of the criteria options will be selected:

- ✓ Mitigate – Identify ways to reduce the likelihood or the severity of the risk.
- ✓ Transfer – Make another party responsible for the risk (buy insurance, outsourcing, etc.).
- ✓ Avoid – Eliminate the risk by eliminating the cause.
- ✓ Accept – Nothing will be done.

Deciding whether a risk can be mitigated involves assessing the nature of the risk, understanding its potential impact, and evaluating potential mitigation measures such as implementing controls, adopting best practices, modifying processes, and using tools that can help reduce the likelihood or severity of the risk.

[52] https://www.pmi.org/learning/library/high-risk-critical-path-projects-7675
[53] "Risk, High Risk, Risk Assessments and Data Protection Impact Assessments under the GDPR", CIPL GDPR Interpretation and Implementation Project, 2016

Not all risks can be fully mitigated. Some risks may be inherent and cannot be entirely avoided. In such cases, the goal is to reduce the risk to an acceptable level or to put in place measures that help manage the severity of the risk effectively.

Presentation of mitigation measure examples/risk treatment options

including an assessment of their practical feasibility and a definition of the criteria to define the level of mitigation obtained.

In our use case we have identified several very high and high level risks. After going through the feasibility analysis and the treatment criteria, we have decided that we cannot **transfer** the risks to any third party, we cannot **avoid** all the risks, and **acceptance** of the risks is an unacceptable option for us. As long as there are measures that we can implement to help us mitigate the risks, resulting in acceptable conditions to go on with the implementation, we choose the treatment option of risk **mitigation**.

We have identified the following risk mitigation measures:

Data Protection and Privacy Risks	**Risk Level**	**Risk Mitigation measures**	**Feasibility Assessment**	**New risk Level after mitigation**
Insufficient protection of personal data what eventually can be the cause of a data breach: safeguards for the protection of personal data are not implemented or are insufficient	Very High	The third-party vendor chosen must have implemented security measures such as secure transmissions, strong access control measures, and data encryption in transit and at rest and sufficient privacy design strategies[54] to protect the data. We will ask certifications[55] and results of a pentest[56] to the vendor. As controller we can also protect the specific sensitive data in the documents by applying pseudonymization or anonymization techniques after the data extraction. Depending on the different needs, we could implement default anonymization or reversible data masking [57]techniques, for instance allowing access to the unmasked data to certain people. If the NER SaaS solution does not offer the possibility to implement these techniques, we could decide to look for a vendor that can offer them or implement them ourselves.	1. Cost of implementation: The implementation of pseudonymization or anonymization techniques after the data extraction would imply additional cost. 2. Impact on purpose of identification of named identities: No 3. Impact on expectations of individuals: No 4. Impact on transparency and fairness of the processing: No	Low
Possible adverse impact on data subjects that could	Very High	If we plan to use the output of the NER model to make decisions about individuals, we will need to make sure that the NER	1. Cost of implementation: They highest cost is the effort that implies doing	Medium

[54] "Privacy Design Strategies" Jaap-Henk Hoepman, 2022
[55] Security certifications such as ISO27001, and SOC2
[56] A penetration test, colloquially known as a pentest or ethical hacking, is an authorized simulated cyberattack on a computer system, performed to evaluate the security of the system.
[57] Vendors like Private Ai and Pangeanic offer the possibility to mask sensitive data. Protected entities can be identified by a NER system and be immediately masked. https://docs.private-ai.com/what-is-privategpt/; https://blog.pangeanic.com/discover-ner-model-data-anonymization

negatively impact fundamental rights: the output of the system could have an adverse impact on the individual if erroneous data are used for important decisions.		system we use offers a high performance guarantee with a F1 score closer to 1[58]. The NER system should have been trained with representative data of our domain and context to avoid bias and risk of misidentifications. Vendors should provide information about the dataset source, the quality of the annotation and the guidelines used for the annotation process. Vendors should also confirm if the dataset includes entities relevant to our required domain. If possible, we should examine a sample of annotations in the dataset and verify if the labeled entities are accurate and correctly categorized based on our context. We can use text analysis tools[59] to analyze the frequency of different words and entities. This can reveal patterns and biases in the data. If the data that we process is domain specific, we might decide to choose the option of using a custom model that we can train with our own data. This will also contribute to an increase in the accuracy. Another important aspect is that currently there are no systems offering 100% accuracy, and the only way to achieve that and avoid any error is by doing a human review and correction of the output.	the human review when needed and providing the human resources for that. Vendors providing custom model solutions and other features could be more expensive. 2. Impact on purpose of identification of named identities: No 3. Impact on expectations of individuals: No for data subjects, but it has an impact on the employees that would be in charge of the human review 4. Impact on transparency and fairness of the processing: No if it is properly implemented and information about how the accuracy of the system works is provided to users and eventually to data subjects.	
Unlawful repurpose of personal data: personal data extracted is used for a different purpose	Very High	Option 1:One of the best mitigation measure would be using a SaaS solution that offers the option to keep the data on premises. This is the case if the data processing takes place at location and the input is automatically deleted and the output data is only stored at the user location. Option 2: If the data extraction takes place at the vendor's location, then a minimum of security measures such as access control, audit trail and encryption together with proper data protection agreements need to be in place.	1. Cost of implementation: If the on-premise option is offered by the vendor this could have an additional cost. It could also imply that we need to make resources available for taking care of the on-premise solution. 2. Impact on purpose of identification of named identities: No 3. Impact on expectations of individuals: No 4. Impact on transparency and fairness of the processing: No	Option 1:Low Option 2:Medium
Lack of compliance with GDPR by not granting data subjects their right to data rectification and erasure: Data subjects' requests to rectify or	High	We could implement a human verification step for at least the category of entities related to personal data. Once the check is done, the extracted data could be stored in searchable format to be ready for data subject access requests. Editable format access could be granted if necessary to answer update and	1. Cost of implementation: This requirements will bring extra implementation costs not only financially but also in terms of human resources.	Low

[58] NER systems based on supervised machine learning for English texts can offer F-scores close to human score that is around 94% (Zhou and Su 2002)
[59] https://www.nltk.org/, https://spacy.io/, https://www.eweek.com/artificial-intelligence/text-analysis-tools/#comparison, https://github.com/neulab/InterpretEval

to erase personal data cannot be completed		corrections requests. This access could be granted to only a controlled number of people and we could also implement access audit trail.[60] Besides access restrictions, we could implement privacy by design strategies to protect the data.	2. Impact on purpose of identification of named identities: No 3. Impact on expectations of individuals: No 4. Impact on transparency and fairness of the processing: No	
Unlawful transfer of personal data: Data are being processed in countries without an adequate level of protection	High	We can implement a NER system from a third-party provider that is located in a country offering an adequate level of protection.	1. Cost of implementation: This measure should in principle not have any additional cost, but it depends on the vendors availability. 2. Impact on purpose of identification of named identities: No 3. Impact on expectations of individual: No 4. Impact on transparency and fairness of the processing: No	Low
Excessive storage of personal data: Input and output data are being stored longer than necessary	High	We could try to implement a NER system in which deletion of data can be configured[61] so that input and output data are deleted from the system immediately after the data extraction or at a scheduled moment (this is by most vendors a period of 24 to 48 hours). We could also implement a retention period for the output data that is already in our own premises. If the vendor is storing the input and/or output data, we should negotiate contractual agreements about the retention period of the data.	1. Cost of implementation: If the option is offered by the vendor this could have an additional cost. 2. Impact on purpose of identification of named identities: No 3. Impact on expectations of individuals: No 4. Impact on transparency and fairness of the processing: No	Low

Residual risk acceptance

After the feasibility assessment has been done and the mitigation measures have been identified and implemented, we should assess again the likelihood and severity of each risk to obtain a new risk classification level and in this way assess if there is any remaining or residual risk.

In our use case, after the assessment, all the risks have been reduced to the lowest risk level ‘low’, and there are two risks with a possible classification of ‘Medium’ depending in the example on the mitigation measure adopted.

[60] Chronological record or log of activities, events, or actions taken within a system, application, or organization.
[61] https://learn.microsoft.com/en-us/legal/cognitive-services/language-service/cner-data-privacy-security

We calculate the residual risk by evaluating the likelihood and severity of the risks that still exists despite the implemented mitigation measures. This residual risk represents the level of risk that remains after taking mitigation actions.
Once residual risk has been identified, we need to decide whether the residual risk is within acceptable levels for our organization. If it is, we can decide to accept it. If it's not, we would need to consider further mitigation strategies.
Some organizations establish criteria for acceptability of residual risks based on elements such us social norms, benefits, harms, similar use cases, etc[62].
Organizations must be able to justify their risk mitigation and acceptance decisions as part of their accountability obligations which also fall under the GDPR principle of accountability (Article 5.2, Recital 74).

Example of general mitigation measures related to risks of NER systems

Choosing appropriate risks mitigation measures should be done on a case-by-case basis. We are going to examine some of the possible mitigation measures that could be implemented to mitigate privacy and data protection risks specific for NER technologies. These measures are general and not related to any specific use case.

Data Protection and Privacy Risks	**Mitigation measures examples**
Insufficient protection of personal data what eventually can be the cause of a data breach	As **user**, **procurement** entity and **developer**, it is important to verify[63] that APIs are securely implemented, transmission of data is protected with the adequate encryption protocols, data at rest are encrypted, there is an adequate access control mechanism implemented, there are measures implemented for protection and identification of insider threats, measures to mitigate supply chain attacks that could give access to the training data and/or the data storage and encryption keys, measures implemented to prevent risks associated to the use of deep learning such as the risk of reprogramming deep neural net attacks[64], membership inference[65], inversion[66] and poisoning attacks[67]. Also access and change logs should be implemented to document who and when has access to the data.
Possible adverse impact on data subjects that could negatively impact fundamental rights	As **user**, implement NER solutions that offer a high performance rate closer to the last available benchmark[68] (an average of 90% / 0.9 F1). The acceptable F1 score may differ based on factors like the types of entities being recognized, the available datasets, the system's intended use, and the potential consequences of false positives and false negatives. Sometimes the systems offer the results of this metric after every data processing. It is important to monitor the values and make the necessary adjustments and corrections to the results. Make sure the system recognizes different conditions applicable to the input data. The quality of input data is important. This is important for **users** of NER systems as well as for **developers** that need to use training data of quality to train their models. As **developer** there are techniques and tools[69] you can use to reduce the low quality of input data.

[62] https://www.sciencedirect.com/topics/engineering/residual-risk
[63] This could be done by performing a pentest and/or requesting pentest results to the vendor.
[64] “Adversarial reprogramming of neural networks”, Elsayed et al, 2018
[65] “Membership Inference Attacks Against Machine Learning Models”, Shokri et al, 2017
[66] “Generative Model-Inversion Attacks Against Deep Neural Networks”, Zhang et al, 2020
[67] “Practical Poisoning Attacks on Neural Networks”, Junfeng Guo, Cong Liu, 2020
[68] http://nlpprogress.com/english/named_entity_recognition.html
[69] See next section under ‘Tools and techniques’

	Context ambiguity is one of the causes of incorrect output, one possible solution is to add more examples of the problematic named entity to the training set or incorporate pattern recognition matching learning algorithms to the process.[70]
Possible adverse impact on data subjects and lack of compliance with GDPR requirement of providing human intervention for processing that can have a legal or important effect on the data subject	**Users** could implement a human review process to verify the correctness of the personal data, especially if this is sensitive, and a process to approve high risk decisions after human verification has been done. It is also important that the system provides with an overview of the accuracy levels achieved after the data extraction with a dashboard or any other type of interface for manual human review and correction. In certain use cases it might be necessary to implement a redress mechanism for data subjects.
Excessive storage of personal data	As **user** and **procurement** entity, make agreements with the third-party supplier about how long the input data and output data should be stored. This can be part of the service contract, product documentation or data processing agreement. If data are being stored on your premises, establish retention rules and /or a mechanism for the deletion of data.
Breach of the data minimization principle	For **users** and **developers**, one possible way to mitigate this risk is by providing documents to the NER model where personal data has been replaced by synthetic data. As user, it is also important to compare the different NER solutions available on the market to understand which systems require less volume of data to train the models and to improve the accuracy levels.
Unlawful transfer of personal data	As **user** and **procurement** entity, verify with the vendor where the data processing is taking place. Make the necessary safeguard diligences and when necessary, perform a Data Transfer Impact Assessment and make the necessary contractual agreements. Consider this risk when making a selection among different vendors.

Once risk mitigation measures have been implemented, it is crucial to continuously monitor their effectiveness. Implementing methodologies like threat modeling for the identification of risks, maintaining a risk register and assigning risk owners are effective strategies for regularly reviewing and reassessing the risk landscape. This ensures that the implemented risk mitigation measures remain relevant and effective in preventing data protection and privacy risks that could adversely impact individuals and organizations.

[70] "Evaluation of Named Entity Recognition in Dutch online criminal complaints", Schraagen, Brinkhuis and Bex, 2017

Reference to specific technologies, tools, methodologies, processes or strategies.

Unless standardised and freely and easily accessible, explanation on how these technologies, tools, methodologies and processes work.

Methodologies for measuring accuracy in NER:[71]

Throughout the years different NER forums have proposed different evaluation metrics:

- CoNLL: Computational Natural Language Learning

The Language-Independent Named Entity Recognition task introduced at CoNLL-2003 measures for the performance in terms of precision, recall and f1-score.

- Automatic Content Extraction (ACE)

The ACE proposes a more complex evaluation metric which includes a weighting schema:
"Automatic Content Extraction 2008 Evaluation Plan (ACE08)"
"The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation"

- Message Understanding Conference (MUC)

MUC proposes detailed metrics considering different categories of errors:
https://github.com/jantrienes/nereval

- International Workshop on Semantic Evaluation (SemEval)

The SemEval'13 proposes four different ways to measure precision/recall/f1-score results based on the metrics defined by MUC.

NER quality standard:

There are currently not specific industry standards developed for NER[72].

Tools and techniques:

- NLTK Natural Language Toolkit is an open-source platform for building Python programs and provides libraries, tools, and resources to facilitate the development of NLP applications:
  - https://www.nltk.org/
- SpaCy is an open-source NLP library for Python. It provides tools and capabilities for various NLP tasks, and it integrates with modern machine learning techniques.
  - https://spacy.io/
- Standford NER, also known as CFRClassifier as part of CoreNLP toolkit, is another open-source natural language processing tool developed by Stanford University. It uses conditional random fields (CRF) as one of its underlying algorithms for NER. It's widely used for research, education, and practical applications in the NLP field
  - https://stanfordnlp.github.io/CoreNLP/ner.html
- CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources.
  - https://www.clarin.eu/resource-families/tools-named-entity-recognition
- Huggingface Bert-base-NER is a fine-tuned BERT model that is ready to use for Named Entity Recognition and achieves state-of-the-art performance for the NER task. It has been trained to recognize four types of entities: location (LOC), organizations (ORG), person (PER) and Miscellaneous (MISC).

[71] https://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/
[72] "Is NLP Ready for Standardization?" Lauriane Aufrant, Inria, Findings of the Association for Computational Linguistics: EMNLP 2022
https://aclanthology.org/2022.findings-emnlp.202.pdf

  - https://huggingface.co/dslim/bert-base-NER
- OpeNER is a project funded by the European Commission under the FP7 (7th Framework Program). Its acronym stands for Open Polarity Enhanced Name Entity Recognition. It provides a set of ready to use tools to perform NLP tasks, free and easy to adapt for Academia, Research and Small and Medium Enterprise to integrate them in their workflow.
  - https://www.opener-project.eu/

Privacy preserving techniques and tools:

These two vendors provide services to mask sensitive data identified with NER systems:

- Self-hosted general-purpose NER in over 30 languages:
  - https://www.private-ai.com/ner/
- Pangeanic has developed a multilingual anonymization kit based on NER applicable to all EU languages.
  - https://pangeanic.com/use-cases/mapa

General resources:

- Evaluation method to interpret the differences in NER models and datasets, as well as the interplay between them, identifying the strengths and weaknesses of current systems:
  - https://virtual.2020.emnlp.org/paper_main.3648.html
  - https://github.com/neulab/InterpretEval
- Scorer is a tool from Spacy to compute evaluation metrics such as precision, recall, and F1 score for different NLP tasks:
  - https://spacy.io/api/scorer
- Microsoft evaluation metrics for custom named entity recognition models:
  - https://learn.microsoft.com/en-us/azure/cognitive-services/language-service/custom-named-entity-recognition/concepts/evaluation-metrics
- This research paper proposes a generalized evaluation methodology to interpret model biases, dataset biases, and their correlation:
  - https://openreview.net/pdf?id=HJxTgeBtDr
- Research paper about the use of NER for the detection of bias:
  - “Dbias: detecting biases and ensuring fairness in news articles”, Shaina Raza, Deepak John Reji & Chen Ding, 2022.

Methodologies and tools for the identification of data protection and privacy risks:

- ❖ Privacy Library of Threats (PLOT4ai) is a threat modeling methodology for the identification of risks in AI systems. It also contains a library with more than 80 risks specific to AI systems: https://plot4.ai/
- ❖ MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems), is a knowledge base of adversary tactics, techniques, and case studies for machine learning (ML) systems: https://atlas.mitre.org/
- ❖ Assessment List for Trustworthy Artificial Intelligence (ALTAI) is a checklist that guides developers and deployers of AI systems in implementing trustworthy AI principles: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

edpb

European Data Protection Board