A Privacy-Preserving Corpus for Occupational Health in Spanish: Evaluation for NER and Classification Tasks

Claudio Aracena[1,5], **Luis Miranda**[2,5], **Thomas Vakili**[3], **Fabián Villena**[4,5], **Tamara Quiroga**[2,5], **Fredy Núñez-Torres**[6], **Victor Rocco**[7], and **Jocelyn Dunstan**[2,5]
[1]Faculty of Physical and Mathematical Sciences, University of Chile
[2]Department of Computer Science, Pontifical Catholic University of Chile
[3]Department of Computer and Systems Sciences, Stockholm University
[4]Department of Computer Science, University of Chile
[5]Millennium Institute Foundational Research on Data (IMFD), Chile
[6]Department of Language Science, Pontifical Catholic University of Chile
[7]Chilean Safety Association (ACHS), Chile
claudio.aracena@uchile.cl, lmirandn@uc.cl, thomas.vakili@dsv.su.se, fvillena@imfd.cl, t.quiroga@uc.cl, frnunez@uc.cl, varoccoc@achs.cl, jdunstan@uc.cl

Abstract

Annotated corpora are essential to reliable natural language processing. While they are expensive to create, they are essential for building and evaluating systems. This study introduces a new corpus of 2,869 medical and admission reports collected by an occupational insurance and health provider. The corpus has been carefully annotated for personally identifiable information (PII) and is shared, masking this information. Two annotators adhered to annotation guidelines during the annotation process, and a referee later resolved annotation conflicts in a consolidation process to build a gold standard subcorpus. The inter-annotator agreement values, measured in F_1, range between 0.86 and 0.93 depending on the selected subcorpus. The value of the corpus is demonstrated by evaluating its use for NER of PII and a classification task. The evaluations find that fine-tuned models and GPT-3.5 reach F_1 of 0.911 and 0.720 in NER of PII, respectively. In the case of the insurance coverage classification task, using the original or de-identified corpus results in similar performance. The annotated data are released in de-identified form.

1 Introduction

Text plays a relevant role in healthcare since it is one of the richest forms of information inside electronic health records (Dalianis, 2018). Therefore, developing tools for processing and analyzing clinical text is an important goal of clinical natural language processing (NLP). However, one of the challenges when processing clinical text is the appearance of PII, such as names, locations, and identification numbers. To develop tools that can help the clinical community process text, researchers and developers need to access clinical text in a privacy-preserving manner for the patients involved. Otherwise, patients' rights are being violated.

A common way to share clinical text without violating patients' rights is to publish a de-identified version of a clinical corpus. Some of the most known clinical datasets are the MIMIC (Multi-parameter Intelligent Monitoring for Intensive Care) databases (Moody and Mark; Saeed et al., 2011; Johnson et al., 2016, 2023). These databases contain not just clinical text from critical care units but also the whole structure and data from their databases.

The previously described datasets are uncommon in languages other than English (Névéol et al., 2018). In particular, for Spanish, few clinical annotated corpus have been released. Some examples are: CANTEMIST (Miranda-Escalada et al., 2020), an annotated corpus of oncology reports; CT-EBM-SP (Campillos-Llanos et al., 2021), an annotated corpus of clinical trials; NUBes (Lima Lopez et al., 2020) an annotated corpus with negation and uncertainty entities in anonymized health records; and the Chilean waiting list corpus (Báez et al., 2020; Báez et al., 2022), an annotated corpus of referrals for the Chilean waiting list.

This work presents a corpus for occupational health in Spanish. Occupational health is an area of work in public health to promote and maintain the highest degree of physical, mental, and social well-being of workers in all occupations (World Health Organization, 2023). Occupational insurance and health providers collect patient data

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 111–121

whenever patients face a work-related accident or disease. This data is used to deliver better treatment and to decide if an occupational insurer will cover a patient.

The corpus presented in this work is similar to MEDDOPROF (Lima-López et al., 2021), an annotated corpus of occupations in clinical texts in Spanish. However, MEDDOPROF focuses on annotating only occupations, while our corpus also annotates PII. In that sense, our corpus is comparable to MEDDOCAN (Marimon et al., 2019), one of the few freely available clinical datasets for PII identification in Spanish. But there are two main differences, MEDDOCAN is a synthetic corpus, while our corpus uses actual data, and PII are masked.

Our team holds an agreement with one of the biggest occupational insurance and health providers in Chile, giving us access to their data for research purposes. Hence, the corpus introduced in this paper is a clinical corpus containing information that must be protected. The annotation procedures outlined in Section 3 describe the precautions taken to minimize the privacy risks described in Section 2. Later, Section 4 reports the experiments run with the original and de-identified corpus, including NER and classification tasks, and Section 5 shows the results and discussion of the experiments. Finally, Section 6 states the main conclusion and future work that can be done.

The main contributions of this work are:

- Publicly available pseudonymous corpus of 2,869 medical and admission reports.
- Performance comparison between fine-tuning in existing synthetic clinical corpus and our corpus for NER of PII.
- Performance comparison of a downstream task between fine-tuning in our corpus with and without PII.

2 Related Research

2.1 Privacy in NLP

With the dominance of data-driven approaches to NLP, state-of-the-art results are attained by relying on large corpora. This tendency has been further compounded with the introduction of transformer models. It is not uncommon to read about models trained using many gigabytes of textual data. However, datasets of that scale are too large to be manually audited. This means they typically contain large amounts of PII, which is a privacy risk. The parameter sizes of modern transformer models compound this risk by providing ample opportunity for training data to be memorized.

The risks of memorization in transformer models have been demonstrated through mounting attacks on pre-trained language models. Carlini et al. (2021) demonstrated that it was possible to extract memorized sequences of PII from the model GPT-2 (Brown et al., 2020). These kinds of training data extraction attacks have been repeated for other models as well, with varying success (Huang et al., 2022). Other researchers have focused on determining whether models are susceptible to membership inference attacks. These attacks are less ambitious, aiming to determine if a given datapoint was used to train a model. Such attacks have been successful even when targeting models for which training data extraction has failed (Lehman et al., 2021; Vakili and Dalianis, 2021), as demonstrated by Mireshghallah et al. (2022).

Although privacy in the context of language models is difficult to measure (Vakili and Dalianis, 2023) or even define (Brown et al., 2022), any risk of training data leakage threatens privacy. While privacy is a right that should always be protected, it is an especially pertinent value when dealing with data from sensitive sources, as is often the case in the clinical domain.

2.2 De-Identification

One way of reducing the privacy risks of using sensitive corpora for training is by de-identifying the data. This entails finding sensitive spans of texts and sanitizing them. When corpora are large, this can be done through automated means. Automatic de-identification is a process that typically relies on NER models to detect sensitive entities and then handle them in various ways. Automatic de-identification has been shown to decrease privacy risks while preserving the utility of the data both for fine-tuning and pre-training purposes (Verkijk and Vossen, 2022; Vakili et al., 2023). However, the impact on utility may vary depending on the task, the sanitization strategy and the quality of the underlying NER model (Berg et al., 2020; Lothritz et al., 2023). Crucially, a well-performing automatic de-identifier needs a high-quality PII dataset to train a sufficiently powerful NER model.

MEDDOCAN (Marimon et al., 2019) is one of

few freely available clinical dataset for PII identification in Spanish. The corpus comprises 1,000 synthetic documents describing fictional patients and is annotated for a wide range of PII. It was created for a shared task in which several systems were able to attain impressive F_1 scores reaching over 0.96. However, the documents were synthetically created for the shared task. A consequence of this is that the documents have certain artifacts that may make classification easier, but that may be absent in data encountered elsewhere. For example, MEDDOCAN documents always begin with a structured list of PII describing the patient. These include the patient's name, address, and the date of their imagined visit.

As with MEDDOCAN, the corpora annotated for PII typically originate from one or a few sources. This means that there is a risk that the models trained using the data overfit to peculiarities found in the specific datasets. Thus, it is not always clear that the NER models will generalize and be as effective at detecting sensitive information in data from other institutions unseen during training. Previous studies (Yang et al., 2019; Bridal et al., 2022) have found that performance may decrease when using data from new sources and that mismatches in annotation guidelines may make results difficult to interpret. Cross-institutional evaluations are challenging because of legal and ethical barriers to data sharing. In this paper, we not only perform such an evaluation but make our data available to other researchers interested in evaluating the cross-institutional validity of their systems. Furthermore, the data are carefully de-identified and audited by humans, meaning there is a high degree of confidence that the data are safe to share.

3 Corpus

In Chile, occupational insurance and health providers actively address work-related health problems during commuting or within the workplace. The core of this procedure involves creating a document known as an admission report, in which an administrative employee compiles a narrative summary of the events surrounding the incident. After this process, a medical report called anamnesis is generated. This new document is a clinical report where healthcare professionals register the clinical details of the affected patient and the specifics of the problem from a medical perspective.

In this work, we compiled a dataset of 3,000 work-related accidents. Typically, each case includes both a medical report and an admission report; however, there are instances where only one of the reports is available. As a result, we constructed an annotated corpus consisting of 2,869 documents, divided into 1,383 medical reports (anamnesis) and 1,486 admission reports. These documents are presented in a free-text format, enabling a rich and diverse range of textual content (it is noteworthy that many contain PII).

Table 1 provides a detailed analysis of corpus statistics, differentiating between medical and admission reports, and drawing a comparison between our comprehensive annotated corpus and the MEDDOCAN annotated corpus. While the MEDDOCAN corpus comprises of a smaller number of documents, it contains over twice the number of tokens and more than three times the quantity of entities compared to our dataset. This disparity can be attributed to the synthetic nature of the MEDDOCAN corpus, intentionally designed to incorporate a substantial volume of PII. Nevertheless, as outlined in Section 2.2, it is important to note that MEDDOCAN is a semi-structured corpus, which is reflected in its comparatively lower lexical diversity in contrast to our corpus.

Conversely, within our dataset, we noted that the admission report typically demonstrates a more pronounced structural organization than to the medical report. This results in a reduced lexical variety, as illustrated in Table 1.

3.1 Annotation Procedure

The annotation process consisted of three distinct stages. We developed a preliminary version of the annotation guidelines in the initial stage by thoroughly reviewing existing guidelines and studies about NER in Spanish or NER of PII (Dalianis and Velupillai, 2010; Báez et al., 2020; Marimon et al., 2019). We also integrated insights from the Health Insurance Portability and Accountability Act (HIPAA) (Office of the Federal Register, National Archives and Records Administration, 1996), a U.S. law defining 18 personal identifiers in Clinical Health Records.

In the second stage, one annotator annotated the entire corpus. This task included continually refining the annotation guidelines by examining encountered scenarios and ongoing discussions.

In the third and final stage, armed with well-

Metric	Total	Med.	Adm.	MEDDOCAN
Documents	2,869	1,383	1,486	1,000
Tokens	243,537	125,404	118,147	508,340
Vocabulary	18,261	14,483	6,018	19,699
Lexical diversity	7.5%	11.5%	5.1%	3.8 %
Tok. per doc.	85± 37	91± 53	79±18	508 ± 47
Ent. per doc.	2.1 ± 1.7	2.3±1.8	1.8±1.7	32.5±1.9
Annotated tokens	8,447	5,194	3,253	42,254
Entities	5,895	3,152	2,743	22,795

Table 1: Corpus statistics divided by medical and admission reports and comparison with MEDDOCAN.

consolidated annotation guidelines, a second annotator successfully annotated 956 documents within the corpus. This comprised 496 admission reports and 460 medical reports. This phase marked a significant milestone in our annotation process, allowing us to refine further and enhance the quality of our annotated data.

After the annotation process, we implemented a consolidation process to resolve disagreements between the first and second annotators. Each annotation underwent a comprehensive review by a team of three researchers: the two annotators and a referee responsible for making the final decision. This team examined and discussed each annotation, engaging in detailed deliberations to reach a consensus. This review process resulted in the creation of a gold standard dataset comprising 956 documents.

3.2 Annotation Scheme

The annotation scheme for this research exclusively encompasses non-overlapping entities. In other words, each token can have at most one associated entity. After careful consideration, we have utilized 11 entities shown in Table 2.

We incorporated all the entities proposed by Dalianis and Velupillai (2010) plus extra ones described in the next paragraphs. The annotators in this study drew inspiration from HIPAA guidelines to shape these entities, making specific modifications through their discussions. However, concerning the *Location* label, the authors unified the tags for *Country*, *Municipality*, *Street address*, and *Town* into a single category called *Location*. In contrast, we decided to preserve the *Organization* as a distinct entity, which we named *Institution*. The fact that our dataset frequently included institution names influenced this choice, usually related to the institution where the person works but does not necessarily correspond to a location.

Given the frequent occurrence of patient occupation data within our dataset, we introduced the *Occupation* label, as outlined in the MEDDOCAN guidelines (Marimon et al., 2019). The primary motive behind its inclusion is the presence of particular occupations in the procedure annotations, suggesting that individuals could be identified based on their occupation.

Furthermore, due to the [country redacted for anonymity] civil registration origin of the data, we introduced the *Personal ID* label, which denotes a unique identification number allocated to individuals and legal entities for tax and spread use for administrative purposes.

Analyzing the annotated entities in Table 2, we detail the number of entities categorized by their respective entity classes across each subcorpus. Notably, the quantity of entities significantly fluctuates depending on their class. For instance, the most frequently occurring entity is *Full Date*, predominantly present in admission reports, and the second most prevalent entity is *Occupation*, primarily sourced from medical reports. In contrast, the *Phone Number* tag is exceptionally rare, appearing only three times throughout the entire corpus.

Furthermore, Figure 1 illustrates the token frequency distribution for each entity within the subcorpus and the distribution of annotated entities per document. Concerning token frequency, it is noteworthy that distinct subcorpora exhibit varying distributions. Generally, entities consist of a single token, but there are multi-token entities. In the admission report corpus, entities like *Location* and *Health Care Unit* are mostly composed of more than one token. Additionally, in the medical report subcorpus, entities such as *Occupation*, *Institution*, and *Location* are multi-token.

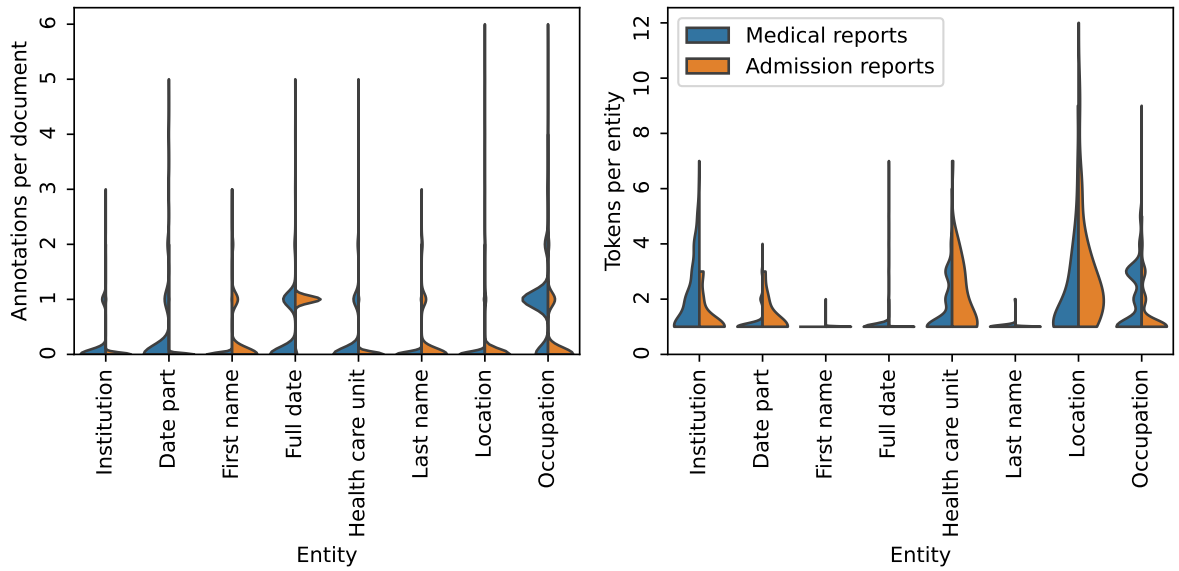

Figure 1: Frequency distribution of (left) annotated entities per document by subcorpus, and (right) tokens per entity across the subcorpus.

Entity	Total	Med.	Adm.
Age	195	194	1
Institution	242	217	25
Health Care Unit	394	348	46
Date Part	485	473	12
Full Date	1981	563	1418
First Name	402	21	381
Last Name	358	53	305
Location	197	63	134
Occupation	1634	1214	420
Phone Number	3	2	1
Personal ID	4	4	0

Table 2: Number of entities by entity class and if they are in the medical or administrative subcorpus.

Conversely, when considering the distribution of annotated entities per document, Figure 1 reveals that, on the whole, documents tend to contain a relatively small number of entities. However, it's worth noting that admission reports, on average, contain one *Full Date* entity per document, while medical reports, on average, feature one *Occupation* entity per document.

3.3 Annotation Guidelines

Three researchers collaboratively drafted a comprehensive document outlining the annotation guidelines: the annotator responsible for the entire corpus, a linguist, and a computer science professor. It resulted from a thorough review of literature (Báez et al., 2020; Marimon et al., 2019; Dalianis and Velupillai, 2010; Office of the Federal Register, National Archives and Records Administration, 1996) and discussions on annotation casuistry, where regular meetings were held to ensure that the guidelines maintained linguistic and syntactic coherence while enhancing privacy protection without undermining the texts' narrative. The current version of the annotation guidelines is freely available[1].

Building upon the framework established by Báez et al. (2020) for annotation guidelines, we categorize the rules into two sections: general rules, which have universal application to all entities, and specific rules customized for each entity. Within the general and specific rule sections, we further distinguish between positive rules (guiding what should be annotated) and negative rules (clearly outlining what should not be tagged or what constitutes an incorrect annotation). Finally, the guidelines provide informative explanations regarding typical scenarios encountered within the dataset.

We elucidated two general rules, refraining from incorporating trailing punctuation marks or white spaces after entities. Furthermore, we emphasized the importance of tagging each entity

[1] https://totoiii.github.io/clinical_deidentification_guideline/

with the utmost specificity to ensure the most comprehensive coverage of the entity.

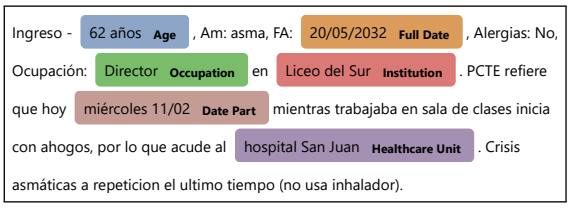

Figure 2: Example of an annotated document where PII has been modified. Translation: *Entry - the 62 years old, PMH: asthma, DOA: 05/20/2032, Allergies: None. Principal at South High School. The patient reports that on Wednesday, 11/02, while working in a classroom, he began experiencing shortness of breath, prompting him to seek care at Saint John Hospital. He has been experiencing recurrent asthma attacks recently and is not using an inhaler.*

Finally, Figure 2 presents an example document with annotations drawn from the existing corpus and modified for explanatory purposes, with all PII appropriately modified.

3.4 Inter-Annotator Agreement

We evaluated the challenge of achieving consistent annotations by assessing inter-annotator agreement (IAA). Specifically, the macro F_1 was employed to assess and compare the annotations. Table 3 depicts the agreements for each comparison within the different subcorpora. These comparisons entail assessments between annotator 1 and annotator 2 and between each annotator and the gold standard corpus.

Furthermore, Figure 3 visualizes the IAA for various entity classes, except for (*Age*, *Phone Number*, and *Personal ID*) that have too few instances. The figure highlights that, in most cases, more favorable agreement results are evident in the admission report as compared to the medical report. This can be attributed to the slightly more structural organization found in the admission report subcorpus when compared to the medical report subcorpus.

3.5 Masking Procedure

A masking process was carried out to share our corpus without private or sensitive information. The masking process adds a mask using the tag `"__entity_name__"` for every entity, where the entity name corresponds to the name of an entity, e.g., First Name.

	Global	Medical	Admission
A1 - A2	0.90	0.86	0.93
GS - A1	0.97	0.94	0.98
GS - A2	0.92	0.90	0.93

Table 3: Macro F_1 agreements for each comparison and subcorpus. Where A1 is annotator 1, A2 is annotator 2, and GS is gold standard corpus.

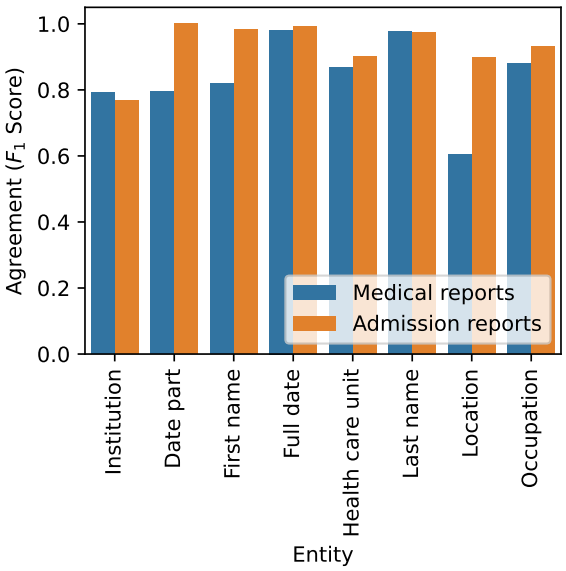

Figure 3: Macro F_1 score for IAA for each entity and subcorpus.

4 Experiments

The three experiments described in this section demonstrated the value of the corpus:

1. GPT-3.5 is used to detect PII and compare it with human annotators.
2. The corpus is used to train and evaluate NER models for privacy-preserving purposes.
3. The corpus is used to perform an insurance coverage classification task.

4.1 NER via Few-Shot In-Context Learning

The utilization and experimentation with Large Language Models (LLMs) in NER tasks hold profound significance in the realm of NLP. These models, equipped with their vast contextual understanding, have the potential to greatly enhance the performance and efficiency of identifying named entities within text.

In this experiment, we employed the gpt-3.5-turbo model (OpenAI, 2023) through Microsoft

Azure to conduct few-shot NER on the entire corpus. The prompt given to the model involved explaining the 11 entities outlined in Section 3.2, along with providing brief descriptions of the content associated with each tag. Additionally, the prompt included the desired output format, which involves annotating the input text using a markup language format, as follows: `...<Entity class>Named entity body</Entity class>...`.

Furthermore, we generated and instructed the model with five distinct examples for each subcorpus. These examples were crafted in alignment with the unique characteristics of their corresponding subcorpus.

4.2 Training and Evaluating NER

As explained in Section 2.2, an important use case for NER is for automatic de-identification of sensitive data. However, assessing the cross-institutional validity of such models is difficult due to data scarcity, which is especially dire in languages other than English. In this experiment, the transferability of performance gained through training models using the MEDDOCAN corpus was evaluated using our new corpus. Models were trained using either MEDDOCAN data or our corpus and then evaluated on the curated gold standard part of the corpus.

A wide range of models for Spanish language modeling were selected. The best base models for Spanish NER suggested by Agerri and Agirre (2023) were fine-tuned for PII detection. The models chosen were the multilingual model mDeBERTaV3 (He et al., 2023) and the monolingual Spanish model IXABERTes-v2[2] that is based on RoBERTa (Liu et al., 2019). Fine-tuned models were created using MEDDOCAN as well as the corpus introduced in this study.

The PII tags defined for MEDDOCAN and our corpus differed in a few ways. Before training the models, the tagsets were harmonized. This involved translating MEDDOCAN tags into their counterparts in our corpus. If a MEDDOCAN tag lacked a counterpart, it was ignored. The procedure also involved collapsing labels found in our corpus that were not distinguished in MEDDOCAN. The distinction between first and last names, and between partial and full dates, was present in our corpus but not in MEDDOCAN. Consequently, they were collapsed into the tags *Name* and *Date*. Both datasets were then converted into the IOB format[3].

After training each model configuration for five epochs, the best checkpoint was selected based on the F_1 score on the validation set. The selected models were evaluated on the gold standard described in Section 3 and a held-out test set from MEDDOCAN. The performance difference when testing models on the unseen dataset indicates how much they generalize to novel data. Our corpus's classification results were compared with those obtained by prompting GPT-3.5.

4.3 Insurance Coverage Classification

The insurance coverage classification task is selected to assess the impact of de-identifying PII on downstream tasks' performance. This task aims to classify the insurance coverage decision of the occupational insurance provider. Following Aracena et al. (2023), the process of building a classifier consists of using the pre-trained model bsc-bio-ehr-es[4] (Carrino et al., 2022) as a base model. Then, a fine-tuning step is carried out, in which the documents from the corpus with their corresponding label for the insurance coverage decision are used for this purpose [5]. Lastly, the fine-tuned model is evaluated in other cases not part of the corpus.

The previous process is implemented for the original and de-identified corpus, and also for the admission subcorpus, the medical subcorpus, and both combined. The de-identification was performed by replacing each sensitive entity with its class name.

5 Results

This section shows the results of the experiments and discusses the implications for NER and downstream tasks.

5.1 NER for De-Identification

Four fine-tuned models were trained based on the pre-trained mDeBERTaV3 and IXABERTes-v2 models. Each model was trained on either the MEDDOCAN data, or our corpus. Table 4 shows

[2] http://www.deeptext.eus/es/node/3

[3] Specifically, we use the version of IOB that reserves *B* for entities spanning multiple tokens.

[4] https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es

[5] The labels for the classification task are not part of the released corpus.

Base model	Training data	Test F_1 score	
		MEDDOCAN	Our corpus
mDeBERTaV3	Our corpus	0.400	0.853
mDeBERTaV3	MEDDOCAN	0.990	0.498
IXABERTes-v2	Our corpus	0.368	0.834
IXABERTes-v2	MEDDOCAN	0.990	0.381
gpt-3.5-turbo	N/A	N/A	0.720

Table 4: F_1 scores for each combination of model, training dataset and testing dataset. gpt-3.5-turbo was not fine-tuned but was accessed through an API and prompted using a few-shot approach targeting the new corpus.

Base model	Fine-tuning data	Test F_1 score	
		Original	De-identified
bsc-bio-ehr-es	Admission	0.726 ± 0.015	0.708 ± 0.004
	Medical	0.738 ± 0.006	0.743 ± 0.008
	Admission+Medical	0.750 ± 0.002	0.763 ± 0.006

Table 5: F_1 scores for classification task in the test set.

the results of evaluating the models on the test version of their training data and the test set of the other dataset. The models perform substantially worse in all four cases when evaluated on novel data. This indicates a clear mismatch between the two datasets, even though the task they represent is ostensibly equivalent.

It is not obvious if the mismatch between our corpus and MEDDOCAN is due to the synthetic nature of MEDDOCAN or stems from an inherent diversity in how PII are represented in the clinical domain. A truly cross-institutional PII tagger for de-identification purposes should perform well on a diverse range of datasets. The new corpus thus functions as a source of training data, and as a benchmark to evaluate the generalizability of NER models trained on other data sources.

Additionally, we show the performance of gpt-3.5-turbo when performing few-shot NER through in-context learning on the new corpus. Even though it does not show the best results, it performs better than the cross-institutional taggers, which is still remarkable considering that just a few examples were given to understand the task. However, similar to a previous experience (Wang et al., 2023), the performance of gpt-3.5-turbo for NER tasks is not state-of-the-art.

The gpt-3.5-turbo outputs sometimes deviated from expectations by altering the original text in various ways. These alterations included fixing misspelled words or introducing punctuation not in the original text. This posed a significant challenge, resulting in misaligning the original annotations with the model-generated ones. To address this issue, we analyzed in detail the disparities between the original text and the model-modified text. We then adjusted the positions of tokens for each annotated entity in the model output, enabling us to make precise comparisons between the annotations.

5.2 Insurance Coverage Classification

Six fine-tuning configurations were used to train models, three with the original corpus and three with the de-identified corpus. Admission subcorpus, medical subcorpus, and both combined were used to fine-tune models in each type of corpus. For every fine-tuning configuration, three random seeds were used to check the variability of the results, and for each run, three epochs were used. Table 5 shows the insurance coverage classification results. The positive class is the decision not to cover a patient, as this is the less frequent class.

Under the described conditions, none to little differences were found between using the original or the de-identified corpus. These results suggest that using a de-identified corpus for downstream tasks should not impact the performance. However, depending on the task under study, this situation may vary. Aracena et al. (2023) reported bet-

ter performance for the same task, reaching 0.963 of AUC. This is due to the amount of data used for fine-tuning, which is more than 200 times bigger than this study.

6 Conclusions

This work introduces a novel corpus of admission and medical reports retrieved from an insurance and health provider specialized in occupational health. The annotation of PII and the subsequent de-identification process highlight the importance of releasing data considering ethical and privacy matters. This corpus is released[6] in de-identified form, where all sensitive entities are replaced with their class names.

Our exploration of the corpus has revealed its inherent value in named entity recognition and classification tasks. The insights gained through these analyses not only contribute to the existing body of knowledge but also hold practical implications for improving information extraction within the specified domain.

As future work, one promising avenue involves exploring synthetic data to replace masked PII entities. This approach has the potential to not only safeguard privacy but also the possibility of building robust models that can be applied in diverse real-world scenarios with synthetic data.

Acknowledgements

This work was funded by ANID Chile: Millennium Science Initiative Program ICN17_002 (IMFD), Basal Funds for Center of Excellence FB210005 (CMM) and FB0008 (AC3E), Fondecyt Regular 1241825 (JD) and National Doctoral Scholarship 21211659 (CA), 21220200 (FV), and 21220586 (TQ), and the 2023 CMM PhD Visiting Scholarship and the DataLEASH project (TV).

References

Rodrigo Agerri and Eneko Agirre. 2023. Lessons learned from the evaluation of Spanish Language Models. *Procesamiento del Lenguaje Natural*, 70(0):157–170. Number: 0.

Claudio Aracena, Nicolás Rodríguez, Victor Rocco, and Jocelyn Dunstan. 2023. Pre-trained language models in Spanish for health insurance coverage. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 433–438, Toronto, Canada. Association for Computational Linguistics.

Pablo Báez, Felipe Bravo-Marquez, Jocelyn Dunstan, Matías Rojas, and Fabián Villena. 2022. Automatic extraction of nested entities in clinical referrals in spanish. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(3):1–22.

Hanna Berg, Aron Henriksson, and Hercules Dalianis. 2020. The Impact of De-identification on Downstream Named Entity Recognition in Clinical Text. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 1–11, Online. Association for Computational Linguistics.

Olle Bridal, Thomas Vakili, and Marina Santini. 2022. Cross-Clinic De-Identification of Swedish Electronic Health Records: Nuances and Caveats. In *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference*, pages 49–52, Marseille, France. European Language Resources Association.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What Does it Mean for a Language Model to Preserve Privacy? In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2280–2292, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Pablo Báez, Fabián Villena, Matías Rojas, Manuel Durán, and Jocelyn Dunstan. 2020. The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 291–300, Online. Association for Computational Linguistics.

Leonardo Campillos-Llanos, Ana Valverde-Mateos, Adrián Capllonch-Carrión, and Antonio Moreno-Sandoval. 2021. A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. *BMC Medical Informatics and Decision Making*, 21(1):69.

[6]The data are available upon request at:https://zenodo.org/records/11035754

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *Proceedings of the 30th USENIX Security Symposium*, pages 2633–2650.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained Biomedical Language Models for Clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland. Association for Computational Linguistics.

Hercules Dalianis. 2018. *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer International Publishing.

Hercules Dalianis and Sumithra Velupillai. 2010. De-identifying swedish clinical text - refinement of a gold standard and experiments with conditional random fields. *Journal of Biomedical Semantics*, 1(1):6.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are Large Pre-Trained Language Models Leaking Your Personal Information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1).

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1).

Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron Wallace. 2021. Does BERT Pretrained on Clinical Notes Reveal Sensitive Data? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 946–959, Online. Association for Computational Linguistics.

Salvador Lima Lopez, Naiara Perez, Montse Cuadros, and German Rigau. 2020. NUBes: A Corpus of Negation and Uncertainty in Spanish Clinical Texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5772–5781, Marseille, France. European Language Resources Association.

Salvador Lima-López, Eulàlia Farré-Maduell, Antonio Miranda-Escalada, Vicent Brivá-Iglesias, and Martin Krallinger. 2021. NLP applied to occupational health: MEDDOPROF shared task at IberLEF 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts. *Procesamiento del Lenguaje Natural*, 67(0):243–256.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Saad Ezzini, Tegawendé Bissyandé, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2023. Evaluating the Impact of Text De-Identification on Downstream NLP Tasks. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 10–16, Tórshavn, Faroe Islands. University of Tartu Library.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic De-identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results. In *IberLEF@ SEPLN*, pages 618–638.

A Miranda-Escalada, E Farré, and M Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the Cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying privacy risks of masked language models using membership inference attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8332–8347, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

G.B. Moody and R.G. Mark. A database to support development and evaluation of intelligent intensive care monitoring. In *Computers in Cardiology 1996*. IEEE.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum.

2018. Clinical Natural Language Processing in languages other than English: opportunities and challenges. *Journal of Biomedical Semantics*, 9(1):12.

Office of the Federal Register, National Archives and Records Administration. 1996. Public law 104 - 191 - health insurance portability and accountability act of 1996.

OpenAI. 2023. Models. https://platform.openai.com/docs/models/gpt-3-5. [Accessed 20-10-2023].

Mohammed Saeed, Mauricio Villarroel, Andrew T. Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H. Kyaw, Benjamin Moody, and Roger G. Mark. 2011. Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine*, 39(5):952–960.

Thomas Vakili and Hercules Dalianis. 2021. Are Clinical BERT Models Privacy Preserving? The Difficulty of Extracting Patient-Condition Associations. In *Proceedings of the AAAI 2021 Fall Symposium on Human Partnership with Medical AI: Design, Operationalization, and Ethics (AAAI-HUMAN 2021)*.

Thomas Vakili and Hercules Dalianis. 2023. Using Membership Inference Attacks to Evaluate Privacy-Preserving Language Modeling Fails for Pseudonymizing Data. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, NEALT Proceedings Series, pages 318–323, Tórshavn, Faroe Islands. University of Tartu Library.

Thomas Vakili, Aron Henriksson, and Hercules Dalianis. 2023. End-to-End Pseudonymization of Fine-Tuned Clinical BERT Models.

Stella Verkijk and Piek Vossen. 2022. Efficiently and Thoroughly Anonymizing a Transformer Language Model for Dutch Electronic Health Records: a Two-Step Method. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1098–1103, Marseille, France. European Language Resources Association.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

World Health Organization. 2023. Occupational health. https://www.who.int/health-topics/occupational-health. [Online; accessed 16-October-2023].

Xi Yang, Tianchen Lyu, Qian Li, Chih-Yin Lee, Jiang Bian, William R. Hogan, and Yonghui Wu. 2019. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, 19(5):232.