Pre-trained language models in Spanish for health insurance coverage

Claudio Aracena[1,2], **Nicolás Rodríguez**[3], **Victor Rocco**[3], **and Jocelyn Dunstan**[2,4,5,6]
[1]Faculty of Physical and Mathematical Sciences, University of Chile
[2]Millennium Institute Foundational Research on Data, Chile
[3]Chilean Safety Association, Chile
[4]Department of Computer Science, Catholic University of Chile
[5]Institute for Mathematical and Computational Engineering, Catholic University of Chile
[6]Center for Mathematical Modeling, University of Chile
claudio.aracena@uchile.cl, {nrodrigueza,varoccoc}@achs.cl, jdunstan@uc.cl

Abstract

The field of clinical natural language processing (NLP) can extract useful information from clinical text. Since 2017, the NLP field has shifted towards using pre-trained language models (PLMs), improving performance in several tasks. Most of the research in this field has focused on English text, but there are some available PLMs in Spanish. In this work, we use clinical PLMs to analyze text from admission and medical reports in Spanish for an insurance and health provider to give a probability of no coverage in a labor insurance process. Our results show that fine-tuning a PLM pre-trained with the provider's data leads to better results, but this process is time-consuming and computationally expensive. At least for this task, fine-tuning publicly available clinical PLM leads to comparable results to a custom PLM, but in less time and with fewer resources. Analyzing large volumes of insurance requests is burdensome for employers, and models can ease this task by pre-classifying reports that are likely not to have coverage. Our approach of entirely using clinical-related text improves the current models while reinforcing the idea of clinical support systems that simplify human labor but do not replace it. To our knowledge, the clinical corpus collected for this study is the largest one reported for the Spanish language.

1 Introduction

Clinical text is one of the most comprehensive data types in electronic health records. Therefore, clinical natural language processing (NLP) has become relevant to extracting helpful information from clinical writing and supporting decision-making. The complexity of human languages makes it difficult to analyze unstructured text. Additionally, the clinical text is complicated because of the heavy use of jargon, unusual spellings, and abbreviations (Dalianis, 2018).

In this complex scenario, there are various tasks that clinical NLP aims to handle. These tasks might be anything from language-related ones like text categorization, relation extraction, and entity extraction to prediction-related ones like estimating patient mortality, length of hospital stay, unplanned readmissions, etc. Several publications have addressed these tasks that have produced specialized models (Dalianis, 2018).

However, since 2017, the NLP field has worked towards creating pre-trained language models (PLMs) that can be fine-tuned for any specific downstream task. These language models are built for a much simpler task, such as next-word or masked-word prediction in a massive amount of text. This process, known as pre-training, allows the language model to acquire language understanding that can be used for any text-related task (Tunstall et al., 2022).

As soon as the NLP field started to work in PLMs, clinical NLP introduced this type of model into its set of techniques to improve performance. Some examples of clinical PLMs are two different versions of ClinicalBERT (Alsentzer et al., 2019; Huang et al., 2020). These models show a significant improvement in language tasks and a moderate improvement in prediction tasks.

Most of the research in clinical NLP has been done for text written in English, but not so much for other languages (Névéol et al., 2018). In Spanish, some publicly available PLMs relevant to clinical NLP are bsc-bio-ehr-es (Carrino et al., 2022) and Spanish Clinical Flair (Rojas et al., 2022). These PLMs were pre-trained heavily in general and biomedical text with minor additions of clinical text. Despite this drawback, they outperform general and biomedical PLMs in language tasks.

In this context, an insurance and health provider aims to analyze their clinical text to apply in a labor insurance coverage process. This provider receives patients who have suffered from a labor-related accident. When a patient is admitted to one of their clinics, admitting staff writes a report detailing the

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 433–438
July 14, 2023 ©2023 Association for Computational Linguistics

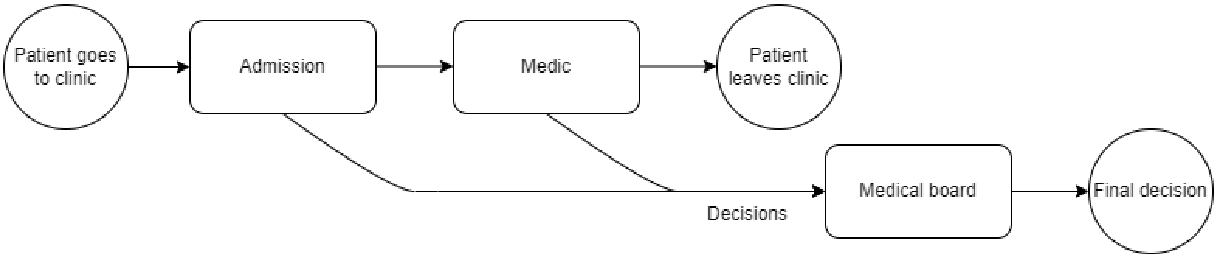

Figure 1: Flow diagram of patients and insurance coverage decisions.

accident. This report includes rich contextual information about what happened in the accident. After admission, a physician checks the patient and writes medical information. Every clinic's medical and administrative board decides a final coverage rating for each accident the next business day, considering both reports.

Currently, the provider has a model that gives a probability of not covering a patient given a specific diagnosis. This model serves as a ranking tool for the medical board to review cases with a high likelihood of no coverage. However, not all diagnoses are included in this model (considered as a categorical variable), and admission and medical reports are not considered to calculate the probability. Additionally, the model can calculate the probability of no coverage just after a physician diagnoses a patient.

This work aims to analyze clinical text from admission and medical reports to give a probability of no coverage. We try three approaches that use clinical PLMs in Spanish to carry out this goal. First, we use a clinical PLM. Second, we do continual pre-training of the previous PLM with text data from admission and medical reports. Finally, we pre-trained a LM from scratch using admission and medical reports. All outcomes will be compared to the current model performance.

2 Problem statement

In Chile, employers must hire an insurance and health provider specialized in labor accidents. These providers should cover all labor accidents. To decide if the insurance will cover a worker, the providers have clinics where they admit and check the workers to make a decision.

In this work, we use data from one of these providers. As this provider is specialized in labor accidents, the data that collects has some features. First, it has a high level of detail because that admission and medical reports are used to justify insurance coverage decisions. Second, many physicians can treat the same patient, requiring information in clinical records to be as complete as possible so that any medical staff can give better continuity to patient care over time. These features make the data cleaner compared to general health provider records.

The Asociación Chilena de Seguridad (ACHS), Chilean Safety Association in English, is a pre-eminent non-profit insurance and health provider. Its principal objective is to conceptualize and administer risk prevention programs alongside providing comprehensive coverage for occupational accidents. As evidence of its influence, ACHS accommodates more than 2.6 million affiliated workers and over 73,000 affiliated employing entities nationwide. Moreover, with a record of the lowest average accident rate, ACHS unequivocally operates as the largest mutual association in Chile.

The stringent regulations under Law No. 16,744 mandate that all Chilean employing entities, regardless of their operational scale, must be affiliated with a Social Security Administration agency. This agency is responsible for safeguarding against the risks of Occupational Accidents and Diseases. As one of three private administrative bodies, ACHS is tasked with formulating risk prevention programs. It also offers health coverage and compensation for occupational accidents, transport mishaps, and professional illnesses.

The type of labor accidents can be of two types, work-related and commuting accidents. Work-related accidents happen at the workplace or as a result of work. Commuting accidents happen on the way to or from work with no stops in between (direct trips). The staff writes an admission report when the patient is admitted in both cases. Later, when a physician receives the patient, a medical report is written.

The medical report is based on three sources. The first includes the patient's anamnesis. The second information is from the physical examination performed on the patient. The third is the medical indication for treatment. Each time a new or old patient passes through this healthcare provider and needs to be seen by a doctor, a new medical report entry is generated.

The admitting staff and the physician give a label (covered or uncovered) classification on the reports. The final classification is made the next business day after the patient is seen by the medic. A board of physicians and administrative heads from each clinic determines a final coverage rating for each case. This final rating takes into account the medical and admitting staff reports.

Most admitting staff's labels will state that patients will be covered, and medics, after clinical examination, have a more robust filter to say whether a patient will be covered or not. The committee of physicians and administrative heads has a reviewer role, and some decisions are finally changed. Figure 1 shows the flow diagram of the described process.

The current model employed by the healthcare provider only makes predictions in 72.1% of cases, basing its predictions on structured diagnoses alone. Unfortunately, this approach results in a lack of predictions for less frequently observed diagnoses. However, the majority of cases come with either admission records or medical reports, making it possible to improve coverage by utilizing these additional resources.

We expect that the use of admission and medical reports can help to take better coverage classification compared to the current model that only uses diagnosis as a categorical variable with the most common ones. Moreover, the classification prediction with the admission report can help the physician consider more information that may have been overlooked.

3 Datasets

For this study, three different types of datasets were built, for fine-tuning, continual pre-training, and pre-training from scratch. Here we list the details of these datasets:

1. For the fine-tuning process, three datasets were created. An admission dataset, which only contains text from admission reports, and a label with the final decision if that case was covered (coverage decision). A medical report, which only contains text from medical reports, and coverage decision labels. Finally, an admission and medical dataset, which concatenate text from admission and medical reports, and coverage decision labels.

2. For the continual pre-training process, also three datasets were created (admission, medical, and admission-medical datasets) similar to the fine-tuning datasets. We do not need a label in this case since these datasets are only used to continue pre-training a pre-existing PLM.

3. For the pre-training process from scratch, only one dataset was created, combining all admission and medical reports available. This dataset does not include a coverage decision label, as it is used for pre-training. However, it is bigger than previous datasets because it is used to pre-trained a PLM from scratch. According to our knowledge, this is the biggest corpus containing only clinical-related text in Spanish.

Table 1 shows details for every dataset.

Datasets	documents	tokens
Fine-tuning		
Admission	300 k	22.5 M
Medical	300 k	26.3 M
Admision+Medical	300 k	57.2 M
Continual Pre-training		
Admission	1.5 M	112.6 M
Medical	1.2 M	154.0 M
Admision+Medical	855 k	164.6 M
Pre-training		
Admision+Medical	7.1 M	1.03 B

Table 1: Number of documents and tokens in every dataset.

4 Methods

This section described the processes of pre-training and fine-tuning using the datasets described in the previous section.

4.1 Fine-tuninig of bsc-bio-ehr-es

Bsc-bio-es and bsc-bio-ehr-es are the first PLMs trained with exclusively biomedical and clinical text in Spanish (Carrino et al., 2022). These PLMs have a RoBERTa architecture and contain around 130 million parameters. Two corpora were built for

Model	bsc-bio-ehr-es	continual PLM	custom PLM
Admission	93.2 ± 0.9	**93.5 ± 0.9**	92.8 ± 0.7
Medical	94.4 ± 0.6	**94.9 ± 0.6**	94.9 ± 0.7
Admission+medical	95.9 ± 0	96.1 ± 0.1	**96.3 ± 0.2**

Table 2: Results in the test set (AUC) for all fine-tuned models.

this purpose, biomedical and clinical. The biomedical corpus consists of 2.5 million documents and 1.1 billion tokens, and the clinical corpus consists of 514k documents and 95 million tokens.

A biomedical corpus refers to medical text from academic sources, such as scientific publications or clinical trials. On the contrary, a clinical corpus is a collection of documents collected from the medical practice. In other words, it is what clinicians write during and/or after the examination of a patient.

Bsc-bio-es was pre-trained only with the biomedical corpus and bsc-bio-ehr-es with the biomedical and clinical corpora. The reason behind this design decision is two-fold; the clinical corpus is too small to create a functional PLM by itself, and to assess if adding a small clinical corpus to a large biomedical corpus positively impacts clinical NLP tasks.

As a first step, fine-tuning processes were carried out with the three fine-tuning datasets using bsc-bio-ehr-es as PLM. As a result, three fine-tuned models were built.

4.2 Fine-tuninig of a continual pre-training of bsc-bio-ehr-es

As a second step, continual pre-training processes were implemented using bsc-bio-ehr-es as a base PLM. For continuing the pre-training, the second type of datasets were used. One T4 GPU (16 GB) was used, and the processes lasted 42 hours for each. After this step, three PLM were built (admission, medical, admission+medical). Then, like the previous step, fine-tuning processes were carried out, and three more fine-tuned models were obtained.

4.3 Fine-tuninig of a PLM pre-trained from scratch

Finally, a pre-training process from scratch was implemented. This process used the same configuration as bsc-bio-ehr-es (RoBERTa) and our clinical-related corpus. Four T4 GPU (16GB) were used, and pre-training lasted 96 hours in 2 epochs. After this process, a new custom PLM was built.

With this new PLM, similar to the previous steps, fine-tuning processes were carried out, and three more fine-tuned models were obtained.

5 Results

Table 2 shows test results for every fine-tuned model. The test set only contains data not included in the fine-tuning or pre-training datasets.

We can notice that the continual PLMs and the custom PLMs are the best performers, but all the models are close performance-wise. Also, as expected, medical models are better than admission models, given that medical models capture more clinical information than admission models. The admission+medical models are the best performers since they combine admission and medical information.

As the metrics of all admission+medical models are close, we could select the least expensive and time-consuming when implementing it. In the case of this task, this process is the fine-tuning of the publicly available PLM, bsc-bio-ehr-es. However, this evidence should not be generalized for other types of tasks like named entity recognition or question answering, which are more complex and may benefit from lexical specificity. In those tasks, a PLM pre-trained with more clinical-related text could be better than a PLM trained with a mix of biomedical and clinical text.

Interestingly, admission models perform 1 to 2% worse than medical models. Therefore, there is an opportunity to make a coverage prediction before physicians check patients, helping the physicians review more medical details when their coverage decision does not match the predictions. Moreover, the admitting staff can have a stronger opinion on a coverage decision. Another benefit of providing a coverage prediction prior to the medical checkup is the possibility that the patient can manage his or her case more effectively. Depending on the likelihood of coverage the model provides, the patient may seek resources to help better justify his or her accident.

Finally, implementing a pre-trained language model could help the healthcare provider increase

Model	AUC	coverage
Current model	95.8	72.7
Custom PLM: admission+medical	**97.7**	**96.1**

Table 3: Results calculated on the accidents that both models have in common in January 2023 (AUC-Coverage).

savings from the correct classification of accidents. Table 3 shows the results of a shadow deployment (a method for simulating the new model's performance in the production environment) comparing the performance of the custom PLM with both administrative and medical reports against the current model in January of 2023. We estimated that with the implementation of this new model into production, more cases would be covered by a model, increasing between 20 to 24%. Considering the increase in coverage and in addition to the increase in the predictive metrics of the continual model, it is estimated that the health care provider could save between 1.5 to 2.5MM US annually. The saving will come from correctly classified cases where the administrative and medical cases were classified as covered, but in reality, they should not be covered.

6 Related work

NLP has been used in applications of the insurance industry in recent years. NLP techniques can be used to analyze vast amounts of unstructured data, such as customer interactions and policy documents, to gain insights and make informed decisions. In several areas within the insurance industry, NLP is being used, including customer service, claims processing, and fraud detection (Ly et al., 2020).

One of the most significant uses of NLP in the insurance sector is customer service (Quarteroni, 2018). To ascertain a client's wants and preferences, NLP techniques can be utilized to evaluate customer interactions such as phone calls and chat chats. Customers may receive more individualized help and recommendations thanks to the utilization of this data. Additionally, regular customer support operations like responding to frequently requested queries, have been automated using NLP algorithms.

Another area where NLP is employed in the insurance sector is claims processing (Popowich, 2005). By automating the analysis and classification of claims, NLP approaches can cut down on the time and resources needed to process claims. To make more educated judgments about claims, NLP algorithms have been employed, for instance, to extract information from claim documents, such as the type of injury and the reason for the accident.

Fraud detection is another area where NLP is being used in the insurance industry (Wang and Xu, 2018). Huge amounts of unstructured data, including policy documents and customer interactions, can be analyzed using NLP approaches to spot probable fraud cases.

7 Conclusion

This work studied the performance of clinical PLMs in a coverage prediction task. Three approaches were implemented, and the best model was compared to the current model used by the health provider. A PLM from scratch was the best-performing model but the most expensive and time-consuming.

Clinical natural language processing has great potential to impact the insurance industry, not only because of the great predictive power they offer but also because it is unnecessary to implement expensive training in the models. As there are no significant differences in performance between the pre-trained model and the fine-tuning with the admission+medical data, by just fine-tuning a PLM we can obtain good results at a lower cost for this downstream task. However, the situation might differ in other NLP tasks that benefit from lexical specificity.

8 Limitations

Some limitations of this work are listed below:

- The architecture and configuration for the custom PLM are the same as bsc-bio-ehr-es. Another architecture and configuration could obtain better results.
- The textual data come from just one provider. Using data from several providers could help with generalization.
- The custom PLM has not been compared with other PLMs in language tasks such as named entity recognition or question answering. This

comparison can help to understand if the custom PLM can outperform available PLMs in other types of tasks.

Ethics Statement

The ethical considerations of this work are related to the data that we used and the models we built. The data was extracted from administrative and clinical records from an insurance and health provider that specialized in labor accidents. Within this data, it is possible to find personal and sensitive information such as personal and company names, addresses, health information, pre-existing conditions, and diagnoses, among others. An anonymization process was not carried out since the model will be used for internal purposes and will not be released. As a process of memorization can occur in the PLM, we believe it is best to keep the model private because privacy attacks can extract personal and sensitive information.

We did not test the models for any bias under any protected field. Therefore, the trained models could benefit certain patients or accidents over others in the insurance decision. If a biased model is deployed in this provider's systems, it could harm patients with their insurance coverage decisions.

Acknowledgements

This work was funded by ACHS and ANID Chile: Basal Funds for Center of Excellence FB210005 (CMM); Millennium Science Initiative Program ICN17_002 (IMFD) and ICN2021_004 (iHealth), Fondecyt grant 11201250, and National Doctoral Scholarships 21211659 (Claudio Aracena).

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission.

Antoine Ly, Benno Uthayasooriyar, and Tingting Wang. 2020. A survey on natural language processing (nlp) and applications in insurance. *arXiv preprint arXiv:2010.00462*.

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.

Fred Popowich. 2005. Using text mining and natural language processing for health care claims processing. *ACM SIGKDD Explorations Newsletter*, 7(1):59–66.

Silvia Quarteroni. 2018. Natural language processing for industry: Elca's experience. *Informatik-Spektrum*, 41(2):105–112.

Matías Rojas, Jocelyn Dunstan, and Fabián Villena. 2022. Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 87–92, Seattle, WA. Association for Computational Linguistics.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural Language Processing with Transformers*. O'Reilly Media, Inc.

Yibo Wang and Wei Xu. 2018. Leveraging deep learning with lda-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, 105:87–95.