ANEXO N°45
INFORME
FINAL

Serie Proyectos de Investigación e Innovación

Superintendencia de Seguridad SocialSantiago - Chile

INFORME FINAL

304-2023 - Análisis masivo de textos clínicos para la construcción de herramientas de lenguaje en salud

Claudio Aracena, Sofía Correa, Jocelyn Dunstan, Luis Miranda y Fabián Villena

Enero 2025

Este trabajo fue seleccionado en la Convocatoria de Proyectos de Investigación e Innovación en Prevención de Accidentes y Enfermedades Profesionales 2023 de la Superintendencia de Seguridad Social (Chile), y fue financiado por la Asociación Chilena de Seguridad con recursos del Seguro Social de la Ley N°16.744 de Accidentes del Trabajo y Enfermedades Profesionales.

SUPERINTENDENCIA DE SEGURIDAD SOCIAL

SUPERINTENDENCE OF SOCIAL SECURITY

La serie Proyectos de Investigación e Innovación corresponde a una línea de publicaciones de la Superintendencia de Seguridad Social, que tiene por objetivo divulgar los trabajos de investigación einnovación en Prevención de Accidentes y Enfermedades del Trabajo financiados por los recursos delSeguro Social de la Ley 16.744.

Los trabajos aquí publicados son los informes finales y están disponibles para su conocimiento y uso. Los contenidos, análisis y conclusiones expresados son de exclusiva responsabilidad de su(s) autor(es), y no reflejan necesariamente la opinión de la Superintendencia de Seguridad Social.

Si requiere de mayor información, sobre el estudio o proyecto escriba a: investigaciones@suseso.cl.

Si desea conocer otras publicaciones, artículos de investigación y proyectos de la Superintendencia de Seguridad Social, visite nuestro sitio web: www.suseso.cl

The Research and Innovation Projects series corresponds to a line of publications of the Superintendence of Social Security, which aims to disseminate the research and innovation work in thePrevention of Occupational Accidents and Illnesses financed by the resources of Law Insurance 16,744.

The papers published here are the final reports and are available for your knowledge and use. Thecontent, analysis and conclusions are solely the responsibility of the author (s), and do not necessarilyreflect the opinion of the Superintendence of Social Security.

For further information, please write to: investigaciones@suseso.cl.

For other publications, research papers and projects of the Superintendence of Social Security, pleasevisit our website: www.suseso.cl.

Superintendencia de Seguridad
SocialHuérfanos 1376
Santiago, Chile.

Análisis masivo de textos clínicos para la construcción de herramientas de lenguaje en salud

Claudio Aracena, Sofía Correa, Jocelyn Dunstan, Luis Miranda y Fabián Villena

ÍNDICE

1.	Resumen ejecutivo	3
2.	Palabras claves	3
3.	Introducción y antecedentes	3
4.	Definición del problema, pregunta de investigación o desafío de innovación, objetivos	5
5.	Revisión de la literatura o experiencias relevantes	6
6.	Descripción de la metodología o etapas de la innovación	11
7.	Resultados	13
8.	Recomendaciones para Sistema de Seguridad y Salud en el Trabajo	19
9.	Conclusiones	20
10.	Referencias	21

1. Resumen ejecutivo

El proyecto se centró en desarrollar herramientas de inteligencia artificial diseñadas para procesar textos clínicos en español a gran escala, optimizando su uso en el contexto de salud ocupacional en Chile. Este trabajo responde a la necesidad de recursos en español para modelos de lenguaje especializados, dado el predominio de desarrollos en inglés.

Primero, se creó un corpus clínico basado en datos de la Asociación Chilena de Seguridad (ACHS), compuesto por 1,802 relatos de admisión equilibrados y rigurosamente anotados. Este corpus incluye información crítica sobre agentes, mecanismos y causas externas de accidentes laborales, lo que permite entrenar y evaluar modelos predictivos de alto impacto. Se desarrolló una guía de anotación detallada para asegurar la consistencia y calidad del corpus, incorporando tanto criterios lingüísticos como normativos.

En paralelo, se entrenaron diversos modelos de lenguaje preentrenados (MLP) utilizando técnicas de preentrenamiento continuo y desde cero, adaptadas al español y al dominio clínico. Los modelos se ajustaron para tareas de reconocimiento de entidades nombradas (NER), logrando resultados competitivos en corpus internacionales y destacándose en datos chilenos.

Finalmente, se exploró la integración de estos modelos en sistemas predictivos prácticos, como la codificación automática de mecanismos de accidentes, demostrando el potencial de estas herramientas para transformar la gestión de datos clínicos y mejorar la prevención de accidentes laborales en Chile.

2. Palabras claves

Seguridad y Salud en el Trabajo. Prevención de Riesgos Laborales. Accidentes del Trabajo. Modelos Predictivos. Documentación de Accidentes. Bases de Datos Etiquetadas. Gestión de Riesgos Laborales. Documentación de Accidentes. Procesamiento del Lenguaje Natural

3. Introducción y antecedentes

El lenguaje natural es una de las principales formas en que los profesionales de la salud documentan, comunican e interpretan información médica. Desde reportes clínicos hasta descripciones de accidentes laborales, estos textos contienen datos valiosos que, si se procesan efectivamente mediante herramientas de inteligencia artificial (IA), pueden transformar procesos clave como la gestión de prestaciones de salud y el diseño de intervenciones preventivas. Sin embargo, el volumen y la complejidad inherentes de los datos textuales clínicos y biomédicos plantean retos significativos al procesamiento automático de esta información.

En los últimos años, los modelos de lenguaje preentrenados (MLP) han revolucionado el campo del procesamiento del lenguaje natural (PLN). Basados en arquitecturas avanzadas como los transformers, estos modelos han demostrado ser herramientas poderosas para interpretar y generar texto en múltiples lenguajes y dominios. En particular, los MLP especializados en salud han permitido avances en tareas como el reconocimiento de entidades nombradas (NER, por sus siglas en inglés), la extracción de relaciones y la codificación automática de datos clínicos. Sin embargo, la mayoría de estos modelos se han desarrollado para textos en inglés, mientras que los recursos en español, especialmente para contextos clínicos

y biomédicos, siguen siendo limitados.

Este informe presenta el trabajo realizado para desarrollar e implementar herramientas de inteligencia artificial centradas en la creación de modelos de lenguaje en español capaces de procesar eficientemente textos clínicos de gran escala, tales como los reportes de admisión y médicos generados en la Asociación Chilena de Seguridad (ACHS). Se busca con ello impactar positivamente la forma en que se gestionan las prestaciones de salud, optimizando procesos como la clasificación de accidentes y la codificación de mecanismos según la normatividad vigente.

A lo largo de este documento, se describen los avances realizados en tres áreas fundamentales: (1) la creación de modelos de lenguaje específicos para el contexto de salud ocupacional en Chile, (2) el desarrollo de un corpus anotado y balanceado que incluye taxonomías relevantes para tareas clínicas, como agentes, mecanismos y causas externas, y (3) la integración de estos modelos en sistemas predictivos de uso práctico, como la codificación automática de datos. Estos esfuerzos no solo buscan mejorar la eficiencia de los procesos actuales, sino también sentar las bases para nuevas capacidades analíticas que contribuyan a prevenir accidentes laborales y mejorar la calidad de la atención.

La innovación de este trabajo radica en la combinación de enfoques locales y globales: desde la construcción de un corpus en idioma español basado en datos específicos del contexto, hasta la adaptación de arquitecturas avanzadas como RoBERTa mediante técnicas como el preentrenamiento continuo y el ajuste fino. A través de estos esfuerzos, se espera habilitar soluciones que no solo estén al nivel del estado del arte en PLN, sino que también respondan a las necesidades particulares de la comunidad de salud laboral en Chile.

En las siguientes secciones, se desarrollan los antecedentes teóricos y metodológicos esenciales, seguidos de los resultados clave y conclusiones derivadas del trabajo realizado, evidenciando el potencial de la inteligencia artificial para transformar la gestión de textos clínicos y promover innovaciones en el sector salud.

4. Definición del problema, pregunta de investigación o desafío de innovación, objetivos.

El procesamiento de textos clínicos y biomédicos en español enfrenta importantes limitaciones debido a la escasez de recursos específicos para este idioma y dominio. Esta falta de herramientas especializadas afecta particularmente a las organizaciones de salud en Chile, como la Asociación Chilena de Seguridad (ACHS), que generan grandes volúmenes de datos textuales valiosos en forma de reportes médicos y de accidentes laborales. Procesar estos textos de manera eficiente podría optimizar la gestión de prestaciones de salud y mejorar la prevención de accidentes laborales, pero la complejidad y el volumen de los datos actuales representan un desafío significativo para los métodos tradicionales de procesamiento del lenguaje natural (PLN).

Lo anterior conduce a la siguiente pregunta de investigación: ¿Cómo desarrollar herramientas de inteligencia artificial, basadas en modelos de lenguaje en español, que sean capaces de procesar textos clínicos masivos para optimizar la gestión de prestaciones de salud y diseñar intervenciones preventivas más efectivas en el ámbito de la salud ocupacional en Chile?

Con relación a los objetivos, se plantean los siguientes:

- **Creación de modelos del lenguaje en el estado del arte:** Desarrollar y adaptar modelos preentrenados que aprovechen arquitecturas avanzadas, como RoBERTa, para abordar las especificidades del lenguaje clínico y ocupacional en español.
- **Creación del corpus y reconocimiento de entidades nombradas:** Diseñar y anotar un corpus balanceado y de alta calidad, enfocado en taxonomías clínicas relevantes (como agentes, mecanismos y causas externas de accidentes) para tareas de reconocimiento de entidades nombradas (NER).
- **Integración modelos en sistemas predictivos:** Incorporar estos modelos en herramientas prácticas, como la codificación automática de datos clínicos, evaluando su rendimiento en escenarios reales y optimizando su aplicabilidad con base en los recursos computacionales disponibles.

5. Revisión de la literatura o experiencias relevantes

Con el auge del aprendizaje profundo, el campo del Procesamiento del Lenguaje Natural (PLN) ha probado varias arquitecturas de redes neuronales para aplicarlas a sus tareas. En los últimos cinco años, el “transformer” ha sido una de las arquitecturas más utilizadas y el uso de modelos previamente entrenados se ha convertido en un enfoque común para resolver una amplia gama de problemas, desde la clasificación de texto hasta la generación de texto.

Desde la creación de transformers [29] y ULMFiT [14], hemos sido testigos del desarrollo de varios modelos de lenguaje pre-entrenados (MLP) que se han convertido en el estado del arte en diferentes tareas. Comenzando con GPT [24] y BERT [8], continuando con RoBERTa [21] y DeBERTa [13,12], entre otros, la comunidad de investigación se ha centrado en encontrar nuevas arquitecturas que puedan superar los rendimientos actuales y aplicarlos en tantas áreas como sea posible.

Uno de los ámbitos donde es posible encontrar varios MLP es el de la salud. Apenas unos meses después de la publicación de BERT, se lanzaron MLP que utilizan texto clínico [1,15]. Además, aunque los MLP están diseñados principalmente para texto en inglés, todavía es posible encontrar MLP clínicos y biomédicos para español [6,27].

Esta sección presenta el concepto de modelos de lenguaje y fundamentos para MLP, describe algunos de los trabajos más relevantes en MLP de dominio general y continúa con MLP en los campos clínico y biomédico. Posteriormente, se describen importantes MLP clínicos y biomédicos para el español. Finalmente, se muestran algunas tareas de evaluación de PLN biomédica y clínica en español.

5.1. Modelos de lenguaje

Un modelo de lenguaje (ML) es un modelo que asigna probabilidades a secuencias de palabras. Este tipo de modelo es útil para tareas como reconocimiento de voz, corrección ortográfica, traducción automática y generación de texto, entre otras. Uno de los enfoques más simples para crear un ML es el modelo de n-gramas. Un n-grama es una secuencia de n palabras. Un n-grama con n=2 se llama bigrama, y con n=3 se llama trigrama [17].

Para representar la probabilidad de una palabra dada una secuencia, el modelo de n-gramas aproxima la historia usando solo las últimas palabras en lugar de la secuencia completa. Por ejemplo, el modelo de bigram aproxima la probabilidad de una palabra dada una secuencia P($w_n|w_1,\ldots, w_{n-1}$) usando la probabilidad condicional de la palabra anterior en la secuencia P(w_n |w_{n-1}).

Aunque el modelo de n-gramas es bastante simple, era el enfoque más común antes del aprendizaje profundo. Con el surgimiento del aprendizaje profundo se implementaron nuevos paradigmas para crear LM, como las redes neuronales recurrentes y los transformers.

5.1.1. Transformers

El transformer se propuso como una arquitectura de codificador-decodificador que no necesita utilizar redes neuronales recurrentes, que fue el enfoque dominante hasta 2017. El problema con redes neuronales recurrentes es que no pueden usar la paralelización dentro de ejemplos de entrenamiento dada su naturaleza secuencial. En cambio, el transformer utiliza un mecanismo de atención para dibujar dependencias globales entre la entrada y la salida, eliminando el problema de paralelización. Este cambio permitió que el transformer alcanzara un rendimiento de última generación en calidad de traducción [29].

El transformer es una de las piezas fundamentales en el rápido crecimiento de los MLP. Otra pieza es el uso de MLP optimizado para resolver tareas posteriores. ULMFiT introdujo este concepto en 2018.

5.1.2. UMLFiT

El aprendizaje por transferencia estaba ampliamente extendido en la visión computacional, pero el PLN aún requería un modelo de tarea específica y entrenamiento desde cero para abordar sus problemas en 2018. ULMFiT cambió este escenario mediante la implementación de MLP, que puede resolver cualquier tarea de PLN mediante un proceso de ajuste o fine-tuning [14]. Aunque la idea de pre-entrenamiento ya se había utilizado en el pasado [9], todavía faltaba conocimiento sobre cómo entrenar a los MLP de manera efectiva.

ULMFiT consiste en tres etapas. Primero, un ML se pre-entrena en un corpus de dominio general para la tarea de predicción del siguiente token. En segundo lugar, el ML se ajusta en los datos de la tarea posterior para la misma tarea que en la etapa uno. En tercer lugar, utilizando el MLP ajustado, se hace fine-tuning de un clasificador en los datos de una tarea específica. Este proceso les permitió superar las seis tareas de clasificación de texto del estado del arte, lo que demuestra que el pre-entrenamiento con texto general y específico de dominio agrega información valiosa para tareas específicas.

Aunque los MLP recientes no utilizan el mismo proceso que ULMFiT, impulsó la creación de los MLP más populares basados en la arquitectura de transformer: GPT y BERT.

5.2. Modelos de lenguaje pre-entrenados generales

5.2.1. GPT

El pre-entrenamiento generativo (GPT, por sus siglas en inglés) es uno de los primeros MLP que incluye la arquitectura de transformer [24]. Al igual que ULMFiT, GPT utiliza un proceso de varios pasos de entrenamiento previo en un gran corpus de texto y ajuste de los datos de las tareas posteriores. Sin embargo, el proceso de fine-tuning combina predicción y clasificación de texto, como se muestra en la imagen izquierda de la Figura \ref{fig4}, en lugar de tener un paso define-tuning de ML y un paso de fine-tuning del clasificador como ULMFiT. La arquitectura GPT utiliza un decodificador de transformer y utiliza tokens pasados para predecir el siguiente token. BookCorpus [31] se utilizó como datos para el pre-entrenamiento.

Aunque GPT logró un rendimiento de estado del arte, tiene problemas como el uso solo de tokens pasados para hacer predicciones, conocido como enfoque unidireccional, y solo capas de decodificador de transfomers. Un grupo de investigación de Google aprovechó estas oportunidades para crear BERT.

5.2.2. BERT

Las representaciones de codificador bidireccional de transformers (BERT, por sus siglas en inglés) es un MLP creado por Google [8]. Al igual que GPT, utiliza la arquitectura de transformer y procesos de pre-entrenamiento y fine-tuning para resolver tareas específicas. Sin embargo, BERT utiliza un enfoque bidireccional en el proceso de pre-entrenamiento y predice tokens enmascarados en una secuencia (ML enmascarado), en lugar de simplemente predecir el siguiente token como GPT. Además, el entrenamiento previo de BERT incluye una tarea de predicción de la siguiente oración que ayuda a mejorar el rendimiento en las tareas específicas de respuesta a preguntas e inferencia del lenguaje natural. Se utilizaron BookCorpus [31] y Wikipedia en inglés como datos de pre-entrenamiento, con un total de 16 GB de texto sin comprimir.

5.2.3. RoBERTa

Un grupo de investigación de Facebook realizó un estudio de replicación del pre-entrenamiento de BERT y descubrió que estaba significativamente sub-entrenado. Introdujeron algunos cambios en el pre-entrenamiento de BERT original para crear un enfoque BERT sólidamente optimizado (RoBERTa, por sus siglas en inglés) [21]. Las modificaciones fueron las siguientes: (1) entrenar el modelo por más tiempo y con más datos, (2) eliminar la tarea predecir la siguiente oración, (3) entrenar en secuencias más largas y (4) cambiar dinámicamente el patrón de enmascaramiento aplicado a los datos de entrenamiento. Estos cambios permitieron a RoBERTa obtener un rendimiento de estado del arte en cuatro de nueve tareas GLUE e igualar el rendimiento de estado del arte en las pruebas comparativas SQuAD [26,25] y RACE [18].

5.3. Modelos de lenguaje pre-entrenados biomédicos y clínicos

Esta subsección presenta algunos trabajos relevantes de MLP biomédicos y clínicos. Los MLP biomédicos se refieren a modelos pre-entrenados en textos médicos de fuentes académicas, como publicaciones científicas, y los MLP clínicos se refieren a modelos pre-entrenados en textos médicos de la práctica médica,

como notas clínicas y prescripciones.

5.3.1. BioBERT

Tras el éxito de BERT, otros campos crearon sus propias versiones de este modelo. BioBERT fue el primer MLP basado en BERT pre-entrenado en un corpus biomédico [19]. Utilizando la arquitectura BERT, BioBERT fue pre-entrenado con Wikipedia en inglés, BookCorpus [31], resúmenes de PubMed y artículos de texto completo de PubMedCentral (PMC), con un total de 21 mil millones de palabras. BioBERT se ajustó en una serie de tareas biomédicas de PLN, como reconocimiento de entidades nombradas (NER, por sus siglas en inglés), extracción de relaciones y respuesta a preguntas. BioBERT logró un mejor rendimiento en la mayoría de las tareas estudiadas y supera significativamente a BERT, lo que demuestra que el pre-entrenamiento en datos de dominios específicos es un paso clave para el desempeño de las tareas específicas.

5.3.2. ClinicalBERT

Uno de los primeros MLP clínicos basados en BERT fue ClinicalBERT [1]. A diferencia de BioBERT, ClinicalBERT utilizó texto clínico como notas médicas y resúmenes de alta, ambos extraídos de la base de datos MIMIC-III [16]. Este tipo de texto es diferente al que podemos encontrar en resúmenes o artículos de científicos en PubMed. Por tanto, tiene sentido disponer de un modelo específico para el texto clínico.

Se implementaron varias versiones de ClinicalBERT utilizando una técnica conocida actualmente como pre-entrenamiento continuo. El pre-entrenamiento continuo significa comenzar desde un MLP existente para continuar el proceso de pre-entrenamiento con datos adicionales. En el caso de ClinicalBERT, se utilizaron BERT y BioBERT como puntos de partida, y notas clínicas y resúmenes de alta como datos adicionales para continuar con el pre-entrenamiento.

Bio+ClinicalBERT (ClinicalBERT a partir de BioBERT) supera a BioBERT en 2 de 5 tareas, lo que demuestra que agregar texto clínico adicional en el pre-entrenamiento continuo puede mejorar el rendimiento, pero no en todos los casos.

5.4. MLP de dominio general, biomédicos y clínicos en Español

5.4.1. BETO

La primera implementación de BERT para español, BETO, se construyó en la Universidad de Chile [5]. BETO utilizó la misma arquitectura que el modelo BERT original, pero para el proceso de pre-entrenamiento utilizó algunos cambios introducidos por RoBERTa, como el enmascaramiento dinámico. Se utilizaron Wikipedia en español y las partes en español del Proyecto OPUS [28] como corpus de preentrenamiento, con un total de 3 mil millones de palabras.

5.4.2. RoBERTA-bne

Se creó un conjunto de tareas de evaluación, GLUES [5], para comparar BETO con una implementación multilingüe de BERT, mBERT [30]. BETO supera la mayoría de los resultados de mBERT, excluyendo algunas tareas de respuesta a preguntas. Estos resultados muestran que el pre-entrenamiento con texto en español mejora el rendimiento para la mayoría de las tareas específicas en comparación con el MLP

multilingüe.

MarIA es una familia de MLP desarrollada en el Centro de Supercomputación de Barcelona [11]. Un MLP de interés en MarIA es la implementación de RoBERTa con texto en español, RoBERTa-bne. RoBERTa-bne fue pre-entrenado con texto extraído de la Biblioteca Nacional de España (BNE). El corpus de pre-entrenamiento consta de más de 200 millones de documentos y 135 mil millones de tokens, con un total de 570 GB de texto sin comprimir. Se construyeron dos versiones de RoBERTa-bne, básica y grande, con fines comparativos. De nueve tareas de dominio general, RoBERTa-bne supera a ocho en comparación con BETO y mBERT.

5.4.3. Bsc-bio-es and bsc-bio-ehr-es

Los MLP anteriores para español fueron pre-entrenados en textos de dominio general. Bsc-bio-es y bsc-bio-ehr-es son los primeros MLP entrenados con texto exclusivamente biomédico y clínico en español [6]. Para ello se construyeron dos corpus, uno biomédico y otro clínico. El corpus biomédico consta de 2,5 millones de documentos y 1.100 millones de tokens, y el corpus clínico consta de 514.000 documentos y 95 millones de tokens.

Ambos PLM utilizan la arquitectura RoBERTa. Sin embargo, bsc-bio-es fue pre-entrenado únicamente con el corpus biomédico y bsc-bio-ehr-es con los corpus biomédico y clínico. La razón detrás de esta decisión de diseño es doble: el corpus clínico es demasiado pequeño para crear un MLP funcional por sí solo y para evaluar si agregar un corpus clínico pequeño a un corpus biomédico grande tiene un impacto positivo en las tareas clínicas de PLN.

Se utilizaron tres tareas clínicas para comparar los MLP con otros como BETO-Galén, mBERT, mBERT-Galén, RoBERTa-bne, entre otros. BETO-Galén y mBERT-Galén son versiones BETO y mBERT con pre-entrenamiento continuo en el corpus Galén Oncología [22], respectivamente. bsc-bio-ehr-es supera a todos los demás MLP, incluído bsc-bio-es, en dos tareas, y bsc-bio-es quedó en segundo lugar. Para la tarea restante, otro PLM, XLM-R-Galén (una versión de pre-entrenamiento continuo de XML-R [7], una versión multilingüe de RoBERTa) supera a todos los demás MLP, pero bsc-bio-es y bsc -bio-ehr-es quedó en segundo y tercer lugar, respectivamente.

Estos resultados muestran que el uso únicamente de corpus biomédicos y clínicos para el proceso de pre-entrenamiento mejora el rendimiento en comparación con textos de dominio general en español. Este resultado es congruente con los hallazgos de PubMedBERT para el inglés. Los resultados siguen siendo válidos incluso cuando se compara con RoBERTa-bne, un modelo previamente entrenado con 100 veces más texto que bsc-bio-es y bsc-bio-ehr-es. Curiosamente, XLM-R-Galén, entrenado en más de 2 TB de datos de dominio general en 100 idiomas, tiene los mejores resultados para una tarea, pero bsc-bio-es y bsc-bio-ehr-es rondan el 0,1% de F-score de distancia 4.

5.5. Tareas de evaluación para modelos de lenguaje biomedicos y clínicos en español

Existen tareas de reconocimiento de entidades nombradas (NER por sus siglas en inglés) clínicas y biomédicas que se utilizan habitualmente para evaluar los PLM creados con corpus españoles. A continuación, se incluye una lista de todas las tareas de interés para este trabajo:

- CANTEMIST [23]: corpus anotado con menciones de morfología tumoral en 1.301 informes de casos clínicos oncológicos.

- PharmaCoNER [10]: corpus anotado con entidades como compuestos químicos y fármacos en 1.000 estudios de casos clínicos.

- CT-EBM-SP [4]: corpus anotado con entidades UMLS en 1.200 textos sobre estudios y anuncios de ensayos clínicos.

- NUBes [20]: corpus anotado con entidades de negación e incertidumbre en historias clínicas anonimizadas (29.682 frases).

- The Chilean Waiting List Corpus [3,2]: corpus anotado con entidades clínicas como hallazgos, procedimientos, medicamentos, etc. en 10.000 derivaciones anonimizadas para consultas de especialidad de la lista de espera de hospitales públicos chilenos.

6. Descripción de la metodología o etapas de la innovación

El desarrollo de este proyecto se estructuró en cuatro etapas principales, cada una diseñada para abordar los objetivos clave de la investigación y garantizar la calidad y utilidad de los resultados obtenidos. Estas etapas abarcaron desde la preparación de los datos hasta la implementación de herramientas predictivas, como se detalla a continuación:

La primera etapa consistió en la recopilación y curación del corpus de textos clínicos provenientes de la Asociación Chilena de Seguridad (ACHS). Se diseñó un proceso de selección y estratificación de los datos que aseguró una representación balanceada de los casos a lo largo del tiempo y diferentes niveles de dificultad en la redacción de los relatos de accidentes. Para garantizar la calidad del corpus, se implementó una guía de anotación exhaustiva que incluyó reglas generales y específicas para cada entidad nombrada. Esta guía fue refinada iterativamente mediante un proceso de revisión colaborativa, asegurando su utilidad para anotadores sin experiencia previa. La anotación del corpus fue realizada en equipos, con revisiones periódicas para mantener la consistencia y precisión. Adicionalmente, se monitoreó la concordancia entre anotadores a lo largo de todo el proceso. Este enfoque permitió la creación de un conjunto de datos anotados de alta calidad, que sirvió como base para las siguientes etapas del proyecto.

La segunda etapa involucró la construcción de modelos de lenguaje preentrenados (MLP) especializados en el dominio clínico y en español. Utilizando arquitecturas avanzadas como BERT y RoBERTa, se implementaron dos metodologías principales de preentrenamiento: continuo y desde cero. Estas técnicas se aplicaron sobre el corpus curado, maximizando la relevancia y especificidad de los modelos para tareas clínicas. Los modelos preentrenados fueron evaluados mediante tareas de reconocimiento de entidades nombradas (NER) en corpus tanto nacionales como internacionales. Se seleccionaron los mejores modelos basados en métricas de desempeño como el puntaje F1, considerando además el equilibrio entre rendimiento y eficiencia computacional. Los modelos seleccionados sirvieron como base para el ajuste fino en tareas específicas del proyecto.

En la tercera etapa, los modelos seleccionados fueron ajustados específicamente

para tareas de codificación automática y NER relacionadas con accidentes laborales. El ajuste fino se realizó utilizando el corpus anotado y se evaluó el desempeño de los modelos en términos de precisión, cobertura y generalización. Se realizaron comparaciones con modelos de lenguaje de gran escala como CLIN-X y RoBERTa-bne-large, evidenciando la competitividad de los modelos desarrollados, especialmente en tareas aplicadas al contexto chileno.

Finalmente, se integraron los modelos ajustados en sistemas predictivos diseñados para facilitar la gestión de información en la ACHS. Se llevaron a cabo pruebas piloto en entornos reales, evaluando la utilidad de las herramientas en tareas como la clasificación de mecanismos y agentes de accidentes laborales, así como la codificación automática de datos. Estas pruebas no solo permitieron identificar áreas de mejora en los sistemas, sino también validar su impacto en la optimización de procesos internos.

7. Resultados

7.1. Crear modelos del lenguaje en el estado del arte

Para construir modelos de lenguaje que puedan aprovechar los datos de la ACHS, se compilo un corpus de reportes de adminsión y médicos. Las estadisticas del corpus compilado se muestran a continuación:

Métrica	**Total**
N. de documentos	7.1 M
N. de tokens	1.030 M
Tokens por documento	146.1 ± 121.4
Vocabulario	2.3 M

Cuadro 1: Estadísticas del corpus de reportes de admisión y médico de la ACHS.

Con el corpus compilado, se procedió a pre-entrenar diversos MLP mostrados en el marco teórico. Este entrenamiento consistió en utilizar el corpus de la ACHS para 2 tipos de metodología de pre-entrenamiento, pre-entrenamiento continuo y pre-entrenamiento desde cero. El pre-entrenamiento continuo consiste en continuar el pre-entrenamiento del un MLP ya existente, mientras que el pre-entrenamiento desde cero, se comienza el pre-entrenamiento desde la configuración inicial de un MLP ya existente. En la siguiente tabla se muestra las diversas configuraciones probadas para generar MLP con el corpus de la ACHS:

Modelo	**Modelo base**	**Pre-entrenamiento continuo**	**Pre-entrenamiento desde cero**
BERTACHS	BETO	X	
BERTACHS-scratch	BETO		X
ROBERTACHS	RoBERTa-bne	X	
ROBERTACHS-scratch	RoBERTa-bne		X
bio-ROBERTACHS	bsc-bio-es	X	
bio-ROBERTACHS-scratch	bsc-bio-es		X
bio-cli-ROBERTACHS	bsc-bio-ehr-es	X	
ROBERTACHS-large	RoBERTa-bne-large	X	
CLINX-ACHS	CLIN-X	X	
ALBERTACHS	ALBETO	X	
ALBERTACHS-scratch	ALBETO		X

Cuadro 2: Lista de MLP utilizando pre-entrenamiento continuo y pre-entrenamiento desde cero utilizando el corpus de la ACHS.

El cuadro 3 muestra el tiempo de ejecución de entrenamiento y la cantidad de épocas para cada implementación de MLP. Todas las versiones preentrenadas continuas se entrenaron con una GPU A5000 con 24 GB, y todas las versiones preentrenadas desde cero, excepto BERTACHS-scratch, se entrenaron con dos

GPU A6000 con 24G de memoria cada una. Como resultado, podemos ver que las versiones preentrenadas desde cero tienen un tiempo de ejecución menor que su contraparte preentrenada continua.

En general, todos los MLP se entrenaron para dos épocas. CLIN-X tiene una cantidad menor de épocas, dado su mayor tiempo de ejecución de entrenamiento. La decisión sobre la cantidad de épocas se basó en los recursos disponibles para ejecutar los procesos de entrenamiento.

Modelo	**Modelo base**	**Tiempo de ejecución**	**Épocas**
BERTACHS	BETO	3d 1h	2
BERTACHS-scratch	BETO	3d 6h	2
ROBERTACHS	RoBERTa-bne	3d 17h	2
ROBERTACHS-scratch	RoBERTa-bne	2d 23h	2
bio-ROBERTACHS	bsc-bio-es	3d 12h	2
bio-ROBERTACHS-scratch	bsc-bio-es	2d 9h	2
bio-cli-ROBERTACHS	bsc-bio-ehr-es	3d 17h	2
ROBERTACHS-large	RoBERTa-bne-large	7d 21h	2
CLINX-ACHS	CLIN-X	9d 6h	1.1
ALBERTACHS	ALBETO	2d 20h	2
ALBERTACHS-scratch	ALBETO	1d 15h	2

Cuadro 3: Lista de versiones pre-entrenadas continuas y pre-entrenadas desde cero de PLM con tiempo de ejecución de entrenamiento y número de épocas.

Para seleccionar el mejor MLP se procedió a realizar el ajuste de los modelos para tareas de NER, en un proceso conocido como fine-tuning. Las tareas NER seleccionadas son las mostradas en el marco teórico del presente trabajo. El cuadro 4 muestra los resultados de los PLM implementados y comparables para todos los corpus NER estudiados (CANTEMIST-NER, PharmaCoNER, CT-EBM-SP y NUBes, y el corpus de lista de espera chilena (CWLC)).

Model	**CANTEMIST**	**PharmaCoNER**	**CT-EBM-SP**	**NUBes**	**CWLC**
Tulio	0,790	0,904	0,877	0,892	0,583
BETO	0,753	0,872	0,851	0,879	0,634
BERTACHS	0,775	0,907	0,875	0,886	**0,679**
BERTACHS-scratch	0,746	0,869	0,843	0,863	0,633
ROBERTACHS	0,771	0,896	0,861	0,890	0,571
ROBERTACHS-scratch	0,711	0,839	0,799	0,829	0,436
RoBERTAa-bne	0,760	0,895	0,867	0,893	0,571
bio-ROBERTACHS	0,797	0,856	0,841	0,871	0,569
bsc-bio-ehr-es	0,807	0,874	0,852	0,865	0,630
bio-cli-ROBERTACHS	0,742	0,745	0,735	0,749	0,418
bsc-bio-es	0,798	0,868	0,860	0,874	0,607
bio-ROBERTACHS-scratch	0,732	0,766	0,727	0,828	0,478
RoBERTa-bne-large	**0,815**	0,922	0,888	0,904	0,626
ROBERTACHS-large	0,683	0,791	0,766	0,776	0,368
CLINX-ACHS	0,720	0,861	0,859	0,868	0,511
CLIN-X	0,805	**0,926**	**0,910**	**0,907**	0,614

XLM-RoBERTa-large	0,809	0,922	0,907	0,905	0,625
ALBERTACHS	0,731	0,851	0,839	0,869	0,607
ALBERTACHS-scratch	0,691	0,801	0,765	0,819	0,539
ALBETO	0,632	0,775	0,825	0,851	0,473
DistilBETO	0,711	0,861	0,824	0,853	0,585

Cuadro 4: Evaluación de los MLP para las tareas NER en corpus clínicos. Se utiliza la puntuación F1 como métrica. Los resultados en negrita son los mejores; los resultados en rojo son los mejores sin tener en cuenta las arquitecturas grandes; los resultados en azul son los mejores para las arquitecturas ligeras

Se observa que CLIN-X y RoBERTa-bne-large tienen el mejor rendimiento para los corpus clínicos internacionales de NER. Ambos modelos utilizan la arquitectura RoBERTa grande como base. Curiosamente, RoBERTa-bne-large no se entrenó previamente con datos biomédicos o clínicos. Adicionalmente, se observa que BERTACHS tiene el mejor resultado para el CWLC (corpus con datos chilenos), a pesar de que su arquitectura no está basada en un modelo grande como los anteriores.

Sin considerar las arquitecturas grandes, ningún modelo claro tiene el mejor rendimiento en los corpus NER de fuentes internacionales. Cuatro modelos (Bsc-bio-ehr-es, BERTACHS, Tulio y RoBERTa-bne) obtienen el mejor rendimiento, uno para cada tarea NER en tareas de fuentes internacionales. Sin embargo, BERTACHS tiene el mejor resultado para el CWLC, de fuente chilena.

Por último, si solo consideramos las arquitecturas ligeras, ALBERTACHS tiene el mejor rendimiento para estas tareas, exceptuando una de ellas.

A partir de este análisis podemos concluir que CLIN-X o BERTACHS son buenas opciones para utilizar como MLP para resolver tareas internas de la ACHS. Si se quiere optimizar por el uso de recursos computacionales, se debería utilizar BERTACHS, ya que su arquitectura no es grande, lo que conlleva a un menor uso de recursos.

7.2. Creación del corpus y reconocimiento de entidades nombradas

Se realizó una etapa de anotación exploratoria de documentos para poder obtener conocimiento general del contenido de los textos de la organización y así poder detectar particularidades que deban ser abordadas en la guía de anotación. La muestra de esta anotación exploratoria, de tamaño 120 datos, se realizó de manera estratificada según la codificación actual existente en la ACHS, para promover la comprensión cabal de las casuísticas en los datos y cómo estas se debieran abordar en la guía de anotación. A su vez, cada dato fue anotado por cada lingüista, y se realizaró una revisión exhaustiva dato por dato y entidad por entidad para unificar el criterio de anotado y resolver cualquier ambigüedad en la guía de anotación, previo a la creación del corpus. Esto promovió la calidad del mismo, y así, aquella del sistema inteligente como producto final.

Se creó una guía de anotación para la tarea de orientar a un sistema inteligente en el discernimiento de los fragmentos del texto que son centrales a la descripción de la causa externa de una lesión a través de la codificación de su Agente y Mecanismo. Esta introduce: la relación de estas taxonomías con los relatos de accidente en las mutualidades chilenas, con la tarea de reconocimiento de entidades nombradas, los criterios de prevención y normativa que deben informar

la anotación, y los criterios lingüísticos que definirán los límites de una entidad nombrada siempre que su naturaleza, como en este caso, sea abstracta.

Las reglas de anotación contenidas en la guía se afinaron seccionando la estructura en reglas generales y reglas por cada entidad. Además, una vez terminada la anotación del corpus y previo al proceso de revisión, se clarificó cada regla agregando ejemplos de cada construcción lingüística prototípica de cada regla, y de cada uno de sus criterios de excepción. El mayor propósito de lo anterior fue promover la facilidad de lectura de la guía para un público sin experiencia de anotación o conocimiento analítico de la lingüística.

Posterior a la creación de la guía de anotación se comenzó la anotación de documentos con información relevante para la organización, siendo esta su(s) Agente, Mecanismo y Causa externa de accidentes laborales. Previo a la anotación, se coordinó una visita a terreno a un centro de atención donde los anotadores puedan ver la interfaz de usuario donde se crean los datos a etiquetar (relato de admisión), y presenciar los procesos y protocolos que permiten entender el dato en su contexto administrativo-institucional. La anotación del corpus contempló chequeos cruzados periódicos entre anotadores para mantener consistente su criterio de anotado.

Se anotó un corpus de 1802 relatos de admisión, balanceado y revisado. El balance de los datos se centra, primero, en la estacionalidad de los mismos. Para cubrir la variedad de accidentes cuya tendencia a ocurrir se ve influenciada por la época del año, como aquellos asociados al clima, el muestreo se estratificó según mes de generación del dato, y se permitió una variación máxima de ±1 dato por mes.

Por otra parte, existen ciertas guías de estructura para el relato de admisión en la plataforma donde este se redacta. Estos son tres encabezados editables como “Al momento del accidente, estaba”. Cuando todas las frases guía cumplen su propósito, la información sobre distintos aspectos del accidente tiende a mantenerse en cierta estructura gramatical y a cierta distancia de las frases. Sin embargo, los usuarios a menudo las eliminan para redactar libremente, o incluso rellenan luego de ellas con elementos vacíos de información, como por ejemplo un punto seguido. Además, luego de la frase guía “El accidente ocurrió con” existe una tendencia en los datos a dar un concepto de una a tres palabras cercanamente relacionado al Agente y/o Mecanismo del accidente. Con tal de evitar que el modelo se sesgue a predecir las entidades en ciertas posiciones relativas a las frases guía, o que dependa del resumen “el accidente ocurrió con” que no siempre está presente, generamos una segunda estratificación equitativa. Esta le asigna al texto una de 5 categorías, que suponen ser una aproximación al ‘nivel de dificultad de automatización’, con las siguientes características:

- Incluye las 3 frases guía, y “el accidente ocurrió con” está seguida de 1 a 3 palabras antes de la puntuación (generalmente un concepto simple relacionado a Agente o Mecanismo).
- Incluye las 3 frases guía, y “el accidente ocurrió con” está seguida de 4 a 9 palabras antes de la puntuación (generalmente una frase compleja relacionada a Agente o Mecanismo).
- Incluye las 3 frases guía, y “el accidente ocurrió con” está seguida de 10 o más palabras antes de la puntuación (generalmente una frase compleja relacionada a Agente o Mecanismo).
- Al menos una de las frases guía fue eliminada o alterada, y al menos una se mantuvo en su formato predeterminado.
- No incluye ninguna de las frases guía.

Para cada dato en el corpus, se realizó la anotación de los fragmentos (spans) asociados al Agente y Mecanismo del accidente descrito, contemplando varios en los casos borde donde existiera más de uno. Además, se anotó el código correspondiente de la taxonomía Agente o Mecanismo, respectivamente, para cada entidad nombrada. Por último, a nivel de documento, se anotó la tarea de clasificación textual de cada relato dentro de la taxonomía Causa Externa CIE-10.

Se monitoreó la concordancia entre anotadores semana a semana, y la precisión de cada anotador, mediante un 10% de los datos (34 semanales), los cuales fueron anotados por ambos y se sometieron a una posterior curación. Los análisis revelan que la calidad y concordancia del corpus se mantuvo estadísticamente consistente a lo largo del proceso de anotación y transversal a la dificultad del texto. Luego, se pasó a un proceso de revisión manual de todos los datos, aplicando los criterios definidos a lo largo de las 5 semanas de curación y consolidados en la versión final de la guía de anotación.

7.3. Integración y evaluación en modelos predictivos

Se realizó un análisis exploratorio del desempeño del modelo gpt-3.5-turbo para la codificación automática de mecanismos de accidentes. En este análisis, se compararon las codificaciones realizadas por el modelo con las del proceso de anotación manual del corpus. El resultado mostró que el modelo obtuvo un puntaje F1 del 49.3% en 1802 relatos analizados.

Para entender mejor los resultados, se utilizó una matriz de confusión que compara las predicciones del modelo con las codificaciones manuales. Una matriz de confusión es una tabla que permite visualizar la precisión de un modelo de clasificación, mostrando en las filas las etiquetas reales (los mecanismos anotados) y en las columnas las predicciones del modelo. Las celdas diagonales indican las predicciones correctas, mientras que las celdas fuera de la diagonal reflejan errores de clasificación. La matriz se presenta en la siguiente figura:

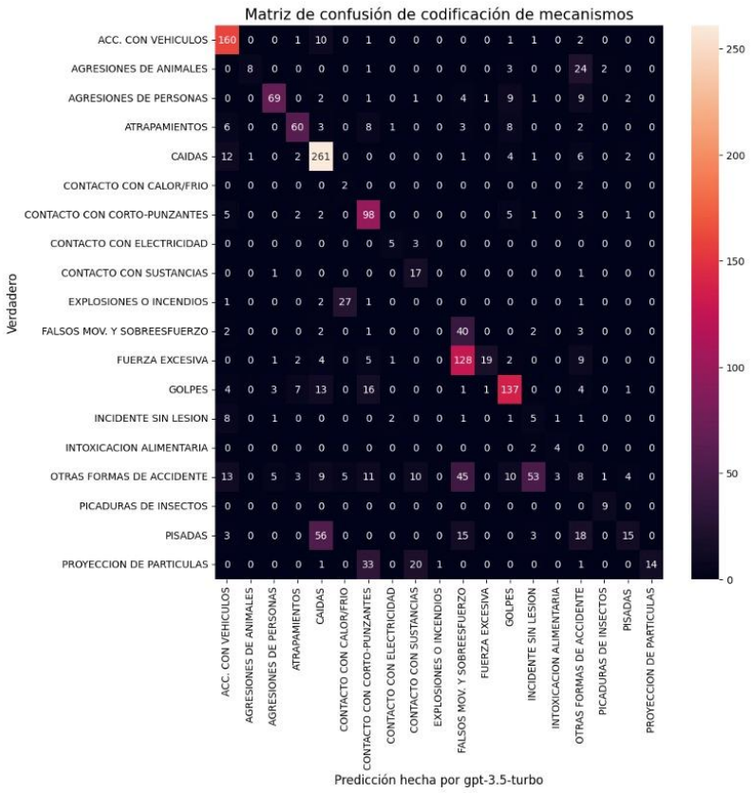

Figura 1: Matriz de confusión para la clasificación de mecanismos hecha por el actual

modelo que utiliza la ACHS (gpt-3.5-turbo) y el gold standard.

Del análisis de la matriz, se observa que el modelo tiene buen desempeño en la clasificación de algunas categorías, como ACC. CON VEHÍCULOS, AGRESIONES DE PERSONAS y CONTACTO CON CORTO-PUNZANTES. Sin embargo, en otras categorías, el modelo comete errores de manera sistemática. Por ejemplo, en la clase FUERZA EXCESIVA, el modelo tiende a etiquetar incorrectamente como FALSO MOVIMIENTO Y SOBREESFUERZO, mientras que EXPLOSIONES O INCENDIOS se clasifica frecuentemente como CONTACTO CON CALOR/FRÍO, y PISADAS como CAÍDAS. Estos resultados sugieren que, aunque el modelo es capaz de codificar correctamente ciertos mecanismos, presenta fallos consistentes en otros. Esto abre la posibilidad de mejorar el desempeño ajustando la forma en que se presentan los datos al modelo para su clasificación.

Se analizó el código empleado para la codificación automática con gpt-3.5-turbo y se observó que utiliza una metodología de one-shot prompting. Esta técnica consiste en proporcionar un único ejemplo al modelo antes de realizar la tarea. Con el objetivo de mejorar los resultados, se propuso e implementó una solución basada en técnicas más avanzadas de prompting y procesamiento de texto, específicamente el few-shot prompting, que implica suministrar múltiples ejemplos al modelo antes de ejecutar la clasificación.

Modelo	**F1**	**Precision**	**Recall**
gpt-3.5-turbo	0.51	0.59	0.56
gpt-3.5-turbo + 1 Similar Example	0.66	0.73	0.67
gpt-4o + 1 Similar Example	0.70	0.69	0.81
gpt-4o + 2 Similar Examples	0.60	0.77	0.62
gpt-4o + 3 Similar Examples	0.60	0.75	0.62

Cuadro 5: Resultados de clasificación utilizando diferentes modelos y configuraciones.

Para optimizar aún más los resultados, se realizó un procesamiento adicional de texto en la base de datos. Este proceso consistió en calcular la similitud semántica a lo largo de la base de datos para identificar los n ejemplos más relevantes para cada entrada (input), donde n representa la cantidad de ejemplos utilizados en el few-shot prompting. Los resultados obtenidos, presentados en el cuadro 5, muestran las métricas de F1-weighted para la tarea de clasificación de mecanismos. En este análisis, el modelo denominado gpt-3.5-turbo corresponde al actualmente implementado en la ACHS para la clasificación de mecanismos, agentes y causas externas, mientras que el modelo gpt-3.5-turbo + 1 Similar Example incorpora las dos técnicas propuestas. Como se observa, este último modelo supera significativamente al que se encuentra en producción.

Adicionalmente, se evaluó una versión más reciente del modelo de OpenAI, gpt-4o, aplicando también la versión n Similar Examples. Los resultados muestran que esta configuración mejora considerablemente los desempeños anteriores, en donde especificamente el mejor resultado en general lo tiene la versión gpt-4o + 1 Similar Example, sugiriendo que la actualización a este modelo y añadir 1 ejemplo similar podría mejorar sustancialmente la precisión de las clasificaciones realizadas.

8. Recomendaciones para Sistema de Seguridad y Salud en el Trabajo

8.1. Implementación de Modelos Predictivos Personalizados

La incorporación de modelos predictivos personalizados en el ámbito de la seguridad y salud en el trabajo puede ser fundamental para mejorar la prevención de accidentes y enfermedades laborales. Estos modelos deben adaptarse a las características específicas de cada empresa y sector, considerando la información proporcionada por las mutualidades. De este modo, se podrían anticipar riesgos de manera más precisa y actuar de forma proactiva para minimizar los incidentes.

Los modelos predictivos pueden analizar patrones históricos de accidentes y enfermedades, identificando variables clave que aumentan la probabilidad de incidentes. La personalización de estos modelos según las particularidades de cada sector laboral permitirá a las organizaciones tomar decisiones informadas basadas en sus propios datos. Es fundamental que estos sistemas no solo detecten riesgos, sino que también sugieran acciones preventivas adaptadas a cada contexto, optimizando así los recursos y maximizando la efectividad de las medidas implementadas.

8.2. Promoción de Estándares Lingüísticos en la Documentación de Accidentes

La calidad y consistencia en la documentación de accidentes laborales es clave para una adecuada gestión de riesgos y para la mejora continua en los sistemas de seguridad y salud. En este sentido, es crucial establecer estándares lingüísticos claros en la redacción de los informes de accidentes. Esto permitirá una mayor uniformidad y calidad en la información recopilada, lo que facilitará su análisis posterior, tanto en términos de prevención como de seguimiento.

Una de las acciones más efectivas es la creación de bases de datos etiquetadas manualmente, donde se clasifiquen de manera precisa los diferentes tipos de accidentes, lesiones, causas y contextos asociados. Estas bases de datos servirían no solo para entrenar modelos predictivos, sino también para asegurar la calidad de la información procesada, lo que redundará en la mejora de la toma de decisiones. El uso de terminología estándar y bien definida en guías de anotación garantizará que los datos sean comparables y entendidos de manera homogénea a lo largo del tiempo, lo que facilitará el análisis y la implementación de medidas correctivas en los sistemas de seguridad y salud laboral.

9. Conclusiones

El desarrollo e implementación de herramientas de inteligencia artificial basadas en modelos de lenguaje preentrenados MLP ofrece una oportunidad única para transformar la gestión de textos clínicos en el ámbito de la salud ocupacional. Este trabajo permitió avanzar significativamente en tres áreas claves: la creación de modelos en el estado del arte adaptados al contexto chileno, la construcción de un corpus anotado y balanceado, y su integración en sistemas predictivos para la codificación automática de entidades relevantes.

Entre los principales logros destaca el diseño de múltiples versiones de MLP, utilizando metodologías de preentrenamiento continuo y desde cero en un corpus clínico especializado. Los experimentos de evaluación demostraron que estos modelos son altamente efectivos para tareas de reconocimiento de entidades nombradas en español. Asimismo, la incorporación de técnicas avanzadas de prompting, como el few-shot prompting con modelos más recientes como GPT-4, permitió mejorar significativamente el desempeño en las tareas de codificación automática, superando al sistema actualmente implementado en la ACHS.

Un aspecto clave de este trabajo ha sido la creación de un corpus cuidadosamente construido, anotado y balanceado, que captura diversidad en la estructura y complejidad de los datos generados por la ACHS. Este recurso no solo fue fundamental para el entrenamiento y evaluación de los modelos, sino que también sienta las bases para futuras investigaciones y desarrollos en el ámbito del procesamiento del lenguaje natural en español, especialmente en dominios clínicos y de salud laboral.

Este estudio demuestra cómo la incorporación de herramientas avanzadas de PLN puede aportar soluciones innovadoras en la automatización y mejora de procesos críticos en la salud ocupacional, optimizando la calidad y eficiencia en la gestión de datos. Los resultados obtenidos refuerzan la importancia de adaptar tecnologías globales al contexto local, donde el uso del español y las particularidades de los datos clínicos chilenos se convierten en elementos centrales para crear soluciones que verdaderamente impacten en el bienestar de las personas trabajadoras. Este esfuerzo constituye un paso decisivo hacia la implementación de sistemas de inteligencia artificial que estén plenamente alineados con los objetivos de la ACHS y las necesidades del sistema de salud.

10. Referencias

[1] Emily Alsentzer et al. “Publicly Available Clinical BERT Embeddings”. En: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, jun. de 2019, págs. 72-78. doi: 10.18653/v1/W19-1909. URL: https://aclanthology.org/W19-1909 (visitado 10-10-2022).

[2] Pablo Báez et al. “Automatic Extraction of Nested Entities in Clinical Referrals in Spanish”. En: *ACM Transactions on Computing for Healthcare (HEALTH)* 3.3 (2022), págs. 1-22. URL: https://dl.acm.org/doi/abs/10.1145/3498324.

[3] Pablo Báez et al. “The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish”. En: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Online: Association for Computational Linguistics, nov. de 2020, págs. 291-300. doi: 10.18653/v1/2020.clinicalnlp-1.32. URL: https://aclanthology.org/2020.clinicalnlp-1.32 (visitado 13-10-2022).

[4] Leonardo Campillos-Llanos et al. “A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine”. En: *BMC Medical Informatics and Decision Making* 21.1 (feb. de 2021), pág. 69. doi: 10.1186/s12911-021-01395-z. URL: https://doi.org/10.1186/s12911-021-01395-z (visitado 17-10-2022).

[5] José Cañete et al. “Spanish pre-trained BERT model and evaluation data”. En: *Practical ML for Developing Countries Workshop @ ICLR 2020*. 2020. URL: https://pml4dc.github.io/iclr2020/program/pml4dc_10.html (visitado 12-10-2022).

[6] Casimiro Pio Carrino et al. “Pretrained Biomedical Language Models for Clinical NLP in Spanish”. En: *Proceedings of the 21st Workshop on Biomedical Language Processing*. Dublin, Ireland: Association for Computational Linguistics, mayo de 2022, págs. 193-199. doi: 10.18653/v1/2022.bionlp-1.19. URL: https://aclanthology.org/2022.bionlp-1.19 (visitado 29-09-2022).

[7] Alexis Conneau et al. “Unsupervised Cross-lingual Representation Learning at Scale”. En: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, jul. de 2020, págs. 8440-8451. doi: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747 (visitado 13-10-2022).

[8] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. En: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, jun. de 2019, págs. 4171-4186. doi: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423 (visitado 10-10-2022).

[9] Dumitru Erhan et al. “Why Does Unsupervised Pre-training Help Deep Learning?” En: *Journal of Machine Learning Research* 11.19 (2010), págs. 625-660. URL: http://jmlr.org/papers/v11/erhan10a.html (visitado 05-10-2022).

[10] Aitor Gonzalez-Agirre et al. “PharmaCoNER: Pharmacological Substances,

Compounds and proteins Named Entity Recognition track". En: *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*. Hong Kong, China: Association for Computational Linguistics, nov. de 2019, págs. 1-10. doi: 10.18653/v1/D19-5701. URL: https://aclanthology.org/D19-5701 (visitado 17-10-2022).

[11] Asier Gutiérrez-Fandiño et al. "MarIA: Spanish Language Models". En: *Procesamiento del Lenguaje Natural* 68.0 (mar. de 2022). Número: 0, págs. 39-60. ISSN: 1989-7553. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405 (visitado 12-10-2022).

[12] Pengcheng He, Jianfeng Gao y Weizhu Chen. "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing". En: *arXiv*, 2021. DOI: 10.48550/ARXIV.2111.09543. URL: https://arxiv.org/abs/2111.09543.

[13] Pengcheng He et al. "DeBERTa: Decoding-enhanced BERT with Disentangled Attention". En: *International Conference on Learning Representations*, 2021.

[14] Jeremy Howard y Sebastian Ruder. "Universal Language Model Fine-tuning for Text Classification". En: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, jul. de 2018, págs. 328-339. DOI: 10.18653/v1/P18-1031. URL: https://aclanthology.org/P18-1031 (visitado 05-10-2022).

[15] Kexin Huang, Jaan Altosaar y Rajesh Ranganath. "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission". En: *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020. URL: https://www.chilconference.org/2020/workshop_20W13.html (visitado 10-10-2022).

[16] Alistair E. W. Johnson et al. "MIMIC-III, a freely accessible critical care database". En: *Scientific Data* 3.1 (mayo de 2016). Número: 1. Editorial: Nature Publishing Group, pág. 160035. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.35. URL: https://www.nature.com/articles/sdata201635 (visitado 10-10-2022).

[17] Daniel Jurafsky y James H. Martin. *Speech and Language Processing*. Dic. de 2021. URL: https://web.stanford.edu/~jurafsky/slp3/.

[18] Guokun Lai et al. "RACE: Large-scale ReAding Comprehension Dataset From Examinations". En: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, sep. de 2017, págs. 785-794. DOI: 10.18653/v1/D17-1082. URL: https://aclanthology.org/D17-1082 (visitado 09-10-2022).

[19] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". En: *Bioinformatics* 36.4 (feb. de 2020), págs. 1234-1240. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz682. URL: https://doi.org/10.1093/bioinformatics/btz682 (visitado 10-10-2022).

[20] Salvador Lima Lopez et al. "NUBes: A Corpus of Negation and Uncertainty in Spanish Clinical Texts". En: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, mayo de 2020, págs. 5772-5781. ISBN: 979-10-95546-34-4. URL:

https://aclanthology.org/2020.lrec-1.708 (visitado 17-10-2022).

[21] Yinhan Liu et al. “RoBERTa: A robustly optimized BERT pretraining approach”. En: *arXiv preprint arXiv:1907.11692* (2019). URL: https://arxiv.org/abs/1907.11692.

[22] Guillermo López-García et al. “Transformers for Clinical Coding in Spanish”. En: *IEEE Access* 9 (2021), págs. 72387-72397. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3080085.

[23] Antonio Miranda-Escalada, Eulàlia Farré y Martin Krallinger. *Cantemist corpus: gold standard of oncology clinical cases annotated with CIE-O 3 terminology*. Versión: 1.6. Tipo: dataset. Abr. de 2020. DOI: 10.5281/ZENODO.3773228. URL: https://zenodo.org/record/3773228 (visitado 17-10-2022).

[24] Alec Radford et al. “Improving Language Understanding by Generative Pre-Training”. En: (2018).

[25] Pranav Rajpurkar, Robin Jia y Percy Liang. “Know What You Don’t Know: Unanswerable Questions for SQuAD”. En: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, jul. de 2018, págs. 784-789. DOI: 10.18653/v1/P18-2124. URL: https://aclanthology.org/P18-2124 (visitado 09-10-2022).

[26] Pranav Rajpurkar et al. “SQuAD: 100,000+ Questions for Machine Comprehension of Text”. En: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, nov. de 2016, págs. 2383-2392. DOI: 10.18653/v1/D16-1264. URL: https://aclanthology.org/D16-1264 (visitado 09-10-2022).

[27] Matías Rojas, Jocelyn Dunstan y Fabián Villena. “Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing”. En: *Proceedings of the 4th Clinical Natural Language Processing Workshop*. Seattle, WA: Association for Computational Linguistics, jul. de 2022, págs. 87-92. DOI: 10.18653/v1/2022.clinicalnlp-1.9. URL: https://aclanthology.org/2022.clinicalnlp-1.9 (visitado 29-09-2022).

[28] Jörg Tiedemann. “Parallel Data, Tools and Interfaces in OPUS”. En: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*. Istanbul, Turkey: European Language Resources Association (ELRA), mayo de 2012, págs. 2214-2218. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf (visitado 12-10-2022).

[29] Ashish Vaswani et al. “Attention is All you Need”. En: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (visitado 05-10-2022).

[30] Shijie Wu y Mark Dredze. “Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT”. En: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, nov. de 2019, págs. 833-844. DOI: 10.18653/v1/D19-1077. URL: https://aclanthology.org/D19-1077 (visitado 12-10-2022).

[31] Yukun Zhu et al. “Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books”. En: *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), págs. 19-27.