ANEXO N°45
INFORME FINAL

Serie Proyectos de Investigación e Innovación

Superintendencia de
Seguridad Social Santiago -
Chile

"Implementación de una taxonomía de clasificación de causas externas de los accidentes por medio de automatización con machine learning utilizando la información del relato de la denuncia"

INFORME FINAL

Autor:
José Miguel Tobar
Ríos

Año de
publicación
2025

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 e innovación en Prevención de Accidentes y Enfermedades del Trabajo financiados por los recursos del Seguro 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 the Prevention 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. The content, analysis and conclusions are solely the responsibility of the author (s), and do not necessarily reflect 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, please visit our website: www.suseso.cl.

Superintendencia de Seguridad
Social Huérfanos 1376
Santiago,
Chile.

“Implementación de una taxonomía de clasificación de causas externas de los accidentes por medio de automatización con machine learning utilizando la información del relato de la denuncia”

José Miguel Tobar Ríos

Julio 2025

Tabla de contenido

I. Introducción

En el contexto actual de transformación digital en salud y seguridad laboral, la automatización de procesos analíticos sobre información no estructurada representa una oportunidad estratégica para mejorar la prevención de accidentes laborales. Uno de los principales desafíos radica en traducir relatos médicos libres (ricos en detalle y subjetividad) en datos estructurados que permitan una codificación estandarizada, comprensible y útil para sistemas de análisis y toma de decisiones.

Este informe presenta el desarrollo e implementación de un sistema de clasificación automatizada de causas externas de accidentes laborales, basado en la integración de taxonomías médicas y modelos avanzados de procesamiento del lenguaje natural. La iniciativa se enmarca dentro de una línea de innovación orientada a fortalecer las capacidades institucionales de análisis preventivo mediante inteligencia artificial, específicamente a través del uso de Grandes Modelos de Lenguaje (LLMs).

El proyecto responde a una necesidad concreta: establecer un mecanismo escalable y confiable que permita clasificar automáticamente relatos médicos provenientes de denuncias de accidentes laborales, manteniendo estándares de precisión, coherencia semántica y representatividad contextual. Para ello, se evaluó la aplicabilidad de diversas taxonomías internacionales, y se diseñó una arquitectura que combina modelos de similitud semántica, agentes especializados y procesos de validación iterativa. La taxonomía seleccionada —CILCE— fue adaptada e implementada como eje central del sistema, dadas sus ventajas estructurales para la codificación multiaxial y su capacidad de representar con granularidad las circunstancias de cada accidente.

El presente documento expone la fundamentación teórica, las decisiones metodológicas y los resultados obtenidos, destacando el valor de esta propuesta no solo como un avance técnico, sino como un insumo estratégico para futuras acciones en prevención, análisis epidemiológico y mejora continua en el ámbito de la salud laboral.

II. Resumen

El presente informe explora la implementación de un sistema automatizado para la detección y clasificación de accidentes laborales, a partir de relatos médicos, utilizando grandes modelos de lenguaje (LLMs) y la taxonomía CILCE. El objetivo central fue identificar una estructura taxonómica adecuada —en términos de granularidad y contextualización— y evaluar la capacidad de los LLMs para realizar tareas de codificación clínica de manera eficiente y precisa.

Se trabajó con relatos médicos recolectados mediante API durante julio de 2024, correspondientes al promedio mensual de registros de una mutual de salud laboral. De un universo de 15.000 relatos, se extrajo una submuestra aleatoria de 1.000 casos, y una muestra reducida de 100 relatos fue utilizada para el etiquetado manual, orientado al entrenamiento y evaluación preliminar del sistema.

El sistema se basó en una arquitectura que integra modelos de similitud semántica, embeddings vectoriales e interacciones multiagente para mejorar tanto el etiquetado como la validación iterativa. Los resultados obtenidos, con una precisión del 75,5%, demuestran la eficacia de los LLMs en la comprensión contextual, selección de etiquetas y coherencia semántica, incluso en entornos clínicos complejos.

Se destaca el valor estructural de la taxonomía CILCE, diseñada con múltiples ejes (lugar, ocupación, actividad, transporte e intención), lo cual permite una codificación granular y situacional del accidente, superando a otras clasificaciones convencionales. Este enfoque no solo facilita la clasificación automatizada, sino que habilita el análisis preventivo y epidemiológico mediante la transformación de relatos en datos estructurados, explotables por sistemas de inteligencia artificial y business intelligence.

En definitiva, el proyecto demuestra que la combinación de taxonomías especializadas como CILCE, modelos de lenguaje de última generación y arquitecturas semánticas avanzadas representa un camino prometedor para la automatización de tareas críticas en salud y seguridad laboral. Además, establece una base sólida para desarrollar sistemas predictivos y estrategias preventivas basadas en datos clínicos codificados.

III. Objetivos

III.I. Objetivo General

- Desarrollar el proceso para definir la taxonomía más adecuada para clasificar las causas externas de los accidentes, y que sea escalable a un modelo que clasifique automáticamente la información contenida en el relato de la denuncia.

III.II. Objetivos Específicos

- Revisar las taxonomías existentes, y su aplicabilidad a los datos de Mutual de Seguridad.
- Validar las clasificaciones en muestras de relatos de accidentes en base a criterio experto.
- Implementar un modelo automatizado de clasificación de causas externas de accidentes.
- Evaluar y validar resultados del modelo.
- Entregar recomendaciones para la implementación y aplicación del modelo para mejorar la estrategia preventiva de Mutual de Seguridad.

IV. Discusión sobre Taxonomías

La clasificación de las causas externas en accidentes laborales constituye un elemento fundamental en el análisis y resolución de casos relacionados con este tipo de eventos, donde la vasta cantidad de información disponible sobre el tema representa un significativo avance en la identificación de agentes causales y factores contextuales, permitiendo una comprensión más profunda de las circunstancias reportadas en las denuncias realizadas. Este enfoque no solo contribuye al entendimiento de las situaciones específicas, sino que también facilita la asignación de códigos precisos a los relatos, optimizando así la documentación y la resolución integral de casos. Durante las últimas décadas se han desarrollados distintos enfoques para en análisis de accidentes de causas externas, dentro de las cuales destacan las siguientes taxonomías médicas: ICD-10 (también CIE-10 en español), la clasificación internacional de enfermedades de la décima edición (con apartado para clasificación por tipo de accidente); la taxonomía de agente material y forma del accidente de la OIT; ICECI, la Clasificación Internacional de Lesiones de Causa Externa (también CILCE, en español) perteneciente a WHO family of international classification; NOMESCO del Nordic Medico Statistical Committee; y las dos últimas versiones de ICD, es decir, ICD-10-CM e ICD-11.

La taxonomía ICD-10 de la World Health Organization (WHO) fue creada en el año 1992 y su estructura es alfanumérica, compuesta por letras y números, los cuales tienen una relación directa con los capítulos, categorías y subcategorías de clasificación. Por ejemplo, el código H04.3, donde la "H" hace referencia al capítulo "Enfermedades del ojo y sus anexos", el n° 4 indica la categoría "Trastornos del aparato lagrimal" y el n° 3 señala la subcategoría de "Inflamación aguda y la no especificada de las vías lagrimales" (Organización Mundial de la Salud, 2008, p. 404).

El propósito principal de la CIE-10 es, por tanto, ofrecer una clasificación completa de enfermedades, síndromes y afecciones médicas. Sin embargo, la CIE-10 también incorpora información sobre causas

externas de lesiones en su capítulo XX, lo que extiende su utilidad a la codificación de accidentes, envenenamientos, agresiones y otros eventos externos que afectan la salud de las personas. Este aspecto representa una ventaja significativa, ya que permite una cierta caracterización de eventos no patológicos que también tienen impacto en el estado de salud.

No obstante, cuando se trata de codificar relatos detallados de accidentes, la CIE-10 presenta limitaciones importantes. Si bien clasifica el tipo de lesión, el agente causante, el lugar del evento y el mecanismo de ocurrencia (por ejemplo, caída, golpe, quemadura, etc.), no logra capturar con precisión otros elementos contextuales fundamentales que suelen estar presentes en los relatos narrativos, como:

- La intención del acto (accidental, autoinfligido, por negligencia de terceros).
- La secuencia de eventos que condujeron al accidente.
- Las condiciones ambientales o sociales implicadas.
- La actividad que realizaba la persona en el momento del accidente.
- Detalles relacionados con la persona involucrada, como su rol (trabajador, peatón, ciclista), o si hubo factores de riesgo específicos (falta de señalización, condiciones laborales inseguras, uso de sustancias, etc.)

Esta falta de granularidad puede conducir a una pérdida de información relevante al momento de transformar relatos ricos en detalles en códigos estadísticos rígidos. De hecho, en situaciones donde se busca comprender mejor la causalidad, la prevención o los factores modificables de un accidente, la CIE-10 no ofrece herramientas suficientes. Su estructura fue diseñada principalmente para diagnosticar y clasificar enfermedades médicas, no para describir en profundidad la complejidad de los eventos accidentales. Aunque la CIE-10 sigue siendo una herramienta indispensable en la práctica clínica y estadística global, su utilidad para la codificación de relatos de accidentes es limitada. Cuando se requiere una descripción más fiel, completa y contextual de un evento lesivo, resulta necesario recurrir a sistemas complementarios que capturen la riqueza narrativa y la multidimensionalidad de estos casos.

En continuidad con esta discusión bibliográfica, el ICD-10-CM también posee una estructura alfanumérica, compuesta por capítulos, categorías y subcategorías, pero con el agregado de apartados vinculados con la modificación clínica de esta versión. La adaptación tiene como propósito el entregar una clasificación más exhaustiva para los diagnósticos clínicos. En su base utiliza las taxonomías de ICD-10, ahora bien ingresa mayores detalles debido a ser una edición CM, es decir, que indica modificaciones clínicas.

Esta versión en particular surge y es utilizada por Estados Unidos, cuya modificaciones también son más específicas, por lo que existen ciertos riesgos en que no pueda ser aplicada en su totalidad internacional. Tal como se indica en el foro de resolución de consultas sobre las clasificaciones internacionales de la Organización Panamericana de la Salud: “la “CIE-10 CM”, adopta un cambio de clasificación basándose en la CIE-10, pero además requiere del conocimiento y manejo de las convenciones y normas que la rigen, una mayor exigencia de conocimientos en anatomía y terminología quirúrgica y, sobre todo, contar con la documentación clínica necesaria” (Organización Panamericana de la Salud, 2018, párr. 5), por lo que, tal como se mencionó, esta versión comprende una codificación mucho más profunda que la versión original.

Luego, el ICD-11, posee la misma estructura de las taxonomías base de las anteriores. Para este caso, la clasificación de enfermedades y problemas de salud se ha digitalizado y, además, se cuenta con un buscador web que ayuda a determinar de forma rápida la codificación para un caso específico de uso, por ejemplo con el ingreso de una palabra clave de la enfermedad la página contiene bloques en los que se arroja el código o los códigos más apropiados según el caso (OMS, 2022). Además, en esta versión la ICD-11 implementó codificaciones agrupadas o clústeres, lo que genera un alto potencial en el ámbito de clasificación automatizada de códigos.

Por un lado, esto puede ser completamente positivo, ya que la documentación, que antes requería una búsqueda exhaustiva y lenta debido a su minuciosidad, ahora está disponible de manera más accesible en formato digital. No obstante, una desventaja podría ser que al ser una transición bastante diferente en términos de desafíos digitales (capacitación, accesos, entre otros factores) a las documentaciones anteriores de ICD-10, podría ocurrir el caso de que la implementación sea lenta o compleja en sus fases iniciales debido a la necesidad de capacitaciones, uso de internet, entre otros.

Tras la revisión de las versiones iniciales y posteriores de ICD-10, a continuación se revisará la Clasificación Internacional de Lesiones de Causa Externa (ICECI), cuya estructura se caracteriza por entregar una codificación multiaxial, modular y alfanumérica. En ese sentido, los códigos de la clasificación se rigen por ejes que se conforman de un modo relacional. Además, al contar con la estructura modular, la organización se vuelve más expedita, puesto que los relatos de los pacientes pueden ser abordados con mayor orden. Luego, en el ámbito de la forma en que se estructuran los códigos, estos al ser alfanuméricos, permiten, más que una mejor búsqueda, una mejor clasificación, puesto que los casos tienen una letra y número asociados según el tipo de módulo en el que ingresen, por ejemplo, el código C4 siempre tiene relación con el lugar donde se encontraba la persona en el momento en que se efectuó su lesión de causa externa.

Uno de los principales propósitos de ICECI, es lograr una clasificación lo más íntegra posible para la codificación de las lesiones. Esto con tal de incidir en la prevención de las afectaciones ocasionadas, pero principalmente para entender todo el escenario del accidente y la performance de la víctima en cuestión. Con el fin de lograr su propósito es que implementó un sistema modular donde se puede clasificar en detalle el contexto del accidente. Para eso se utilizan los siguientes módulos: el módulo básico; el módulo de violencia; el módulo de transporte; el módulo de lugar; el módulo de deportes; y el módulo ocupacional, apartados que, a su vez, cumplen con ejes que permiten la consideración de elementos necesarios para clasificar el accidente.

Dentro de sus principales ventajas se encuentra su extensión en la estandarización de los códigos en el ámbito internacional, lo que indica una investigación exhaustiva de los posibles casos de lesiones. Al mismo tiempo, este formato permite una amplia adaptabilidad no solo en la implementación en un centro de salud, sino también en el caso de una adaptación a otra taxonomía (caso de NOMESCO); ahora bien, esta ventaja también puede volverse todo lo contrario debido a la especificidad de los códigos, puesto que requiere una capacitación previa al uso, esto debido a la complejidad de las agrupaciones, terminología médica y, por supuesto, por el riesgo que implica la asertividad y responsabilidad ética en la codificación.

Más tarde, otras de las taxonomías propuestas a revisar es NOMESCO Classification of external causes of injuries (NCECI), la que posee una clasificación multiaxial por módulos y alfanumérica, en

la cual, además, se pueden encontrar descripciones más profundas de ciertos códigos. En ese sentido, es un buen manual para el registro de causas externas de accidente no intencionales.

Su propósito es la búsqueda de prevención de muertes evitables, clasificación de las lesiones de mayor gravedad, aunque no precisamente mortales y también los casos de incapacitación duradera tras las lesiones. Además, la presente taxonomía busca indicar los daños o pérdidas de los materiales que se asocian a los causantes principales de las respectivas lesiones, puesto que también pueden repercutir en una gravedad mayor, es decir, en un problema social (Nordic Medico-Statistical Committee, 2007, p. 9).

Debido a que NOMESCO mantiene una relación directa con ICECI, ambas clasificaciones comparten un enfoque estructurado y detallado en torno a la codificación de lesiones y sus circunstancias. NOMESCO organiza su sistema mediante códigos que incluyen apartados específicos, similares a los de ICECI, los cuales abarcan desde secciones hasta módulos que permiten registrar información contextual relevante sobre la situación en que ocurrió la lesión. Ejemplos de estos apartados incluyen: accidente de transporte, accidente vehicular, actividad industrial, práctica deportiva o autolesión intencional. La relación entre ambas clasificaciones no se limita a su estructura, sino que también se extiende a sus perspectivas complementarias. Mientras ICECI se enfoca en describir las circunstancias externas y causas del evento lesivo (como el lugar, actividad, mecanismo y agente implicado), NOMESCO aporta información desde el ámbito clínico y de intervención médica, permitiendo registrar los procedimientos realizados tras el accidente. Esta complementariedad resulta especialmente útil en estrategias de prevención y análisis epidemiológico, ya que permite correlacionar el tipo de intervención médica con el contexto causal del accidente, facilitando la identificación de patrones y la planificación de acciones preventivas centradas en los tipos de eventos más recurrentes o de mayor impacto.

En uno de sus apartados introductorios, se presenta el método de Mecanismos de lesión, mediante el recurso de preguntas clave, las cuales son tres. En primer lugar, “¿Cuál era la actividad de la víctima?”, es decir, si se encontraba cocinando, corriendo, caminando. Esto ayuda a clasificar la acción del paciente a modo previo del accidente. En segundo lugar, “¿Cuál fue el problema que ocurrió?”, pregunta que apunta a la “desviación”, pues busca dar respuesta al motivo por el cual la víctima no pudo continuar realizando su actividad, qué fue lo que desvió su actividad inicial, por lo que, como se indica en el documento (Nordic Medico-Statistical Committee, 2007, p. 16), la presión de una olla, el calentamiento excesivo de una mezcla química, entre otros factores sería la respuesta prototípica. En tercer lugar, “¿Cómo ocurrió la lesión?” busca resolver el detalle exacto de cómo y con qué elemento la víctima se vio afectada: olla, cuchillo, entre otros, y de qué modo. En ese sentido, este método no solo es útil para la clasificación general de las especificaciones por accidentes, sino que, por ejemplo, para casos en que haya existido violencia de terceros, la información se vuelve mucho más íntegra.

Este método aporta un valor importante, puesto que contiene una especie de narración de los hechos, es decir, indica una secuencia de acciones, desde una actividad inicial, un factor detonante y una consecuencia final, lo que es muy útil puesto que es lo más fidedigno a la ocurrencia de la lesión.

Una de las diferencias más destacables entre ICECI y NOMESCO radica en su alcance geográfico y propósito institucional. ICECI fue diseñada con una visión global, promovida por la OMS para facilitar la descripción, medición y monitoreo de lesiones en diversos contextos y regiones del mundo.

En contraste, NOMESCO se originó en los países nórdicos, con la primera edición publicada en 1984 y revisiones sucesivas en 1990, 1997 y 2007, adaptada específicamente para las necesidades de esos países. Aunque NOMESCO es técnicamente exportable y ha sido presentado en conferencias internacionales, su principal intención es estandarizar datos y promover la comparabilidad estadística en los países nórdicos, facilitando la generación de informes sobre lesiones por causas externas dirigidos por los sectores de prevención en la región. Por otra parte, además, la ICECI tiende a brindar mayor especificidad en sus descriptores de códigos, lo que se traduce como una detección fina de casos de accidentes de cada paciente; en cambio, NOMESCO, tiene un enfoque ligado a la estandarización de datos para reportes estadísticos sobre las lesiones de causa externa. Por otro lado, ICECI se encuentra disponible de versiones originales, en inglés, una en francés como borrador y la versión en español está bajo discusión, según se indica en la página oficial de la World Health Organization; Nomesco en la versión NCECI se encuentra en inglés.

Por último, la taxonomía Forma del accidente y Agente material del accidente (1997) son parte de la Organización Internacional del Trabajo. Según el *Registro y notificación de accidentes del trabajo y enfermedades profesionales*, el término agente: “puede utilizarse para clasificar los accidentes de trabajo ya sea según el agente material en relación con la lesión o según el agente material en relación con el accidente” (Organización Internacional del Trabajo, 1996, p. 69), en ese sentido, agente sería el elemento causante del accidente; mientras que forma: “refiere a las características del acontecimiento que ha tenido como resultado directo la lesión, es decir, la manera en que el objeto o la sustancia en cuestión ha entrado en contacto con la persona afectada.” (Organización Internacional del Trabajo, 1996, p.67), por lo que apunta al modo en que el agente incidió en el accidente.

Ahora bien, respecto a su estructura, esta destaca por ser capitular (anexos) cuyas agrupaciones inician desde una categoría mayor a una más específica, la cual posee un nombre y un código numérico. Por ejemplo el código "1.1.2 Máquina de combustión interna" del anexo I, donde el número 1.1. corresponde a la subcategoría "Generadores de energía, excepto motores eléctricos" y el 1. Hace referencia a "Máquinas" y el Anexo como tal a la Clasificación de accidentes de trabajo según agente material. En ese sentido, esta estructura brinda un aporte específico respecto del agente, lo que también puede verse aplicado en el caso del Anexo H: "Clasificación de los accidentes de trabajo según la forma del accidente".

Es importante destacar las ventajas y desventajas presentes en la totalidad de las taxonomías revisadas. Por un lado, la amplia cantidad de documentación existente en la actualidad brinda un gran aporte en la clasificación de lesiones de causa externa, puesto que también refleja el gran trabajo e investigación respecto de la dimensión de los tipos de accidentes posibles. Además, el hecho de que diversos países y comités hayan o estén trabajando en las versiones de taxonomías mejoradas y mayormente específicas, quizá por país o por regiones, también pueden ser probablemente aplicables a otros espacios, y el tener una investigación que aporte con códigos específicos ya preparados, indica la escalabilidad de las taxonomías en cuestión.

Por otro lado, esta cualidad de diversidad puede ralentizar los procesos de agilidad en la codificación, puesto que la lectura de los relatos de pacientes junto a la asociación de códigos y descriptores acertados no es una tarea sencilla y corta, mucho menos si se dispone de una amplia gama de taxonomías. Además, debido a la cantidad relevante de información a codificar, existe la constante demanda por actualizar los documentos. Por estos motivos, es menester buscar el medio más automatizable para la selección de las taxonomías más aptas según los casos médicos, con la finalidad

no solo de codificar y entregar soluciones efectivas, sino también de buscar un proceso expedito en que existe un balance entre rapidez y calidad.

Un nuevo método que se ha buscado incluir en el proceso de clasificación de códigos para los registros médicos es el del proceso automatizado, el que contemple la utilización de herramientas sustentadas bajo inteligencia artificial, en la cual los grandes modelos de lenguaje (LLMs) cumplen un rol fundamental. Casos como estos se pueden ver evidenciados en el artículo “Do Large Language Models understand Medical Codes?” (Lee & Lindsey, 2024) por Simon A. Lee y Timothy Lindsey, donde se utilizan LLMs para lograr codificar casos médicos de taxonomías, dentro de las cuales destaca ICD. La investigación, si bien plantea que habría una agilidad operativa debido a la casi nula presencia de codificación manual que implicaría, este tipo de modelos en algunos casos, puede incurrir en alucinaciones. En estos casos, el modelo tiene a resolver utilizando su conocimiento generalizado, provocando así falsos positivos, es decir, asignando una clasificación como correcta cuando en realidad es errada. Esto puede expresarse en que el modelo etiquete con un código incorrecto, ya sea que no asigne ninguna etiqueta a un relato médico que sí debería tener una categoría, o que clasifique un caso con un categoría cuando el relato en sí no citicontieneene información suficiente para justificar, algo que, en una clasificación manual, probablemente no ocurriría.

Con el propósito de evitar este tipo de problemáticas, es fundamental que en la etapa previa, donde se procesan los datos y el modelo adquiere conocimiento, se realicen pruebas que cumplan con estrictos porcentajes mínimos en las métricas de validación. En el caso del artículo, los investigadores realizaron pruebas con diversos modelos de lenguaje para analizar cuál de estos respondía correctamente frente a la solicitud de la asociación de código y el descriptor de este: "Prompt: "Give me the corresponding Medical names to these Medical Codes"" (Lee & Lindsey, 2024, p. 4). Los resultados del estudio, si bien presentaron logros para los LLMs en su capacidad de grandes datos de conocimiento, presentan alucinaciones que pueden verse reflejadas como riesgos. Por lo que el trabajo concluye que la inserción de este tipo de automatizaciones sigue a la espera de mejoras en sus resultados, por lo que el desafío sigue presente. Ahora bien, aunque el uso de LLMs puede aportar una automatización significativa y agilidad operativa en cuanto a la codificación, también conlleva riesgos graves en cuanto a las alucinaciones en la asignación de códigos, por lo que frente a este tipo de desafíos, la literatura ofrece soluciones parciales tales como procesos de validación rigurosos, junto a pruebas previas a implementación; evaluación comparativa de modelos; supervisión y revisión humana; y, mejoras continuas según actualización de modelos o actualización de formato de data (relato médico) utilizada para la clasificación.

Otra de las investigaciones vinculadas con la automatización de las taxonomías es “Automated Generation of ICD-11 Cluster Codes for Precision Medical Record Classification” (Feng et al., 2024), en la cual se detalla un gran logro en ámbito de la precisión en la clasificación de registros médicos, la cual obtiene hasta un F1 de 0.91 de codificación. Lo anterior, ocurre mediante las nuevas agrupaciones de códigos (OMS, 2022) aportados por la ICD 11, los cuales son utilizados junto a corpus de textos médicos que se representan en lenguaje natural como vectores y la utilización de similitud de coseno para hallar la correspondencia en los códigos a clasificar (Feng et al., 2024). Esta investigación junto a la anterior son de las más actuales que permiten comprender el avance respecto a las automatizaciones en la codificación de registros médicos y, sobre todo, el aporte interdisciplinario que se puede lograr al realizar el vínculo entre las ciencias computacionales-matemáticas y las ciencias de la salud.

A modo cierre, se ha podido observar que a lo largo de las décadas han existido diferentes tipos de modos de clasificar la información recopilada por altas cantidades de registros médicos, dentro de las cuales se han creado taxonomías específicas para la codificación de enfermedades, otras para riesgos de accidentes y otras sobre lesiones específicas de causas externas. También se ha destacado la importante y extensa labor no solo en el registro médico, sino también el proceso de clasificación y, en consecuencia, de resolución de casos por paciente. Por ello, el cierre de este apartado presentó uno de los enfoques más actuales sobre automatizaciones para el proceso de codificación, pero con grandes modelos de lenguaje los cuales, por supuesto, se sustentan en base a la Inteligencia Artificial, lo que permite vislumbrar un prometedor camino frente al desarrollo de nuevas herramientas que logren conseguir resultados esperados en los que la calidad, rapidez y ética estén unificados.

IV.I. Alcances y desafíos entre taxonomías

Una de las principales virtudes de las taxonomías médicas es su estructura detallada y organizada, lo cual constituye una de sus características más valiosas. Este tipo de formato permite entregar la información de manera sistemática a los especialistas encargados de llevar a cabo procesos de codificación médica, facilitando así tareas complejas que requieren precisión y constancia. No obstante, esta misma especificidad también puede representar un desafío al momento de clasificar, ya que exige una comprensión profunda de relatos clínicos que, en muchos casos, involucran múltiples componentes como: agentes causales, formas de ocurrencia, o elementos que no se encuentran explícitamente indicados en el texto, lo que añade un nivel considerable de complejidad.

Taxonomías como CILCE, CIE-10 y el sistema de Agente y Forma de la Organización Internacional del Trabajo presentan características que las hacen especialmente adecuadas para el desarrollo de herramientas automatizadas destinadas a la codificación rápida y eficiente de relatos médicos. Sin embargo, también poseen particularidades que representan desafíos importantes para su implementación en sistemas basados en modelos de lenguaje o herramientas que emplean métricas de similitud textual. Por ejemplo, la amplitud de categorías en CILCE como: su módulo básico, intención, ocupación, violencia, transporte, deporte y lugar, o la polisemia presente en algunas categorías de la CIE-10, pueden dificultar la clasificación automática, requiriendo estrategias específicas de tratamiento y ajuste en dichos modelos.

A continuación, se detalla cómo fueron trabajadas las distintas taxonomías, utilizando como base una arquitectura sustentada en Grandes Modelos de Lenguaje (LLMs). El conjunto de datos fue proporcionado por la Mutual de Seguridad mediante el acceso a una API, que permitió extraer denuncias de accidentes de trabajo y trayecto correspondientes a junio de 2024. Estos datos se encontraban estructurados en formato tabular e incluían información relevante sobre los antecedentes de cada siniestro.

CILCE. Para el caso de la presente taxonomía, en una fase inicial solo se utilizó el módulo básico C, el cual incluía sus respectivos submódulos, cuyo *dataset* fue utilizado no solo para este caso sino también para las taxonomías posteriores. Cabe destacar que este sistema de categorías médicas es ideal para casos en que se clasifiquen accidentes de pacientes que deban recibir algún tipo de indemnización, porque incorpora bastante información sobre el caso. Si bien se comentó el módulo C, y este a su vez comprende submódulos, también existen otros, tales como el de transporte, violencia, entre otros, por lo que el relato puede ser clasificado de forma granular, la que puede ser

complementada con taxonomías tales como NOMESCO (para integrar el tipo de procedimiento quirúrgico posterior, por ejemplo) o CIE 10 para complementar o reforzar detalles de módulos tales como el de Lugar o el Básico, debido a la composición sintáctica de sus categorías organizadas en causantes, modo de ocurrencia y lugar.

CIE 10. Se trabajó con la lista de 2870 categorías que contemplan los agentes causantes, modo en que ocurrió el accidente y, en algunos casos, el lugar. Esta taxonomía es de gran utilidad para la clasificación de casos en que se busque la prevención futura de accidentes en el ámbito laboral, aunque también, como se comentaba en el punto anterior, puede complementarse con CILCE. Esto es debido a que, si bien entrega información relevante para clasificar un caso, su granularidad no es su principal característica, en comparación con la taxonomía anteriormente mencionada.

Agente y Forma. Para este caso se utilizaron dos archivos de la OIT, uno de agente y otro de forma, donde en cada uno se encontraban las listas numeradas de categorías posibles. Esta taxonomía posee características similares a la de CIE 10, pero su especificidad y, en consecuencia, granularidad de descripción es inferior y más condensada en cuanto a la categorización, pues en esta taxonomía se pueden observar categorías compuestas y simples, en lugar de solo compuestas, como la mayoría de otras taxonomías (una sola palabra u oraciones cortas), por lo que en lugar de recomendar su uso para casos de extrema urgencia, indemnización por parte de pacientes, su utilidad podría estar ligada con procesos de investigaciones iniciales de casos de prevención, en ese sentido, podría complementarse con CIE 10.

A continuación se despliega una tabla con los principales alcances de las iteraciones realizadas en el marco del presente proyecto, junto a sus principales desafíos. Se comparan las tres taxonomías trabajadas: CILCE, CIE 10 y Agente y Forma de OIT:

Tabla 1: Tabla de alcances y desafíos de las taxonomías médicas.

Taxonomía	Alcances	Ejemplo	Desafíos	Ejemplo
CILCE	• La alta cantidad de grupos del módulo C: Básico, que posee CILCE (intención, lugar, entre otros) pueden ser útiles debido a que representan **grandes opciones** a los modelos para que puedan escoger de una gama superior el caso que mejor pueda clasificar. • **Preciso** en términos de **agentes** e incluso en términos de **modos** en que se producen los accidentes.	**Relato:** TRABAJO HABITUAL. bodeguero. ID UNICO. 1234567. CENTRO ATENCION. centro de at. quilicura. LUGAR ACCIDENTE. camarin. QUE HACIA. boeguero. paciente estaba saliendose de la ducha.. COMO OCURRIO. paciente refiere que al salir de la **ducha, el piso s eencontraba mojado y paciente resbala, cae contra el suelo** (...) el dia de hoy 01/07/2024, siendo las 16:40, paciente refiere que al salir de la ducha en el trabajo, el piso se encontraba mojado y paciente resbala, cae contra el suelo golpeandose ambas manos. (...) conciente, orientado en las tres esferas de tiempo, lugar y persona, colaborador, con adecuada interaccion con el examinador, hemodinamicamente estable (...) **Clasificación automatizada** correcta en la mayoría de los submódulos: C1.1 **No intencional** C2.7.1 Sobre **esfuerzo agudo**, hiper extensión C3.98.98.98 Otro objeto/sustancia especificado C4.8.3 **Fábrica/planta** C5.1.8 **Otro trabajo remunerado** especificado C6.2 Sin sospecha o evidencia de uso de alcohol por cualquier persona	• La **diversidad** de módulos (básico, ocupacional, lugar, violencia) puede generar la existencia de casos en que los submódulos del módulo básico no logren coincidir entre agentes causales y modo en que ocurre un accidente. • **Impreciso** en cuanto a los lugares en que se producen los accidentes. • Ha sido la taxonomía de mayor complejidad, debido a la granularidad que desempeña respecto a su clasificación. Pues si bien esta taxonomía entrega mayor especificidad, este factor también puede indicar el aporte de datos que son similares al relato médico en cuestión, pero no exactamente el mismo	**Relato:** TRABAJO HABITUAL. cargas de helio. ID UNICO. 1234567. CENTRO ATENCION. centro de at. la florida. LUGAR ACCIDENTE. centro medico. QUE HACIA. cargas de helio; relata que al estar trabajando. COMO OCURRIO. al mover contenedor con ruedas por subida de autos, al trabarse hace fuerza para moverlo y siente crujido.. presenta dolor y dificultad de caminar. ANAMNESIS. alergias -patologias (...) **Clasificación automatizada** C1.1 No intencional C2.98.8 Otro mecanismo de lesión especificada, C2.98.8 Otro mecanismo de lesión especificado C3.11.01.25 Maquinaria de montacargas C4.8.3 Fábrica/planta C5.1.8 Otro trabajo remunerado especificado C6.2 Sin sospecha o evidencia de uso de alcohol por cualquier persona involucrada en el evento de la lesión

		involucrada en el evento de la lesión C7.2 **Sin sospecha o evidencia de uso de droga** u otra sustancia psicoactiva por cualquier persona involucrada en el evento relacionado con la lesión.		C7.2 Sin sospecha o evidencia de uso de droga u otra sustancia psicoactiva por cualquier persona involucrada en el evento relacionado con la lesión
CIE 10	• **Conceptos utilizados globalmente**, lo que facilita la comprensión de LLMs, debido al reconocimiento de altas cantidades de palabras. • Las categorías suelen contener **agentes causantes claros**, lugares en que se provocan los accidentes e incluso los modos, lo que otorga un contexto completo. • Taxonomía de complejidad media.	**Relato:** trabajo habitual. capitan, patron de nave. id unico. 1234567. centro atencion. puerto montt. lugar accidente. cubierta embarcacion. que hacia. capitan, patron de nave, refiere habian sufrido una pane hidraulica de la embarcacion. como ocurrio. menciona se habia reventado una manguera la cual habia salpicado aceite por toda la empbarcacion, tuvo que salir a la embarcacion (...) y cae sobre la baranda, sufriendo lesion en brazo y pecho lado der.. at de urgencia en hospital quellon,. anamnesis. am: niega **Clasificación automatizada** accidente en un **barco mercante**, sin accidente del **barco**, que no causa ahogamiento ni sumersión.	• Puede presentar **ambigüedad** y polisemia en sus categorías. • La clasificación jerárquica puede presentar dificultades de **organización** • Los agentes causantes y los modos en que se producen los accidentes están unidos en algunos casos y en otros no, lo que puede provocar **confusiones** a nivel de funcionalidad de los sintagmas. • Puede ser **impreciso** en cuanto a los lugares en que se producen los accidentes.	**Relato:** trabajo habitual. operaria. id unico. 1234567. centro atencion. hospital santiago. lugar accidente. **en la calle**. que hacia. relata operaria ty de su trabajo hacia su hogar, iba viajando en la motocicleta como pasajero. como ocurrio. manifiesta que sufre caida en motocicleta, recibe lesion en la pierna izquierda se dirigia a su hogar. (...) diagnostico. contusion leve de rodilla. contusion leve de pierna. . **Clasificación automatizada** agresión con disparo de rifle, escopeta o arma larga, **ocurrida en un área industrial o de la construcción**
Agente y Forma OIT	• La taxonomía posee la cualidad de contar con una **división clara de causantes y de modos** en que estos causantes actúan. Lo que favorece la comprensión de los datos en Agentes y en Formas. • **No existe un exceso de categorías de difícil comprensión**, lo que facilita su usabilidad en términos de conteo.	**Relato:** TRABAJO HABITUAL. maestro de terminaciones. (...) QUE HACIA. maestro de terminaciones mientras pinta casa. COMO OCURRIO. refiere se come **durazno luego empieza con alergia y picazon** en cuerpo extremidades ,cuello. ANAMNESIS. am: htaalergias: **Clasificación automatizada**	• La **cantidad** de categorías de agente y forma puede ser baja, en algunos casos de menor complejidad, lo que puede generar que herramientas como LLMs no encuentren una categoría ideal, sino más bien una parecida o incluso puedan alucinar.	**Relato:** QUE HACIA. guardia de seguridad maritimo se dirigia a su trabajo. COMO OCURRIO. se dirigia a su trabajo y sufre asalto por un individuo, sufre agresion fisica en cabeza. ANAMNESIS. (...) paciente refeire que hoy a las 6:30 cuando se dirigia a su trabajo sufre asalto y es **golpeado con puño en region**

	• Ha sido la **taxonomía** más sencilla y de **mejores resultados**.	Contacto por absorción de sustancias nocivas de agentes químicos o **biológicos** (a través de la piel y de los ojos)		**izquierda de craneo** **Clasificación automatizada** Golpeado con/contra **objetos fijos** (inmóviles)

A modo de cierre del presente apartado, es posible evidenciar la complejidad no solo de las taxonomías, sino también de la composición de los relatos médicos, la cual va más allá de elementos de puntuaciones y pausas, e incluso redundancias u agramaticalidades, también existen vacíos de información, o exceso de esta. Estos elementos pueden provocar otra complejidad, que es la de generar una arquitectura con un entorno apto para la codificación precisa de cada caso, pero este, tal como se ha enseñado en la tabla anterior, solo es un desafío, no una desventaja.

V. Metodología

En el marco del presente proyecto, orientado a fortalecer las capacidades institucionales de prevención y análisis en salud laboral, se planteó como objetivo central establecer una taxonomía adecuada para la clasificación automatizada de causas externas de accidentes, utilizando relatos médicos como fuente principal de datos. Este esfuerzo responde a la necesidad de transformar información no estructurada en insumos codificados y comparables, que permitan no solo mejorar la trazabilidad de los eventos, sino también habilitar herramientas de análisis predictivo y diseño de intervenciones preventivas más eficaces.

Bajo este propósito, se realizó una revisión crítica de las principales taxonomías disponibles, considerando tanto su aplicabilidad técnica como su valor analítico en contextos de alta complejidad. En ese análisis, la Clasificación Internacional de Lesiones de Causa Externa (ICECI/CILCE) destacó por su nivel de granularidad, su diseño multiaxial y su enfoque preventivo, lo que la convierte en una opción adecuada para abordar no solo la codificación clínica, sino también el entendimiento contextual del accidente. Así, su selección se fundamentó en dos líneas clave: por un lado, su estructura jerárquica permite representar las relaciones entre módulos (como lugar, actividad, agente y mecanismo) de manera coherente y detallada; por otro, su riqueza categorial facilita un análisis profundo y comparativo, incluso a nivel internacional.

Este enfoque cobra especial sentido en un proyecto como este, que busca superar las limitaciones de los sistemas tradicionales de clasificación textual (basados en reglas rígidas) mediante el uso de modelos avanzados de lenguaje natural. Tal como se ha discutido en la literatura reciente (Lee & Lindsey, 2024), los enfoques basados en reglas tienden a fallar en dominios donde el lenguaje es ambiguo, el contexto es crítico y la codificación exige una interpretación semántica profunda. En el caso de ICECI, con más de 800 categorías posibles y una alta complejidad conceptual, las metodologías clásicas demostraron baja precisión y una elevada tasa de errores, lo que refleja su incapacidad para abordar adecuadamente las variaciones lingüísticas y la riqueza narrativa de los relatos clínicos.

Estos resultados concuerdan con estudios como el de Feng et al. (2024), que evidencian cómo los sistemas basados en reglas tienden a fragmentar la información o pasar por alto relaciones implícitas, como la secuencia temporal de eventos o la causalidad. En este contexto, optar por una arquitectura apoyada en Grandes Modelos de Lenguaje se presentó como una solución más robusta y adaptativa para capturar la complejidad inherente a los relatos médicos y traducirla en codificaciones útiles para la toma de decisiones en prevención. Debido a esto último, para este proyecto se optó por la explotación de grandes modelos de lenguaje de carácter *Open Source*.

Ante este escenario, y con el objetivo de diseñar una solución escalable, interpretable y de alto rendimiento, el proyecto adoptó una arquitectura basada en agentes distribuibles e independientes, capaces de colaborar en el proceso de codificación textual. Esta arquitectura se sustenta en tres principios de diseño fundamentales: especialización funcional, control semántico y validación iterativa.

Desde una perspectiva metodológica, la adopción del enfoque multiagente se justificó por su capacidad de descomponer tareas complejas en subprocesos coordinados, facilitando así la

interpretación modular del rendimiento, el aislamiento de errores y la trazabilidad de decisiones. Esto responde no solo a una necesidad técnica, sino también investigativa, ya que permite evaluar de manera diferenciada el impacto de cada componente en la calidad final del etiquetado.

Para maximizar el potencial de los LLMs en esta arquitectura, se diseñó un sistema compuesto por tres capas funcionales principales:

Agentes Especializados por Dimensión Taxonómica

Cada agente está entrenado o ajustado para operar sobre una dimensión específica de la taxonomía CILCE (como C1: Intencionalidad, C2: Mecanismo de la lesión, C3: Objeto causante, etc.). Este diseño permite que cada agente opere con un conjunto reducido y optimizado de categorías, facilitando una clasificación más precisa y coherente. La especialización reduce la sobrecarga cognitiva del modelo base y mejora su capacidad de generalización dentro de dominios temáticos bien definidos.

Ejemplo: El Agente C2 opera exclusivamente sobre el subconjunto de 113 categorías del módulo "modo de la lesión", aplicando criterios semánticos propios de ese campo para su decisión.

Modelo de Similitud Semántica para Reducción de Categorías

Antes de que cada agente tome una decisión, un modelo de similitud semántica basado en embeddings (SentenceTransformer) calcula la cercanía contextual entre el relato médico y todas las posibles categorías. De este análisis se extraen las Top-K categorías más relevantes, que luego son presentadas al LLM como opciones restringidas. Esta etapa no solo mejora la precisión al reducir la ambigüedad del espacio de búsqueda, sino que también reduce los riesgos de alucinación típicos en modelos generativos.

Esta fase cumple una doble función investigativa: (1) controla el espacio de inferencia del modelo, y (2) permite evaluar el impacto de distintos valores de K en la precisión final del sistema.

Agente Validador y Control de Calidad

Una vez que los agentes especializados asignan sus etiquetas preliminares, entra en operación un Agente Validador independiente. Este componente revisa tanto el relato médico original como la etiqueta sugerida, y decide si confirma, corrige o descarta la categoría propuesta. El validador utiliza criterios de consistencia semántica y plausibilidad contextual, incorporando una lógica de control cruzado entre módulos.

Este componente fue diseñado no sólo como filtro de errores, sino también como mecanismo de explicabilidad: al registrar los cambios realizados y los criterios de descarte, permite una retroalimentación cualitativa sobre el funcionamiento del sistema.

Este diseño responde tanto a requerimientos técnicos como de negocio. Desde el punto de vista computacional, la segmentación por agentes permite paralelizar tareas y optimizar tiempos de respuesta, clave para la escalabilidad futura del sistema. Desde el punto de vista investigativo, la división por dimensiones facilita una evaluación independiente de cada componente, permitiendo identificar cuellos de botella, analizar errores por dimensión y mejorar progresivamente la arquitectura.

Además, este enfoque favorece la interpretabilidad del sistema, ya que cada decisión de clasificación puede ser auditada por dimensión y agente. En un entorno como el de la codificación clínica automatizada, donde la confianza institucional y la trazabilidad son fundamentales, contar con una arquitectura explicable es una ventaja comparativa sustantiva frente a modelos monolíticos de caja negra.

A continuación, se puede observar un diagrama con el flujo de este sistema actual, tras la implementación del agente validado:

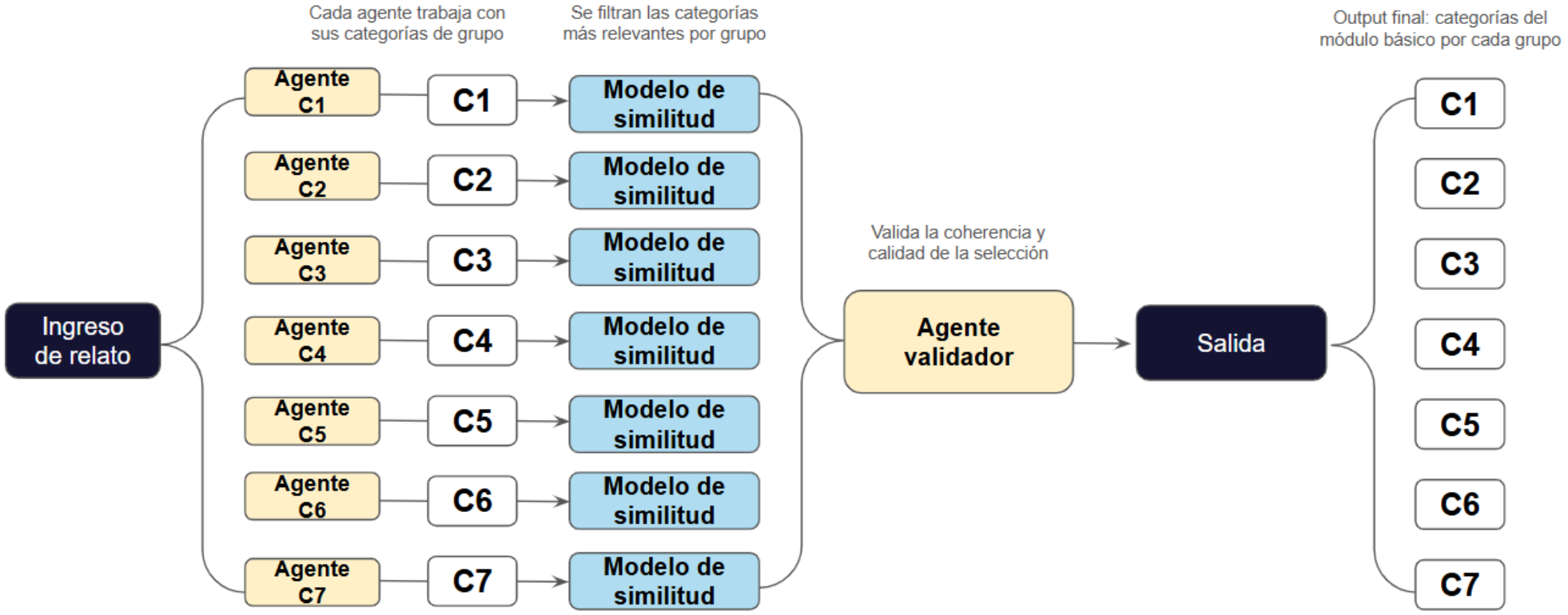

Diagrama 1. *Flujo de CILCE para módulo básico.*

Este enfoque integrador permite abordar la complejidad del problema optimizando, a la vez, el rendimiento del sistema. Al reducir el espacio de búsqueda de categorías mediante filtrado semántico y aplicar un doble control en la asignación de etiquetas, se incrementa la precisión y se minimizan los errores de clasificación.

Previo a la implementación de la arquitectura definitiva, se llevó a cabo un análisis preliminar con el fin de determinar cuál modelo de lenguaje ofrecía el mejor desempeño de base. Esta evaluación inicial se realizó sin incorporar ni el modelo de similitud semántica ni la arquitectura multiagente, lo que permitió medir directamente el rendimiento bruto de distintos LLMs en un escenario de máxima exigencia: procesar relatos médicos completos y asignar etiquetas sin asistencia contextual o filtros previos.

El objetivo de esta fase fue identificar el modelo con mayor precisión y capacidad de comprensión contextual, incluso bajo condiciones desfavorables. Los resultados detallados de este análisis, así como las métricas asociadas, se presentan en las secciones siguientes.

V.I. Modelo de reglas lingüísticas

Para el desarrollo de un sistema que facilite una identificación de relatos médicos en distintas categorías, bajo un enfoque taxonómico, se realizó un proceso de etiquetado de relatos médicos basado en la taxonomía CILCE, utilizando una muestra inicial de 100 datos, para posteriormente ser etiquetados bajo el consumo de modelos lingüísticos diseñada para identificar y etiquetar los módulos principales de la taxonomía: C, O, P, S, T y V, junto con sus respectivos subcampos y categorías específicas.

V.I.I. Descripción y Análisis de data

La muestra de relatos corresponde al mes de julio de 2024, elegido porque coincide con la fecha en que el proveedor de datos habilitó un servicio API para la recopilación automatizada de relatos médicos. Esta API proporcionó datos vinculados a siniestros por accidentes de trayecto y trabajo, incluyendo los antecedentes registrados por el médico revisor. Entre los campos recogidos destacan: número de siniestro, centro de atención, lugar del accidente, actividad realizada, mecanismo del accidente, examen físico, diagnóstico, anamnesis y nota clínica.

Durante dicho mes se generaron cerca de 15.000 relatos médicos de pacientes que indican un caso de accidente laboral o de trayecto, este volumen representa el promedio mensual de registros que maneja la Mutual en esta categoría. A partir de este universo, se aplicó un muestreo aleatorio simple para seleccionar una submuestra de 1000 casos (6,67%), de los cuales se extrajo una muestra reducida de 100 relatos para el proceso de etiquetado inicial. Esta última selección respondió a consideraciones prácticas, ya que el etiquetado manual de textos narrativos bajo taxonomías complejas es una tarea intensiva, que demanda alta precisión, supervisión humana y una inversión significativa de tiempo.

Esta estrategia metodológica combina una base probabilística amplia (1000 casos) con una muestra acotada para análisis intensivo (100 relatos). Si bien desde una perspectiva estadística se estima que una muestra de aproximadamente 385 casos sería necesaria para realizar estimaciones poblacionales con un margen de error de ±5% y un nivel de confianza del 95%, dicha cifra responde a contextos de validación cuantitativa. En esta etapa, sin embargo, el enfoque es exploratorio, técnico y cualitativo,

orientado a probar la viabilidad semántica y funcional del sistema de etiquetado. Por ello, se optó por un conjunto manejable pero suficientemente representativo para iniciar el entrenamiento, ajuste y evaluación inicial del modelo lingüístico basado en la taxonomía CILCE.

Los relatos médicos presentan una estructura semi-estructurada en forma de texto libre, que incluye distintos campos clínico-administrativos organizados bajo una secuencia estandarizada. La estructura de los relatos contempla varios campos evaluados por el médico tratante, campos tales como: trabajo habitual, id único, centro de atención, lugar de accidente, qué hacía, cómo ocurrió, anamnesis, examen físico, nota clínica y diagnóstico. Este conjunto de campos se completa y da forma a un relato del paciente. Algunos campos quedan vacíos. No existen otros datos que permitan caracterizar el accidente en dimensiones como sector, tipo de accidente u otros datos nominales, ya que el trabajo se realizó directamente con los relatos e información no estructurada.

En términos lingüísticos, este tipo de relato presenta desafíos en cuanto a relación de términos, sintaxis y elisión de palabras u oraciones. Al ser el relato médico una representación de la oralidad, este puede presentar elementos que dificultan la totalidad de la comprensión del texto o relaciones más delimitadas entre palabras. En ese sentido, el relato tiene una complejidad media.
Tras la extracción de una muestra de 100 relatos médicos de accidente de trabajo y trayecto, realizó un análisis del formato de salida, tipo de texto en términos semánticos, y, por supuesto, un análisis de la taxonomía CILCE.

En primer lugar, se comprendieron los campos y la utilización de estos con la finalidad de buscar una tecnología apta para el desarrollo de una arquitectura capaz de comprenderla como un humano. En ese sentido, los campos tales como: “Trabajo habitual”, “Lugar de accidente”, “Cómo ocurrió”, fueron el foco de interés, siendo el resto, elementos complementarios a revisar. A continuación, se detalla el análisis en fases:

a. Visualización de clasificación actual por modelo de reglas. En esta fase se extrae una muestra previamente etiquetada con la finalidad de comprender cuáles son los módulos mejor clasificados por la tecnología, con la finalidad de comprender el desafío léxico-semántico de la taxonomía y, así, adaptar una solución ideal al comportamiento de los datos.
b. Comprensión de los campos semánticos. Debido a la completa gama de clasificación que posee la taxonomía, en cuanto a sus módulos de lugar, objetos de afectación, mecanismo de lesión, ocupación, termina siendo un sistema ideal para la comprensión total del relato del paciente. Ahora bien, esto también sitúa a la clasificación, sobre todo basada en inteligencia artificial, a encontrar la tipificación exclusiva para el caso en cuestión, lo que, debido a la alta cantidad de categorías por grupo, se puede tornar complejo sin un estudio detallado.
c. Análisis de entidades lingüísticas. Este trabajo consideró a las palabras de mayor frecuencia e impacto a nivel semántico dentro de las oraciones (categorías). Es un análisis útil cuando se utilizan LLMs y modelos de similitud de semántica, debido a la curación previa de datos, en que las estructuras oracionales pueden recibir modificaciones con tal de que el modelo las capte con una precisión alta, pues en este proceso importa conocer cuáles son las entidades de las categorías que no deben ser eliminadas o modificadas, debido a su importancia en términos de significancia.

V.I.I.I. Validación de Modelo de Reglas

Una vez completado el etiquetado manual, se procedió a evaluar la capacidad del modelo para clasificar los relatos médicos de manera precisa, utilizando un enfoque basado en reglas. Este modelo de reglas se diseñó para identificar patrones textuales específicos asociados con las categorías del módulo básico (C) y sus subcampos.

El conjunto de datos previamente revisado de 100 relatos se utilizó como base para esta evaluación. Cada relato fue procesado mediante las reglas definidas, y los resultados obtenidos se compararon con las etiquetas corregidas manualmente. La comparación permitió medir métricas clave como precisión, sensibilidad y especificidad del modelo de reglas. Este análisis proporcionó información valiosa sobre las limitaciones del modelo actual y áreas de mejora, estableciendo un marco para iteraciones futuras con modelos más sofisticados y robustos.

V.I.I.II. Curado manual de datos

Si bien la arquitectura logró identificar los módulos principales, se observó una baja precisión en la detección de categorías específicas, la cual fue del 59% de un total de 100 casos, lo que expuso la necesidad de un proceso de curado adicional para asegurar la calidad de los datos.

El proceso de curado manual se llevó a cabo como una estrategia de validación y mejora de los datos generados automáticamente por el modelo lingüístico. Para ello, se seleccionó una muestra aleatoria de 100 datos representativos del conjunto inicial de 1000 relatos médicos. El etiquetado manual se centró específicamente en las categorías del módulo básico (C), verificando la precisión de las etiquetas asignadas por el modelo en comparación con las interpretaciones esperadas según la taxonomía CILCE.

Cada dato fue evaluado bajo un sistema binario: los casos correctamente clasificados fueron marcados con un valor de 1, mientras que los errores de clasificación se identificaron con un valor de 0. Los datos con errores fueron revisados y corregidos manualmente por expertos en el dominio, asegurando que las etiquetas reflejaran fielmente el contenido de los relatos médicos. Este proceso permitió depurar la muestra y generar un conjunto de datos validado que sirve como referencia confiable para la evaluación de modelos futuros.

V.II. LLMs Open Source

V.II.I. Selección de Modelo

Para implementar una solución robusta y confiable en la categorización de relatos médicos, el primer paso fue seleccionar un modelo de lenguaje adecuado. Esta elección fue crucial, dado que el modelo debía cumplir con las siguientes características:

1. **Capacidad de Comprensión Semántica**:
    - Ser capaz de interpretar relatos médicos, que a menudo incluyen términos especializados, descripciones contextuales complejas y detalles relevantes para la categorización.
2. **Precisión en el etiquetado**:

    - Garantizar una asignación coherente de las categorías en función de los criterios predefinidos.
3. **Eficiencia Computacional**:
    - Operar dentro de los recursos de hardware disponibles, evitando modelos excesivamente costosos en términos de memoria RAM y VRAM.
4. **Consistencia y Escalabilidad**:
    - Ofrecer resultados consistentes en pruebas con datos reales y tener la capacidad de escalar para manejar un mayor volumen de relatos médicos.

V.II.II. Dataset de Evaluación

Para realizar la selección, se utilizó un **dataset de 100 relatos médicos etiquetados manualmente**. Este conjunto de datos se construyó cuidadosamente para reflejar la diversidad y complejidad de los relatos reales, abarcando una variedad de categorías en cada uno de los ámbitos definidos.

Los modelos evaluados fueron probados en este dataset con el fin de medir su precisión en la asignación de etiquetas. Cada modelo produjo predicciones que luego se compararon con las etiquetas manuales utilizando métricas de evaluación estándar.

V.II.III. Modelos Evaluados

Para este estudio, se llevaron a cabo dos iteraciones de evaluación con modelos disponibles en el entorno de Ollama[1], seleccionados por su potencia computacional, accesibilidad y optimización para tareas de instrucción.

Iteración 1

Se evaluaron cuatro modelos que combinan eficiencia y rendimiento:

- **llama3.2-3b-instruct-fp16** (Ollama, s.f.)
    - **Tamaño**: 4.5 GB
    - **Descripción**: Modelo de mediano tamaño optimizado para tareas de instrucción con pesos en punto flotante de 16 bits (FP16).
    - **Ventaja**: Ligero y eficiente en términos computacionales.
- **gemma2-9b-instruct-q4_0** (Ollama, s.f.)
    - **Tamaño**: 5.0 GB
    - **Descripción**: Modelo robusto con 9 billones de parámetros, optimizado para tareas de instrucción. Los pesos cuantizados a 4 bits (Q4_0) ofrecen un balance entre rendimiento y eficiencia.
    - **Ventaja**: Alto rendimiento en tareas contextuales complejas.

- **llama3.1-8b-instruct-q4_0** (Ollama, s.f.)
    - **Tamaño**: 5.1 GB
    - **Descripción**: Modelo más grande que el anterior, con 8 billones de parámetros y optimización en pesos cuantizados a 4 bits.

[1] Ollama es una herramienta de código abierto que te permite descargar y ejecutar modelos de inteligencia artificial (como Llama 3, DeepSeek, Qwen, Gemma, entre otros) de forma completamente local

    - **Ventaja**: Capacidad mejorada en tareas complejas gracias a su tamaño mayor.

- **mistral7b-instruct-q5_0** (Ollama, s.f.)
    - **Tamaño**: 4.7 GB
    - **Descripción**: Modelo de 7 billones de parámetros, enfocado en tareas de instrucción con pesos cuantizados a 5 bits (Q5_0).
    - **Ventaja**: Excelente equilibrio entre tamaño y rendimiento computacional.

Iteración 2

Con el objetivo de explorar modelos de mayor capacidad y especialización, se incluyeron dos modelos adicionales:

- **falcon3-10b-instruct-q4_K_M (Hugging Face, 2024)**
    - **Tamaño**: 5.7 GB
    - **Descripción**: Modelo avanzado con 10 billones de parámetros, optimizado para tareas de instrucción con pesos cuantizados a 4 bits (Q4_K_M). Diseñado para destacar en ciencia, matemáticas y programación.
    - **Ventaja**: Alto rendimiento en tareas técnicas y de razonamiento complejo.

- **qwen2.5-14b-instruct-q4_K_M (Ollama, s.f.)**
    - **Tamaño**: 8.1 GB
    - **Descripción**: Modelo robusto con 14 billones de parámetros, optimizado para tareas de instrucción y cuantizado a 4 bits (Q4_K_M). Ofrece soporte multilingüe y capacidades mejoradas en codificación y matemáticas.
    - **Ventaja**: Gran versatilidad y capacidad para manejar contextos largos y tareas especializadas.

V.II.IV. Resultados de la Evaluación

Aspecto técnicos y benchmark

La arquitectura diseñada para este sistema tiene como objetivo resolver la tarea de categorización automática de relatos médicos utilizando una solución basada en un modelo de lenguaje extenso (LLM), apoyada por técnicas avanzadas de procesamiento de lenguaje natural. El modelo seleccionado, *qwen2.5-14b-instruct-q4_K_M,* es el núcleo de la solución y se complementa con una infraestructura multiagente y un modelo de similitud semántica para maximizar la precisión, escalabilidad y confiabilidad.

Este enfoque permite abordar la complejidad intrínseca de los relatos médicos, que a menudo incluyen lenguaje técnico, descripciones contextuales y relaciones implícitas, mediante una combinación de procesamiento lingüístico avanzado y técnicas de validación robustas.

Los modelos fueron evaluados en función de tres métricas principales:

1. Precisión:
    - Proporción de etiquetas correctas asignadas por el modelo en comparación con las etiquetas manuales.
2. Cohesión Semántica:
    - Capacidad del modelo para comprender correctamente el contexto del relato médico, identificando categorías relevantes incluso en situaciones ambiguas.
        - Baja. El modelo clasifica de forma irregular en general, el relato no suele poseer características lingüísticas similares, no hay presencia, inclusive de sinónimos o términos relacionados entre categoría asignada y relato médico.
        - Media. El modelo clasifica de forma regular en varios casos, es decir, sí se aprecia una relación léxica entre relato y categoría asociada pero no en la mayoría de los casos.
        - Alta. El modelo clasifica de forma regular y correcta un relato con su respectiva categoría, generando relaciones léxicas tanto entre sinónimos, como con términos exactos de la taxonomía y del texto del paciente.
3. Consistencia:
    - Capacidad del modelo para asignar una gama diversa de etiquetas relevantes según la variabilidad de los relatos médicos, evitando sesgos hacia unas pocas categorías predominantes.
        - Baja. El modelo tiende a asignar un conjunto limitado de etiquetas, mostrando poca diversidad en la clasificación y sesgo hacia categorías específicas.
        - Media. El modelo asigna una variedad moderada de etiquetas, capturando parcialmente la diversidad de los relatos, pero con cierta preferencia por categorías más frecuentes.
        - Alta. El modelo asigna una amplia gama de etiquetas relevantes, reflejando adecuadamente la diversidad de los relatos médicos sin sesgos significativos.

Los elementos 2 y 3 fueron analizados bajo una perspectiva cualitativa bajo criterio experto.

Iteración 1

Modelo	Precisión (sobre 100 datos)	Cohesión Semántica	Consistencia
Llama3.2-3b-instruct-fp16	15.1%	Baja	Alta
Gemma2-9b-instruct-q4_0	**21.5%**	**Moderada**	**Moderada**
Llama3.1-8b-instruct-q4_0	13.5%	Baja	Alta
Mistral7b-instruct-q5_0	7.1%	Baja	Alta

Tabla 2. Evaluación de Iteración 1 para un N = 100.

Iteración 2

falcon3-10b-instruct-q4_K_M	55.5%	Alta	Moderada
qwen2.5-14b-instruct-q4_K_M	**75.7%**	**Alta**	**Baja**

Tabla 3. Evaluación de Iteración 2 para un N=100.

Selección del Modelo: **qwen2.5-14b-instruct-q4_K_M**

De acuerdo con los resultados, se seleccionó el modelo **qwen2.5-14b-instruct-q4_K_M** debido a su desempeño consistentemente superior en las métricas clave:

1. **Precisión Alta**:
    - Este modelo destacó por su capacidad de asignar etiquetas correctas en categorías complejas o ambiguas.
2. **Cohesión Semántica**:
    - Fue especialmente notable en relatos con múltiples elementos contextuales, interpretando adecuadamente las intenciones del texto.
3. **Consistencia**:
    - Mostró una baja repetibilidad en sus resultados, crucial para garantizar la diversidad en la categorización.

El modelo **qwen2.5-14b** fue elegido como la base para la solución debido a su equilibrio óptimo entre precisión, comprensión semántica y consistencia. Además, su integración en la arquitectura de Ollama permitió aprovechar plenamente su potencial en tareas de categorización complejas.

Aunque otros modelos ofrecieron un rendimiento competitivo, **qwen2-5-14b** se destacó como el mejor candidato para abordar la categorización de relatos médicos en un entorno de alta exigencia.

V.II.V. Componentes Principales de la Arquitectura

La arquitectura está compuesta por tres pilares fundamentales:

1. **Modelo de Lenguaje Qwen2.5-14b a través de Ollama**
    - **Qwen2.5-14b-instruct-q4_K_M** es un modelo LLM con aproximadamente 14 billones de parámetros. Su capacidad para comprender el lenguaje natural especializado, como el usado en los relatos médicos, lo hace ideal para esta tarea.
    - El modelo es accesible a través de **Ollama**, una infraestructura que permite realizar consultas al LLM de manera eficiente y gestionar los flujos de trabajo necesarios para integrar sus capacidades en el sistema multiagente.

2. **Infraestructura Multiagente**
    - Cada relato es procesado por un conjunto de agentes especializados. La arquitectura multiagente incluye dos tipos de agentes:
        - **Agentes Etiquetadores**: Un agente por cada categoría (C1: Intencionalidad, C2: Modo de la lesión, etc.). Cada agente asigna una etiqueta relevante al relato, utilizando las capacidades lingüísticas de qwen2.5-14b y un conjunto reducido de categorías prefiltradas.
        - **Agente Validador**: Un agente independiente que verifica la coherencia y validez de las etiquetas asignadas. En caso de que una etiqueta no sea coherente con el contenido del relato, este agente la reemplaza por "N/A". Dicho agente se basa en un modelo de mayor capacidad, con una precisión promedio del 95% en tareas de clasificación, y tiene como función principal verificar la corrección de las etiquetas proporcionadas por el agente inicial, conforme a las definiciones establecidas para cada submódulo.
    - Este diseño asegura un control de calidad robusto, minimizando los errores en las etiquetas finales. Sin embargo, el análisis de rendimiento sólo se aplicó a nivel de módulo general y su grupo, pero no a nivel de granular dentro de la-categoría, ya que para ello, se requiere de un trabajo manual mucho más profundo y una representación más granular de cada sub-categoría.
3. **Modelo de Similitud Semántica**
    - Antes de la intervención de los agentes, un modelo de similitud basado en embeddings generados por un modelo tipo **SentenceTransformer**, selecciona las **Top K categorías** más relevantes para cada dimensión. Esto reduce significativamente el espacio de búsqueda y mejora la precisión, al enfocar los esfuerzos del LLM en opciones de alta probabilidad.
    - El criterio utilizado en esta fase es netamente automatizado bajo modelos que combinan el entendimiento semántico entre las palabras y un sistema de vectores. En ese sentido, cuando se menciona “categorías más relevantes” se hace referencia a que

el modelo, mediante su proceso de vectorización, logra detectar un top K de categorías (de la taxonomía) que tienen una mayor relación semántica con el relato médico, es por esto que se las denomina como relevantes. Estas al ser seleccionadas, el agente validador verifica su clasificación.

V.II.VI. Flujo de Trabajo

El sistema sigue un flujo de procesamiento estructurado que combina las capacidades de cada componente. Para el flujo de trabajo se utilizó datos del el periodo del segundo semestre del año 2024:

1. **Entrada**:
    - El sistema recibe un relato médico en texto libre que contiene el lugar, diagnóstico, anamnesis, lugar, y actividad realizada.
2. **Filtrado Inicial: Top K Categorías**:
    - El modelo de similitud analiza el texto del relato y calcula los embeddings correspondientes.
    - A partir de estos embeddings, se identifican las **Top K categorías** más relevantes de una lista predefinida para cada dimensión. Cabe destacar que esta lista son las categorías oficiales de los grupos de los módulos de la respectiva taxonomía.
3. **Asignación de Categorías**:
    - Cada **Agente Etiquetador** toma las categorías Top K y el relato médico como entrada. Utilizando qwen2.5-14b-instruct-q4_K_M, asigna una etiqueta adecuada basada en las opciones prefiltradas.
4. **Validación de Categorías**:
    - El **Agente Validador** analiza el relato médico y la etiqueta asignada por el etiquetador.
    - Si considera que la etiqueta es coherente con el relato, la confirma; de lo contrario, la reemplaza por "N/A".
5. **Salida**:
    - El sistema genera un conjunto de etiquetas finales para las dimensiones del relato.

V.II.VII. Categorización por Dimensiones

El sistema clasifica cada relato en las siguientes siete dimensiones, con las correspondientes cantidades de categorías:

- **C1**: Intencionalidad del accidente (7 categorías).
- **C2**: Modo de la lesión (113 categorías).
- **C3**: Objeto o sustancia causante (609 categorías).
- **C4**: Lugar de la lesión (66 categorías).
- **C5**: Actividad durante la lesión (3 categorías).
- **C6**: Uso de alcohol (2 categorías).
- **C7**: Uso de drogas (2 categorías).

Cada dimensión tiene un conjunto limitado de opciones que facilita la clasificación. Este diseño estructurado es fundamental para garantizar un análisis detallado de las circunstancias de la lesión.

V.II.VIII. Uso de Recursos

La solución fue diseñada teniendo en cuenta los requerimientos computacionales de **qwen2.5-14b-instruct-q4_K_M** y los modelos auxiliares:

- **Memoria RAM**:
    - El sistema utiliza aproximadamente **5.7 GB** de RAM de manera constante durante la ejecución.
- **VRAM (memoria de GPU)**:
    - El consumo de VRAM asciende a **15.3 GB**, considerando las consultas al modelo LLM y el cálculo de embeddings para la similitud semántica.

Este uso de recursos es eficiente considerando la complejidad del modelo y la naturaleza intensiva del procesamiento lingüístico que requiere.

V.II.IX. Desempeño

- **Tiempo promedio por relato**:
    - El tiempo total de procesamiento por relato es de aproximadamente **18.3 segundos**.
    - Este tiempo incluye:
        - La extracción de las categorías Top K.
        - El etiquetado inicial realizado por los agentes.
        - La validación de cada etiqueta por el Agente Validador.
        - La categorización de C1 a C7.

A pesar de los tiempos relativamente altos, el diseño modular y escalable permite optimizaciones futuras para mejorar la eficiencia sin comprometer la calidad del análisis.

V.II.X. Ventajas Técnicas

1. **Modularidad**:
    - Los componentes del sistema son independientes y fácilmente intercambiables. Por ejemplo, se pueden añadir nuevas categorías o reemplazar el modelo de similitud sin necesidad de rediseñar todo el flujo.
2. **Alta Precisión y Fiabilidad**:
    - La combinación del modelo de similitud, el etiquetado y la validación asegura resultados consistentes y coherentes, reduciendo la tasa de errores.
3. **Optimización del Espacio de Búsqueda**:
    - El filtrado previo mediante embeddings y la selección de Top K categorías optimizan los recursos computacionales, al enfocar el procesamiento en opciones relevantes.
4. **Capacidad de Escalabilidad**:
    - La arquitectura multiagente permite distribuir las cargas de trabajo, haciendo que el sistema sea capaz de manejar mayores volúmenes de datos con ajustes mínimos.
5. **Soporte para Lenguaje Técnico**:
    - El uso de **qwen2.5-14b-instruct-q4_K_M** proporciona una capacidad avanzada para entender el lenguaje médico especializado, garantizando que las etiquetas sean relevantes y contextualmente precisas.

La arquitectura multiagente, en combinación con **qwen2.5-14b-instruct-q4_K_M** y el modelo de similitud, permite un sistema de categorización robusto, escalable y confiable. Este enfoque equilibra precisión, modularidad y eficiencia, convirtiéndo a la arquitectura en una herramienta óptima para el análisis de información médica no estructurada. Con futuras optimizaciones en paralelización y reducción de tiempo por relato, la solución está bien preparada para satisfacer el flujo esperado.

VI. Metodología de modularización final

VI.I. Módulos y subcategorías CILCE

La taxonomía médica de CILCE comprende características de alta complejidad, tales como su jerarquía estructurada por módulos y subgrupos; los códigos alfanuméricos que deglosan y acompañan a las categorías de cada subgrupo de módulos; el lenguaje estandarizado pero de dificultad media para público no especializado en áreas de la salud, pero fácilmente comprensible en un entorno de labores relacionadas con ese entorno profesional o de sistemas sanitarios; su compatibilidad con taxonomías externas, como lo es CIE 10.

En ese sentido, en el siguiente apartado se indican la totalidad de módulos utilizados en la integración de la modularización y etiquetado final, junto a la cantidad de categorías encontradas y abarcadas por la última versión de la prueba, es decir, utilizando el modelo *qwen2.5-14b-instruct-q4_K_M***:**

- Módulo básico. Se utilizó una cantidad de 802 categorías médicas, las cuales son parte de los siguientes grupos: Intencionalidad, Mecanismo de lesión, Objeto/sustancia que produce la lesión, Lugar de ocurrencia, Actividad en caso de lesión, Uso de alcohol y Uso de droga o sustancias psicoactivas
- Módulo de violencia. Se trabajó con una cantidad de 212 categorías dentro las cuales se presentaron los siguientes grupos: Factores de riesgo precipitantes para daño intencional autoinfligido, Historia de intento de suicidio, Relación víctima/agresor, Sexo del agresor, Contexto de la agresión, Tipo de intervención legal, Tipo de conflicto.
- Módulo de transporte. Este módulo utilizó 149 categorías dentro de las cuales se encontraron los grupos de Modo de transporte, Papel de la persona lesionada, Contraparte y Tipo de evento relacionado con la lesión de transporte.
- Módulo de lugar. Para este caso se incluyeron 45 categorías, cuyos grupos fueron: P2 - Parte de inmueble o de predio, P3 - Tipo de vivienda, P4 - Residente de la vivienda, Tipo de lugar de atención médica y el Tipo de escuela
- Módulo de deportes. Se utilizaron 292 categorías claves dentro de las cuales se encuentran tres grupos: Tipo de actividad de deporte/ejercicio, Medidas de control individual, Medidas de control ambiental.
- Módulo ocupacional. Este último caso utilizó 27 categorías médicas de dos grupos: Actividad económica y Ocupación.

VI.II. Desafíos evidenciados

Durante la implementación del sistema de etiquetado automatizado basado en módulos de la taxonomía CILCE, se identificaron una serie de desafíos técnicos, semánticos y operativos que influyeron en el diseño de los agentes y en la estructura de la lógica de clasificación:

- Ambigüedad y solapamiento entre categorías: Algunas categorías, especialmente en módulos como el básico o el de violencia, presentan descripciones similares a nivel de sub-categorías, lo que puede dificultar la clasificación precisa sin una contextualización clínica más profunda.
- Desbalance de datos entre módulos: Algunos módulos como el ocupacional o el de lugar contienen muchas menos categorías que otros como el básico o el de violencia, lo que genera

un sesgo de atención en los modelos LLM hacia los módulos con más datos disponibles.

- Limitaciones computacionales en la clasificación semántica: La comparación de similitud semántica y la generación de respuestas por parte del LLM pueden verse afectadas por cuellos de botella en el procesamiento, especialmente cuando se deben analizar cientos de categorías por iteración.

- Problemas de conectividad o latencia con el servicio Ollama: Al depender de un modelo servido de manera local, los tiempos de respuesta dependerá directamente del stock de hardware disponible para consumir el modelo.

VI.III. Flujo de CILCE con integración completa de módulos

La arquitectura del sistema de etiquetado automático de CILCE se diseñó con un enfoque modular y secuencial, permitiendo aplicar distintas funciones de clasificación sobre el mismo input sobre el relato médico.

A continuación, se detalla el flujo operativo:

1. Preprocesamiento: Se limpia el relato médico eliminando identificadores irrelevantes y se normaliza el texto para mejorar la comprensión semántica por parte del modelo. Se realizó un concatenado de columnas, y se eliminaron ciertos campos tales como el ID del paciente y centro médico. Este proceso ocurre con la finalidad de entregar datos que aporten con la información necesaria para la clasificación de la solución automatizada y que esta no reciba contenidos no atingentes.

2. Clasificación por módulos:
    - El sistema invoca una función *process_modulo* para cada módulo, utilizando ya sea lógica semántica por embeddings o agentes LLM.
    - Cada módulo retorna un conjunto de etiquetas que se agregan al DataFrame original como columnas nuevas (e.j. o1, o2, p3, t2...).

3. Integración de resultados:
    - Se concatenan las salidas de los distintos módulos y se preserva la estructura original del dataset.
    - Las columnas añadidas permiten tanto análisis por módulo como análisis transversal entre etiquetas.

4. Exportación y validación:
    - El DataFrame enriquecido se exporta a formatos estándar (Excel, CSV) para posterior revisión.
    - Se pueden aplicar scripts adicionales de auditoría o validación manual según las necesidades del equipo clínico.

Este flujo permite escalar el sistema con nuevos módulos o versiones de modelos LLM sin modificar la estructura central, y garantiza reproducibilidad y trazabilidad en cada iteración del etiquetado.

Tras el desarrollo del flujo mencionado con anterioridad, a continuación se indica una tabla que enseña resultados del análisis de precisión por etiquetado de módulos. Cabe destacar que el total de la muestra para cada caso fue de una muestra de 100 datos:

Letra	Módulo	Precisión
C	Básico	77,5%
O	Ocupacional	69%
P	Lugar	90,8%
S	Deportes	84%
T	Transporte	57,1%
V	Violencia	25%
	Global	67,7%

Tabla 4. Precisión de etiquetado de módulos *CILCE.*

Letra	Módulo	Precisión
C1	Intencionalidad	71,4%
C2	Modo	75,6%
C3	Objeto o Sustancia	54.2%
C4	Lugar	77.9%
C5	Actividad Realizada	98.9%
C6	Uso de Alcohol	100%
C7	Uso de Drogas	100%

Tabla 4. Precisión Submodulos (C) *CILCE.*

Respecto a los resultados obtenidos se logró evidenciar una variabilidad considerable en algunos de los casos de la precisión de ciertos módulos, esto nos ofrece datos importantes acerca del sistema y de oportunidades de futuras mejoras.

- En primer lugar, si se considera al grupo de alta precisión, el módulo P de Lugar es el que alcanzó un porcentaje del 90,8%, seguido del módulo S de Deportes que obtuvo un 84% y, el básico C, con un 77,5%. Este logro en el desempeño puede deberse a diversos factores, dentro de los cuales unos podrían ser la característica de patrones lingüísticos mayormente homogéneos de las categorías, lo que puede ser útil para modelos con los que se ha trabajado.
- En segundo lugar, se presentan los módulos con un desempeño de precisión moderada, en este caso se encuentran el módulo ocupacional O, el cual logró un porcentaje del 69, el que si bien

cumple con una calidad aceptable, ofrece posibilidades de mejora. Esta bajada puede deberse a la posible ambigüedad de términos que se pueden presentar en ciertas categorías y, además, puede deberse a la falta de especificidad en algunos casos en que las labores son más bien particulares. En esta sección, se encuentra también el módulo de transporte T, el cual obtuvo un 57% de precisión, donde se observa la caída de más de un 10%.

- En tercer lugar, se encuentra el módulo de bajo rendimiento: Violencia V, con la menor precisión evidenciada del 25%. Este resultado puede deberse a la complejidad con la que se etiquetan casos relacionados con violencias, los que ameritan un nivel de análisis con una cantidad de datos relevantes en términos de contexto y factores indicativos de los casos en cuestión.

A modo de cierre del presente apartado, se consideró un rendimiento global del 67,7% de precisión, lo que nos indica un sistema cuyo punto inicial es bastante robusto y que, a su vez, permite dilucidar pasos claros de análisis y detalles tanto de la arquitectura como de los datos.

V.II. Conclusión y recomendaciones

A modo de cierre, el presente proyecto buscó generar una comprensión sobre las taxonomías que de mejor forma pueden describir accidentes y lograr de forma automatizada la codificación de accidentes basados en relatos médicos de pacientes. Esto, mediante la utilización de la taxonomía CILCE y grandes modelos de lenguaje, los cuales han sido validados y han sido preparados para ser ejecutados en arquitecturas que consideran modelos de similitud semántica y los multiagentes, para casos de etiquetado y de validación.

El proyecto concluye resaltando el valor central de CILCE, una taxonomía diseñada para una codificación granular y contextualizada de accidentes basada en relatos médicos. Gracias a su estructura jerárquica y multi‑eje, CILCE integra diversos módulos (como lugar, ocupación, actividad, transporte e intención) que permiten describir con detalle las circunstancias previas al episodio lesivo, brindando una visión integral que pocas clasificaciones alcanzan. Esta riqueza es especialmente eficaz cuando se complementa con grandes modelos de lenguaje optimizados por arquitecturas que combinan similitud semántica e interacciones multi‑agente: los embeddings vectoriales permiten identificar coincidencias contextuales en los relatos, mientras que los agentes distribuidos facilitan procesos coordinados de etiquetado y validación iterativa, mejorando la precisión del reconocimiento automático de módulos narrativos.

Además, el uso de CILCE como base garantiza que el sistema no solo detecte la intervención médica, sino que comprenda el contexto situacional del accidente. Esto habilita una capa de análisis preventivo más sofisticada: al articular el tipo de evento (por ejemplo, accidente vehicular en ambiente laboral) con la intervención clínica realizada, se potencia la capacidad de anticipar riesgos, diseñar medidas de seguridad específicas y optimizar protocolos de atención. En última instancia, esta combinación de codificación precisa + IA semántica + agentes cooperativos convierte al sistema en una herramienta poderosa tanto para el análisis epidemiológico como para la prevención activa de lesiones laborales.

La incorporación de LLMs ha representado un avance significativo, al aportar una capacidad superior para la comprensión contextual, la selección precisa de etiquetas y la coherencia en la clasificación de textos médicos. En particular, la utilización del modelo **qwen2.5-14b-instruct-q4_K_M**, junto con mecanismos de reducción de opciones basados en similitud semántica, ha demostrado ser clave para lograr una clasificación eficiente y precisa.

Los resultados obtenidos, con una precisión del 75,5 % en la última iteración, sólo considerando el Módulo básico (C), así como una alta cohesión semántica entre categorías y relatos, evidencian el potencial de esta tecnología en contextos de alta complejidad y exigencia, como lo es el entorno clínico. No obstante, persisten desafíos importantes, especialmente en cuanto a la optimización del tiempo de procesamiento y la evaluación de nuevos modelos emergentes que puedan mejorar aún más el rendimiento.

Derivado de los hallazgos y orientado a la mejora de la estrategia preventiva de Mutual de Seguridad, se presentan las siguientes recomendaciones para la implementación y aplicación del modelo:

- Implementación progresiva y escalonada: Se sugiere iniciar la integración del modelo mediante un piloto en un subconjunto controlado de relatos médicos, lo cual permitirá validar

operativamente y en la práctica el sistema, generar retroalimentación en contexto real y realizar ajustes antes de una eventual ampliación institucional.

- Capacitación y participación interdisciplinaria: La implementación del sistema requiere la colaboración activa de equipos clínicos, de datos y de prevención. Se recomienda realizar talleres de capacitación y sesiones de análisis conjunto para maximizar el entendimiento de las capacidades del modelo y fomentar su adopción.

- Optimización del pipeline de procesamiento: Considerando los tiempos actuales de respuesta, es pertinente evaluar mejoras técnicas tales como paralelización, optimización del preprocesamiento textual o la utilización de distinto hardware para lograr una mayor escalabilidad en el proceso de codificación.

- Monitoreo y evaluación continua del desempeño: Se propone establecer métricas estables de seguimiento que permitan evaluar de manera continua la precisión del modelo, la coherencia semántica y su contribución a la calidad de la información estructurada. Este monitoreo debe incluir comparativas con codificaciones manuales y análisis de error.

- Revisión y fortalecimiento de la taxonomía CILCE: La experiencia obtenida permite vislumbrar áreas de mejora dentro de la taxonomía, tales como categorías ambiguas o insuficientemente representadas. Se recomienda una revisión iterativa de la CILCE, guiada por evidencia empírica obtenida del modelo y del análisis clínico.

- Aprovechamiento estratégico de los datos estructurados: La correcta codificación de accidentes representa un insumo valioso para sistemas de análisis predictivo, identificación de patrones de riesgo y generación de alertas tempranas. Se recomienda articular este modelo con sistemas de business intelligence y prevención, permitiendo retroalimentar directamente la estrategia preventiva institucional.

Uno de los aportes más relevantes de este proyecto, más allá de la mejora en los procesos de codificación automática, es el potencial que ofrece para el aprovechamiento estratégico de los datos clínicos estructurados como base para la toma de decisiones preventivas. La transformación de relatos médicos en categorías normalizadas, mediante el uso de modelos de lenguaje, permite consolidar un repositorio coherente y analíticamente robusto de eventos asociados a accidentes laborales, enriquecido semánticamente y preparado para su explotación por parte de sistemas analíticos avanzados.

En este contexto, la aplicación de modelos de análisis estadístico y de inteligencia artificial sobre esta información codificada abre múltiples oportunidades. Por ejemplo, al identificar patrones de ocurrencia en ciertos tipos de accidentes (según sector económico, tipo de actividad, perfil ocupacional o causas predominantes), es posible generar indicadores clave de riesgo que informen directamente a las unidades de prevención. Estos indicadores pueden ser utilizados para priorizar intervenciones, ajustar protocolos de seguridad, o focalizar campañas formativas en grupos específicos de trabajadores o sectores productivos. Se recomiendan 3 estrategias en este contexto: (1) Integrar el sistema de codificación con plataformas de business intelligence, de modo que los datos estructurados alimenten en tiempo real dashboards, reportes analíticos y tableros de seguimiento preventivo. (2) Desarrollar modelos de segmentación y clasificación de riesgo basados en los datos

históricos codificados, lo que permitirá definir perfiles de riesgo ocupacional con mayor precisión. (3) Implementar algoritmos de detección temprana de anomalías, que analicen flujos de relatos en tiempo real para identificar desviaciones significativas respecto de los patrones esperados, y activar alertas preventivas para análisis por parte de expertos.

En síntesis, este trabajo demuestra que las tecnologías generativas aplicadas a datos no estructurados en salud abren nuevas posibilidades para la gestión y análisis de información clínica. La arquitectura propuesta no solo responde eficazmente a las demandas actuales, sino que establece una base sólida para futuras mejoras. En particular, el tratamiento de la taxonomía CILCE permitió alcanzar un equilibrio entre precisión y coherencia semántica, siendo esta última una de las cualidades más difíciles de garantizar en dominios altamente especializados.

V.III. Bibliografía

Artículos:

1. Lee, S. A., & Lindsey, T. (2024). Do large language models understand medical codes?. Arxiv., Volumen (1). 1-21. DOI:10.48550/arXiv.2403.10822

2. Feng, J., Zhang, R., Chen, D., Shi, L & Li, Z. (2024). Automated Generation of ICD-11 Cluster Codes for Precision Medical Record Classification. International Journal of Computer Communications & Control. Volumen (9). 1-14. https://univagora.ro/jour/index.php/ijccc/article/view/6251

3. Feng, Z., Singh, V., Qian, C., & Wang, Y. (2024). *Clinical Text Classification with Large Language Models: A Comparative Study*. arXiv. https://arxiv.org/abs/2504.08040

4. Lee, S. A., & Lindsey, T. (2024). *Do Large Language Models Understand Medical Codes?* arXiv. https://arxiv.org/abs/2503.01159

5. ICECI Coordination and Maintenance Group. (2004).International Classification of External Causes of Injuries (ICECI), version 1.2. Consumer Safety Institute & AIHW National Injury Surveillance Unit.

6. Nordic Medico-Statistical Committee (NOMESCO). (2007). NOMESCO classification of external causes of injuries (4th rev. ed.). Nordic Medico-Statistical Committee.

7. Organización Mundial de la Salud. (2008). Clasificación estadística internacional de enfermedades y problemas relacionados con la salud (CIE-10) (10ª rev.). OMS.

8. Organización Mundial de la Salud. (2022). Clasificación internacional de enfermedades: 11ª revisión (CIE-11). OMS.

9. Organización Panamericana de la Salud. (2018, 4 de septiembre). Diferencias con CIE-10 de CIE10-CM. Organización Panamericana de la Salud. Recuperado de https://www3.paho.org/relacsis/index.php/es/foros-relacsis/foro-becker-fci-oms/61-foros/consultas-becker/974-diferencias-con-cie-10-de-cie10-cm/#:~:text=La%20CIE%2D10%20CM%20es,las%20estad%C3%ADsticas%20en%20los%20pa%C3%ADses

10. Organización Internacional del Trabajo. (1996a). Anexo H: Clasificación de los accidentes de trabajo según la forma del accidente. En Registro y notificación de accidentes del trabajo y enfermedades profesionales. Oficina Internacional del Trabajo.

11. Organización Internacional del Trabajo. (1996b). Anexo I: Clasificación de los accidentes de trabajo según el agente material. En Registro y notificación de accidentes del trabajo y enfermedades profesionales. Oficina Internacional del Trabajo.

Repositorios web de LLMs:

1. **Ollama – LLaMA 3.2 3B Instruct**

   Ollama. (s.f.). *llama3.2:3b-instruct-fp16*. Ollama. https://ollama.com/library/llama3.2:3b-instruct-fp16

2. **Ollama – Gemma 2 9B Instruct**

   Ollama. (s.f.). *gemma2:9b-instruct-q4_0*. Ollama. https://ollama.com/library/gemma2:9b-instruct-q4_0

3. **Ollama – LLaMA 3.1 8B Instruct**

   Ollama. (s.f.). *llama3.1:8b-instruct-q4_0*. Ollama. https://ollama.com/library/llama3.1:8b-instruct-q4_0

4. **Ollama – Mistral 7B Instruct**

   Ollama. (s.f.). *mistral:7b-instruct-q5_0*. Ollama. https://ollama.com/library/mistral:7b-instruct-q5_0

5. **Hugging Face Blog – Falcon 3**

   Hugging Face. (2024). *Introducing Falcon 3*. GitHub. https://github.com/huggingface/blog/blob/main/falcon3.md

6. **Ollama – Qwen 2.5 14B Instruct**

   Ollama. (s.f.). *qwen2.5:14b-instruct-q4_K_M*. Ollama. https://ollama.com/library/qwen2.5:14b-instruct-q4_K_M