Modelo combinado de co-training y aprendizaje por transferencia para clasificación de documentos, a partir de un análisis comparativo de modelos de aprendizaje semi-supervisados

Alex Santiago Cevallos Culqui
Directora: Dra. Claudia Pons | Co-director: Dr. Gustavo Rodríguez
Doctorado en Ciencias Informáticas - Facultad de Informática - UNLP
12 de Noviembre 2025

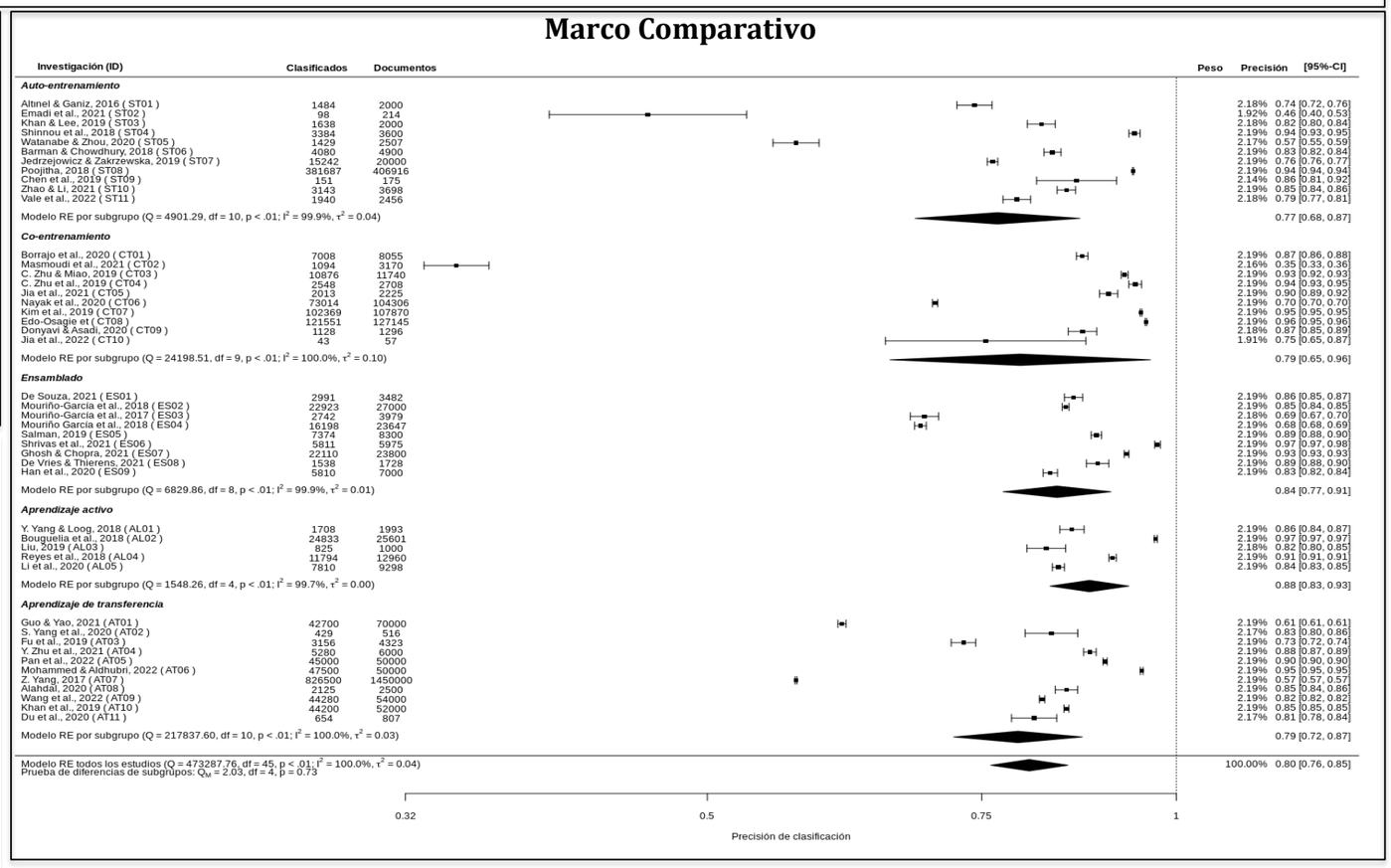
MOTIVACIÓN

La creciente cantidad de documentos digitales ha generado desafíos importantes en su clasificación y organización. Esto impacta directamente en la búsqueda de información y en la toma de decisiones en distintos ámbitos. Las técnicas de procesamiento de lenguaje natural buscan responder a esta problemática de manera eficiente. Sin embargo, los modelos supervisados dependen de datos etiquetados, cuya disponibilidad suele ser limitada. Por otro lado, los modelos no supervisados carecen de precisión para clasificaciones más detalladas. Ante ello, los enfoques semi-supervisados (SSL) surgen como una alternativa que combina ambos métodos, sin embargo, se identifican las siguientes limitantes: La *escasez de análisis comparativos* de modelos semi-supervisados limita la comprensión de su desempeño en la clasificación de documentos, dificultando la selección de técnicas adecuadas; las *limitaciones de los modelos SSL*, especialmente en contextos con pocos datos etiquetados, afectan su precisión y eficacia, reduciendo la calidad de los resultados obtenidos; la *adaptación de dominio* evidencian dificultades en la generalización de estos modelos hacia nuevos contextos de datos, la falta de técnicas robustas de transferencia de aprendizaje limita la capacidad de aprovechar conocimiento previo; finalmente, el *límite de decisión* que introduce ambigüedad en el etiquetado de documentos, particularmente en aquellos ubicados en los bordes de las agrupaciones, esto afecta negativamente la precisión e incrementa la incertidumbre en la asignación de categorías. En respuesta a estas limitaciones, se propone el desarrollo de un *marco comparativo* que evalúe el desempeño de modelos SSL en distintos contextos de clasificación documental. Se plantea la implementación de un *modelo combinado* denominado COTRA que integre las mejores prácticas de distintos enfoques SSL, en este modelo se propone la aplicación fusionada de técnicas de *transferencia de aprendizaje* y *redundancia en la toma de decisiones* de clasificación para mitigar la adaptación de dominios y límite de decisión respectivamente.

Marco comparativo de modelos SSL

La solución propuesta presenta un marco comparativo que evalúa cómo funcionan los modelos SSL en diferentes contextos de clasificación de documentos, considerando variables relevantes como el tipo de documento, número de clases, cantidad de datos etiquetados y sus niveles de precisión. Esta solución permite a los usuarios identificar las ventajas y desventajas de cada modelo en contextos específicos, fortaleciendo su capacidad para tomar decisiones informadas. Al proporcionar un análisis contextualizado, se mejora la selección de modelos y se optimiza la clasificación de documentos

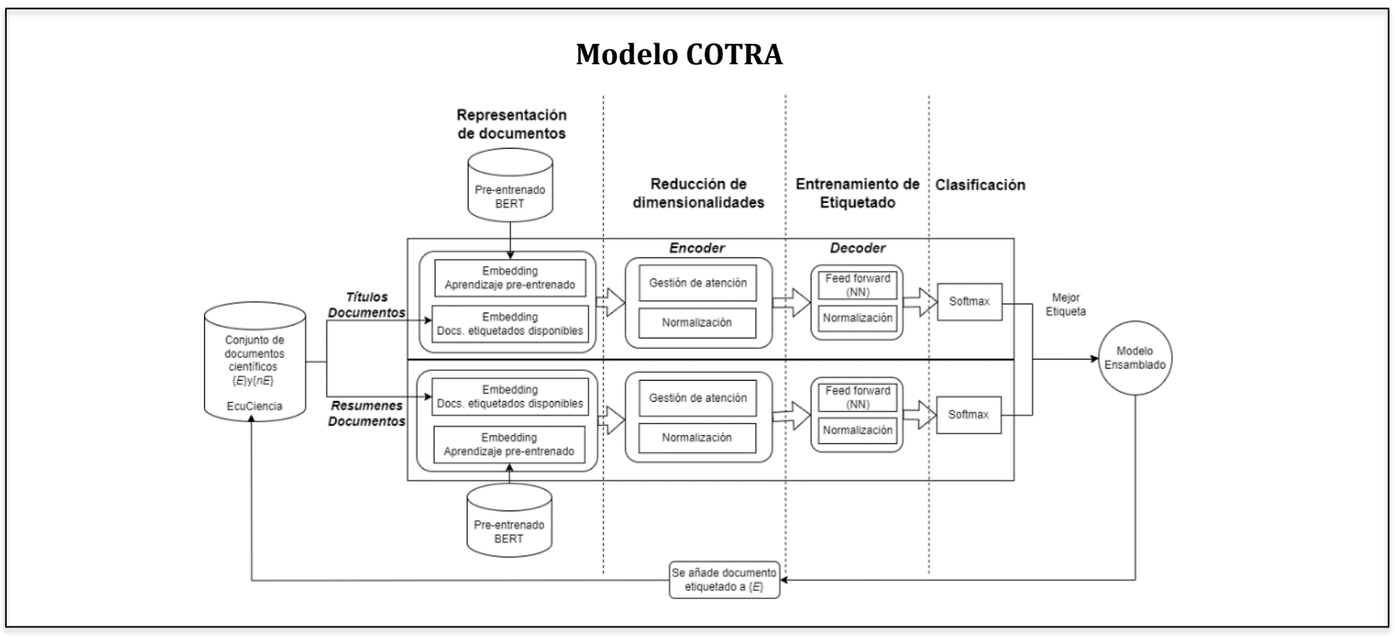
Marco Comparativo



Modelo combinado COTRA

Se ha desarrollado un modelo combinado que integra las mejores prácticas y enfoques de los diferentes modelos SSL analizados. Este modelo está diseñado para optimizar la clasificación de documentos, mejorando tanto la precisión como la eficacia, en contextos con conjuntos de datos etiquetados limitados. La solución propuesta aborda el problema combinando fortalezas de diferentes enfoques SSL, permitiendo que el modelo aproveche tanto los datos etiquetados como los no etiquetados y pre-entrenados. Esto aumenta la capacidad del sistema para clasificar documentos de manera más efectiva, lo que se traduce en una mejora en la calidad de la información y en la toma de decisiones.

Modelo COTRA

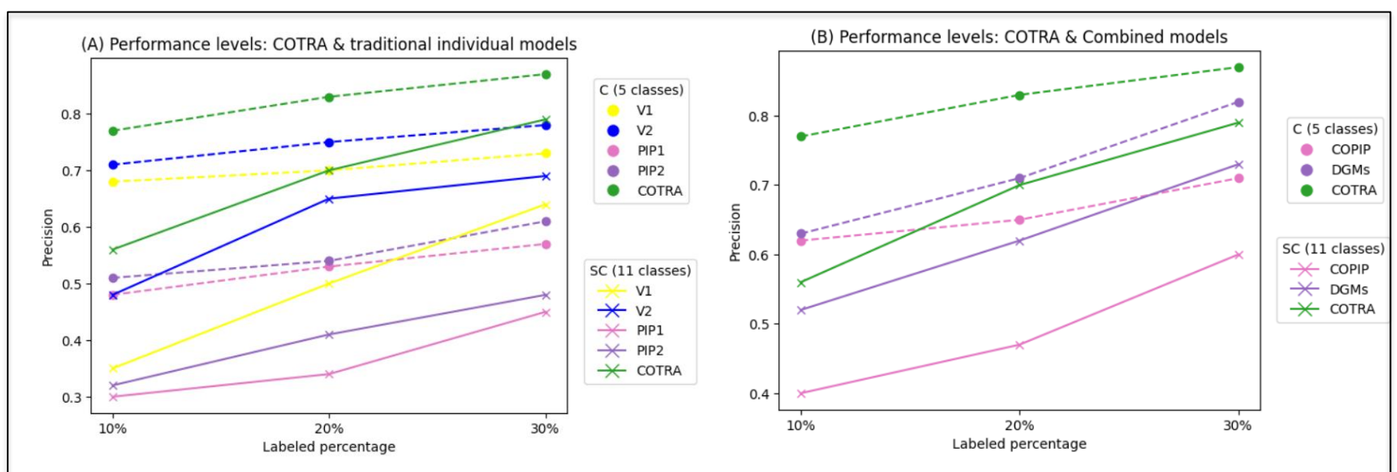


Gestión de adaptación de dominio

La propuesta es la implementación de un modelo que utiliza técnicas de transferencia de aprendizaje. Esta técnica permite que el modelo aproveche el conocimiento adquirido en un dominio fuente para mejorar su rendimiento en un dominio objetivo. Este modelo incluye el uso de redes neuronales pre-entrenadas, se ha planteado una arquitectura del modelo para que sea más flexible y facilite el ajuste fino de parámetros específicos del nuevo dominio. La implementación de técnicas de transferencia de aprendizaje ayuda a mitigar las falencias de generalización de los modelos SSL, permitiendo que estos se beneficien de datos y características de otros dominios. Esto mejora la capacidad de los modelos para aprender de manera eficiente en contextos diferentes, incrementando así su precisión y efectividad. La adaptabilidad del modelo mejora, lo que puede conducir a una mayor satisfacción del usuario al ofrecer resultados más precisos y relevantes.

Gestión del límite de decisión

Para abordar el problema de clasificación efectiva por límite de decisión, se propone la estructura de un modelo combinado, que fusiona dos modelos destinados a mejorar la robustez de las decisiones de clasificación, así se consigue una redundancia y mejora en la identificación y el tratamiento de documentos en los bordes de las agrupaciones. Esta solución resuelve el problema planteado al permitir que el modelo disponga de redundancia en la decisión de clasificación, así gestiona mejor los documentos que se encuentran en el límite de decisión. Así se minimiza la ambigüedad en la clasificación, aumentando la precisión en la asignación de categorías.



LÍNEAS FUTURAS

El modelo COTRA presenta ciertos desafíos que podrían abordarse en futuras investigaciones. Una de sus principales limitaciones radica en la necesidad de un ajuste fino adecuado, ya que un entrenamiento ineficiente podría llevar al modelo a sobreajustarse a las características del dominio fuente, lo que afecta su capacidad de generalización en el dominio destino. Otro desafío radica en la complejidad computacional generada por la estructura combinada utilizada para gestionar los documentos etiquetados entre las distintas vistas y los conjuntos preentrenados. La integración del transformer y el co-entrenamiento simultáneo aumentan los requerimientos de memoria y procesamiento, lo cual puede dificultar la implementación en entornos con recursos limitados. También, se podrían explorar distintos enfoques para analizar la eficiencia en la clasificación de documentos científicos mediante la estructura del modelo COTRA. Por ejemplo, en lugar de las técnicas de co-entrenamiento, sería posible evaluar el rendimiento utilizando técnicas de ensamblado de modelos, combinando múltiples clasificadores para gestionar las predicciones. Asimismo, se podría experimentar con técnicas de aprendizaje activo, permitiendo que el modelo seleccione proactivamente los ejemplos más informativos para su etiquetado. Además, sería relevante extender la aplicación del modelo COTRA a otros dominios que presenten características similares a los documentos científicos, como artículos de prensa, literatura médica, documentos legales, informes técnicos u otros. La capacidad del modelo para manejar datos no estructurados y transferir conocimiento entre dominios podría resultar beneficiosa en tareas de análisis de contenido, recuperación de información y generación de resúmenes automáticos.