Combined model for document classification based on the comparative evaluation of semi-supervised models

Alex Cevallos-Culqui[1], Claudia Pons[1,2,3], Gustavo Rodriguez[4]

[1] LIFIA - Facultad de Informática, Universidad Nacional de La Plata, La Plata, Argentina.
alex.cevallosc@info.unlp.edu.ar
[2] CAETI, Universidad Abierta Interamericana, Buenos Aires, Argentina.
[3] CIC - Comisión de Investigaciones Científicas de Buenos Aires, Buenos Aires, Argentina.
claudia.pons@uai.edu.ar
[4] Departamento de Tecnologías de Información y Comunicación, UTC, Latacunga, Ecuador.
gustavo.rodriguez@utc.edu.ec

Abstract. The limited availability of labeled documents limits automatic document classification and increases labeling costs. In this context, semi-supervised learning (SSL) constitutes a viable alternative; however, selecting appropriate models remains challenging due to the lack of comparative analyses that clearly identify their strengths and limitations. This work presents two main contributions. The first is a Systematic Literature Review (SLR), based on PICOC and PRISMA, which systematizes the state of the art of SSL models applied to document classification and identifies two recurring problems: domain adaptation and the definition of decision boundaries. The second contribution is COTRA, a semi-supervised multi-view model for scientific document classification that integrates co-training and transfer learning techniques through pre-trained BERT representations. Compared with individual models, COTRA jointly addresses domain adaptation and decision boundary issues through a multi-view representation that combines complementary information from documents. This reduces uncertainty in label assignment and improves generalization in scenarios with limited labeled data. The proposed model is implemented and experimentally evaluated on the EcuCiencia dataset, which belongs to a university scientific platform, and its performance is compared with individual and combined models under different percentages of labeled documents. The results show that COTRA achieves an accuracy of up to 0.81, outperforming the reference models and demonstrating the relevance of integrating a multi-view strategy with transfer learning for document classification.

Keywords: co-training, transfer learning, document classification.

Modelo combinado para clasificación de documentos, a partir de la evaluación comparativa de modelos semi-supervisados

Resumen. La escasez de documentos etiquetados limita la clasificación automática de documentos y eleva los costos de etiquetado. En este contexto, el aprendizaje semi-supervisado (SSL) constituye una alternativa viable; sin embargo, la selección de modelos adecuados se dificulta por la ausencia de análisis comparativos que evidencien sus fortalezas y limitaciones. Este trabajo presenta dos aportes principales. El primero es una Revisión Sistemática de la Literatura (SLR), basada en PICOC y PRISMA, que sistematiza el estado del arte de los modelos SSL aplicados a la clasificación de documentos e identifica dos problemas recurrentes: la adaptación de dominio y la definición de límites de decisión. El segundo aporte es COTRA, un modelo semi-supervisado multivista para la clasificación de documentos científicos que integra técnicas de co-entrenamiento y aprendizaje por transferencia, mediante representaciones preentrenadas de BERT. Frente a modelos individuales, COTRA aborda de manera conjunta los problemas de adaptación de dominio y límites de decisión mediante una representación multivista que combina información complementaria de los documentos. Esto reduce la incertidumbre en la asignación de etiquetas y mejora la generalización en escenarios con pocos datos etiquetados. El modelo propuesto se implementa y evalúa experimentalmente sobre el conjunto de datos EcuCiencia, perteneciente a una plataforma científica universitaria, y su desempeño se compara con modelos individuales y combinados bajo diferentes porcentajes de documentos etiquetados. Los resultados muestran que COTRA alcanza una precisión de hasta 0,81, superando a los modelos de referencia y evidenciando la pertinencia de integrar una estrategia multivista con aprendizaje por transferencia para la clasificación de documentos.

Palabras clave: co-entrenamiento, aprendizaje de transferencia, clasificación de documentos.

1 Introducción

La creciente cantidad de documentos digitales disponibles en repositorios académicos e institucionales ha generado un desafío significativo en la clasificación y categorización de los documentos, afectando tanto la facilidad de búsqueda de documentos como la calidad de la toma de decisiones en diversos ámbitos, desde la investigación científica hasta la gestión empresarial (Macgregor, 2023). La dificultad de realizar un análisis de texto efectivo del significado de un documento generalmente provoca errores en su clasificación, limitando así la capacidad de las organizaciones para extraer conocimiento valioso de su información y facilitar su intercambio (Kuckartz, 2019).

En este ámbito, se han implementado algunas estrategias para abordar el desafío de la clasificación y categorización de documentos. Un importante enfoque es el uso de

modelos de aprendizaje supervisado y no supervisado, para la clasificación de texto. Por ejemplo, estudios han demostrado que los modelos de clasificación automática, como los basados en entrenamientos supervisados (Ginde et al., 2016) y no supervisados (McNie et al., 2016) (Ibáñez, 2015) (Maridueña et al., 2016), pueden alcanzar niveles de precisión considerables en la categorización de documentos. Estos modelos analizan características por medio de técnicas de procesamiento de lenguaje natural (Natural Language Processing, NLP), lo que permite automatizar la clasificación.

Sin embargo, los modelos supervisados dependen en gran medida de conjuntos de datos etiquetados, cuya disponibilidad puede ser limitada, lo que restringe su aplicabilidad en contextos donde las etiquetas son escasas o costosas de obtener (Jiang et al., 2020). Por otro lado, los modelos no supervisados, aunque útiles para descubrir patrones en datos no etiquetados, a menudo carecen de la precisión necesaria para clasificaciones más finas (Almuqati et al., 2024). Esta limitación ha llevado a un creciente interés en los modelos semi-supervisados, que combinan las ventajas de ambos enfoques al utilizar un pequeño número de etiquetas junto con grandes volúmenes de datos no etiquetados (Van Engelen & Hoos, 2020). Estos modelos permiten una mayor flexibilidad y eficacia en la clasificación, facilitando el aprendizaje en escenarios donde la disponibilidad de datos etiquetados es un obstáculo (Padmanabha et al., 2018). No obstante, son escasos los estudios focalizados en una valoración de los modelos semi-supervisado (Su et al., 2021).

Con este preámbulo los problemas que se abordan en este estudio son los siguientes:

Escasez de análisis de modelos semi-supervisados.- Existe dificultad en acceder a información comparativa sobre el desempeño de los modelos semi-supervisados en la clasificación de documentos (Shrutika & Prabukumar, 2017). Actualmente, hay limitados estudios que aborden esta área, y las revisiones existentes no se enfocan específicamente en la clasificación de documentos (Cevallos-Culqui et al., 2024), lo que dificulta la capacidad de los usuarios para tomar decisiones informadas. Sin disponer de un análisis comparativo claro, investigadores y profesionales enfrentan dificultades para elegir el modelo más adecuado para sus necesidades. Esto puede llevar a la implementación de técnicas ineficaces, resultando en una clasificación subóptima de documentos que afecta la gestión de información en las organizaciones. Las consecuencias incluyen baja eficiencia, costos elevados y decisiones erróneas que impactan negativamente en el acceso a información relevante (Altınel & Ganiz, 2016).

Limitaciones de los modelos Semi-supervisados.- A través de la revisión de los diferentes modelos semi-supervisados, se ha identificado que algunos métodos no logran satisfacer adecuadamente las necesidades de clasificación (Khan & Lee, 2019), especialmente en contextos donde los conjuntos de datos etiquetados son escasos o inexistentes, lo que resulta en limitaciones del modelo en precisión y eficacia (Barman & Chowdhury, 2018) (Emadi et al., 2021). Este problema es significativo, ya que la clasificación de documentos es fundamental en diversas aplicaciones, desde la gestión de información hasta la automatización de procesos. Una clasificación ineficaz puede comprometer la calidad de la información y afectar la toma de decisiones, impactando directamente en la operación y competitividad de las organizaciones (Li et al., 2020). La falta de un modelo robusto que combine las mejores prácticas identificadas en la revisión de modelos semi-supervisados, limita la capacidad de las organizaciones para clasificar documentos de manera efectiva (Yang & Loog, 2018). Esto puede traducirse

para las empresas en pérdidas operativas, decisiones basadas en información incorrecta y oportunidades no aprovechadas.

Desafíos de la adaptación de dominio en modelos semi-supervisados.- Existe dificultad para adaptar modelos semi-supervisados a nuevos dominios de datos. Estos modelos, que combinan información etiquetada, no etiquetada y pre-entrenada de otros dominios a menudo tienen falencias de generalización cuando se aplican a contextos diferentes, lo que resulta en un rendimiento subóptimo (Liu, 2019). La falta de técnicas efectivas para transferir el conocimiento entre dominios limita la capacidad de estos modelos para aprender de manera eficiente en diversas situaciones (Guo & Yao, 2021).

Límite de decisión y clasificación efectiva.- Existe dificultad en el etiquetado óptimo de documentos sin etiquetar, causada por la presencia de documentos ubicados en los bordes de las agrupaciones a lo que se denomina Límite de decisión (Mohammed et al., 2022). Estos documentos, que presentan etiquetas difusas, afectan negativamente la precisión del proceso de clasificación. A pesar de las técnicas de etiquetado disponibles, el fenómeno del límite de decisión provoca incertidumbre en la asignación de categorías, lo que resulta en etiquetados incorrectos y un deterioro del rendimiento del modelo, especialmente cuando se manejan múltiples categorías (Beltagy et al., 2020). La importancia de este problema radica en que el etiquetado adecuado de documentos es crucial para el éxito de los modelos de clasificación (Ruder et al., 2019) y, en consecuencia, para la efectividad de los procesos organizacionales que dependen del análisis de datos (Ouali et al., 2020). Un etiquetado inexacto no solo puede llevar a decisiones erróneas, sino que también afecta la calidad del servicio al usuario. En un entorno donde la información se genera y consume a gran velocidad, la capacidad de clasificar documentos de manera precisa es fundamental para mantener la competitividad y responder a las necesidades del mercado.

Con estas consideraciones, el presente estudio tiene como objetivo diseñar y evaluar un modelo combinado de clasificación semi-supervisada orientado a optimizar el proceso de etiquetado y mejorar la precisión en la clasificación de documentos científicos. La contribución del trabajo se estructura en dos niveles. En primer lugar, se desarrolla una Revisión Sistemática de la Literatura que permite sistematizar el estado del arte de los modelos de aprendizaje semi-supervisado aplicados a la clasificación de documentos, identificando sus principales fortalezas, limitaciones y desafíos recurrentes. En segundo lugar, a partir de los hallazgos obtenidos, se propone COTRA, un modelo semi-supervisado multivista que integra co-entrenamiento y aprendizaje por transferencia mediante representaciones preentrenadas de BERT. A diferencia de enfoques previos centrados en modelos individuales, COTRA combina dos vistas complementarias de los documentos para reforzar progresivamente el proceso de etiquetado y mejorar la generalización del clasificador en escenarios con baja disponibilidad de datos etiquetados. De esta manera, la originalidad de la propuesta radica en articular una revisión comparativa previa con el diseño de una arquitectura multivista específica para documentos, orientada a abordar conjuntamente los problemas de adaptación de dominio y definición de límites de decisión. La experimentación del modelo se implementa y evalúa sobre documentos científicos

recopilados en la plataforma EcuCiencia de la Universidad Técnica de Cotopaxi (UTC), disponible en el dominio https://ecuciencia.utc.edu.ec/. De esta manera, la originalidad de la propuesta radica en articular una revisión comparativa previa con el diseño y validación experimental de una arquitectura multivista específica para documentos científicos, orientada a abordar conjuntamente los problemas de adaptación de dominio y definición de límites de decisión.

2 Método

Inicialmente se plantea una Revisión Sistemática de Literatura (RSL), en este estudio se presenta el detalle de una revisión que amplía el conocimiento de los modelos de aprendizaje semi-supervisados aplicados a la clasificación de documentos. En dicha revisión se analizaron 332 investigaciones, filtrando 46 estudios relevantes publicados entre 2017 y 2022. El análisis sigue un protocolo sistemático basado en guías reconocidas como PRISMA, utilizando el enfoque PICOC para estructurar las búsquedas, los detalles sobre los criterios de selección se presentan en la revisión elaborada (Cevallos-Culqui et al., 2023). Se recopilaron datos de repositorios como SCOPUS, IEEE Xplore, Springer y Elsevier, organizando los modelos en categorías según el tipo de aprendizaje semi-supervisado (Ver Fig. 1).

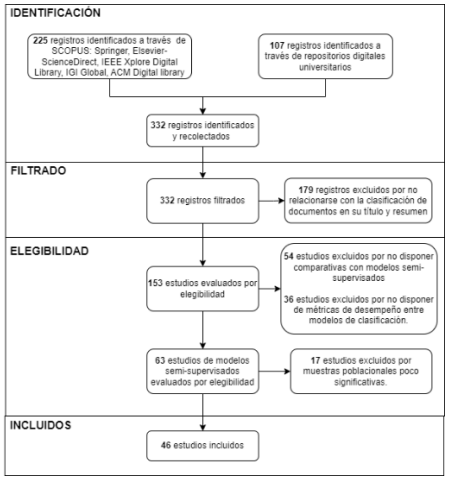

Fig. 1. Diagrama de flujo PRISMA.
Tomado de (Cevallos-Culqui et al., 2023)

La evaluación del rendimiento de los modelos se realizó considerando diferentes variables asociadas al proceso de clasificación. Entre ellas se incluyeron las técnicas empleadas en cada etapa del modelo: representación de documentos, reducción de dimensionalidad, etiquetado y clasificación. Asimismo, se analizaron determinadas variables de los conjuntos de datos utilizados (Dataset), tales como el número de

documentos (Docs.), el número de clases (Clases) y la proporción de documentos etiquetados (E), no etiquetados (NE) y destinados a prueba (Prueba).

Además, se evaluó la precisión de los modelos (P%), mediante análisis estadísticos en R-Studio, utilizando gráficos de meta-análisis (forestplot) para interpretar resultados. La precisión de los modelos fue el principal criterio, calculada a partir de la proporción de verdaderos positivos (TP) frente a falsos positivos (FP): TP/(TP+FP).

Este marco comparativo que evalúa cómo funcionan los modelos semi-supervisados en diferentes contextos de clasificación de documentos, considerando variables relevantes, ha permitido identificar las ventajas y desventajas de cada modelo en contextos específicos, fortaleciendo la capacidad para tomar decisiones informadas. Al proporcionar un análisis contextualizado, se mejora la selección de modelos y se optimiza la clasificación de documentos.

Considerando las ventajas y desventajas identificadas se ha planteado la estructura de un modelo denominado COTRA, el modelo combina el co-entrenamiento y el aprendizaje por transferencia. Un modelo de co-entrenamiento también se lo conoce como modelo multi-vista, por la razón de que plantea una estructura de entrenamiento con distintos enfoques de sus características (Masmoudi et al., 2021) . Los documentos etiquetados pueden ser tratados de forma específica o compartida entre las vistas con el fin de reforzar el proceso de entrenamiento (Jia et al., 2021). La evaluación de la predicción, se convierte en un proceso más complejo por la diversidad de clasificadores existentes y las estrategias de decisión (Jia et al., 2022).

Por otro lado, los modelos de transferencia permiten aprovechar el conocimiento adquirido de un dominio fuente, es decir se dispone de un aprendizaje pre-entrenado que puede ser transferido y adaptado a un dominio destino, para el entrenamiento de un nuevo modelo con un conjunto de etiquetados reducido (Cevallos-Culqui et al., 2023). Este aprendizaje de transferencia se la lleva a cabo utilizando estructuras Transformers, que reciben texto de entrada para ser procesado con embeddings, encoders, decoders y los datos pre-entrenados para generar una salida.

A partir de este análisis, se estructura una propuesta de modelo cuya contribución frente a enfoques previos se sintetiza en la Tabla 1, donde se identifican sus principales elementos diferenciales. Esta comparación evidencia que la originalidad de la propuesta no se limita a la integración de co-entrenamiento y aprendizaje por transferencia, sino que también radica en la revisión sistemática previa que fundamenta el diseño del modelo, una representación multivista basada en distintas vistas textuales de los documentos y una validación experimental sobre documentos científicos recopilados en una plataforma científica universitaria.

Tabla 1. Diferenciación de la propuesta COTRA frente a enfoques previos

Aspecto comparativo	Enfoques previos	Propuesta COTRA
Base metodológica	Aplicación de modelos SSL sin un análisis comparativo previa específica para clasificación de documentos.	Parte de una RSL que identifica fortalezas, limitaciones y desafíos recurrentes de los modelos SSL.
Tipo de modelo	Uso de modelos individuales o estrategias SSL evaluadas de forma independiente.	Integra co-entrenamiento y aprendizaje por transferencia en una arquitectura SSL.
Problema abordado	Enfoque parcial sobre precisión, etiquetado o clasificación.	Aborda conjuntamente la adaptación de dominio y la definición de límites de decisión.
Proceso de etiquetado	Dependencia de conjuntos etiquetados o estrategias de pseudoetiquetado individuales.	Aplica un proceso de etiquetado apoyado en vistas complementarias y representaciones preentrenadas de BERT.

La estructura del modelo propuesto combina dos vistas que refuerzan su entrenamiento con un aprendizaje de transferencia soportado en una estructura transformer con datos pre-entrenados de BERT. La primera vista es entrenada con los títulos de documentos y la segunda vista es entrenada con los resúmenes de los documentos. Los entrenamientos de las dos vistas son asociadas para seleccionar la mejor predicción. En la Fig. 2 se pude apreciar un resumen del modelo.

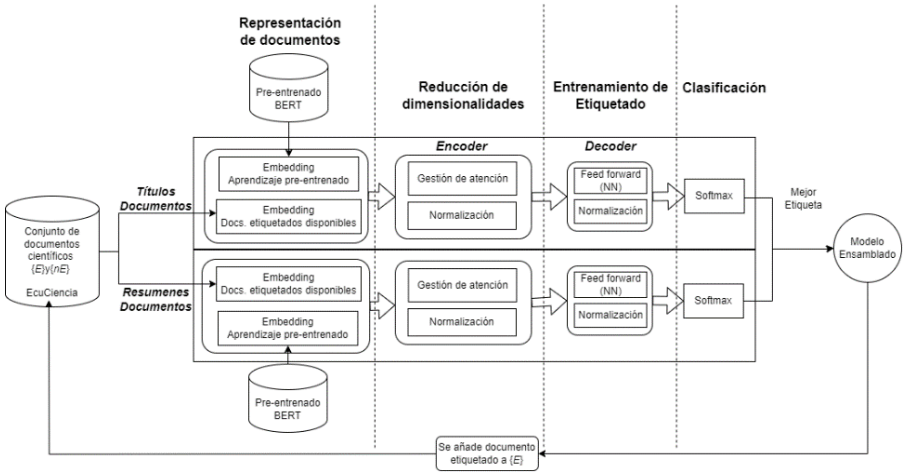

Fig. 2. Estructura del Modelo COTRA

Las dos vistas planteadas siguen el proceso de clasificación de los modelos de co-entrenamiento, al considerar también el conjunto de datos pre-entrenados de BERT, el transformer ofrece las siguientes fortalezas para las dos vistas:

Representación de documentos.- Las representaciones textuales se obtienen mediante LaBSE/BERT. El tokenizador asociado transforma títulos y resúmenes en secuencias de entrada compuestas por identificadores de tokens. Posteriormente, dichas entradas son procesadas por BertModel.from_pretrained, del cual se extrae la representación semántica del texto. Esta representación se utiliza como entrada para la capa de clasificación de cada vista, de este modo, BERT actúa como codificador preentrenado para títulos y resúmenes, mientras que la contribución específica de COTRA se encuentra en la integración de dichas representaciones dentro de un esquema de co-entrenamiento multivista.

Reducción de dimensionalidades. - Por medio de un proceso de normalización, con los encoders se adecua las dimensionalidades de representación de las características para mantener un rango de compatibilidad de entrada-salida y mejorar su entrenamiento.

Etiquetado: Para el proceso de entrenamiento se hace uso de los documentos etiquetados disponibles en el TD, a estos documentos se agrega el conocimiento pre-entrenado de BERT a través de decoders, ambos conjuntos de conocimiento son fusionados y son entrenados en una red neuronal.

Clasificación.- Los decoders poseen como última capa a softmax que tiene la función de distribuir probabilidades en un conjunto de etiquetas para definir cuál es la categoría más idónea para el texto de entrada y definir la tarea de clasificación.

Proceso de entrenamiento multivista de COTRA

El proceso de entrenamiento de COTRA se basa en una arquitectura multivista compuesta por dos clasificadores independientes. El modelo se estructura a partir de dos vistas que representan diferentes componentes textuales del documento; en este estudio, dichas vistas se implementan mediante los títulos y resúmenes del documento. Ambas vistas tienen como objetivo predecir la misma clase final del documento, aunque pueden generar predicciones diferentes debido a la naturaleza y extensión de la información textual procesada. Por esta razón, cada clasificador se entrena de forma independiente, con parámetros e hiperparámetros ajustados según la longitud del texto de entrada, considerando secuencias más cortas para títulos y secuencias más extensas para resúmenes.

En cada iteración, los documentos no etiquetados son procesados por los clasificadores de ambas vistas. Las predicciones generadas se evalúan mediante la probabilidad asignada a la clase estimada, y únicamente se incorporan al conjunto de

entrenamiento aquellas pseudoetiquetas cuya confianza sea mayor o igual a 0,80. Este umbral permite reducir la incorporación de etiquetas inciertas y controlar la propagación de errores durante el proceso de co-entrenamiento. Cuando ambas vistas producen predicciones distintas, se selecciona la clase asociada al clasificador que presenta menor pérdida de validación, priorizando la vista con mejor comportamiento de generalización. Los documentos aceptados como pseudoetiquetados se integran progresivamente al conjunto de entrenamiento, reforzando el aprendizaje de los clasificadores en las siguientes iteraciones.

El criterio de parada del entrenamiento combina un número máximo de épocas con un mecanismo de parada temprana basado en la pérdida de validación. Para la vista basada en títulos se establecen 10 épocas, mientras que para la vista basada en resúmenes se consideran 15 épocas. Adicionalmente, se monitorea la pérdida de validación en cada época; si esta disminuye, se actualiza el mejor valor registrado y se reinicia el contador de paciencia. En cambio, si la pérdida aumenta respecto al mejor valor observado, el contador se incrementa. Dado que la paciencia se configura en 1 y el margen mínimo de mejora en 0, el entrenamiento se detiene cuando la pérdida de validación no mejora en una evaluación consecutiva. Este mecanismo busca reducir el sobreajuste y evitar entrenamientos innecesarios.

Como entrada COTRA recibe: las dos vistas (título y resúmenes) $\{V\}=\{(V_i)|i=1,2\}$, el conjunto $\{Esd\}=\{(d_i)|i=1,\ldots,n\}$ que contiene documentos etiquetados con pre-entrenamiento de BERT (SD), el conjunto $\{Etd^{(v1)}\}=\{(d_j)|j=1,\ldots,n\}$ que contiene la vista 1 con los títulos de los documentos etiquetados del TD, el conjunto $\{Etd^{(v2)}\}=\{(d_j)|j=1,\ldots,n\}$ que contiene la vista 2 con los resúmenes de los documentos etiquetados del TD, el conjunto $\{nEtd^{(v1)}\}=\{(d_k)|k=n+1,\ldots,n+m\}$ que contiene documentos no etiquetados de la vista 1, el conjunto $\{nEtd^{(v2)}\}=\{(d_l)|l=n+1,\ldots,n+m\}$ que contiene documentos no etiquetados de la vista 2, el conjunto $\{Cl\}=\{(c_k)|k=1,\ldots,n\}$ que posee las clases de distribución que corresponden a las líneas de investigación de los documentos, dos estrategias de representación de documentos dr_1(SD), dr_2(TD) y el clasificador C1.

El detalle del método de desempeño de COTRA se presenta en el Algoritmo 1. En primera instancia se realiza la representación de documentos acorde a su tipo de texto (título o resumen) de *{Esd}*, *{Etd}* y*{nEtd}* en dos conjuntos de características representados por embeddings. Definida la representación, el clasificador entrena en función del conjunto de documentos etiquetados y el conjunto de datos pre-entrenado de BERT. Posteriormente los documentos no etiquetados son sometidos al clasificador entrenado para predecir su etiqueta de clase de distribución *{Cl}*. El entrenamiento de las dos vistas es compartido para volver a entrenar el modelo con un conjunto más amplio de documentos etiquetados, así el modelo paulatinamente dispone de un mayor número de etiquetados para mejorar su rendimiento de predicción.

Algoritmo 1. Modelo de co-entrenamiento COTRA

Entrada:
Documentos etiquetados SD $\{Esd\}=\{(d_i)|i=1,\ldots,n\}$;
Documentos (títulos) etiquetados TD $\{Etd^{(v1)}\}=\{(d_j)|j=1,\ldots,n\}$;
Documentos (resúmenes) etiquetados TD $\{Etd^{(v2)}\}=\{(d_j)|j=1,\ldots,n\}$;
Documentos no etiquetados TD$\{nEtd^{(v1)}\}=\{(dj)|j=n+1,\ldots,n+m\}$;
Documentos no etiquetados TD$\{nEtd^{(v2)}\}=\{(dj)|j=n+1,\ldots,n+m\}$;
Clases de distribución $\{Cl\}=\{(c_k)|k=1,\ldots,n\}$;
Para $\{Esd\}$y$\{Etd\}$ se establece dos representaciones de documentos dr_1 y dr_2;
Para $\{Etd^{(v1}\}$y$\{Etd^{(v2)}\}$ se establece dos clasificadores C_1 y C_2;
Instancias de entrenamiento por vista: *V1 y V2*;
$\{dr\}$={embeddings};
$\{C\}=\{NN\}$;

1: **repeat**
2: Representación de documentos $dr^{(v1)}$ y $dr^{(v2)}$ en cada vista para los documentos de SD → $Esd^{(V)}$, TD → $Etd^{(V)}$ y $nEtd^{(V)}$;
3: Entrena el clasificador de cada vista $C^{(v1)}$ y $C^{(v2)}$ con los documentos etiquetados $Esd^{(v1)}$, $Etd^{(v1)}$ y $Esd^{(v2)}$, $Etd^{(v2)}$;
4: Clasifica $nEtd^{(v1)}$ y $nEtd^{(v2)}$ con los clasificadores C_{v1} y C_{v2} de su clase Cl;
5: Los documentos que se van etiquetando clasificados por C_{v1} y C_{v2} se añaden al conjunto de etiquetados $Etd_{cl}^{(v1)}$ y $Etd_{cl}^{(v2)}$;
6: Se retira de $nEtd^{(v1)}$ y $nEtd^{(v2)}$ los documentos clasificados;
7: **until** $\{nEtd\}=\emptyset$
8: Se comparte $Etd^{(v1)}$ y $Etd^{(v2)}$ para incrementar el conjunto de entrenamiento;

Salida: Documentos etiquetados $\{E\}$

3 Experimentación y Resultados

Se analizaron 46 estudios sobre la clasificación de documentos mediante modelos de aprendizaje SSL, distribuidos de la siguiente manera: 11 estudios enfocados en modelos de auto-entrenamiento (ST), 10 en co-entrenamiento (CT), 9 en modelos ensamblados (ES), 5 en aprendizaje activo (AL) y 11 en aprendizaje por transferencia (AT). Para el meta-análisis, se consideró como referencia la precisión obtenida en la clasificación de documentos.

La Tabla 2 incluye un ForestPlot que permite comparar los modelos de los estudios identificados. En esta tabla, los estudios se agrupan en cinco tipos de modelos de SSL y se presentan datos como el número de documentos clasificados correctamente, el total de documentos analizados, el peso asignado a cada estudio según el modelo ajustado, el porcentaje de precisión en la clasificación, el intervalo de confianza.

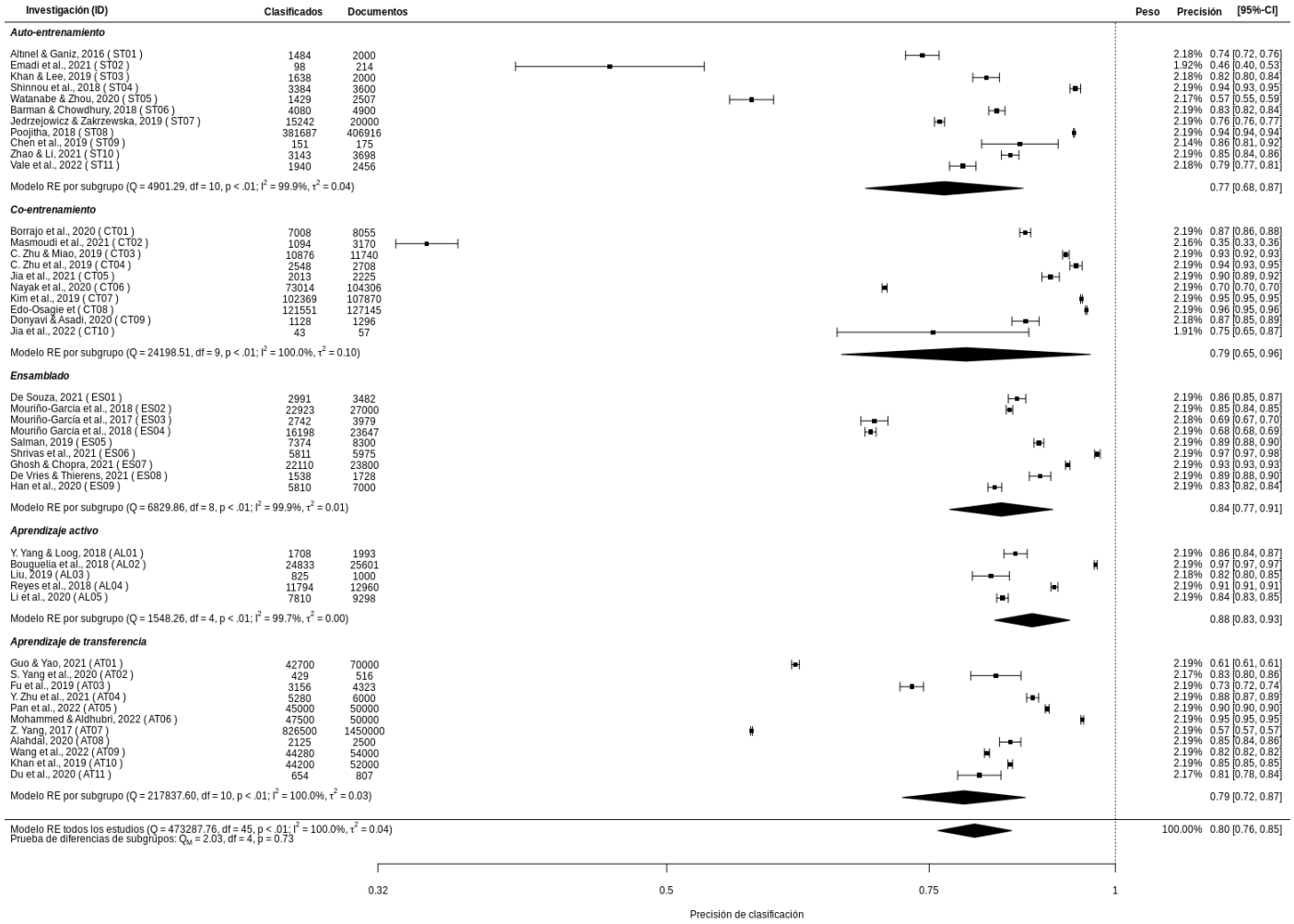

Tabla 2. ForestPlot de niveles de precisión de SSL. Tomado de (Cevallos-Culqui et al., 2023)

Así también, se identifican y definen las fortalezas y limitaciones más relevantes y recurrentes de los modelos de aprendizaje semi-supervisado (ver Fig. 3.). Este análisis se llevó a cabo mediante una evaluación de una lista de casos de estudio asociados a cada tipo de modelo SSL, integrando además los resultados obtenidos a partir de un meta-análisis comparativo. Este enfoque permitió no solo agrupar las principales características y restricciones inherentes a cada tipo de modelo, sino también establecer un marco de referencia que facilita la comprensión de su desempeño en diferentes escenarios aplicativos. La combinación de análisis cualitativos y cuantitativos asegura una visión integral, abarcando aspectos técnicos, contextuales y prácticos de los modelos SSL.

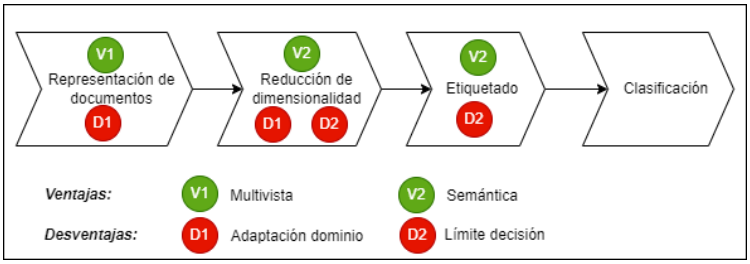

Fig. 3. Ventajas y desventajas del proceso de clasificación SSL.

Para la evaluación de COTRA se prepara un trabajo experimental diseñado con Python y el framework PyCharm, esta propuesta es un segmento de la plataforma científica EcuCiencia que recopila documentos científicos de los investigadores de la UTC. En el presente estudio se evalúa la eficiencia de clasificación del modelo propuesto con métricas de rendimiento de precisión y un conjunto de datos formado por los resúmenes de los documentos científicos recopilados de la plataforma EcuCiencia.

En la Tabla 3 se proporciona información sobre el conjunto de datos EcuCiencia, posee 898 documentos científicos recopilados del repositorio. El 90% de los datos fueron seleccionados para entrenamiento, el 5% para validación y el 5% para pruebas. Además, se seleccionaron diferentes porcentajes de documentos etiquetados: 10%, 20% y 30% de los datos de entrenamiento, para realizar diversas pruebas (al calcular el porcentaje de documentos, se aproxima al número entero superior o inferior más cercano).

Tabla 3. Características y parámetros del conjunto de datos

Fuente	Conjunto de datos	Clases	10%	20%	30%
EcuCiencia	808 docs. (90% entrenamiento), 45 docs. (5% validación).	5	80	165	245
	45 docs. (5% pruebas).	11	77	165	242

En la Tabla 4 se presenta la distribución de las clases para el conjunto de datos, estos documentos pertenecen a cinco departamentos de la UTC, que representan las clases de primer nivel, las etiquetas para este primer nivel son: (C1) Facultad de Ciencias Agropecuarias, (C2) Facultad de Recursos Naturales, (C3) Facultad de Ingeniería y Ciencias Aplicadas, (C4) Departamento de Ciencias Administrativas y Económicas, y (C5) Departamento de Ciencias Sociales, Artes y Educación. Cada departamento abarca diversas líneas de investigación, sumando un total de once categorías que representan las clases de segundo nivel (SC). La experimentación de este modelo implica la utilización de ambos conjuntos de clases (C y SC) para el entrenamiento individual de los modelos de co-entrenamiento y transferencia con cada vista.

Tabla 4. Clases o líneas de investigación.

C		**SC**
C1	SC1	Desarrollo y seguridad alimentaria
	SC2	Salud animal
C2	SC3	Análisis, conservación y aprovechamiento de la biodiversidad local
	SC4	Planificación y gestión del turismo sostenible
C3	SC5	Energías alternativas y renovables, eficiencia energética y protección ambiental
	SC6	Procesos industriales
	SC7	Tecnologías de información y comunicación.
C4	SC8	Administración y economía para el desarrollo humano y social
	SC9	Gestión de la calidad y seguridad laboral
C5	SC10	Educación, comunicación y diseño gráfico para el desarrollo humano y social
	SC11	Cultura, patrimonio y saberes ancestrales

Con el fin de analizar la distribución de documentos entre clases, la Tabla 5 presenta la cantidad de documentos disponibles para cada una de las 5 clases y 11 subclases consideradas en la experimentación. Esta información permite verificar el grado de balance del conjunto de datos y contextualizar la interpretación de los resultados de clasificación.

Tabla 5. Distribución de documentos por clase y subclase.

C	Nº Documentos	SC	Nº Documentos
C1	156	SC1	78
		SC2	78
C2	152	SC3	74
		SC4	78
C3	248	SC5	98
		SC6	78
		SC7	72
C4	188	SC8	81
		SC9	107
C5	154	SC10	80
		SC11	74

Como se observa en la Tabla 5, la distribución de documentos presenta un desbalance leve a nivel de subclases y moderado a nivel de clases principales. En las subclases, las frecuencias oscilan entre 72 y 107 documentos, con una relación aproximada de 1,49:1 entre la subclase mayoritaria y la minoritaria. En las clases principales, la distribución varía entre 152 y 248 documentos, con una relación aproximada de 1,63:1. Si bien el conjunto de datos no es perfectamente balanceado, no se evidencia un desbalance severo. Además, la mayor concentración de documentos en C3 responde a que esta clase agrupa tres subclases, mientras que las demás clases agrupan dos.

La Tabla 6 presenta las características experimentales para los diferentes modelos de clasificación planteados. El rendimiento de COTRA es comparado con el rendimiento individual de las vistas por título (V1) y por resumen (V2), así como también con un modelo de transferencia que utiliza la abstracción de pipeline con el transformer BART, el cual dispone de un conjunto de datos pre-entrenados que complementa su entrenamiento con los títulos (PIP1) y resúmenes (PIP2) de los documentos científicos. Finalmente, se compara con el co-entrenamiento de PIP1 y PIP2 (COPIP).

Tabla 6. Características de los modelos de experimentación.

Id	Modelo	Tipo	Repr. Docs.	Clasificador
V1	Individual	Entrenamiento-Títulos	Embeddings	NN
V2	Individual	Entrenamiento-Resúmenes	Embeddings	NN
PIP1	Pipeline	Pre-entrenado / Títulos	Embeddings	NN
PIP2	Pipeline	Pre-entrenado / Resúmenes	Embeddings	NN
COPIP	Co-entrenamiento	Pre-entrenado / Títulos&Resúmenes	Embeddings	NN
COTRA	Co-entrenamiento	Pre-entrenado / Títulos&Resúmenes	Embeddings	NN

La comparación experimental se estructuró con el propósito de aislar progresivamente el efecto de cada componente del modelo propuesto. El modelo V1 permite evaluar el desempeño de una vista individual basada únicamente en títulos, mientras que V2 evalúa una vista individual basada únicamente en resúmenes. Ambos modelos emplean representaciones preentrenadas basadas en LaBSE/BERT, por lo que permiten observar el comportamiento de cada fuente textual por separado. Por su parte, PIP1 y PIP2 corresponden a otros clasificadores basados en la estrategia de clasificación zero-shot con BART, aplicada respectivamente a títulos y resúmenes. COPIP combina dichas salidas en un esquema de co-entrenamiento basado en los clasificadores BART.

Finalmente, COTRA integra las dos vistas textuales (título y resumen) mediante co-entrenamiento y aprendizaje por transferencia con representaciones LaBSE/BERT. Esta organización experimental permite comparar modelos individuales y combinados, así como analizar si la mejora de desempeño se relaciona con la integración multivista y el proceso progresivo de pseudoetiquetado, más allá del uso de representaciones preentrenadas.

En la Fig. 4 se presenta el flujo de trabajo inicial de preprocesamiento y representación de los títulos y resúmenes de los documentos científicos. La figura muestra cómo se lleva a cabo el proceso de preparación de los datos, que incluye diversas etapas como la tokenización de los textos, la eliminación de caracteres no deseados, la eliminación de palabras irrelevantes y la normalización de los términos. Posteriormente, se muestra la etapa de representación de documentos, donde se utilizan técnicas de embeddings para convertir los títulos, resúmenes y datos pre-entrenados (BERT) en vectores numéricos. Después de representar los documentos es necesario que el modelo identifique la posición de los tokens y la relación entre ellos, esto se lo

realiza con la codificación posicional. Además, previo al proceso de entrenamiento se emplea el mecanismo de atención para ponderar el nivel de relevancia existente entre los tokens para fortalecer la semántica del modelo. Esto es una representación vectorial esencial para poder aplicar algoritmos de clasificación y lograr la categorización de los documentos científicos en las diferentes clases establecidas.

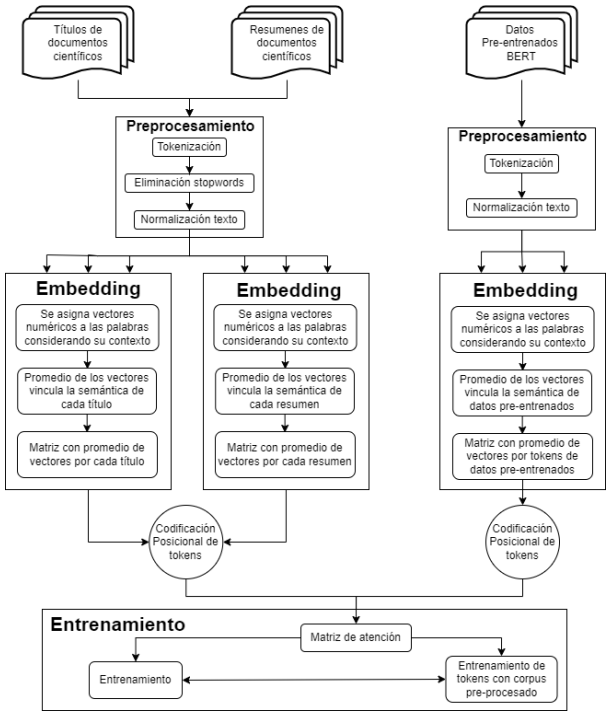

Fig. 4. Preprocesamiento y representación de documentos del modelo.

La Tabla 7 presenta los valores de los parámetros de configuración de los algoritmos utilizados en los modelos evaluados. En los modelos V1, V2 y COTRA, los parámetros de configuración son compartidos. Para evitar el sobreajuste del modelo, se establece una tasa de dropout del 50%. En la capa Linear, las características de salida del transformer BERT, que cuentan con 768 dimensiones, se configuran para adaptarse a 5/11 dimensiones correspondientes al número de clases. Se define la función de activación softmax para que la probabilidad de cada clase se distribuya en el factor 1. Por otro lado, en los modelos PIP1, PIP2 y COPIP, sus parámetros están configurados para utilizar la pipeline de clasificación zero-shot con el transformer que emplea el conjunto de datos pre-entrenados BART.

Tabla 7*1.* Configuración de parámetros de los algoritmos.

Algoritmo	Categoría	Configuración
V1	Repr. Docs. y Modelo	*Parámetros:* dropout=0.5; nn.Linear(768, [5/11]); nn.LogSoftmax(dim=1). *Hiperparámetros:* padding='max_length'; max_length= 20; truncation=True; return_tensors='pt'; batch_size=16; Epochs=10; LR = 5e-5.
V2	Repr. Docs. y Modelo	*Parámetros:* dropout=0.5; nn.Linear(768, [5/11]); nn.LogSoftmax(dim=1). *Hiperparámetros:* max_length= 150; truncation=True; return_tensors='pt'; batch_size=16; Epochs=15; LR = 5e-5.
PIP1	Modelo	*Parámetros:* task=classification; model="bart-large-mnli". *Hiperparámetros:* padding='max_length'; max_length= 20; truncation=True; return_tensors='pt'.
PIP2	Modelo	*Parámetros:* task=classification; model="bart-large-mnli". *Hiperparámetros:* padding='max_length'; max_length= 150; truncation=True; return_tensors='pt'.
COPIP	Modelo	*Parámetros:* dropout=0.5; nn.Linear(768, [5/11]); task=classification; model="bart-large-mnli". *Hiperparámetros:* max_length= 150; truncation=True; return_tensors='pt'; batch_size=16; Epochs=7; LR = 5e-5.
COTRA	Repr. Docs. y Modelo	*Parámetros:* dropout=0.5; nn.Linear(768, [5/11]); nn.LogSoftmax(dim=1). *Hiperparámetros:*max_length= 150; truncation=True; return_tensors='pt'; batch_size=16; Epochs=5; LR = 5e-5.

En lo que respecta a los hiperparámetros de los modelos son presentados en la Tabla 7. Los modelos V1, V2, PIP1, PIP2 y COTRA administran una configuración semejante, en lo que respecta a su tipo de tensor utilizado (return_tensors='pt') después de la tokenización o procesamiento se define el uso de PyTorch; en cuanto al tamaño del lote de datos que se utilizará para el entrenamiento de los modelos es de batch_size=16 con una tasa de aprendizaje de LR = 5e-5. Para la tokenización se establece procesar el texto por tamaños de secuencia (padding='max_length') si excede de su límite se trunca (truncation=True), para el modelo que procesa los títulos de los documentos (V1 y PIP1) se utiliza un max_length= 20 y para el modelo que procesa resúmenes (V2 y PIP2) un max_length= 150; en el caso de las épocas para la red neuronal en V1 se obtiene un entrenamiento óptimo con Epochs=10, para V2 por procesar mayor cantidad de texto en los resúmenes se define en Epochs=15 mientras que COTRA dispone de un texto mejor procesado y ha sido necesario definir Epochs=5.

La medida de rendimiento utilizada para evaluar la clasificación del conjunto de datos es Precisión. Esta métrica se obtiene relacionando los verdaderos positivos (TP) y los falsos positivos (FP) de la siguiente manera: TP / (TP + FP). Se ha evaluado el

desempeño de clasificación de documentos científicos de la plataforma EcuCiencia en 5 y 11 líneas de investigación mediante esta métrica.

En la Tabla 8 se presenta los valores de precisión de los modelos de clasificación V1, V2, PIP1, PIP2, COPIP y COTRA entrenados con diferentes porcentajes (10%, 20%, 30%) de documentos etiquetados, para la clasificación en 5 y 11 clases. La precisión del modelo V1, que utiliza la arquitectura BERT y se entrena con los títulos de los documentos, varía de 0,66 (10%) a 0,72 (30%) para cinco clases (C), y de 0,38 (10%) a 0,59 (30%) para once clases (SC). El modelo V2, que también utiliza BERT y se entrena con los resúmenes de los documentos, muestra la tercera precisión más alta, que varía entre 0,71 y 0,78 para C, y entre 0,47 y 0,70 para SC.

Tabla 8. Valores de precisión de los modelos

	C (5 clases)			**SC (11 clases)**		
Modelo	10%	20%	30%	10%	20%	30%
V1	0,66	0,68	0,72	0,38	0,50	0,59
V2	0,71	0,75	0,78	0,47	0,63	0,70
PIP1	0,48	0,53	0,57	0,30	0,34	0,45
PIP2	0,51	0,54	0,61	0,32	0,41	0,48
COPIP	0,62	0,65	0,71	0,4	0,47	0,6
COTRA	**0,75**	**0,77**	**0,81**	**0,54**	**0,68**	**0,75**

Con respecto al modelo PIP1, entrenado con los títulos de los documentos en una estructura BART, presenta una precisión que varía entre 0,48 y 0,57 para C y entre 0,30 y 0,45 para SC. Por otro lado, el modelo PIP2, también entrenado con BART utilizando los resúmenes de los documentos, muestra una precisión de 0,51 a 0,61 para C y de 0,32 a 0,48 para SC. En cuanto a los modelos combinados, COPIP alcanza valores que oscilan entre 0,62 y 0,71 para C, y entre 0,40 y 0,60 para SC. Finalmente, se observa que el modelo de co-entrenamiento COTRA exhibe la mayor precisión entre todos los modelos, con valores que varían entre 0,75 y 0,81 para C y entre 0,54 y 0,75 para SC.

La evaluación del rendimiento de COTRA se presenta mediante una comparación con varios modelos individuales tradicionales (ver Fig. 5) y modelos combinados (ver Fig. 6), considerando diferentes porcentajes de documentos etiquetados y cantidades de clases. Se identifica que la precisión de todos los modelos aumenta a medida que incrementa el tamaño del conjunto de documentos etiquetados. También se observa que la precisión mejora cuando el número de clases es menor. Además, la curva correspondiente al modelo COTRA muestra la mayor precisión en todos los porcentajes de etiquetado, seguida de las curvas del V2. En contraste, la curva correspondiente a los modelos PIP presenta la menor precisión en comparación con el resto de modelos.

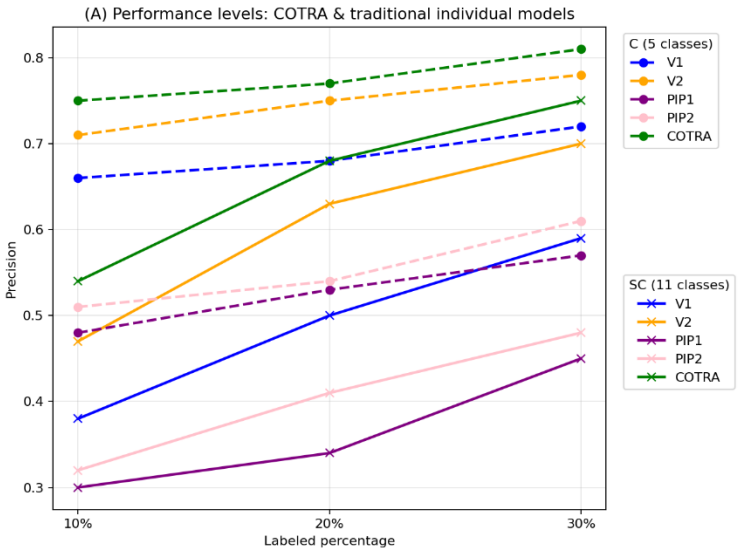

Fig. 5. Niveles de rendimiento con modelos tradicionales individuales.

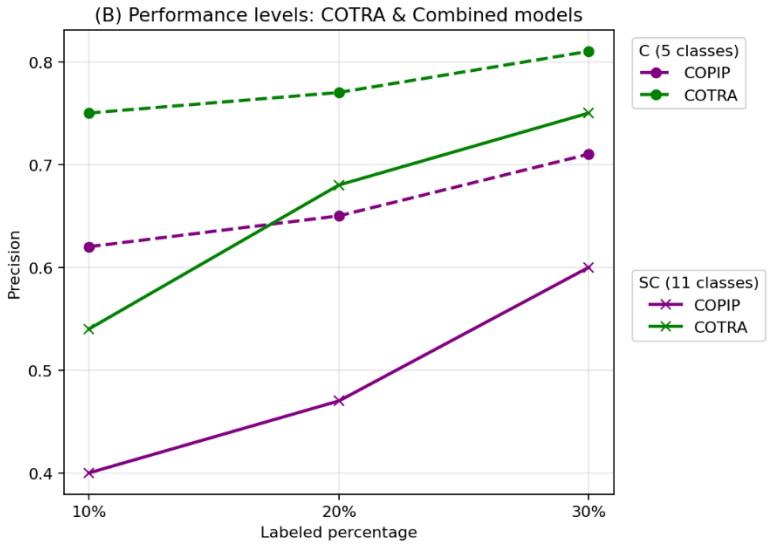

Fig. 6. Niveles de rendimiento con modelos combinados.

Para complementar el análisis descriptivo de los resultados, se realizó un análisis estadístico de las diferencias observadas en la precisión de los modelos evaluados. Se aplicó un ANOVA de dos vías considerando como factores el porcentaje de documentos etiquetados (10 %, 20 % y 30 %) y el nivel de clasificación (5 clases y 11

subclases). Los resultados muestran que el porcentaje de documentos etiquetados tiene un efecto significativo sobre la precisión de los modelos ($p = 0,0159$), lo que indica que el incremento de documentos etiquetados mejora el desempeño de clasificación. Este comportamiento se observa en COTRA, cuya precisión aumenta de 0,75 a 0,81 en la clasificación de 5 clases, y de 0,54 a 0,75 en la clasificación de 11 subclases.

Asimismo, el nivel de clasificación presenta un efecto significativo sobre la precisión de los modelos ($p = 0,00013$). En general, los modelos alcanzan mejores resultados cuando la clasificación se realiza en 5 clases principales que cuando se consideran 11 subclases, lo que evidencia que el incremento del número de categorías aumenta la complejidad de la tarea de clasificación. Por otro lado, la interacción entre el porcentaje de documentos etiquetados y el nivel de clasificación no fue significativa ($p = 0,4287$), lo que sugiere que el efecto positivo del aumento de documentos etiquetados sobre la precisión se mantiene de forma similar tanto en la clasificación de 5 clases como en la de 11 subclases.

Adicionalmente, se aplicó un ANOVA de medidas repetidas para analizar la variación de la precisión de los modelos a través de los distintos porcentajes de documentos etiquetados. Para este análisis, se consideraron como medidas repetidas los niveles de etiquetado del 10 %, 20 % y 30 %, tomando como unidades de observación las combinaciones entre modelo y nivel de clasificación. Los resultados muestran un efecto significativo del porcentaje de documentos etiquetados sobre la precisión, $F(2, 22) = 44,40$; $p < 0,001$. Esto indica que existen diferencias estadísticamente significativas entre los valores de precisión obtenidos en los niveles de etiquetado evaluados. En conjunto, estos resultados respaldan que la disponibilidad de documentos etiquetados influye significativamente en el rendimiento de los modelos y refuerzan la pertinencia de estrategias semi-supervisadas como COTRA en escenarios con disponibilidad limitada de etiquetas.

La Tabla 9 resume los resultados del análisis estadístico aplicado sobre los valores de precisión reportados en la Tabla 8. El ANOVA de dos vías permite evaluar el efecto del porcentaje de documentos etiquetados y del nivel de clasificación sobre la precisión de los modelos, mientras que el ANOVA de medidas repetidas analiza la variación del desempeño a través de los distintos porcentajes de etiquetado.

Tabla 9. Análisis estadístico de la precisión de los modelos evaluados

Análisis estadístico	Factor evaluado	gl	F	p-valor	Interpretación
ANOVA de dos vías	Porcentaje de documentos etiquetados	2, 30	4,77	0,0159	Efecto significativo
ANOVA de dos vías	Nivel de clasificación: 5 clases / 11 subclases	1, 30	19,29	0,00013	Efecto significativo
ANOVA de dos vías	Interacción porcentaje × nivel de clasificación	2, 30	0,87	0,4287	No significativo
ANOVA de medidas repetidas	Variación entre porcentajes de etiquetado	2, 22	44,40	< 0,001	Efecto significativo

Finalmente, la comparación experimental se implementó y evaluó en la plataforma científica EcuCiencia de la Universidad Técnica de Cotopaxi, feel clasificador se encuentra disponible en https://ecuciencia.utc.edu.ec/grafica/clasificacion-doc y la evaluación en http://3.84.217.107:5000/. En este entorno se ejecutaron los modelos V1, V2, PIP1, PIP2, COPIP y COTRA bajo las mismas condiciones experimentales, considerando distintos porcentajes de documentos etiquetados y dos niveles de clasificación: clases principales y subclases.

4 Conclusiones

En este trabajo se ha expuesto el diseño y la evaluación de un modelo combinado que integra el enfoque de co-entrenamiento y el aprendizaje por transferencia (COTRA) para la clasificación de documentos científicos en función de sus áreas de investigación. Esta combinación aprovecha la complementariedad de múltiples vistas y la capacidad de generalización proporcionada por modelos pre-entrenados. Se ha detallado la estructura del modelo, adoptando un enfoque de representación multivista, distribuyendo la representación de los documentos en dos vistas: títulos y resúmenes. Lo que ha permitido capturar distintas perspectivas del contenido textual, enriqueciendo la representación semántica y optimizando el proceso de clasificación.

Al emplear características complementarias en cada vista, el modelo mejora su capacidad de generalización y reduce la incertidumbre en la asignación de etiquetas, alineándose con enfoques previos en clasificación de textos mediante representaciones heterogéneas. En el caso del enfoque del aprendizaje por transferencia, la integración de modelos pre-entrenados, como BERT y BART, ha permitido transferir conocimiento lingüístico previamente adquirido, optimizando la representación de los documentos y mejorando la precisión en la clasificación. Este enfoque ha demostrado ser efectivo para mitigar las limitaciones derivadas de conjuntos de datos reducidos, facilitando el aprendizaje en escenarios con escasez de etiquetas. Asimismo, la transferencia de aprendizaje ha contribuido a mejorar la estabilidad del modelo frente a variaciones en los datos, asegurando una mayor robustez en la predicción de las etiquetas del modelo.

La experimentación realizada con el conjunto de datos EcuCiencia ha permitido evaluar la eficiencia de COTRA en comparación con otros modelos individuales y combinados. Los resultados muestran que el modelo propuesto supera a los modelos tradicionales en términos de precisión, logrando un rendimiento superior al aprovechar datos pre-entrenados y optimizar el proceso de etiquetado mediante co-entrenamiento. Además, se ha evidenciado que el incremento en el porcentaje de documentos etiquetados mejora la precisión del modelo, resaltando la importancia de contar con datos de entrenamiento representativos.

Finalmente, se ha confirmado que la combinación de co-entrenamiento y aprendizaje por transferencia mejora la generalización del modelo, reduciendo la dependencia de grandes volúmenes de datos etiquetados y optimizando la clasificación de documentos en escenarios con información limitada. Estos hallazgos sientan las bases para futuras investigaciones enfocadas en la optimización del modelo y su aplicación en otros contextos académicos y organizacionales.

5 Referencias

Almuqati, M. T., et al. (2024). Challenges in supervised and unsupervised learning: A comprehensive overview. *International Journal of Advanced Science and Engineering Information Technology, 14*(4).

Altınel, B., & Ganiz, M. C. (2016). A new hybrid semi-supervised algorithm for text classification with class-based semantics. *Knowledge-Based Systems, 108*, 50–64.

Barman, D., & Chowdhury, N. (2018). A novel semi supervised approach for text classification. *International Journal of Information Technology*.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Cevallos-Culqui, A., Pons, C., & Rodriguez, G. (2023). Semi-supervised learning models for document classification: A systematic review and meta-analysis. *Inteligencia Artificial, 26*(72), 81–111.

Cevallos-Culqui, A., Pons, C., & Rodríguez, G. (2024). A co-training model based in learning transfer for the classification of research papers. In *IEEE 12th International Conference on Intelligent Systems*.

Emadi, M., et al. (2021). A selection metric for semi-supervised learning based on neighborhood construction. *Information Processing & Management, 58*(2).

Ginde, G., et al. (2016). ScientoBASE: a framework and model for computing scholastic indicators of non-local influence of journals via native data acquisition algorithms. *Scientometrics, 108*(3), 1479–1529.

Guo, S., & Yao, N. (2021). Document vector extension for document classification. *IEEE Transactions on Knowledge and Data Engineering, 33*(8), 3062–3074.

Ibáñez, A. (2015). *Machine learning in scientometrics*. Universidad Politécnica de Madrid. http://oa.upm.es/36488/

Jia, W., et al. (2022). Semi-supervised learning via axiomatic fuzzy set theory and SVM. *IEEE Transactions on Cybernetics, 52*(6), 4661–4674.

Jia, X., et al. (2021). Semi-supervised multi-view deep discriminant representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(7), 2496–2509.

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised machine learning: a brief primer. *Behavior Therapy, 51*(5), 675–687.

Khan, J., & Lee, Y. K. (2019). LeSSA: A unified framework based on lexicons and semi-supervised learning approaches for textual sentiment classification. *Applied Sciences, 9*(24).

Kuckartz, U. (2019). Qualitative text analysis: A systematic approach. *Compendium for Early Career Researchers in Mathematics Education*, 181–197.

Li, J., et al. (2020). An effective framework based on local cores for self-labeled semi-supervised classification. *Knowledge-Based Systems, 197*.

Liu, M. (2019). *Weak supervision and active learning for natural language processing*. Monash University.

Macgregor, G. (2023). Digital repositories and discoverability: definitions and typology. In *Discoverability in Digital Repositories* (pp. 11–31). Routledge.

Maridueña, M., Leyva, M., & Febles, A. (2016). Modelado y análisis de indicadores de ciencia y tecnología mediante mapas cognitivos difusos. *Ciencias de la Información, 47*(1), 17–24.

Masmoudi, A., et al. (2021). A co-training-based approach for the hierarchical multi-label classification of research papers. *Expert Systems, 38*(4).

McNie, E. C., Parris, A., & Sarewitz, D. (2016). Improving the public value of science: A typology to inform discussion, design and implementation of research. *Research Policy, 45*(4), 884–895.

Mohammed, M., et al. (2022). Study on sentiment classification strategies based on fuzzy logic with crow search algorithm. *Research Square*.

Ouali, Y., Hudelot, C., & Tami, M. (2020). An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*.

Padmanabha, Y., Viswanath, P., & Eswara, B. (2018). Semi-supervised learning: a brief review. *International Journal of Engineering & Technology, 7*(1.8), 81.

Ruder, S., et al. (2019). Transfer learning in natural language processing. In *NAACL Tutorials* (pp. 15–18).

Shrutika, S., & Prabukumar, M. (2017). Semi-supervised techniques based hyperspectral image classification: A survey. In *International Conference on Innovations in Power and Advanced Computing Technologies* (pp. 1–8).

Su, J.-C., Cheng, Z., & Maji, S. (2021). A realistic evaluation of semi-supervised learning for fine-grained classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12966–12975).

Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning, 109*(2), 373–440.

Yang, Y., & Loog, M. (2018). A benchmark and comparison of active learning for logistic regression. *Pattern Recognition, 83*, 401–415.