
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

Autor

Alex Santiago Cevallos Culqui

Director

Dra. Claudia Pons

Codirector

Dr. Gustavo Rodríguez

Titulación

Doctorado en Ciencias Informáticas

Lugar y fecha de presentación

La Plata - Marzo de 2025



Facultad de Informática
Universidad Nacional de La Plata (UNLP)

Resumen

La escasez de documentos etiquetados en la mayoría de los conjuntos de datos en distintos dominios dificulta la correcta clasificación de documentos y la generación de aprendizaje, además de implicar altos costos en términos de recursos para su proceso de etiquetado. En este contexto, los modelos de aprendizaje semi-supervisados (Semi-Supervised Learning, SSL) surgen como una alternativa para mitigar esta limitación, sin embargo, la falta de un análisis comparativo que evidencie las fortalezas y debilidades de los distintos tipos de modelos dificulta su selección. Así, se plantea una Revisión de Literatura Sistemática (Systematic Literature Review, SLR) que identifica que las principales limitaciones de los modelos están relacionadas con los límites de decisión y la adaptación de dominio, factores que afectan sus niveles de rendimiento medidos en precisión. Es por esta razón que en la presente tesis se propone diseñar un modelo combinado de clasificación SSL que optimiza el proceso de etiquetado y la clasificación de documentos, mejorando su eficacia y niveles de precisión. Para ello, se desarrolla un marco comparativo que evalúa los distintos tipos de modelos y se implementa una estructura que integra las mejores prácticas identificadas. La metodología de trabajo para el análisis de los modelos se fundamenta en el enfoque PICOC para la estrategia de búsqueda y en la guía PRISMA para la definición de los criterios de exclusión. El modelo propuesto se estructura empleando una combinación de técnicas de co-entrenamiento y transferencia de aprendizaje (COTRA) para el procesamiento y entrenamiento de datos respectivamente, su entrenamiento se refuerza mediante el uso del conjunto de datos pre-entrenado de BERT. A diferencia de modelos previos, la estructura de COTRA fortalece el estado del arte al abordar de manera conjunta los desafíos de adaptación de dominio y límite de decisión. Esto se logra mediante una representación multivista optimizada que integra diversas representaciones de documentos con información complementaria proveniente de fuentes pre-entrenadas. Esta estrategia permite reducir la incertidumbre en la asignación de etiquetas y mejorar la capacidad de generalización en escenarios con datos etiquetados limitados, proporcionando un modelo más robusto y adaptable para la clasificación de textos en contextos con restricciones de datos. Para la evaluación de COTRA, se llevaron a cabo experimentos con documentos científicos clasificados en cinco y once categorías correspondientes a sus áreas de estudio. El modelo se comparó con modelos SSL individuales basados en auto-entrenamiento, así también con modelos que incorporan co-entrenamiento, algoritmos genéticos y aprendizaje por transferencia a través de pipelines de clasificación como enfoques zero-shot. COTRA ha logrado los mejores niveles de rendimiento en comparación con el resto de modelos, alcanzando una precisión máxima de 0,87 entre los modelos de co-entrenamiento, frente a la mejor métrica de 0,78 obtenida por los modelos individuales de auto-entrenamiento en la clasificación de cinco categorías. Estos resultados indican que el co-entrenamiento representa una estrategia efectiva para mejorar el desempeño predictivo en la clasificación de documentos.

Palabras claves

co-entrenamiento, aprendizaje transferencia, semi-supervisado, SSL, clasificación documentos, procesamiento de texto

Abstract

The limited availability of labeled documents in most datasets across diverse domains constitutes a critical obstacle to accurate document classification and the development of effective learning models, while also incurring high resource costs for labeling. In this context, Semi-Supervised Learning (SSL) models emerge as an alternative to mitigate this limitation. However, the absence of a comprehensive comparative analysis that highlights the strengths and weaknesses of different model types poses a significant challenge to informed model selection. This work proposes a Systematic Literature Review (SLR), identifying that the main limitations of SSL models are related to decision boundaries and domain adaptation, which impact their performance levels measured in accuracy. Accordingly, the present thesis aims to design a combined SSL classification model that optimizes the labeling process and document classification, improving both efficiency and accuracy levels. A comparative framework is developed to evaluate different types of models, and a structure is implemented to integrate the best identified practices. The methodology for models analysis is based on the PICOC framework for search strategy and the PRISMA guidelines for defining exclusion criteria. The proposed model is structured using a combination of co-training and transfer learning techniques (COTRA) for data processing and training, respectively. Its training is reinforced with the use of the pre-trained BERT dataset. In contrast to previous models, COTRA contributes to the state of the art by jointly addressing the issues of domain adaptation and decision boundary. This is achieved through an optimized multi-view representation that integrates various document representations with complementary information from pre-trained sources. This strategy reduces uncertainty in label assignment and improves generalization capability in scenarios with limited labeled data, providing a robust and adaptable model for text classification in contexts with a limited number of labeled data. To evaluate COTRA, experiments were conducted using scientific documents classified into five and eleven categories corresponding to their research fields. The model was compared with individual SSL models based on self-training, as well as approaches that integrate co-training, genetic algorithms, and transfer learning through classification pipelines under zero-shot conditions. In this context, COTRA achieved the highest performance levels among the models, reaching a maximum accuracy of 0.87 among co-training models, compared to the best metric of 0.78 obtained by individual self-training models in the five-category classification task. These outcomes demonstrate that co-training represents an effective strategy for improving predictive performance in documents classification.

Keywords

co-training, transfer learning, semi-supervised learning, SSL, document classification, text processing

Dedicatoria

A Andreina, por ser mi compañera incondicional, por su amor, apoyo y paciencia en cada etapa de este proceso. Su confianza en mí ha sido mi mayor impulso.

A Adela y Marco, por ser mi ejemplo de esfuerzo, dedicación y valores. Gracias por enseñarme con su vida que la perseverancia y el compromiso son la clave para alcanzar los sueños.

A las personas que acompañan mis pasos, por su respaldo incondicional, sus palabras de aliento y su confianza en mí. Su presencia ha sido fundamental en este camino.

A todos ustedes, con gratitud y amor.

Reconocimientos

A mi directora de tesis Claudia Pons, por su invaluable guía, paciencia y dedicación en cada etapa de este proceso. Su compromiso y orientación han sido esenciales para la culminación de este trabajo.

A mi co-director, Gustavo Rodríguez, por su apoyo a lo largo de este proceso. Su disposición ha sido fundamental para el desarrollo y culminación de esta investigación.

Al personal administrativo de la secretaría de posgrado de la Facultad de Informática de la Universidad Nacional de la Plata, cuyo esfuerzo y dedicación permiten el adecuado funcionamiento de la Facultad, facilitando el desarrollo de nuestras actividades académicas.

A mis colegas, por compartir conocimientos, experiencias y reflexiones que han contribuido significativamente a mi crecimiento académico y profesional.

A mis alumnos, quienes con su entusiasmo y curiosidad me motivan cada día a seguir aprendiendo.

A todos ustedes, mi más sincero reconocimiento y gratitud.

Índice General

RESUMEN	2
ABSTRACT	3
DEDICATORIA	4
RECONOCIMIENTOS	5
ÍNDICE GENERAL	6
ÍNDICE DE FIGURAS	8
ÍNDICE DE TABLAS	9
CAPÍTULO 1	10
1. Introducción	10
1.1. Problemas y soluciones	12
1.2. Objetivo del trabajo	15
1.3. Transferencia de los resultados obtenidos.	16
1.4. Contribuciones principales	18
1.5. Trabajos presentados vinculados con la tesis	20
1.6. Estructura General de la tesis	20
CAPÍTULO 2	22
2. SSL en clasificación de documentos	22
2.1. Introducción a SSL	22
2.2. Auto-entrenamiento	27
2.3. Co-entrenamiento	28
2.4. Ensamblado	29
2.5. Aprendizaje Activo	30
2.6. Aprendizaje por transferencia	32
2.7. Conclusiones del capítulo	33
CAPÍTULO 3	35
3. Comparación de enfoques en modelos SSL	35
3.1. Análisis de modelos de auto-entrenamiento	36
3.2. Análisis de modelos de co-entrenamiento	40
3.3. Análisis de modelos de ensamblados	44
3.4. Análisis de modelos de aprendizaje activo	47
3.5. Análisis de modelos de aprendizaje de transferencia	50

3.6. Meta-analisis comparativo de los diferentes tipos de modelos SSL.....	53
3.7. Ventajas y desventajas de los modelos SSL	55
3.8. Conclusiones del capítulo	58
CAPÍTULO 4.....	60
4. Explorando documentos científicos por áreas de investigación	60
4.1. Introducción	61
4.2. Estructura del modelo COTRA.....	63
4.3. Caso de estudio	66
4.4. Evaluación y resultados	70
4.5. Conclusiones del capítulo	73
CAPÍTULO 5.....	75
5. Conclusión y trabajo futuro	75
5.1. Conclusiones	75
5.2. Trabajo futuro	79
ACRÓNIMOS O SIGLAS	81
REFERENCIAS	82

Índice de Figuras

Figura 1. Tipos de entrenamiento en modelos semi-supervisados.....	26
Figura 2. Etapas del modelo de auto-entrenamiento para clasificación de documentos	26
Figura 3. Flujo del modelo de aprendizaje por auto-entrenamiento.....	27
Figura 4. Estructura del aprendizaje por co-entrenamiento.....	28
Figura 5. Estructura del aprendizaje por ensamblado.	30
Figura 6. Estructura del aprendizaje por aprendizaje activo.	31
Figura 7. Estructura del aprendizaje por aprendizaje de transferencia.....	32
Figura 8. Diagrama de flujo PRISMA.....	36
Figura 9. Ventajas y desventajas del proceso de clasificación.....	55
Figura 10. Estructura de transformer.....	64
Figura 11. Estructura del modelo COTRA.....	64
Figura 12. Preprocesamiento y representación de documentos del modelo.....	69
Figura 13. Niveles de rendimiento con modelos tradicionales individuales.....	71
Figura 14. Niveles de rendimiento con modelos combinados.....	72

Índice de Tablas

Tabla 1. Estudios y técnicas usadas en etapas del modelo de auto-entrenamiento.....	37
Tabla 2. Estudios y técnicas usadas en etapas del modelo de co-entrenamiento.....	41
Tabla 3. Estudios y técnicas usadas en etapas del modelo de ensamblado.....	45
Tabla 4. Estudios y técnicas usadas en etapas del modelo de aprendizaje activo.....	48
Tabla 5. Estudios y técnicas usadas en etapas del modelo de aprendizaje de transferencia	51
Tabla 6. Forest plot agrupado de niveles de precisión de modelos semi-supervisados	54
Tabla 7. Características y parámetros del conjunto de datos	67
Tabla 8. Clases o líneas de investigación.....	67
Tabla 9. Características de los modelos de experimentación.....	68
Tabla 10. Configuración de parámetros de los algoritmos.	69
Tabla 11. Valores de precisión de los modelos.....	71

Capítulo 1

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 [1]. 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 [2].

En este contexto, la búsqueda e implementación de técnicas eficientes de procesamiento de lenguaje natural para la clasificación automatizada de documentos fortalecen este escenario. Los beneficios de esta propuesta incluyen mejoras en la organización de los documentos, lo que facilita el acceso y la recuperación de información relevante, promoviendo el intercambio de conocimientos entre diferentes sectores, potenciando la innovación y la colaboración interdisciplinaria, lo cual es crucial en un mundo interconectado y en constante evolución. Además, conseguir una clasificación con niveles de precisión más altos, permite una toma de decisiones más informada y basada en evidencia, reduciendo así la probabilidad de error.

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 [3] y no supervisados [4][5][6], 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 [7]. 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 [8]. 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 [9]. 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 [10]. No obstante, son escasos los estudios focalizados en una valoración de los modelos semi-supervisado [11].

Por tal razón, en el presente trabajo se propone un análisis comparativo de los diferentes tipos de modelos semi-supervisados en la clasificación de documentos, así como el desarrollo de un modelo combinado que integra los enfoques más eficientes extraídos de dicho análisis. La propuesta del análisis comparativo de diferentes modelos semi-supervisados en la clasificación de documentos permite identificar cuáles son los enfoques más eficaces para contextos específicos. Esto facilita a investigadores y profesionales tomar decisiones informadas sobre las técnicas que deben implementar en sus sistemas de gestión documental. Al mejorar la precisión y efectividad de la clasificación, la propuesta no solo optimiza el proceso de categorización, sino que también contribuye a la creación de herramientas más accesibles y eficientes para la organización de la información. Esto se traduce en una mejor visibilidad y recuperabilidad de los documentos, lo que potencia el intercambio de conocimiento y la colaboración entre usuarios, optimizando así el acceso a información relevante y actualizada.

La propuesta presentada ofrece varias características que la distinguen de enfoques previos en la clasificación y categorización de documentos. A continuación, se describen estas características y se contrastan con soluciones anteriores:

La propuesta de creación de un marco comparativo se distingue por su evaluación contextualizada, que analiza cómo los modelos semi-supervisados funcionan en diversas situaciones de clasificación de documentos, teniendo en cuenta variables como el tipo de documento, número de clases y la cantidad de etiquetados. Esta característica permite a los usuarios identificar las ventajas y desventajas de cada modelo en diversos contextos, lo que refuerza su capacidad para tomar decisiones informadas al elegir el modelo más adecuado según sus necesidades específicas. Escasos son los estudios del desempeño de los modelos semi-supervisados, entre estas escasas revisiones, se encuentra el estudio [10], que ofrece una conceptualización breve de los tipos de aprendizaje, definiendo la clasificación semi-supervisada y la agrupación (clustering) semi-supervisada. Por otro lado, el trabajo [12] revisa los modelos de clasificación semi-supervisada en el contexto del análisis de imágenes. En el estudio [9], se presenta una recopilación de los últimos métodos de aprendizaje semi-supervisados. Sin embargo, estas revisiones no se centran en el análisis de la clasificación de documentos ni incluyen una comparación del desempeño de los diferentes modelos en este ámbito.

Un modelo combinado que integre las mejores estructuras analizadas en el estudio de diferentes tipos de modelos semi-supervisados refuerza la eficiencia y precisión en la clasificación de documentos, incluso cuando el conjunto de datos etiquetados es escaso o inexistente [13]. Se han identificado estudios previos que llevan a cabo la clasificación de documentos utilizando enfoques supervisados y no supervisados; sin embargo, estos métodos no han logrado abordar el problema de manera efectiva, ya que dependen de la disponibilidad de datos etiquetados en el caso de los modelos supervisados [14] [15] [16] [17][18], o carecen de la precisión necesaria para clasificaciones más detalladas en el caso

de los modelos no supervisados [4] [5] [6]. En el caso de los modelos semi-supervisados se identifica estudios mayoritariamente focalizados en análisis de sentimientos (dos etiquetas) [15] [19] [20] y la clasificación de textos cortos [21] [22] [15]. Las limitaciones de memoria y problemas de fragmentación de contexto en los modelos estándar, son factores que provocan la carencia de experimentaciones de clasificación de documentos con textos largos [23].

1.1. Problemas y soluciones

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 [24]. Actualmente, hay pocos estudios que aborden esta área, y las revisiones existentes no se enfocan específicamente en la clasificación de documentos [25], lo que limita 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 [26].

La solución propuesta presenta un marco comparativo que evalúa cómo funcionan los modelos semi-supervisados 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. Sin embargo, la solución propuesta puede tener un alcance limitado, ya que el análisis comparativo podría dejar algunas técnicas relevantes sin evaluar. Así como contextos que no se consideran en el análisis, lo que provocaría que la solución pueda ser menos efectiva.

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 [27], especialmente en contextos donde los conjuntos de datos etiquetados son escasos o inexistentes, lo que resulta en limitaciones del modelo en precisión y eficacia [28] [29]. 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 [30]. 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 [31]. Esto puede traducirse para las empresas en pérdidas operativas, decisiones basadas en información incorrecta y oportunidades no aprovechadas.

Se propone el desarrollo de un modelo combinado que integra las mejores prácticas y enfoques de los diferentes modelos semi-supervisados 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 semi-supervisados, 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.

Sin embargo, la solución también presenta algunas limitaciones. Una de ellas es la complejidad y robustez de ajuste del modelo, se puede requerir recursos técnicos significativos, así como posibles requerimientos computacionales más altos que podrían limitar su aplicabilidad en ciertas infraestructuras. A partir de estas limitaciones, el estudio de este trabajo puede ampliarse en la realización de, pruebas en diferentes dominios y contextos, se podría evaluar la versatilidad y adaptabilidad del modelo combinado, ajustando su diseño según las necesidades específicas de cada aplicación. También, se podría investigar técnicas de aprendizaje activo que permitan al modelo identificar y solicitar etiquetas para ejemplos específicos, mejorando así la calidad de los datos disponibles.

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 [32]. 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 [33].

Este problema es de importancia, ya que, en muchos casos, los datos etiquetados son escasos o costosos de obtener. La efectividad de los modelos semi-supervisados radica en su capacidad para aprovechar grandes volúmenes de datos no etiquetados o pre-entrenados. Sin embargo, si no pueden adaptarse a variaciones en los dominios, se pierde el valor potencial de estos datos. La incapacidad de los modelos para la adaptabilidad y generalización puede resultar en una disminución de su utilidad y en la insatisfacción de los usuarios.

La incapacidad de adaptar modelos a nuevos dominios puede conducir a un rendimiento deficiente, afectando la calidad de los resultados y las decisiones basadas en estos modelos. Además, la inversión en el desarrollo de soluciones que no se adaptan correctamente puede traducirse en un uso ineficiente de recursos financieros y humanos. Esta falta de adaptabilidad también puede limitar la capacidad de la organización para entrar en nuevos mercados, restringiendo así su crecimiento.

La propuesta para abordar los desafíos de adaptación de dominio en modelos semi-supervisados 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 semi-supervisados, 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.

Una de las limitaciones del modelo propuesto radica en el proceso de adaptabilidad entre el dominio fuente y el dominio objetivo, el proceso de entrenamiento de ajuste fino puede ser complejo y requerir tiempo y recursos significativos, especialmente si se necesitan ajustar múltiples parámetros. Existe el riesgo de que el modelo se sobreajuste a los datos del nuevo dominio si no se implementan técnicas adecuadas de regularización, todo esto depende de la calidad de los datos del dominio fuente. Para ampliar el trabajo presentado, se podría investigar la aplicación de enfoques de aprendizaje activo, donde el modelo selecciona de manera proactiva los ejemplos más informativos para etiquetar en el nuevo dominio, mejorando así la calidad de los datos etiquetados.

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 [34]. 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 [35].

La importancia de este problema radica en que el etiquetado adecuado de documentos es crucial para el éxito de los modelos de clasificación [36] y, en consecuencia, para la

efectividad de los procesos organizacionales que dependen del análisis de datos [37]. 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.

La necesidad de realizar ajustes o reentrenamientos frecuentes de los modelos debido a errores en el etiquetado puede traducirse también en costos adicionales en términos de tiempo y recursos humanos[38]. Asimismo, si los productos o servicios de la organización dependen de clasificaciones precisas y estas fallan, la satisfacción del usuario puede verse comprometida, afectando la reputación de la organización.

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.

La limitante del modelo combinado es que su estructura aumenta la complejidad del modelo y potencialmente los tiempos de procesamiento, debido a su redundancia en el proceso de entrenamiento. Además, si las técnicas de agrupamiento no están bien diseñadas, existe el riesgo de continuar generando errores en la clasificación, lo que provocaría disponer una robusta arquitectura sin capacidad de gestionar documentos en límite de decisión. Para ampliar la propuesta a partir de las limitaciones establecidas, se puede profundizar el estudio en técnicas de regularización y enfoques de entrenamiento incremental. Crear un marco de evaluación para medir el rendimiento del modelo bajo diferentes configuraciones, aplicar estas técnicas en proyectos del mundo real, y documentar los resultados permitiría obtener información valiosa, para identificar los más efectivos.

1.2. Objetivo del trabajo

La clasificación de documentos en entornos semi-supervisados enfrenta varios retos, como es la escasez de documentos etiquetados para fortalecer su entrenamiento o la dificultad para adaptar modelos a nuevos dominios. Para abordar estos problemas, es fundamental establecer un enfoque sistemático que permita desarrollar un modelo eficaz y robusto. Los siguientes objetivos han sido definidos para guiar el proceso del presente estudio, asegurando que se aborden las limitaciones identificadas, para la obtención de resultados que mejoran significativamente la calidad de la clasificación de documentos.

Objetivo general

Diseñar un modelo combinado de clasificación semi-supervisada, que optimice tanto el etiquetado como la clasificación de documentos, mejorando su precisión y eficacia, fundamentado en una estructura que se derive de un análisis de la evaluación de los distintos tipos de modelos semi-supervisados.

Objetivo específico #1

Diseñar un marco comparativo que evalúe el desempeño de los diferentes modelos semi-supervisados en la clasificación de documentos, considerando variables como tipo de documento, número de clases y niveles de precisión.

Objetivo específico #2

Implementar un modelo combinado que integre las mejores prácticas de los modelos semi-supervisados analizados, utilizando técnicas de transferencia de aprendizaje para mejorar la adaptabilidad y generalización en nuevos dominios de datos.

Objetivo específico #3

Desarrollar un modelo que combine diferentes entrenamientos de clasificación semi-supervisada, enfocándose en la redundancia y mejora del manejo de documentos situados en los límites de decisión, con el fin de aumentar la precisión del etiquetado y reducir la ambigüedad en la asignación de categorías.

1.3. Transferencia de los resultados obtenidos.

Los resultados de esta investigación tienen aplicaciones prácticas significativas más allá del problema específico abordado. La capacidad de clasificar documentos de manera eficiente en entornos semi-supervisados, con un mínimo de datos etiquetados o incluso sin ellos, es fundamental en diversos sectores que dependen del análisis de grandes volúmenes de información.

En el sector académico, este estudio ha sido aplicado en [13] para optimizar la clasificación de documentos científicos en el repositorio institucional de la Universidad Técnica de Cotopaxi (UTC), categorizándolos según líneas de investigación. Este enfoque facilita la gestión del conocimiento, permitiendo a investigadores y estudiantes acceder más fácilmente a publicaciones relevantes dentro de su área de estudio. Además, la implementación de modelos semi-supervisados en bibliotecas digitales y repositorios científicos puede mejorar los sistemas de recomendación, automatizar la indexación de publicaciones y agilizar el descubrimiento de información relevante.

En el ámbito de la salud, la clasificación automatizada de registros clínicos y reportes médicos representa un desafío debido a la cantidad de datos no estructurados generados diariamente. COTRA puede aplicarse en hospitales y centros de salud para categorizar historias clínicas, informes de diagnóstico, estudios epidemiológicos y prescripciones médicas, mejorando la organización de la información y permitiendo un acceso más

rápido a datos relevantes para la toma de decisiones médicas. Asimismo, su integración con sistemas de vigilancia epidemiológica permitiría identificar patrones en enfermedades emergentes mediante la categorización de reportes clínicos, lo que ayudaría a los profesionales de la salud a responder de manera más eficiente ante brotes y crisis sanitarias.

En el sector empresarial, la clasificación de documentos es crucial para la gestión eficiente de la información y la toma de decisiones estratégicas. COTRA puede ser empleado en departamentos de atención al cliente para clasificar automáticamente solicitudes y reclamos, mejorando el tiempo de respuesta y optimizando los procesos de resolución de problemas. En áreas como el análisis de riesgos financieros, el modelo puede categorizar contratos, reportes contables y documentos regulatorios, reduciendo el esfuerzo manual en auditorías y asegurando el cumplimiento normativo. Además, en la gestión de recursos humanos, COTRA puede ser utilizado para clasificar currículums y filtrar candidatos según las competencias requeridas, optimizando los procesos de reclutamiento.

En el ámbito de la seguridad y el cumplimiento normativo, la creciente necesidad de analizar grandes volúmenes de documentos legales y normativos hace que la implementación de modelos como COTRA sea altamente beneficiosa. Su capacidad para clasificar documentos jurídicos, informes regulatorios y políticas empresariales podría ayudar a departamentos legales y gubernamentales a organizar información y detectar inconsistencias en normativas. También podría utilizarse en la detección de fraudes mediante la categorización de reportes financieros y la identificación de patrones en documentos contables.

Más allá de estas aplicaciones específicas, la solución propuesta en esta investigación puede ser extendida a contextos más amplios, como el análisis de redes sociales, donde la clasificación automática de publicaciones puede contribuir a la identificación de tendencias, análisis de sentimientos y detección de contenido perjudicial en plataformas digitales. Asimismo, en la gestión de correos electrónicos, COTRA podría integrarse en sistemas de clasificación de mensajes para filtrar automáticamente correos no deseados, detectar fraudes y priorizar comunicaciones urgentes.

Para maximizar su aplicabilidad, la implementación del modelo en distintos dominios puede lograrse a través de técnicas de transferencia de aprendizaje, que permiten adaptar la estructura de COTRA a conjuntos de datos específicos sin requerir grandes volúmenes de etiquetado manual. Esto facilita su integración en organizaciones con necesidades particulares, proporcionando una solución flexible y escalable que optimiza la clasificación de documentos en diversos sectores.

Con este enfoque, la presente investigación no solo contribuye al desarrollo del estado del arte en modelos semi-supervisados, sino que también genera un impacto práctico en múltiples disciplinas, permitiendo la adopción de estrategias más eficientes para la organización y gestión de información en distintos entornos.

1.4. Contribuciones principales

Este trabajo se enfoca en mejorar la clasificación de documentos en entornos semi-supervisados, abordando desafíos críticos como la escasez de datos etiquetados y la adaptación a nuevos dominios. Las contribuciones del estudio ofrecen soluciones que no solo optimizan el rendimiento de los modelos existentes, sino que también sientan las bases para futuras investigaciones en el área. A través de un análisis comparativo y el desarrollo de un modelo combinado, este trabajo busca enriquecer la práctica de la clasificación de documentos y abrir nuevas oportunidades de aplicación. Estas contribuciones son presentadas en cuatro partes.

Evaluación de modelos semi-supervisados: La presente investigación aborda la escasez de análisis comparativos de modelos semi-supervisados en la clasificación de documentos, una problemática que ha limitado la capacidad de los usuarios para tomar decisiones informadas. La contribución central radica en el desarrollo de un marco comparativo que evalúa el desempeño de diversos modelos en contextos específicos, teniendo en cuenta variables como el dominio, el tipo de documento, el número de clases y la cantidad de datos etiquetados. Esta herramienta no solo proporciona una visión de las ventajas y desventajas de cada enfoque, sino que también brinda herramientas a investigadores y profesionales para seleccionar el modelo más adecuado según sus necesidades de optimización en su gestión documental.

Además, el trabajo contribuye a establecer un estándar para futuros estudios en el área, fomentando un enfoque sistemático en la evaluación de modelos semi-supervisados. Aunque el marco puede tener limitaciones en su alcance, su implementación representa un avance significativo en el campo. En última instancia, esta contribución promueve una mayor comprensión y un uso efectivo de los modelos semi-supervisados, lo que puede transformar la manera en que las organizaciones gestionan y clasifican su información.

Adaptabilidad de dominios: Cuando los documentos etiquetados son escasos o costosos de obtener, el enfoque de adaptación de dominio mejora la capacidad de un modelo para generalizar su rendimiento al aplicarse en un nuevo dominio diferente al de su entrenamiento, es decir permite que el conocimiento de un dominio fuente se transfiera al dominio objetivo [39]. En este contexto, en el presente trabajo se estructura un modelo que utiliza técnicas de transferencia de aprendizaje para abordar los desafíos de adaptación de dominio en modelos semi-supervisados para clasificación de documentos. El modelo propone generar diferentes representaciones de características de los documentos, de modo que estas representaciones puedan ser sometidas a un entrenamiento por transferencia utilizando diccionarios pre-entrenados, así se incrementa la capacidad de generalización en nuevos dominios, esto ha permitido que el modelo se adapte mejor a variaciones en los datos, mejorando así su rendimiento en contextos diferentes.

Al integrar diversas representaciones de las características del documento, se obtiene un mejor contexto de las palabras en el dominio, lo que incrementa su semántica. De esta manera, el modelo se adapta mejor a los diferentes estilos y lenguajes entre el dominio de origen y el de destino. La optimización del uso de datos pre-entrenados y no etiquetados, reduce la necesidad de inversión en la obtención de documentos etiquetados, lo que permite un uso más eficiente de los recursos. La adaptabilidad del modelo crea oportunidades para abordar nuevos dominios y entornos con mayor eficacia, lo que fomenta su crecimiento, ya que puede ajustarse a las necesidades específicas de diferentes contextos.

Optimización de límites de decisión: En el proceso de categorización existen fronteras que separan las diferentes clases, estas fronteras conocidas como límite de decisión determinan la asignación de etiquetas a nuevos documentos, su flexible y correcta identificación es crucial para la una adecuada clasificación. Los documentos cercanos a este límite, representan casos importantes ya que una ligera variación en sus características puede cambiar su clasificación. Manejar adecuadamente este límite es esencial, especialmente en situaciones donde la precisión es crítica y los datos pueden ser ruidosos.

Se propone un modelo combinado que utiliza diferentes perspectivas o vistas de las características del documento, el modelo de clasificación mejora la precisión al capturar información diversa, lo que resulta en una clasificación más exacta, especialmente cerca del límite de decisión. La redundancia en las representaciones reduce errores al permitir que el modelo verifique y valide decisiones, mientras que la combinación de diferentes perspectivas aumenta la robustez frente al ruido en los datos. Además, gestionar vistas facilita el manejo de casos límite y mejora la confianza en las clasificaciones.

Clasificando artículos universitarios por áreas de investigación: La clasificación de documentos por líneas de investigación en las universidades es fundamental para organizar el conocimiento y facilitar el acceso a información relevante, lo que permite a investigadores y estudiantes encontrar recursos en áreas específicas. Una adecuada categorización de documentos, promueve la colaboración entre académicos que trabajan en temas similares, apoya la toma de decisiones estratégicas en la planificación académica y mejora la visibilidad de la investigación institucional, lo que puede atraer a nuevas investigaciones y financiamiento. En muchos casos, los repositorios universitarios almacenan documentos científicos sin una organización que contemple las áreas de investigación específicas de la universidad. Como resultado, los documentos se dispersan, algunas áreas de conocimiento quedan incompletas y se restringen las oportunidades de colaboración, además de dificultar la identificación de los principales focos de investigación en cada disciplina.

En este contexto, se ha desarrollado un modelo combinado de aprendizaje semi-supervisado con el objetivo de clasificar documentos científicos de los investigadores de la Universidad Técnica de Cotopaxi. Este modelo ha sido diseñado para adaptarse a las áreas de investigación de la universidad, para su entrenamiento, se utilizan conjuntos de

datos pre-entrenados, lo que facilita el aprendizaje de patrones de clasificación a partir de ejemplos previos y mejora la precisión en la asignación de los documentos. Además, el modelo procesa los documentos desde diferentes perspectivas, considerando diferentes elementos del documento como el título, resumen y palabras clave. Este enfoque multidimensional permite una clasificación más precisa y ajustada a las particularidades de cada disciplina, optimizando la organización del conocimiento. De este modo, se mejora la visibilidad de las áreas de investigación y se generan mayores oportunidades de colaboración entre los investigadores.

1.5. Trabajos presentados vinculados con la tesis

En el marco de la investigación desarrollada en esta tesis, se han presentado varios trabajos de investigación que contribuyen directamente al tema central de la clasificación de documentos científicos mediante modelos de aprendizaje semi-supervisado. Estos trabajos han sido publicados en revistas científicas y expuestos en congresos, siendo fundamentales para el avance del estudio y para la validación de los enfoques propuestos.

Uno de los trabajos publicados es "Semi-supervised learning models for document classification: A systematic review and meta-analysis", publicado en la revista Iberamia Journal (<http://journal.iberamia.org/>) [27]. Este artículo ofrece una revisión sistemática y un meta-análisis sobre los modelos de aprendizaje semi-supervisado aplicados a la clasificación de documentos, proporcionando un análisis de las técnicas más relevantes y su desempeño en diferentes contextos de clasificación. La publicación sirvió como base para la implementación del modelo combinado de aprendizaje semi-supervisado propuesto en la tesis, al identificar las mejores prácticas y enfoques existentes en la literatura.

El segundo trabajo presentado es "A co-training model based in learning transfer for the classification of research papers", publicado en IEEE Xplore (<https://ieeexplore.ieee.org>) [13]. Este estudio propone un modelo combinado basado en la transferencia de aprendizaje para la clasificación de artículos de investigación. El artículo destaca por su enfoque de integración en técnicas de transferencia de aprendizaje para mejorar la eficiencia y precisión de la clasificación, especialmente cuando los conjuntos de datos son limitados o desbalanceados. Esta investigación complementa la tesis, ofreciendo una perspectiva alternativa sobre cómo optimizar los modelos de clasificación en el contexto académico. Ambos trabajos no solo enriquecen los aportes teóricos de la tesis, sino que también fortalecen su aplicabilidad práctica en la organización y gestión de documentos científicos.

1.6. Estructura General de la tesis

El manuscrito está compuesto por cinco capítulos que abordan desde los fundamentos teóricos hasta los resultados y conclusiones del presente trabajo de tesis.

Capítulo 1. Introducción

Presenta el contexto y los objetivos de la investigación, proporcionando una visión general de la tesis. Se exponen los problemas abordados y las soluciones propuestas, se definen los objetivos del trabajo y la transferencia de los resultados obtenidos, se destacan las contribuciones principales, se resumen los trabajos previos relacionados con la tesis y se describe la estructura general del documento.

Capítulo 2. Modelos SSL en clasificación de documentos

Introduce el aprendizaje SSL y sus aplicaciones en la clasificación de documentos. Se explican los fundamentos teóricos del aprendizaje semi-supervisado y se presentan diferentes enfoques de SSL, como auto-entrenamiento, co-entrenamiento, modelos ensamblados, aprendizaje activo y aprendizaje por transferencia.

Capítulo 3. Comparación de enfoques en modelos SSL

Este capítulo recopila experimentos de diversas investigaciones para analizar los distintos modelos SSL aplicados a la clasificación de documentos. Se examinan variables como la cantidad de documentos etiquetados y no etiquetados, así como los niveles de rendimiento en cada caso de estudio, abarcando modelos de auto-entrenamiento, co-entrenamiento, ensamblados, aprendizaje activo y aprendizaje por transferencia. Además, se realiza un meta-análisis comparativo de los modelos evaluados, identificando sus ventajas y desventajas para comprender su impacto en el rendimiento de la clasificación.

Capítulo 4. Exploración de documentos científicos por áreas de investigación

Este capítulo presenta la estructura del modelo SSL COTRA y su aplicación en la clasificación de documentos científicos en diversas áreas de investigación. Se describe en detalle su configuración, incluyendo los componentes y enfoques utilizados, y se analizan los principios que sustentan su funcionamiento. Además, se presentan análisis prácticos y experimentales que permiten evaluar su desempeño en relación con otros modelos. Finalmente, se presentan los resultados obtenidos, destacando sus niveles de precisión.

Capítulo 5. Conclusión y trabajo futuro

El apartado expone los hallazgos más relevantes de la investigación y plantea perspectivas para trabajos futuros. Se resumen los resultados obtenidos, se analizan sus implicaciones y se sugieren enfoques para mejorar el desarrollo de modelos SSL en la clasificación de documentos.

Además, la tesis incluye secciones complementarias con acrónimos y siglas, así como las referencias bibliográficas.

Capítulo 2

2. SSL en clasificación de documentos

2.1. *Introducción a SSL*

En la era digital, la cantidad de datos generados diariamente ha crecido de manera exponencial, impulsada por avances tecnológicos como dispositivos móviles, plataformas de redes sociales y transacciones digitales. En este contexto, los repositorios digitales, tanto académicos como empresariales, desempeñan un papel crucial al recopilar y organizar esta información masiva. Los repositorios académicos almacenan datos provenientes de publicaciones científicas, tesis, y proyectos de investigación, mientras que los empresariales gestionan información relacionada con transacciones, comportamiento de clientes y operaciones internas. Esta integración de datos facilita su análisis y uso estratégico, transformando la información en conocimiento valioso para la toma de decisiones en ambos sectores.

La mayoría de estos datos son no estructurados, lo que significa que no están organizados en formatos predeterminados, como bases de datos relacionales, y suelen presentarse en forma de texto, imágenes, videos o registros de eventos. Este tipo de información plantea desafíos importantes para su análisis y aprovechamiento, especialmente cuando los métodos tradicionales de aprendizaje automático requieren grandes volúmenes de datos etiquetados, los cuales no siempre están disponibles.

En este escenario, los modelos de aprendizaje SSL emergen como una solución eficiente para abordar estos retos. Estos modelos son capaces de aprovechar los datos no etiquetados, que suelen estar disponibles en abundancia, y combinarlos con pequeñas cantidades de datos etiquetados para entrenar sistemas de clasificación robustos. Este enfoque no solo reduce la necesidad de etiquetar manualmente grandes conjuntos de datos, una tarea que puede ser costosa y demandante en tiempo, sino que también mejora la capacidad del modelo para generalizar varias situaciones de similar contexto [40].

SSL es un enfoque del aprendizaje automático que combina las fortalezas del aprendizaje supervisado y no supervisado. Este paradigma se utiliza en escenarios donde se dispone de una pequeña cantidad de datos etiquetados y una gran cantidad de datos no etiquetados, lo que permite optimizar el uso de los recursos disponibles al reducir la necesidad de anotaciones manuales extensivas[41]. Los modelos SSL son especialmente útiles en tareas donde etiquetar los datos puede ser costoso, lento o impráctico, como el análisis de textos, imágenes y clasificación de documentos [42].

A diferencia del aprendizaje supervisado, que depende completamente de datos etiquetados, y del aprendizaje no supervisado, que no utiliza etiquetas en absoluto, SSL emplea tanto datos etiquetados como no etiquetados en el proceso de entrenamiento. Este enfoque aprovecha patrones inherentes en los datos no etiquetados para complementar el conocimiento derivado de los datos etiquetados, mejorando así la precisión y robustez del modelo.

Los modelos SSL se basan en diversas técnicas y principios, como la consistencia regularizada, y la propagación de etiquetas [43]. Estas metodologías buscan maximizar la información obtenida de los datos no etiquetados al mismo tiempo que refuerzan la generalización del modelo en nuevas instancias [44]. Por ejemplo, en el contexto de la clasificación de documentos, SSL puede identificar relaciones semánticas y estructurales en textos no etiquetados para mejorar el desempeño del modelo.

SSL puede aplicarse en una amplia variedad de problemas prácticos en los que la disponibilidad de datos etiquetados es limitada, pero se cuenta con un gran volumen de datos no etiquetados. A continuación, se presentan varios campos en los que SSL pueden ser usado:

- **Procesamiento del lenguaje natural (NLP):** SSL permite que los modelos de NLP aprovechen grandes volúmenes de datos no etiquetados para mejorar su rendimiento. Estos datos facilitan la identificación de patrones lingüísticos subyacentes, enriqueciendo el entrenamiento de los modelos y fortaleciendo su capacidad de generalización. Como resultado, los modelos no solo reducen los costos asociados a la obtención de datos etiquetados, sino que también incrementan su precisión y robustez en entornos prácticos. Esto es especialmente útil en tareas complejas como la desambiguación semántica, donde el contexto proporcionado por los datos no etiquetados aporta información clave [45]. Por ejemplo, en la clasificación de opiniones, SSL puede identificar patrones emocionales en reseñas de productos o publicaciones en redes sociales utilizando solo una pequeña cantidad de datos etiquetados, logrando un análisis de sentimientos más preciso y eficiente [46].
- **Descubrimiento de asociaciones:** Los modelos SSL identifican relaciones no evidentes al integrar de manera estratégica las ventajas de los datos etiquetados y no etiquetados durante el proceso de entrenamiento. Los datos etiquetados proporcionan un punto de partida estructurado, mientras que los no etiquetados permiten explorar patrones más sutiles y difíciles de detectar. Mediante técnicas como la propagación de etiquetas, SSL aprovecha las características inherentes de los datos no etiquetados para expandir el conocimiento adquirido y refinar la comprensión de las interacciones entre las variables [47]. Por ejemplo, en un sistema de análisis de preferencias de consumo, los datos etiquetados pueden incluir información explícita como "producto comprado" o "calificación del cliente". Sin embargo, al incorporar datos no etiquetados, como el historial de navegación o los tiempos de visualización, los modelos de SSL pueden identificar tendencias implícitas y correlaciones menos evidentes [48], como la relación entre compras realizadas en ciertas épocas del año y

categorías específicas de productos. Esto no solo mejora la capacidad de hacer predicciones más precisas, sino que también proporciona una visión más profunda de las interacciones complejas en grandes bases de datos.

- **Procesamiento de imágenes:** SSL optimiza el procesamiento de imágenes mediante técnicas como la consistencia regularizada, que asegura que las predicciones del modelo se mantengan estables ante pequeñas perturbaciones, como variaciones de iluminación o rotaciones, mejorando su capacidad de generalización [43]. También utiliza la extracción de características no supervisadas, que permite identificar patrones visuales importantes, como bordes, texturas y formas, a partir de imágenes no etiquetadas mediante el uso de representaciones latentes. Por ejemplo, en la identificación de objetos en imágenes médicas, como tumores en escaneos radiológicos. SSL permite al modelo aprender características importantes de las imágenes no etiquetadas, como las texturas, formas y patrones característicos de un tumor, utilizando las pocas imágenes etiquetadas disponibles como guía. Esto no solo mejora la precisión del modelo, sino que también acelera el diagnóstico y facilita la detección temprana de enfermedades, impactando directamente en la calidad de la atención médica.
- **Clustering con guía parcial:** SSL puede mejorar los resultados de agrupamiento mediante la incorporación de representaciones latentes, utilizando técnicas como autoencoders [49]. Estas transforman los datos no etiquetados en un espacio donde las relaciones y patrones ocultos se vuelven más evidentes, lo que facilita una segmentación más precisa [50]. Esta transformación permite al modelo identificar características relevantes que podrían no ser visibles en el espacio original de características, y ayuda a formar clústeres más significativos. Además, la regularización de consistencia garantiza que los datos cercanos en este nuevo espacio latente se agrupen correctamente, aplicando restricciones que estabilizan las predicciones frente a pequeñas variaciones en los datos. Esta estabilidad refuerza la estructura interna de los clústeres, permitiendo que el modelo capture de manera más precisa la estructura subyacente del conjunto de datos, lo que mejora la coherencia y precisión del agrupamiento. Por ejemplo, en una institución financiera con pocos datos etiquetados sobre sus clientes, pero con abundante información no etiquetada de transacciones, los modelos SSL pueden identificar patrones ocultos en los comportamientos de compra, como la similitud entre clientes que adquieren productos de alto valor, incluso sin etiquetas. La regularización de consistencia asegura que los clientes con comportamientos similares se agrupen adecuadamente, mejorando la segmentación y permitiendo una personalización más precisa de ofertas y estrategias de marketing.
- **Detección de anomalías:** SSL, mediante técnicas como gráficos de similitud, permite propagar etiquetas para identificar puntos de datos con comportamientos similares a los casos confirmados de anomalías [51]. Complementariamente, las representaciones latentes enriquecidas transforman los datos no etiquetados utilizando modelos generativos, creando un espacio en el que las diferencias entre comportamientos

normales y anómalos se hacen más evidentes, lo que facilita la identificación de patrones inusuales. Por ejemplo, en la detección de accesos no autorizados, SSL aprende patrones como horarios inusuales o ubicaciones sospechosas a partir de pocos datos etiquetados. La regularización de consistencia refuerza estas características, mejorando la identificación de nuevas amenazas en tiempo real y fortaleciendo la seguridad del sistema.

- **Clasificación de documentos:** SSL puede utilizar representaciones latentes enriquecidas, mediante técnicas como los embeddings de texto, para convertir los documentos en vectores que capturan relaciones semánticas [52]. Esto permite al modelo identificar patrones incluso cuando los textos presentan diferencias estilísticas o léxicas [53]. Posteriormente, se pueden propagar etiquetas mediante gráficos de similitud, aprovechando las conexiones entre documentos etiquetados y no etiquetados. De esta forma, las etiquetas de categorías conocidas pueden extenderse a textos similares, basándose en características comunes como temas o palabras clave. Esta combinación de técnicas permite a SSL maximizar la información extraída de los documentos disponibles, logrando una clasificación más eficiente, incluso con un conjunto limitado de etiquetas. Por ejemplo, en la clasificación automática de correos electrónicos según su nivel de prioridad. A partir de un pequeño conjunto de correos etiquetados con su prioridad, el sistema utiliza esta información para identificar patrones en otros correos no etiquetados. Así, los correos sin etiqueta se agrupan en las categorías correctas, basándose en su similitud con los correos previamente clasificados, lo que permite organizar automáticamente todos los mensajes de manera eficiente.

Cómo se evidencia, el aprendizaje semi-supervisado se presenta como un enfoque versátil y eficiente para abordar desafíos en una amplia variedad de dominios. Su capacidad para integrar múltiples técnicas y modelos permite desarrollar soluciones que optimizan el análisis y procesamiento de datos, especialmente en contextos donde la complejidad de los patrones subyacentes exige herramientas eficientes. Además, su flexibilidad facilita la adaptación a problemas específicos, maximizando el rendimiento en tareas críticas como la clasificación y el descubrimiento de asociaciones. Esta combinación de adaptabilidad y precisión posiciona al SSL como una técnica clave en la evolución de la inteligencia artificial y sus aplicaciones prácticas.

En el ámbito de la clasificación de documentos, el aprendizaje semi-supervisado se destaca por su capacidad para integrar diversas fuentes de información y técnicas. Al combinar métodos de propagación de etiquetas con la generación de representaciones latentes, SSL es capaz de identificar relaciones semánticas y patrones subyacentes en los documentos, mejorando la exactitud de las clasificaciones.

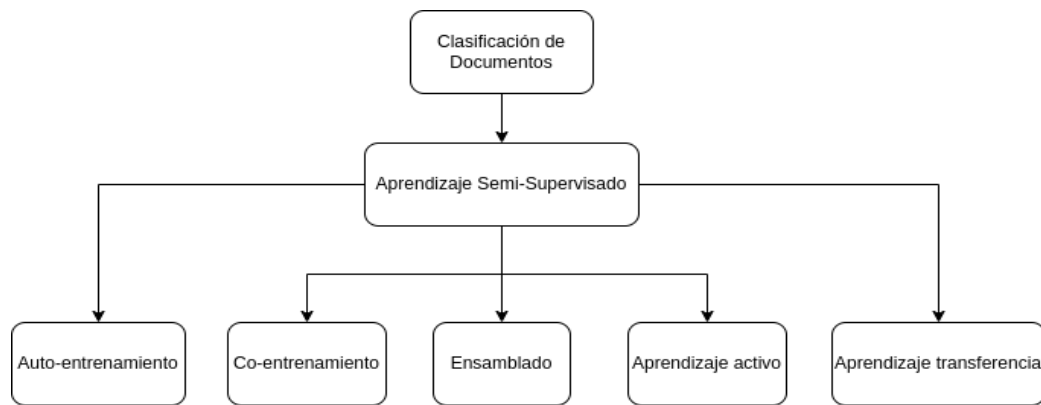


Figura 1. Tipos de entrenamiento en modelos semi-supervisados.

(Tomado de [27])

La presente sección se enfoca en identificar y analizar la estructura de los modelos de aprendizaje semi-supervisado con mejor desempeño en la clasificación de documentos, priorizando aquellos prototipos que se adapten de la manera más adecuada a las necesidades y características de esta tarea. Se exploran diferentes técnicas (Ver Figura 1) y su capacidad para maximizar la precisión y eficiencia en el procesamiento de textos, entre estas técnicas se encuentran el autoentrenamiento, co-entrenamiento, ensamblado, aprendizaje activo y aprendizaje por transferencia, técnicas que destacan por su capacidad para abordar diferentes desafíos y optimizar los procesos de clasificación en diversos contextos aplicativos.

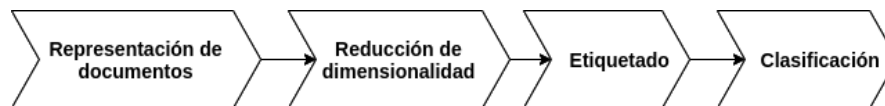


Figura 2. Etapas del modelo de auto-entrenamiento para clasificación de documentos

(Adaptado de [9])

En el análisis realizado se identifican etapas comunes en los modelos utilizados para la clasificación de documentos. En la Figura 2 se ilustran las diferentes fases que componen este tipo de modelos. Inicialmente, la representación de documentos tiene como objetivo procesar el texto de cada documento, generando una representación específica para cada palabra que facilite su análisis. A continuación, dado que estos conjuntos de palabras suelen tener una alta dimensionalidad, se emplean técnicas de reducción de dimensionalidad para seleccionar únicamente las palabras más relevantes. Con este preprocesamiento completo, los datos están preparados para el etiquetado. En esta etapa, se aplican algoritmos que entrenan el modelo utilizando un reducido número de datos etiquetados y una gran cantidad de datos no etiquetados, con el objetivo de incrementar el conjunto de etiquetados. Finalmente, con un volumen adecuado de datos etiquetados, el modelo se entrena para clasificar documentos de prueba de manera eficiente. En el presente capítulo se ofrece una visión general de los enfoques de aprendizaje semi-supervisado.

2.2. Auto-entrenamiento

El auto-entrenamiento es uno de los enfoques más utilizados en el aprendizaje semi-supervisado para la clasificación de documentos. Este modelo se basa en un proceso iterativo donde un clasificador inicial, entrenado con un pequeño conjunto de documentos etiquetados, predice etiquetas para documentos no etiquetados [54]. Los documentos con predicciones más confiables se incorporan posteriormente al conjunto de entrenamiento, permitiendo que el modelo se refine y mejore progresivamente su desempeño (Ver Figura 3). Este enfoque es particularmente útil en escenarios donde las etiquetas disponibles son limitadas, ya que maximiza el uso de los datos no etiquetados para aumentar la precisión de las clasificaciones.

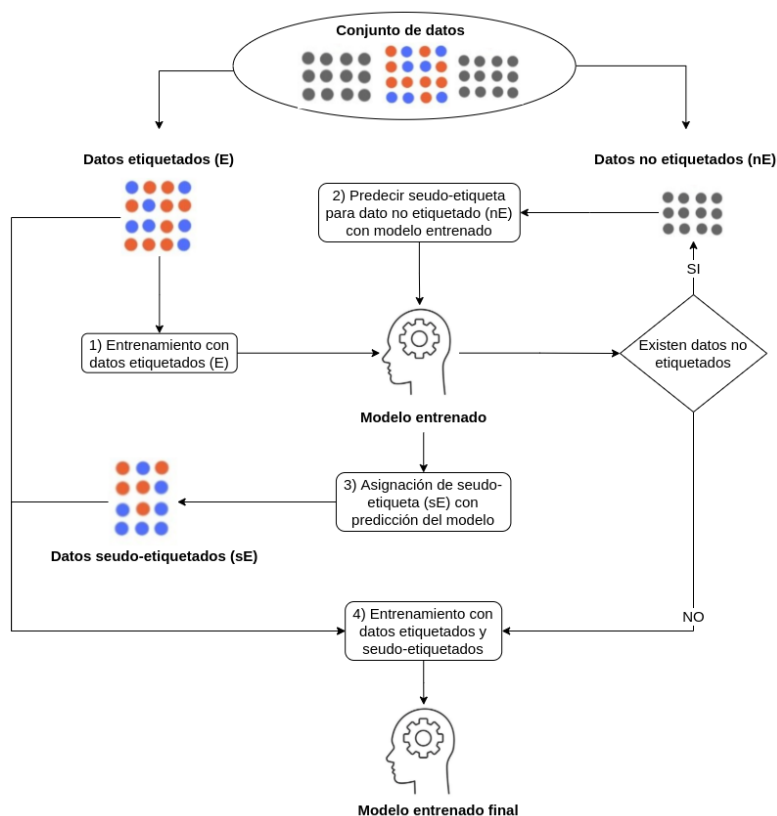


Figura 3. Flujo del modelo de aprendizaje por auto-entrenamiento

Una de las ventajas del auto-entrenamiento en la clasificación de documentos es su facilidad de implementación. Dado que solo requiere un clasificador base y un mecanismo para evaluar la confianza de las predicciones, puede integrarse rápidamente en flujos de trabajo existentes. Además, el modelo tiene la capacidad de adaptarse a cambios graduales en los patrones de los datos, lo que resulta clave en dominios dinámicos, como la categorización de noticias o el filtrado de correos electrónicos. Sin embargo, su éxito depende de la calidad del clasificador inicial y de los criterios establecidos para seleccionar las predicciones confiables, ya que errores en estas etapas pueden propagarse y afectar el rendimiento final.

Por último, el auto-entrenamiento también es compatible con técnicas más avanzadas, como el uso de representaciones latentes enriquecidas o embeddings de texto, que

mejoran la capacidad del modelo para capturar relaciones semánticas en los documentos. Esto permite abordar problemas complejos, como la clasificación de textos con variaciones estilísticas o lingüísticas significativas, ampliando aún más las aplicaciones potenciales de esta técnica en la gestión y análisis de grandes volúmenes de datos textuales.

2.3. Co-entrenamiento

El co-entrenamiento es un enfoque que se basa en entrenar dos o más modelos separados, cada uno con una vista diferente de los datos. Este enfoque es especialmente útil en la clasificación de documentos cuando los datos pueden dividirse en características complementarias que representan diversas perspectivas del contenido, como el texto y los metadatos asociados [55]. Los modelos trabajan de manera colaborativa, mientras uno aprende y etiqueta nuevos datos, esa información se utiliza para mejorar el entrenamiento del otro. Esta interacción permite que el sistema aproveche al máximo las relaciones subyacentes en los datos, incluso cuando solo una fracción de ellos está inicialmente etiquetada (ver Figura 4).

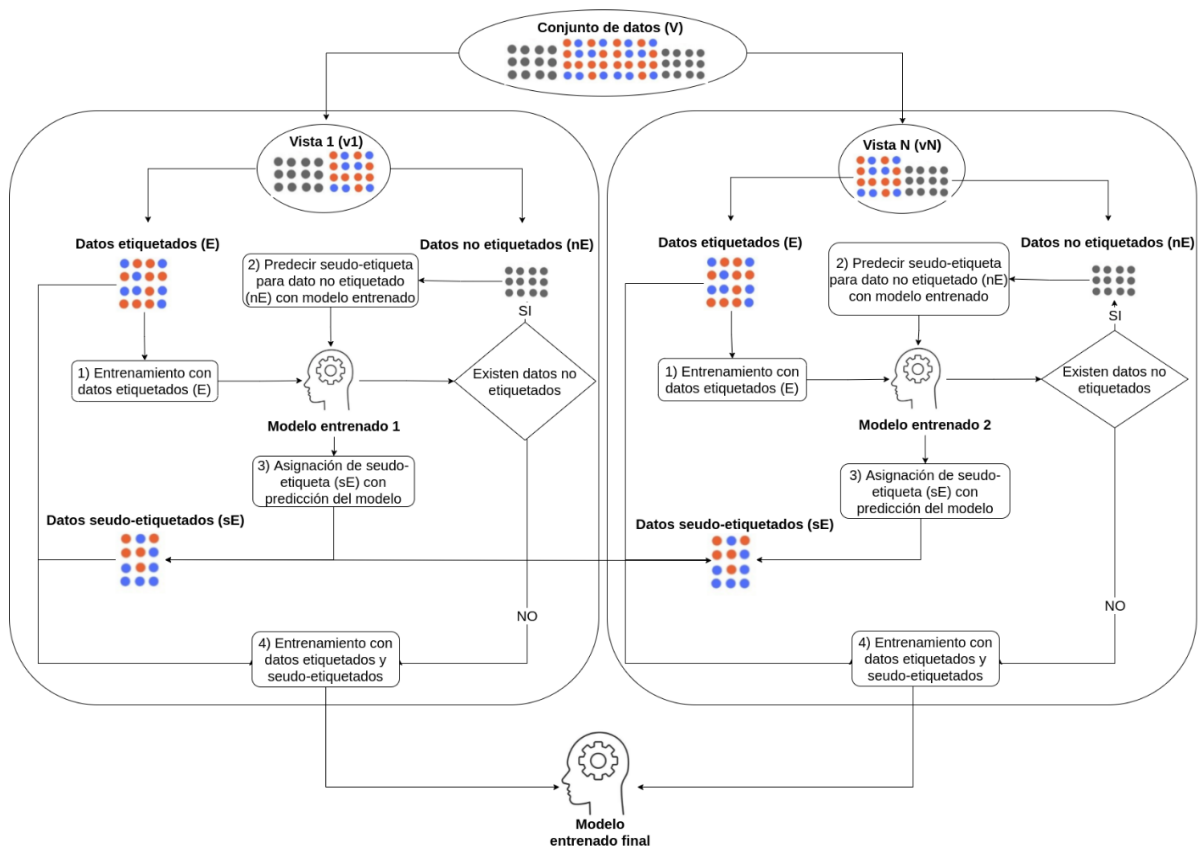


Figura 4. Estructura del aprendizaje por co-entrenamiento.

Una de las principales fortalezas de estos modelos en la clasificación de documentos es su capacidad para manejar información redundante y correlacionada entre las vistas. Por ejemplo, al clasificar páginas web, una vista podría analizar el contenido textual, mientras que otra se centra en la descripción de sus URLs. Cada modelo utiliza estas características

específicas para generar predicciones que el otro modelo puede usar como pseudoetiquetas, mejorando así la calidad del conjunto de entrenamiento. Este enfoque no solo incrementa el volumen de datos etiquetados, sino que también refuerza la precisión en la clasificación al aprovechar diversas fuentes de información.

Además, la técnica es efectiva para reducir el ruido en los datos y mejorar la robustez del modelo frente a errores de etiquetado inicial. La colaboración entre los modelos permite filtrar inconsistencias y concentrarse en patrones confiables, haciendo que este enfoque sea ideal para tareas complejas como la clasificación temática, la detección de spam o la categorización de correos electrónicos. Al integrar diversas vistas del documento, el co-entrenamiento maximiza el aprovechamiento de los datos disponibles y logra una clasificación más precisa y contextualizada.

2.4. Ensamblado

Los modelos ensamblados pueden combinar múltiples modelos base, cada modelo individual puede aportar una perspectiva única del problema, y al unir sus predicciones, se consigue un resultado más robusto y confiable [56]. Al momento de clasificar documentos, este método es particularmente útil para manejar la diversidad en los datos y mejorar la precisión, incluso en condiciones de datos escasos o inconsistentes.

Los modelos ensamblados incluyen técnicas como el bagging (bootstrap), donde múltiples modelos entrenados en subconjuntos de los datos hacen predicciones que luego se combinan, mediante un promedio o votación mayoritaria. En la clasificación de documentos, esto podría aplicarse para categorizar grandes volúmenes de texto, minimizando el impacto de errores de un único modelo. Otra técnica de ensamblado es el boosting, que se enfoca en corregir iterativamente los errores de modelos débiles, fortaleciendo la capacidad del conjunto para identificar patrones difíciles de clasificar (Ver Figura 5).

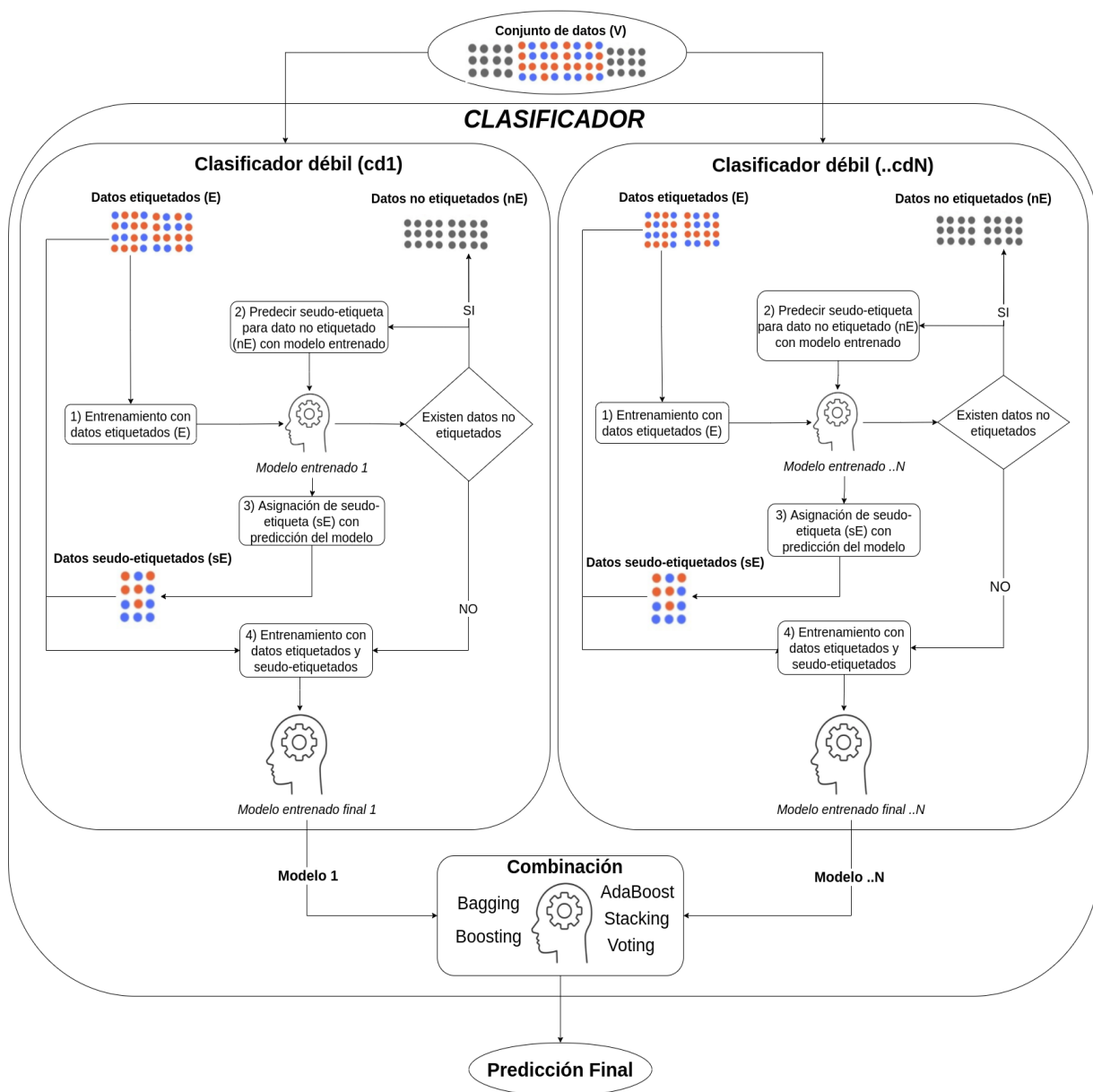


Figura 5. Estructura del aprendizaje por ensamblado.

La principal ventaja de los modelos ensamblados radica en su capacidad para reducir la varianza y el sesgo al combinar modelos con características complementarias. También son efectivos para mitigar el impacto del ruido en los datos, ya que la combinación de múltiples predicciones tiende a cancelar los errores individuales. Esto los hace ideales para tareas donde las diferencias estilísticas y léxicas pueden dificultar el análisis de texto.

2.5. Aprendizaje Activo

El aprendizaje activo busca maximizar la eficiencia del modelo al seleccionar de manera estratégica los datos más informativos para su etiquetado (ver Figura 6). En el contexto de la clasificación de documentos, esta metodología resulta especialmente útil, ya que no todos los documentos no etiquetados tienen el mismo impacto en el desempeño del modelo. A través de criterios específicos, como la incertidumbre en las predicciones o la

representatividad del conjunto de datos, el aprendizaje activo identifica cuáles documentos deberían ser etiquetados manualmente por un experto para optimizar el proceso de entrenamiento[57].

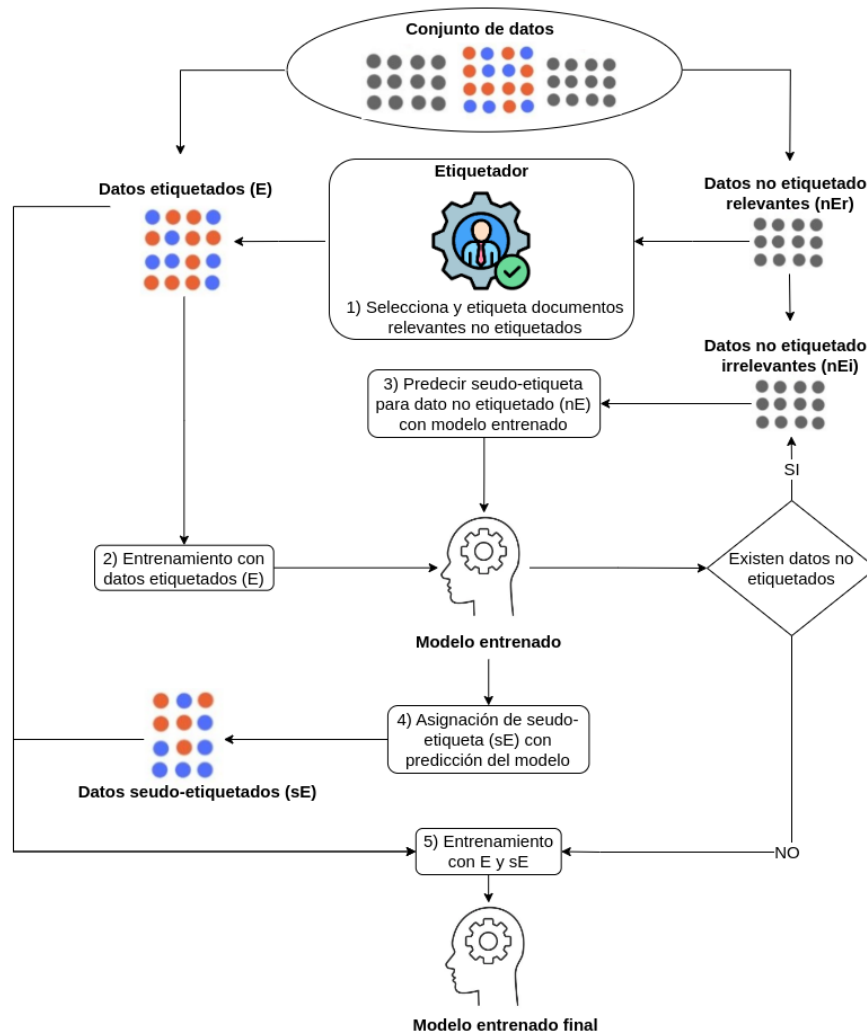


Figura 6. Estructura del aprendizaje por aprendizaje activo.

Una de las principales ventajas del aprendizaje activo en la clasificación de documentos es su capacidad para reducir significativamente el costo de etiquetado. En lugar de etiquetar grandes volúmenes de datos de forma aleatoria, el modelo selecciona aquellos documentos que contienen información clave o que representan casos límite en la clasificación. Esto permite al sistema mejorar su precisión, incluso con un conjunto de datos etiquetados relativamente pequeño.

Además, esta técnica se adapta bien a escenarios donde los documentos pertenecen a clases complejas o tienen características ambiguas. Por ejemplo, en el análisis de opiniones de clientes, el aprendizaje activo puede priorizar documentos con expresiones lingüísticas ambiguas o contradictorias para ser etiquetados primero. Esto ayuda a refinar la capacidad del modelo para manejar casos difíciles, mejorando su rendimiento en el análisis de textos futuros. La incorporación del aprendizaje activo en los sistemas de clasificación de documentos no solo incrementa la eficacia, sino que también mejora la adaptabilidad del modelo. A medida que nuevos documentos se incorporan al sistema,

este puede identificar continuamente cuáles son los más útiles para actualizar y ajustar el modelo, garantizando que se mantenga relevante y preciso en entornos dinámicos.

2.6. Aprendizaje por transferencia

El aprendizaje por transferencia permite reutilizar el conocimiento adquirido por un modelo en un dominio fuente (Source Domain, SD) para mejorar su desempeño en un dominio destino (Target Domain, TD). En el contexto de la clasificación de documentos, el dominio fuente puede estar constituido por conjuntos de textos con características generales o similares, como diccionarios, bibliotecas o bases de datos de documentos, mientras que el dominio destino podría enfocarse en tareas específicas con datos más especializados o limitados.

Este enfoque es eficiente porque el modelo, previamente entrenado en el dominio fuente, ya ha aprendido patrones fundamentales como estructuras lingüísticas, relaciones semánticas o estilísticas. Estos conocimientos son adaptados al dominio destino a través de técnicas como el ajuste fino, donde el modelo es calibrado utilizando un conjunto reducido de datos etiquetados relacionados con la nueva tarea (ver Figura 7). Por ejemplo, un modelo entrenado en un corpus extenso como Wikipedia puede ser ajustado para clasificar documentos de diferentes dominios, logrando precisión con menor esfuerzo computacional y reduciendo la necesidad de grandes volúmenes de datos etiquetados [58].

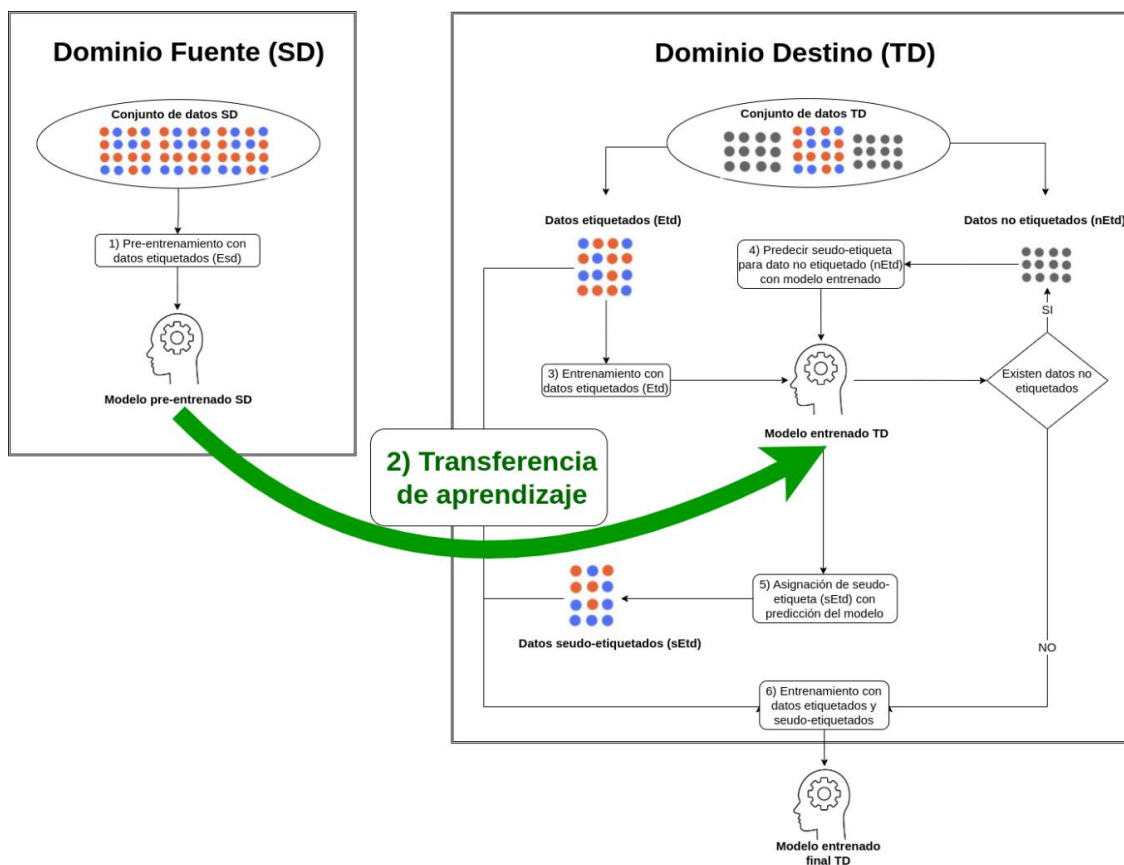


Figura 7. Estructura del aprendizaje por aprendizaje de transferencia.

Además, este tipo de modelos facilita la superación de barreras lingüísticas. Por ejemplo, un modelo inicialmente entrenado en textos en inglés puede ser adaptado para clasificar documentos en español, aprovechando la similitud estructural entre los lenguajes. Esto es particularmente valioso en escenarios multilingües o globales donde los recursos etiquetados en ciertos idiomas son limitados.

El proceso de este modelo está focalizado en seleccionar un modelo adecuado previamente entrenado en el dominio fuente, adaptar sus parámetros mediante técnicas como el ajuste fino o la congelación de capas, y optimizar su rendimiento para el dominio destino. Este enfoque no solo maximiza el aprovechamiento de los datos disponibles, sino que también permite abordar tareas complejas con eficiencia y precisión.

Estudios recientes han demostrado la efectividad del aprendizaje por transferencia en tareas de clasificación, especialmente en el análisis de textos y emociones. Por ejemplo, en [59] demostraron que la aplicación de aprendizaje por transferencia con modelos preentrenados como RoBERTa y XLNet mejoró significativamente la clasificación multietiqueta de emociones en publicaciones de redes sociales, utilizando para ello mecanismos de atención múltiple. De manera similar, un estudio sobre clasificación de textos relacionados con enfermedades mentales [60] utilizó aprendizaje por transferencia con modelos como RoBERTa y BigBIRD en datos de Reddit, demostrando la efectividad de estos modelos en la clasificación de textos complejos. Estos estudios validan el potencial del aprendizaje por transferencia para mejorar el rendimiento en tareas complejas de clasificación, incluso con datos limitados etiquetados, consolidándose como una herramienta valiosa para diversas aplicaciones en minería de textos y más allá.

2.7. Conclusiones del capítulo

En este capítulo se ha explorado el papel fundamental del aprendizaje semi-supervisado (SSL) en la clasificación de documentos, destacando su capacidad para combinar datos etiquetados y no etiquetados con el fin de optimizar el rendimiento de los modelos. Se han descrito las principales técnicas utilizadas en SSL, como la propagación de etiquetas, la consistencia regularizada y la incorporación de representaciones latentes, demostrando su aplicabilidad en diversos ámbitos como el procesamiento de lenguaje natural, el análisis de imágenes, la detección de anomalías y la agrupación de datos.

Asimismo, se han examinado los distintos enfoques de modelos SSL aplicados a la clasificación de documentos, entre ellos el auto-entrenamiento, co-entrenamiento, modelos ensamblados, aprendizaje activo y aprendizaje por transferencia, evaluando su proceso de entrenamiento, ventajas y limitaciones en contextos de clasificación de documentos. Se ha evidenciado que estos modelos pueden optimizar la precisión en la clasificación de documentos al reforzar conjuntos limitados de datos etiquetados a través de la identificación y explotación de patrones latentes en los datos no etiquetados, lo que contribuye a una representación más robusta y a un aprendizaje más eficiente del modelo, reduciendo la dependencia de grandes volúmenes de datos previamente categorizados.

Finalmente, se concluye que el aprendizaje semi-supervisado representa una alternativa eficiente para la clasificación de documentos en entornos con disponibilidad limitada de datos etiquetados. Su implementación permite optimizar el uso de los datos disponibles, mejorar la adaptabilidad de los modelos a nuevos dominios. Estos hallazgos sirven como base para profundizar en la comparación y evaluación de modelos SSL en capítulos posteriores.

Capítulo 3

3. Comparación de enfoques en modelos SSL

Numerosos estudios han destacado la importancia de determinar las condiciones bajo las cuales los datos no etiquetados pueden mejorar la precisión en la clasificación por aprendizaje semi-supervisado [10] [12] [14] [61]. El presente capítulo se enfoca en analizar modelos de aprendizaje semi-supervisado aplicados a la clasificación de documentos, evaluando su rendimiento y explorando sus ventajas y limitaciones en las distintas etapas del procesamiento de texto. Aunque existen investigaciones relacionadas con el aprendizaje semi-supervisado en diversas áreas, las revisiones sistemáticas sobre modelos específicos para la clasificación de documentos son escasas. Estudios previos han abordado la clasificación en otros contextos, como el análisis de imágenes [12], pero no se han focalizado en los documentos ni han comparado el desempeño de diferentes modelos.

Para la presente tesis se plantea una Revisión Sistemática de Literatura, cuyos resultados fueron publicados en [27], 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 [27]. Se recopilieron datos de repositorios como SCOPUS, IEEE Xplore, Springer y Elsevier, organizando los modelos en categorías según el tipo de aprendizaje semi-supervisado (Ver Figura 8).

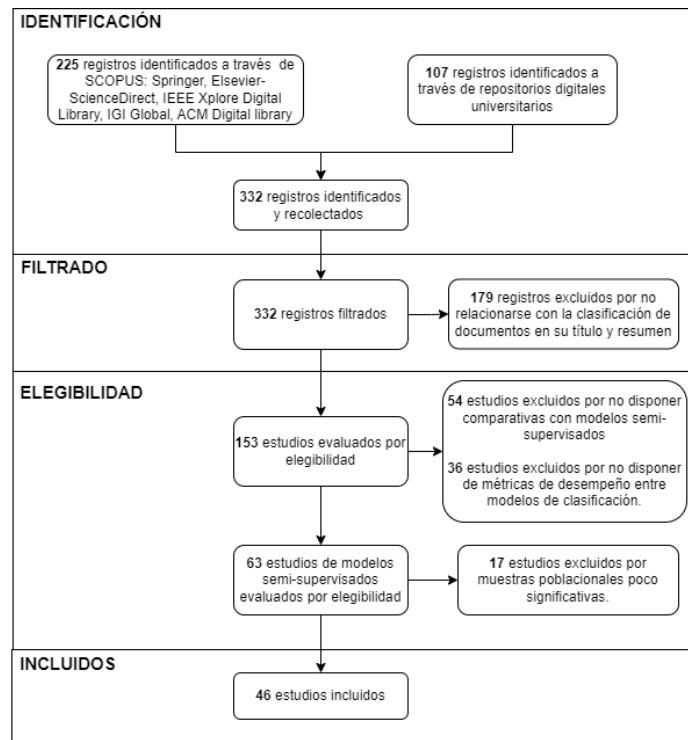


Figura 8. Diagrama de flujo PRISMA.
(Tomado de [27])

La evaluación del rendimiento de los modelos se realizó considerando diferentes variables tales como las técnicas empleadas en cada etapa del modelo, desde la Representación de documentos (E1), la Reducción de dimensionalidad (E2), el Etiquetado (E3), hasta la Clasificación (E4). conjunto de datos utilizados (Dataset), número de documentos (Docs.) y clases (Clases), proporción de documentos etiquetados (E), no etiquetados (NE) y de 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): $TP/(TP+FP)$.

3.1. Análisis de modelos de auto-entrenamiento

En esta sección se presenta un análisis basado en diversos estudios que implementan modelos de auto-entrenamiento. Este análisis sintetiza las técnicas y estrategias más utilizadas, identificando las ventajas y desventajas asociadas a cada estudio, lo que permite comprender de manera integral los avances y las limitaciones en este campo.

La Tabla 1 organiza de manera estructurada los hallazgos de once investigaciones, cada registro de la tabla detalla las técnicas empleadas en la implementación de los modelos, ofreciendo un panorama que facilite la identificación de patrones y tendencias en los enfoques utilizados para el auto-entrenamiento. Este análisis contribuye a establecer un marco de referencia útil para el diseño de nuevos modelos y para la selección de técnicas que optimicen la eficiencia y la eficacia en tareas de clasificación de documentos y etiquetado semi-supervisado, fortaleciendo así las bases para futuras investigaciones en esta área.

Tabla 1. Estudios y técnicas usadas en etapas del modelo de auto-entrenamiento

(Tomado de [27])

Id	Autor	E1	E2	E3	E4	Dataset	Docs.	Clases	E	NE	Prueba	P(%)	Ventajas	Desventajas
ST01	Altinel & Ganiz, 2016 [14]	BoW	IG	Algoritmo HCSC (No iterativo)	SVM	UCI/Mini news	2000	20	200 (10%)	1400 (70%)	400 (20%)	71.65	Rica representación de documentos con mecanismo de búsqueda por significado; efectividad en su etiquetado; utilizado con un reducido número de documentos etiquetados.	Su proceso de etiquetado no es iterativo; las palabras comunes afectan la clasificación.
ST02	Emadi et al., 2021 [17]	s/reg	s/reg	Algoritmo DPC (Iterativo)	SVM	UCI/Glass	214	6	21 (10%)	129 (60%)	64 (30%)	50.07	Permite un ajuste a las predicciones de los pseudoetiquetados; adecuada selección de puntos de información con métricas en el muestreo de documentos no etiquetados.	Baja precisión de clasificación
ST03	Khan & Lee, 2019 [15]	PoS	TF-IDF	Framework MMSL (Iterativo)	SVM	Reseñas/ Amazon	2000	2	200 (10%)	1400 (70%)	400 (20%)	81.9	Analiza los mejores candidatos para etiquetar; Amplio léxico con su significado; selecciona instancias más representativas y con un alto grado de confianza para entrenamiento; elimina características ruidosas.	Su arquitectura multimodelo genera un procesamiento complejo y pesado.
ST04	Shinnou et al., 2018 [62]	s/reg	STFW	EM	NB	UCI/ Newsgroup	3600	6	600 (17%)	1800 (50%)	1200 (33%)	94	Estructura abierta a vincular pre-entrenamientos; adecuada precisión de clasificación con pocos etiquetados.	Complejidad en la configuración de metaparámetros de etiquetado por apertura de pre-entrenamiento.
ST05	Watanabe & Zhou, 2020 [61]	s/reg	LDA	Newsmap	NB	Institucionales/ Debates	2507	5	0	2507 (100%)	2507 (100%)	57	Dispone de un diccionario con palabras semilla que clarifica palabras confusas; análisis de texto con teoría driven; puede trabajar sin etiquetados.	Dependiente de diccionario de palabras que puede tener una producción costosa; con textos cortos se pierde precisión; Baja precisión de clasificación.
ST06	Barman & Chowdhury, 2018 [16]	SOM	SOM	Kohonen	SVM	UCI/Reseñas	4900	4	466 (10%)	4189 (85%)	245 (5%)	83.27	El modelo pseudoetiqueta con la menor cantidad de documentos etiquetados.	Utiliza técnicas de clusterización para agrupar documentos en límite de decisión.
ST07	Jedrzejowicz & Zakrzewska, 2019 [63]	W2V	LDA	LDA-W2V	LDA-W2V	UCI/ Newsgroup	20000	6	2000 (10%)	10000 (50%)	8000 (40%)	76.21	La representación de documentos considera el significado de las palabras; estructura brinda apertura a usar diccionarios pre-entrenados	Un etiquetado por agrupación que no controla grupos con palabras de significado similar.
ST08	Poojitha, 2018 [64]	BoW	LDA	Cross validation	NB	UCI/News	406916	4	284841 (70%)	0	122075 (30%)	93.8	Eficiencia en la técnica de agrupación LDA para la separación de temáticas; el proceso de etiquetado es iterativo.	Modelo eficiente únicamente con alto número de etiquetados; características con alto peso se replican en varias agrupaciones.
ST09	Chen et al., 2019 [28]	SOM	SOM	mLVQb	NN	UCI/Wine	175	9	9 (5%)	114 (65%)	52 (30%)	85.8	El modelo clasifica documentos multi-etiqueta; las etiquetas de alta confianza las incorpora como clase suave (flexible).	Falta de automatización en la definición de los rangos de confianza; el etiquetado es poco eficiente en seleccionar las instancias de mayor confianza.
ST10	Zhao & Li, 2021 [65]	s/reg	s/reg	STDPNaN	NN	UCI/Pendigits	3698	10	370 (10%)	2958 (80%)	370 (10%)	84.79	El modelo permite determinar la distribución de los etiquetados iniciales con distribuciones espirales y no espirales; su ensamblado de clasificadores mejora la predicción de etiquetas.	Las distribuciones sin parámetros tienen un margen de error en la agrupación de determinados etiquetados; el tiempo de respuesta de las predicciones es alto.
ST11	Vale et al., 2022 [66]	s/reg	s/reg	FlexCon-C2	KNN	UCI/Pishing	2456	3	246 (10%)	1964 (80%)	246 (10%)	78.94	Dispone de un mecanismo automático para etiquetar; administra el etiquetado de documentos considerando un rango de confianza.	La automatización de etiquetado no puede ser utilizado en otros modelos semi-supervisados.

Los estudios analizados implementan diversas técnicas distribuidas en las diferentes etapas del proceso de aprendizaje semi-supervisado, cada una diseñada para abordar aspectos específicos del modelo. En la etapa E1, las técnicas predominantes incluyen BoW (Bag of Words), TF-IDF (Term Frequency-Inverse Document Frequency) y W2V (Word2Vec), estas herramientas son fundamentales para la representación vectorial de textos, permitiendo transformar datos textuales en estructuras numéricas comprensibles para los algoritmos de clasificación. La elección de estas técnicas responde a su capacidad para capturar características clave de los textos, como la frecuencia de términos o relaciones semánticas básicas.

En las etapas E2 y E3, se identifican métodos semánticos, entre los que destacan LDA (Linear discriminant analysis) y algoritmos iterativos como el Framework MMSL (Multi-model Sentiment Learning Layer). Estas estrategias no solo permiten una mejor representación semántica de los datos, sino que también optimizan la selección de documentos no etiquetados que serán utilizados para el entrenamiento del modelo. La iteratividad de estos enfoques resulta particularmente valiosa, ya que facilita un refinamiento continuo de las predicciones mediante la incorporación progresiva de nuevas etiquetas.

Finalmente, en la etapa E4, los clasificadores más empleados incluyen SVM (Support Vector Machine), NB (Naive Bayes) y NN (Neural network). Estos modelos son reconocidos por su capacidad para procesar grandes volúmenes de datos, logrando un equilibrio entre precisión y eficiencia computacional. La elección de estos clasificadores refleja su idoneidad para manejar conjuntos de datos no estructurados y variados, asegurando resultados consistentes incluso en escenarios de alta variabilidad temática.

Los modelos de auto-entrenamiento demuestran su adaptabilidad al aplicarse en una variedad de conjuntos de datos, desde contextos amplios como noticias y reseñas (Mini News, Amazon, Newsgroup) del repositorio de aprendizaje automático de la Universidad de California, Irvine (UCI), que destacan por su alta variabilidad temática, hasta dominios específicos como datos institucionales y debates utilizados por [61]. Esto evidencia su capacidad para manejar tanto diversidad de contenidos y patrones lingüísticos como estructuras más definidas, resaltando su flexibilidad en escenarios especializados que requieren una representación semántica precisa y, en algunos casos, apoyo de diccionarios temáticos.

Asimismo, el tamaño de los conjuntos de datos utilizados influye significativamente en los resultados obtenidos. Por ejemplo, en el caso de ST08, que incluye más de 400.000 documentos, se logra una precisión del 93.8%, uno de los valores más altos reportados. Este hallazgo respalda la idea de que datasets más grandes y diversos permiten entrenar modelos con mayor solidez y capacidad de generalización. Sin embargo, también se identifican casos como ST05, que, a pesar de trabajar con un dataset completamente no etiquetado, logra una precisión del 57%, lo que pone de manifiesto la importancia de las técnicas complementarias, como los diccionarios de palabras clave, en contextos de datos escasos.

Así, el análisis de los conjuntos de datos revela que los modelos de autoentrenamiento son versátiles y pueden ajustarse tanto a contextos generales como específicos. Además, resalta la correlación positiva entre el tamaño del dataset y la precisión del modelo, aunque también pone de manifiesto que la calidad y las técnicas empleadas para el procesamiento son determinantes en el éxito del entrenamiento del modelo.

Los resultados en términos de precisión presentan una amplia variación entre los estudios analizados, lo que refleja la diversidad en las técnicas aplicadas y los contextos de los datasets. Los valores más altos se observaron en ST04, con una precisión del 94%, y en ST08, con un 93.8%. En ambos casos, los modelos emplearon enfoques iterativos y aprovecharon técnicas de diccionarios preprocesados, estas características destacan la importancia de integrar procesos iterativos y representaciones semánticas robustas para maximizar el rendimiento.

En contraste, estudios como ST02 y ST05 reportaron precisiones significativamente más bajas, de 50.07% y 57%, respectivamente. En el caso de ST02, la baja precisión se atribuye a las limitaciones en la selección de pseudoetiquetas y en la representatividad de los puntos de información seleccionados, estrategias que no han brindado eficiencia. Por su parte, ST05, aunque destaca por la utilización de diccionarios de palabras semilla, enfrenta desafíos relacionados con su dependencia de estos recursos y la pérdida de precisión en textos más breves.

Estos resultados reflejan la necesidad de equilibrar los enfoques utilizados en los modelos, considerando no solo la sofisticación de las técnicas empleadas, sino también las características de los conjuntos de datos y los recursos adicionales requeridos para optimizar la precisión.

El análisis de los estudios revela ventajas significativas asociadas con los modelos de auto-entrenamiento, algunas están relacionadas con sus capacidades de representación semántica y selección de datos. Por ejemplo, el modelo LDA-W2V implementado en el estudio ST07 destaca por ofrecer representaciones semánticas, lo que permite capturar el significado contextual de las palabras y mejorar la selección de instancias representativas. De manera similar, el enfoque iterativo utilizado por [64] en ST08 resulta particularmente eficaz para la agrupación temática, lo que garantiza una separación clara y coherente entre las diferentes categorías de datos.

Además, algunos modelos demuestran eficiencia en el uso de un número limitado de documentos etiquetados. Por ejemplo, el modelo de ST06 logra realizar pseudo-etiquetados efectivos con una cantidad reducida de datos etiquetados, optimizando así los recursos disponibles, esta característica es relevante en contextos donde los datos etiquetados son escasos o costosos de obtener.

No obstante, también se identifican importantes limitaciones entre los estudios presentados. Una de las desventajas más recurrentes es la alta demanda computacional, como se observó en el Framework MMSL de ST03, su arquitectura multimodelo requiere un procesamiento intensivo, con un uso significativo de recursos, de manera similar pasa con los estudios ST04 y ST10. También, se detectaron limitaciones asociadas a la falta

de iteratividad en algunos modelos. Un caso representativo es el estudio de ST01, donde la ausencia de un proceso iterativo restringe la capacidad del modelo para optimizar las pseudo-etiquetas, lo que deriva en una clasificación menos eficiente.

Estos problemas subrayan la importancia de diseñar modelos que puedan equilibrar la complejidad técnica y la eficiencia, optimizando tanto los resultados como los recursos necesarios para su implementación.

3.2. Análisis de modelos de co-entrenamiento

En este segmento se analiza un conjunto de investigaciones que implementan modelos de co-entrenamiento en el aprendizaje semisupervisado, una técnica que explota múltiples vistas de los datos para mejorar el rendimiento clasificatorio en escenarios con información limitada. La Tabla 3 sintetiza diez estudios representativos, destacando las técnicas aplicadas en cada etapa del proceso de clasificación, así como el número de vistas o perspectivas utilizadas. Este análisis ofrece una visión integral del estado del arte en modelos de co-entrenamiento, permitiendo identificar fortalezas, limitaciones y patrones recurrentes, al tiempo que evalúa su aplicabilidad en contextos diversos, desde problemas de clasificación binaria hasta sistemas de etiquetado múltiple.

Tabla 2. Estudios y técnicas usadas en etapas del modelo de co-entrenamiento.

(Tomado de [27])

Id	Autor	Vistas	E1	E2	E3	E4	Dataset	Docs.	Clases	E	NE	Test	P(%)	Ventajas	Desventajas
CT01	Borrajó et al., 2020 [29]	2	BoW	TF-IDF	HMM	SVM	UCI/Reuters	8055	8	40 (0.5%)	8005 (99.38%)	10 (0.12%)	86.9	Los clasificadores de cada vista aprenden uno de otro; puede trabajar con pocos documentos etiquetados.	Su reducción de dimensionalidad entrega menor peso a términos frecuentes y mayor peso a términos no comunes.
CT02	Masmoudi et al., 2021 [31]	2	Pos	BoW	MLSMOTE	Random Forest	Institucionales/ACM	3170	5	792 (25%)	1585 (50%)	793 (25%)	34.5	Adecuado para entrenar con pocos documentos etiquetados; el esquema de evaluación de predicción tiene baja consistencia.	Tiene un límite de dos vistas de trabajo; no considera el peso de las características en su reducción dimensional.
CT03	C. Zhu & Miao, 2019 [67]	5	s/r	s/r	SOMVfV	s/r	UCI/Reuters	11740	6	22348 (20%)	78218 (70%)	11174 (10%)	92.64	Modelo abierto al procesamiento de datos de alta escala sin obviar los datos de corta escala.	Modelo posee una estructura compleja que afecta su rendimiento; en el proceso se pierden características y vistas.
CT04	C. Zhu et al., 2019 [68]	2	s/r	WMVC	SSOPMV	s/r	UCI/Cora	2708	2	541 (20%)	1896 (70%)	271 (10%)	94.1	Refuerza entrenamiento con la generación de instancias de documentos no etiquetados; si los datos tienen actualizaciones en tiempo real la estructura las considera.	Alto tiempo de rendimiento por estructura abierta a grandes volúmenes de datos; límite de dos vistas y no simultáneas.
CT05	Jia et al., 2021 [69]	3	OC	ASC	SMDDRL	Cross entropy	UCI/BBC	2225	5	222 (10%)	890 (40%)	1113 (50%)	96.15	Con similaridad y ortogonalidad identifica las características específicas y compartidas; crea espacio común para entrenamiento simultáneo; reduce la redundancia; alta precisión.	El esquema no separa automáticamente las características compartidas de las específicas.
CT06	Nayak et al., 2020 [70]	2	s/r	s/r	MIL	NN	UCI/Reseñas	104306	2	200 (1.92%)	10000 (95.82%)	236 (2.3%)	70	Estructura soporta múltiples números de vistas; uso de características en diferentes vistas sin redundancia; concepto de atención para la predicción de etiquetas.	El modelo tiene tendencia de overfit de entrenamiento; las numerosas finas resoluciones generan la pérdida de un buen rendimiento.
CT07	Kim et al., 2019 [71]	2	s/r	V1: TF-IDF V2: LDA	MCT	NB	UCI/Reuters	107870	10	2157 (2%)	75533 (70%)	30180 (28%)	94.9	Trabaja con diferentes técnicas de reducción de dimensionalidad en cada vista; eficiente precisión de clasificación con gran número de clases.	Los conjuntos de características son independientes en cada vista; la alta redundancia genera demasiada carga al algoritmo.
CT08	Edo-Osagie et [72]	2	N-gram	IG	EM	MLP	Red social/ Tweets asma	127145	2	3500 (2.75%)	85501 (67.25%)	38144 (30%)	95.6	Identificación de características relevantes del documento; estructura con un entrenamiento profundo e iterativo.	Demanda de muchos recursos por el uso de técnicas de entrenamiento profundo.
CT09	Donyavi & Asadi, 2020 [73]	3	NSGA-II	NSGA-II	DTGMO-SSC	C4.5 NN	UCI/nursery	12960	5	1296 (10%)	10368 (80%)	1296 (10%)	87.26	Posee un algoritmo evolutivo de auto-etiquetado con una buena gestión de precisión y densidad de datos; modelo idóneo en escasez de etiquetados elimina datos atípicos y realiza buena distribución.	No dispone de técnicas para medir la densidad y diversidad de los datos; conflictos cuando los datos sean desequilibrados.
CT10	Jia et al., 2022 [74]	2	HTF	HTF	Semantic SSL	SVM	UCI/diabetes	768	2	57 (10%)	519 (65%)	192 (25%)	75	Obtención de índices de etiquetado; evalúa semántica por medio de técnicas difusas; construcción de una estructura de distribución con etiquetados y no etiquetados.	Experimentación únicamente con clases binarias; la evaluación de las descripciones tiene considerables márgenes de error.

Los estudios recopilados en la Tabla 2 destacan una variedad de técnicas aplicadas en las diferentes etapas del modelo de co-entrenamiento, una observación recurrente es el uso de métodos clásicos para la representación y transformación de datos, como BoW y TF-IDF, en estudios como CT01 y CT07, combinados con clasificadores como SVM y NB. Estas técnicas son efectivas para datasets con características bien definidas, considerando sus limitaciones en cuanto a la representación de relaciones semánticas complejas.

Por otro lado, investigaciones más recientes, como CT05 y CT08, integran modelos avanzados como SMDDRL (Semi-supervised multi-view deep discriminant representation learning) y técnicas de entrenamiento profundo basadas en redes neuronales, permitiendo identificar tanto características específicas como compartidas entre vistas. Sin embargo, estos modelos tienden a requerir una mayor cantidad de recursos computacionales, lo que limita su aplicabilidad en entornos con recursos restringidos. Por último, algunos enfoques, como los de CT03 y CT09, destacan por su capacidad para manejar datos de gran escala mediante el uso de estructuras complejas y algoritmos evolutivos. No obstante, estos métodos enfrentan desafíos relacionados con la pérdida de características y la gestión de datos desbalanceados, lo que puede afectar la precisión del modelo en ciertos escenarios.

La mayoría de los estudios analizados utilizan dos vistas CT01, CT02, CT04, CT06, CT07 y CT10, lo que permite una colaboración entre perspectivas complementarias de los datos. Sin embargo, algunos estudios exploran escenarios más complejos, como el caso de CT03 con cinco vistas, CT05 y CT09 con tres vistas. Este incremento en el número de vistas permite una representación más rica de los datos, pero también introduce desafíos adicionales, como la demanda de recursos computacional y la gestión de redundancia entre las vistas.

Los tamaños de los datasets varían significativamente, desde conjuntos pequeños como UCI/diabetes (768 documentos en CT10) hasta grandes volúmenes como UCI/Reseñas (104,306 documentos en CT06). Los estudios que emplean conjuntos de datos más extensos tienden a reportar mejores niveles de precisión, como en CT07 (94.9%) y CT08 (95.6%), debido a la mayor diversidad de datos que mejora la capacidad de generalización del modelo.

El balance entre documentos etiquetados (E) y no etiquetados (NE) es un factor clave en los modelos SSL. En estudios como CT01 y CT07, los documentos no etiquetados constituyen más del 95% del dataset, lo que demuestra la capacidad de los modelos de co-entrenamiento para aprovechar información limitada. Sin embargo, cuando esta proporción es más equilibrada, como en CT05, los modelos pueden lograr mejores precisiones (96.15%) al reducir la incertidumbre en las predicciones.

Los conjuntos de datos presentan una amplia diversidad en términos de dominio, abarcando desde noticias y datos generales, como en UCI/Reuters (CT01, CT03, CT07), hasta contextos especializados, como redes sociales (tweets sobre asma en CT08) y aplicaciones biomédicas (UCI/diabetes en CT10). Se aprecia que el contexto de los datos influye en los resultados, por ejemplo, en CT08, el uso de datos de redes sociales se asocia con un rendimiento notable (95.6%) debido a la eficiencia de métodos como los N-gramas

y el entrenamiento iterativo profundo, que son efectivos para manejar ruido. En contraste, CT10, que utiliza datos biomédicos, alcanza una precisión moderada del 75%, evidenciando la necesidad de técnicas más refinadas para abordar datos sensibles con clases binarias.

En conjunto, la Tabla 2 muestra cómo las características de los datasets, el número de vistas y la proporción de datos etiquetados influyen directamente en el rendimiento de los modelos de co-entrenamiento. La elección de un conjunto de datos adecuado y una configuración balanceada son esenciales para optimizar los resultados en escenarios específicos. Además, se destaca la necesidad de continuar explorando combinaciones de vistas y técnicas que puedan adaptarse mejor a las demandas de los dominios emergentes.

La precisión de los modelos varía acorde a las técnicas empleadas, la cantidad de vistas y las características de los conjuntos de datos. Los estudios con los mejores resultados, como CT04 (94.1%), CT05 (96.15%) y CT07 (94.9%), se destacan por combinar métodos de transformación de datos, como ASC (Adversarial Similarity Constraint) y SSOPMV (Semi-supervised one pass multi view), con clasificadores robustos y redes neuronales. Además, estos estudios se benefician de conjuntos de datos bien distribuidos entre datos etiquetados y no etiquetados, lo que mejora el aprendizaje semi-supervisado.

En contraste, los estudios con niveles de precisión bajos, como CT02 (34.5%), enfrentan limitaciones significativas. En este caso, la capacidad de Random Forest para manejar desbalances y capturar relaciones complejas en los datos es limitada, lo que se ve agravado por un esquema de evaluación inconsistente que reduce la fiabilidad al seleccionar la mejor predicción entre las vistas. Aunque el conjunto de datos institucionales empleado tiene una proporción razonable de datos etiquetados (25%), no es suficiente para garantizar un modelo eficaz. De manera similar, estudios como CT01 (86.9%) y CT09 (87.26%) logran precisiones moderadas, pero están limitados por la falta de técnicas que capturen relaciones semánticas complejas o manejen adecuadamente datos desequilibrados.

Este tipo de modelos destacan por su flexibilidad para manejar datos no etiquetados y su eficiencia en escenarios con datos escasos. Por ejemplo, CT01 aprovecha clasificadores en vistas independientes que aprenden mutuamente, permitiendo trabajar con pocos documentos etiquetados y maximizando el aprovechamiento de datos no etiquetados, enfoque que también se observa en CT04 con la generación de instancias no etiquetadas para fortalecer el entrenamiento. Por otro lado, CT09 demuestra una notable eficiencia al utilizar un algoritmo evolutivo de auto-etiquetado que optimiza la precisión y la densidad de los datos, eliminando atípicos y mejorando la distribución en conjuntos desequilibrados, lo que refuerza su capacidad de adaptación en contextos complejos.

Sin embargo, los modelos de co-entrenamiento enfrentan desafíos relacionados con la complejidad de su estructura y los altos costos computacionales. Por ejemplo, CT03 presenta una estructura compleja que puede provocar la pérdida de características, impactando en su rendimiento, mientras que CT06 muestra una tendencia al sobreajuste, lo que limita su capacidad de generalización. Además, modelos como CT08 requieren

una alta inversión de recursos debido al uso de técnicas de entrenamiento profundo, restringiendo su aplicabilidad en entornos con capacidades computacionales limitadas.

3.3. Análisis de modelos de ensamblados

La capacidad de combinar múltiples clasificadores es la cualidad por la cual los modelos ensamblados han obtenido relevancia en los ambientes semi-supervisados, la variedad de clasificadores fortalece la precisión y la generalización de las predicciones. La Tabla 3 presenta nueve estudios que emplean diferentes técnicas en distintas etapas del modelo, aplicadas a datos de diferentes dominios. Este análisis explora las principales ventajas y desventajas de cada modelo, destacando sus características en la eficiencia del procesamiento, la gestión de datos no etiquetados y el manejo de conjuntos de datos complejos, al tiempo que identifica los retos asociados a su implementación, como la complejidad estructural y los costos computacionales.

Los modelos de ensamblado aplican diversas técnicas adaptadas a las necesidades del problema y del conjunto de datos. En preprocesamiento, destacan métodos como BoW, TF-IDF, N-gramas, y técnicas avanzadas como Word2Vec y Spacy. Para clasificación, se emplean algoritmos robustos como SVM, NB, NN y transformers como BERT (Bidirectional Encoder Representations from Transformers) y RoBERTa (Robustly Optimized BERT pre-training Approach), complementados con validación cruzada y métodos de autoetiquetado para aprovechar datos no etiquetados.

Tabla 3. Estudios y técnicas usadas en etapas del modelo de ensamblado.

(Tomado de [27])

Id	Autor	CsD	E1	E2	E3	E4	Dataset	Docs.	Clases	E	NE	Test	P(%)	Ventajas	Desventajas
ES01	De Souza, 2021 [75]	2	BoW	Spacy	1: BERT 2: RoBERTa	CSW	UCI/Tobacco	3482	10	348 (10%)	2786 (80%)	348 (10%)	85.91	Utiliza la eficiencia de los modelos transformer con su análisis de datos secuenciales; buen rendimiento de clasificación con pocos etiquetados y muchas clases.	Modelo complejo por iniciar su funcionalidad con la extracción de texto de imágenes, para posteriormente procesar texto.
ES02	Mouriño-García et al., 2018 [76]	2	N-gram	cd1: BoW cd2: WM	Hybrid-WikiBoc	NB	UCI/Reuters	27000	6	5000 (18%)	20400 (76%)	1600 (6%)	84.9	Realiza una representación de documentos por significado; la reducción de dimensionalidades refuerza el significado de las características con transferencia de aprendizaje de Wikipedia.	Arquitectura multimodal robusta, la apertura al tratamiento de videos, imágenes y textos reduce el rendimiento del modelo.
ES03	Mouriño-García et al., 2017 [77]	2	N-gram	cd1: BoW cd2: WM	Hybrid-WikiBoc	SVM	Institucionales/ UvigoMed	3979	26	500 (12%)	2856 (72%)	623 (16%)	68.9	Análisis semántico de las características; entrenamiento con vistas de datos en diferentes idiomas.	El rendimiento de clasificación se reduce cuando el conjunto de documentos de entrenamiento es grande.
ES04	Mouriño García et al., 2018 [78]	2	N-gram	cd1: BoW cd2: WM	Hybrid-WikiBoc	SVM	Institucionales/ UvigoMed	23647	22	5000 (21%)	6126 (26%)	12521 (53%)	68.5	El proceso de entrenamiento para los diferentes clasificadores es multilingüe.	Existe características que no se consideran durante la interacción entre lenguajes.
ES05	Salman, 2019 [79]	20	W2V	LDA	K-fold Cross validation (5)	WELM	UCI/WebKB	8300	4	2756 (33%)	4169 (50%)	1375 (17%)	88.84	Utiliza una arquitectura de clasificadores débiles para su entrenamiento; técnicas de Adaboost para el consenso de predicción; es multclasificador.	Las características resultantes son de alta dimensionalidad; la diversidad de clasificadores débiles genera un alto costo computacional.
ES06	Shrivastava et al., 2021 [80]	3	s/r	s/r	K-fold Cross validation (10)	cd1: MLP cd2: NB cd3: RF	Spam/e-mails	5975	2	1793 (30%)	3585 (60%)	597 (10%)	97.25	Divide su entrenamiento en capas con cross-validation; fortalece su consenso de etiquetado con bagging, adaboosting y gradient boosting.	El rendimiento del modelo es pesado por lo cual ha sido adecuado para trabajar con dos clases.
ES07	Ghosh & Chopra, 2021 [81]	4	cd1: s/r cd2: s/r cd3: s/r cd4: N-gram	cd1: s/r cd2: LDA cd3: Spacy cd4: TF-IDF	BERT	1: SVM	UCI/Spdra	23800	7	11200 (47%)	5600 (24%)	7000 (29%)	92.9	Esquema de trabajo brinda apertura al pre-entrenamiento; permite la clasificación por multiclases.	No existe un pre-procesamiento adecuado de los documentos por tal razón existe alta dimensionalidad; modelo robusto.
ES08	de Vries & Thierens, 2021 [82]	5	s/r	s/r	RESSELL	1: GNB 2: SVM 3: KNN 4: RDT 5: LR	UCI/cars	1728	4	1123 (65%)	173 (10%)	432 (25%)	88.69	Adecuada automatización del auto-entrenamiento con varios clasificadores; combina clasificadores y unifica la predicción.	Se configuran demasiados parámetros para el desempeño de los clasificadores; el consenso de combinación es poco inteligente.
ES09	Han et al., 2020 [83]	5	BoW	TF-IDF	SSDTM	NN	Reseñas/ Películas	7000	2	1000 (15%)	5000 (70%)	1000 (15%)	82.69	El algoritmo de umbral dinámico pseudoetiqueta documentos evaluando la calidad de predicción de la etiqueta de mayor a menor; sus clasificadores entrenan de forma independiente y son evaluados acorde a su brecha de rendimiento.	Las pruebas de rendimiento se realizan únicamente con dos etiquetas; la independencia de los clasificadores genera complejidad y tiempo en su entrenamiento

El análisis de los estudios en la Tabla 4 destaca la variedad de técnicas empleadas en las etapas del modelo de ensamblado, desde el preprocesamiento hasta la clasificación. En preprocesamiento, se usan métodos tradicionales como BoW y N-gramas, así como las herramientas Word2Vec, TF-IDF y LDA que ofrecen representaciones más complejas, aunque enfrentan desafíos como la alta dimensionalidad. En estudios como ES01 y ES07, la integración de modelos pre-entrenados como combina precisión y eficiencia, pero a costa de mayor complejidad computacional.

En la clasificación, los modelos ensamblados sobresalen por su capacidad de integrar clasificadores mediante técnicas como Adaboost, bagging y gradient boosting. Estos enfoques, vistos en estudios como ES05 y ES06, mejoran la precisión al combinar predicciones, adaptándose a escenarios con datos escasos o no etiquetados. Sin embargo, también resaltan los clasificadores tradicionales como SVM y NB, utilizados en estudios como ES03 y ES04, que ofrecen soluciones más simples, aunque con limitaciones en conjuntos grandes o multilingües. Entre los casos destacados, ES06 logra una precisión del 97.25% con un sistema basado en validación cruzada y boosting, mientras ES05 alcanza un 88.84% combinando Word2Vec y Adaboost, a pesar de su alto costo computacional. No obstante, estos modelos enfrentan retos comunes, como la alta dimensionalidad de las características, la complejidad en el entrenamiento y las dificultades al trabajar con datos multilingües.

Entre los conjuntos de datos más relevantes para clasificación de documentos de estos modelos, se identifica el estudio ES01 que utiliza el conjunto de datos UCI/Tobacco, compuesto por 3482 documentos distribuidos en 10 clases, y emplea modelos de transformers. Este enfoque resalta por su rendimiento en clasificación con pocos documentos etiquetados y varias clases, aunque el modelo es complejo debido a su necesidad de extraer texto de imágenes antes de procesarlo. En el estudio ES02 se usa el conjunto UCI/Reuters con 27000 documentos y 6 clases, aplicando técnicas como BoW y WM (Wikipedia Miner) para representar documentos por semántica, sin embargo, se enfrenta a la reducción de rendimiento al integrar videos, imágenes y textos en una arquitectura multimodal (ES02). En otro caso, el estudio ES05 utiliza UCI/WebKB con 8300 documentos y 4 clases, empleando W2V y LDA, y creando una arquitectura de clasificadores débiles que mejoran el rendimiento mediante Adaboost, no obstante, este enfoque genera una alta dimensionalidad y un alto costo computacional (ES05). Estos estudios demuestran cómo los conjuntos de datos y los contextos de aplicación influyen en las técnicas de clasificación y sus respectivos desafíos.

Se analiza también los valores de precisión obtenidos, evaluando sus técnicas y la naturaleza de los conjuntos de datos, los resultados reflejan el impacto de sus estrategias en la clasificación de documentos, desde sistemas simples hasta arquitecturas complejas. Entre las mejores métricas de precisión se obtiene a ES06 que logra la mejor precisión (97.25%) utilizando un enfoque de aprendizaje en capas con validación cruzada y técnicas de consenso como bagging y boosting. Asimismo, ES07 obtiene una precisión del 92.9% al emplear conjuntos de datos pre-entrenados en su aprendizaje, a pesar de que la ausencia de un preprocesamiento adecuado genera una alta dimensionalidad. Por otro lado, ES05

obtiene una precisión del 88.84%, al emplear clasificadores débiles con técnicas de Adaboost y manejo multiclases.

En contraste, estudios como ES04 y ES03 muestran precisiones más bajas (68.5% y 68.9%, respectivamente). A pesar de integrar enfoques multilingües y análisis semánticos, el rendimiento se ve limitado por el tamaño de los datos de entrenamiento y las características no optimizadas en entornos multilingües. Estos resultados evidencian cómo las estrategias utilizadas y las características de los datos impactan significativamente en los niveles de precisión alcanzados.

Los modelos analizados destacan por su capacidad para integrar técnicas como transformers, aprendizaje multimodal y métodos de boosting, lo que les permite realizar clasificaciones eficientes incluso con datos limitados o en escenarios de múltiples clases. Además, sobresalen por su adaptabilidad a diferentes lenguajes y su capacidad para optimizar la representatividad de las características mediante la reducción de dimensionalidades en el caso de los transformers.

Sin embargo, enfrentan retos significativos como la alta complejidad computacional y el aumento de la dimensionalidad, en el caso de combinar múltiples técnicas sin un preprocesamiento adecuado, lo que impacta negativamente su rendimiento. Además, su desempeño puede disminuir con grandes volúmenes de datos o escenarios complejos.

3.4. Análisis de modelos de aprendizaje activo

Estos tipos de modelos se distinguen por su capacidad para seleccionar de manera estratégica los datos más informativos para el entrenamiento, permitiendo maximizar la eficiencia en escenarios con recursos limitados o escasez de etiquetas. Este enfoque integra en determinados casos la participación de expertos para etiquetar instancias críticas, garantizando mayor precisión en dominios donde el etiquetado manual es costoso o impracticable. En esta sección, en la Tabla 4 se analizan las técnicas empleadas, los contextos de aplicación de los conjuntos de datos y los resultados en términos de precisión.

Tabla 4. Estudios y técnicas usadas en etapas del modelo de aprendizaje activo

(Tomado de [27])

Id	Autor	E1	E2	E3	E4	Dataset	Docs.	Clases	E	NE	Test	P(%)	Ventajas	Desventajas
AL01	Y. Yang & Loog, 2018 [19]	TF-IDF	PCA	MMC	LR	UCI/Baseball	1993	2	2 puntos de datos	s/r	s/r	85.7	Alta sensibilidad al costo y significado de las palabras; estructura con apertura a pre-entrenamiento y multi-etiqueta.	Experimentación solo con regresión lineal; algunos etiquetadores son menos eficientes que una predicción aleatoria; esquema con alto costo computacional.
AL02	Bouguelia et al., 2018 [84]	BoW	s/r	WD1	SVM	UCI/Dígitos de pluma	25601	10	50 puntos de datos	7425	3517	97.2	Identifica instancias con mayor influencia en el modelo y determina etiqueta; mide probabilidad de etiqueta errada	La etiqueta identificada con alto ruido no puede ser re-etiquetada; alto costo computacional.
AL03	Liu, 2019 [20]	NBoW	BoW	SD-TD	WSAL	UCI/Dispositivos electrónicos	1000	2	2700 de SD	900 de TD	100 de TD	82.5	Esquema abierto a pre-entrenamiento y documentos multilingües.	En el etiquetado automático existe dificultad para identificar las características más significativas.
AL04	Reyes et al., 2018 [85]	s/r	s/r	MS	SVM	UCI/Guardería	12960	5	100	6480	s/r	91	Permite la comparación genérica de su rendimiento de clasificación con algún otro método de Active Learning; bajos costos de entrenamiento.	Centrado más en la comparación de rendimientos de clasificación que el afinar su rendimiento de clasificación.
AL05	Li et al., 2020 [18]	s/reg	SSKMS	STDP	SVM	UCI/USPS	9298	10	93 (1%)	912 (98%)	93 (1%)	83.65	Posee un algoritmo de autoetiquetado basado en núcleos con apertura a predecir etiquetas por etiquetado activo y co-etiquetado; núcleos detallan distribución de documentos; utilizado en situaciones con escasez de etiquetados.	Considerables porcentajes de error del algoritmo por núcleos en casos de co-etiquetados.

Los modelos estudiados en la Tabla 4 emplean distintas técnicas para abordar las distintas etapas del aprendizaje activo. Herramientas como BoW y TF-IDF destacan como enfoques iniciales para la representación de texto. Por ejemplo, en AL01 y AL02, estas técnicas se combinan con métodos como PCA (Principal Component Analysis) y WD1 (Weighted disagreement 1) para reducir dimensionalidad y mejorar la identificación de características significativas.

Se identifican modelos con estrategias de reducción de dimensionalidad, como PCA o MMC (Maximum model change) en AL01, las cuales mejoran la sensibilidad hacia características relevantes, permitiendo un enfoque más eficiente en los datos. Sin embargo, esta ventaja se ve contrarrestada por el alto costo computacional asociado a estas operaciones, lo que es evidente en AL01 y AL05, donde los núcleos y algoritmos de co-etiquetado generan mayor complejidad en la etapa de entrenamiento. Del mismo modo, técnicas como WSAL (Warm Start Active Learning) en AL03 destacan por su eficiencia en escenarios multilingües, mientras que enfoques más básicos, como los empleados en AL04, priorizan la facilidad para comparar distintos métodos de clasificación, aunque con un enfoque menos profundo en la optimización del modelo.

Los conjuntos de datos empleados en el listado de estudios varían en tamaño, estructura y dominios, en AL02 se utiliza el conjunto de datos de "Dígitos de pluma" con 25,601 documentos distribuidos en 10 clases, mientras que AL03 opera con un conjunto reducido de 1,000 documentos relacionados con dispositivos electrónicos para 2 clases. En este contraste se registra mejores niveles de precisión en AL02 que alcanzó un 97.2%, deduciendo que los modelos tienden a funcionar mejor en conjuntos con una mayor disponibilidad de datos.

En contextos específicos como los analizados en AL01 (UCI/Baseball) o AL05 (UCI/USPS), los datos presentan características menos complejas, lo que facilita el etiquetado y el entrenamiento inicial. Sin embargo, estudios como AL04, que trabaja con datos más diversificados (guarderías), demuestran que los modelos enfrentan desafíos adicionales cuando los datos incluyen ruido o clases menos representadas. Este entorno sugiere que el diseño del modelo debe adaptarse al tipo y volumen de datos, con el fin de conseguir un desempeño robusto y eficiente.

La precisión obtenida por los modelos varía desde el 82.5% (AL03) hasta el 97.2% (AL02), lo cual permite identificar un patrón en el que los modelos que integran técnicas de reducción de dimensionalidad y clasificadores robustos, como SVM, alcanzan mejores resultados. AL02, con BoW y SVM, logra un buen desempeño debido a su capacidad para identificar instancias influyentes y etiquetarlas de manera precisa.

En contraposición, modelos como AL03, muestran una disminución en precisión (82.5%) debido a la dificultad para manejar características significativas en escenarios multilingües o con una gran cantidad de datos no estructurados. Así también, AL05, con un 83.65% de precisión, destaca por su enfoque en escenarios con escasez de etiquetados, pero muestra un margen de error considerable en los co-etiquetados, lo que subraya la necesidad de estrategias más robustas para minimizar errores en este contexto.

En términos generales, los modelos que priorizan el balance entre representación eficiente y clasificación robusta logran un mejor rendimiento. El patrón observado indica que, aunque las técnicas avanzadas de etiquetado y reducción de dimensionalidad pueden mejorar la precisión, su éxito depende de un diseño óptimo que minimice el costo computacional y el ruido en los datos.

Entre las ventajas destacadas, los modelos de aprendizaje activo muestran su capacidad para identificar instancias influyentes en los datos, como se observa en AL02, donde se mide la probabilidad de error en las etiquetas, permitiendo una selección más precisa de instancias clave para el entrenamiento. Además, la apertura al pre-entrenamiento y el manejo de etiquetas múltiples son características sobresalientes en modelos como AL05, cuyo algoritmo de auto-etiquetado basado en núcleos es especialmente útil en situaciones con escasez de datos etiquetados, ajustando la distribución de los documentos de manera efectiva.

Por otro lado, los modelos enfrentan retos significativos. Por ejemplo, en AL02, la incapacidad para corregir etiquetas con ruido limita su adaptabilidad en escenarios más dinámicos. Asimismo, en AL03, se destaca la dificultad para identificar características significativas durante el etiquetado automático, lo que puede afectar la precisión del modelo en contextos más complejos. El alto costo computacional también es un factor recurrente en modelos como AL01 y AL05, especialmente en esquemas de etiquetado activo que requieren un procesamiento intensivo. Finalmente, la dependencia de configuraciones específicas en modelos más simples, como AL04, limita su capacidad de adaptación en aplicaciones más generales, aunque son útiles para evaluaciones rápidas de métodos.

En síntesis, los modelos de aprendizaje activo son una solución prometedora para tareas de clasificación con datos limitados, aunque su efectividad depende de un diseño cuidadoso y un equilibrio entre costo computacional y precisión en el etiquetado.

3.5. Análisis de modelos de aprendizaje de transferencia

El aprendizaje por transferencia (AT) permite reutilizar conocimiento previamente adquirido en nuevos contextos, optimizando el rendimiento de los modelos, especialmente en dominios con datos escasos o características complejas. A continuación, se analizan las técnicas empleadas, los contextos de aplicación, los resultados en términos de precisión, y las ventajas y desventajas de los modelos presentados en los estudios de la Tabla 5.

Tabla 5. Estudios y técnicas usadas en etapas del modelo de aprendizaje de transferencia

(Tomado de [27])

Id	Autor	E1	E2	E3	E4	Dataset	Docs.	Clases	Pre-E	Test	P(%)	Ventajas	Desventajas
AT01	Guo & Yao, 2021 [21]	CBoW	BoW	DVEM	K-means	Reseñas/Yelp	70000	5	650000	50000	60.56	Eficiente representación de documentos y entrenamiento por clusterización; al esquema de trabajo se puede aplicar redes neuronales para su entrenamiento.	Cuando los documentos conforman grandes cantidades de texto la reducción de dimensionalidad no es óptima y se pierde información semántica.
AT02	S. Yang et al., 2020 [86]	HowNet	SSC/SCM	SNN	NN	Institucionales/Publicaciones	516	2	8551	516	83.1	Análisis semántico mejora clasificación; adecuada reducción de dimensionalidad controlando sinónimos y polisemia; la correlación semántica reduce ambigüedad de palabras.	Diccionario de documentos únicamente con léxico chino para evaluación; el diseño y experimentación del modelo, está pensado y probado con textos de documentos chinos.
AT03	Fu et al., 2019 [87]	Conjunto de terminales GP	Árboles basados en GP	SD y TD	SLLM	UCI/ News group	4323	5	2361	1962	73	Transferencia de conocimiento con algoritmos genéticos generando clasificadores débiles para su procesamiento; puede clasificar sin definición de etiquetas.	El consenso de decisión entre clasificadores débiles es por votos; dificultad en identificar índices de evaluación.
AT04	Y. Zhu et al., 2021 [88]	s/n	BoW	VAE	SDGMs	UCI/ Multilingua	6000	4	4128	1000	88.2	Permite una transferencia de aprendizaje multilingüe; su estructura dispone un modelo generativo profundo.	El modelo está diseñado para trabajar con un dominio de documentos en específico.
AT05	Pan et al., 2022 [89]	N-gram	BoW	BART	LR	Institucionales/ Artículos	50000	3	40000	10000	90	Reducción de dimensionalidad basado en semántica; clasificación jerárquica y por ontología.	El rendimiento se deteriora según las categorías vayan incrementando.
AT06	Mohammed & Aldhubri, 2022 [22]	NLTK	NLTK	Fuzzy Logic CSA	FRBS	Reseñas/ IMDB	50000	2	25000	25000	94.98	Transfiere aprendizaje de sentimientos por medio de diccionarios y grados de pertenencia difusos.	Si las reglas difusas no son claras el rendimiento de clasificación es bajo.
AT07	Z. Yang, 2017 [90]	N-gram	LSTM	VAE	CNN	Reseñas/ Yahoo	1450000	10	100000	10000	57.4	Entrena por aprendizaje de transferencia de vocabularios disponibles; el modelo dispone de mecanismos de atención en las características.	Cuando los documentos son escasos existe dificultad para el entrenamiento.
AT08	Alahdal, 2020 [91]	NLTK	BoW	SD y TD	K-means	Personales/ Diario	2500	5	2B tweets 1.2M diccio.	2500	84.7	Utiliza técnicas de semilla para la transferencia de aprendizaje.	La experimentación se la realiza con textos de contenidos incompletos.
AT09	Wang et al., 2022 [92]	TF-IDF	TF-IDF	ssSCL-ST	SVM	Reseñas/ Amazon	54000	2	12000	12000	82.2	Modelo aprovecha la riqueza de un lenguaje para generar conocimiento y transferir a otro; reduce la pérdida de conocimiento entre lenguajes con un mapeo de uno a muchos en su conexión de pivotes.	Transferencia de lenguaje únicamente entre inglés y chino, es necesaria la apertura a otros lenguajes; los dominios de los lenguajes deben estar vinculados con dominios divergentes el entrenamiento no es eficiente.
AT10	F. H. Khan et al., 2019 [93]	CSWE	POS	SSMT	SVM	Reseñas/ Películas	52000	2	50000	2000	85.3	Modelo de extracción de conocimiento sencillo apoyado en SentiWordNet; posee una configuración de adaptación de dominio flexible de SD simple a TD múltiple o de SD múltiple a TD simple.	La gestión del desequilibrio de características en clases no es adecuada; la adaptación de dominio no considera pesos de características destino.
AT11	Du et al., 2020 [94]	W2V	W2V	TrAdaBoost	GBC	Institucionales/ Bugs	920	2	807	113	81.2	Se rompe el mito de que el conjunto de entrenamiento y prueba deben tener una misma distribución; su aprendizaje por transferencia es más eficiente que incluir ese aprendizaje como etiquetado.	No se dispone métricas para evaluar el rendimiento de la predicción; depende de un solo clasificador.

Entre las técnicas de los estudios revisados en la Tabla 5, se identifica que en AT01 se utiliza una combinación de CBoW y BoW para representar documentos, complementado con K-means para agrupar instancias relevantes. En AT07, la integración de N-gram, LSTM (Long short-term memory) y VAE (Variational autoencoder) demuestra el potencial del aprendizaje profundo y los mecanismos de atención para extraer características complejas. Por su parte, AT05 recurre a N-gram para clasificaciones jerárquicas basadas en ontologías.

Técnicas como la combinación de modelos generativos (VAE en AT04 y AT07) con conjuntos de datos pre-entrenados evidencian que los enfoques híbridos tienden a ser efectivos en contextos multilingües o multiclase. Sin embargo, en modelos como AT02 y AT06, se observa que las técnicas basadas en semántica SSC/SCM (Semantic similarity computation / Strong correlation method) y Fuzzy Logic CSA (Crow search algorithm) son especialmente útiles para reducir la ambigüedad y aumentar la representatividad semántica, lo que puede ser crucial en análisis lingüísticos específicos.

Los modelos registran diversos contextos en sus conjuntos de documentos, desde análisis de sentimientos en reseñas (AT06 en IMDB y AT01 en Yelp) hasta clasificación de documentos institucionales (AT05 y AT11). Se evidencia que los conjuntos de datos más grandes, como el caso de AT07 con reseñas de Yahoo (1,450,000 documentos), tienden a presentar mayores desafíos para la optimización y el manejo semántico, mientras que conjuntos más pequeños como AT11 (920 documentos de bugs) muestran la efectividad de la transferencia de aprendizaje para resolver problemas específicos.

La adecuación del modelo pre-entrenado al dominio mejora en función de la calidad de las etiquetas iniciales presentes en los datos de destino, como el uso de diccionarios específicos para textos en chino en AT02, que logra un alto nivel de precisión (83.1%) en un contexto altamente especializado, pero presenta limitaciones de generalización.

Los niveles de precisión presentan valores destacados como 94.98% en AT06, que utiliza lógica difusa para analizar sentimientos, y 90% en AT05, que aplica datos pre-entrenados en clasificación jerárquica. Sin embargo, modelos como AT07 (57.4%) evidencian desafíos significativos en la representación y el manejo de grandes volúmenes de datos.

Se identifica modelos que combinan técnicas de reducción de dimensionalidad (como TF-IDF en AT09 o BoW en AT01) con métodos de clasificación robustos, como SVM o LR (Logistic Regression), tienden a mantener un equilibrio entre precisión y eficiencia computacional. Sin embargo, en contextos con alta diversidad de datos, como AT08, la precisión puede verse afectada por contenidos incompletos o distribuciones poco consistentes.

Entre los hallazgos más significativos, destacan los modelos que aprovechan técnicas de adaptación semántica y multilingüe. Por ejemplo, AT02 (83.1%) y AT09 (82.2%) logran mejorar la precisión mediante el control de sinónimos, polisemia y la reducción de ambigüedad semántica. De igual forma, AT04 (88.2%) sobresale en contextos multilingües al permitir transferencia de aprendizaje en una estructura generativa profunda. AT05 (90%) para su reducción de dimensionalidad combina BoW con una

clasificación jerárquica basada en semántica, mejorando la organización por categorías de problemas complejos. AT10 (85.3%), por su parte, utiliza una configuración flexible para la adaptación de dominio, ajustándose a diferentes relaciones entre dominio fuente y objetivo.

Sin embargo, también se identifican desventajas en algunos modelos. Por ejemplo, AT01 enfrenta pérdidas de información semántica al trabajar con grandes volúmenes de texto debido a las limitaciones en la reducción de dimensionalidad. De manera similar, AT05 experimenta un deterioro en el rendimiento al aumentar las categorías, lo que evidencia problemas de escalabilidad en su enfoque jerárquico. Restricciones de dominio y lenguaje son evidentes en modelos como AT02 y AT09, que dependen de recursos específicos como léxicos en chino o mapeos entre lenguajes. Esto limita su capacidad de generalización a otros idiomas o dominios divergentes. Por otro lado, AT04 está diseñado exclusivamente para dominios específicos de documentos, restringiendo su aplicabilidad en nuevos contextos.

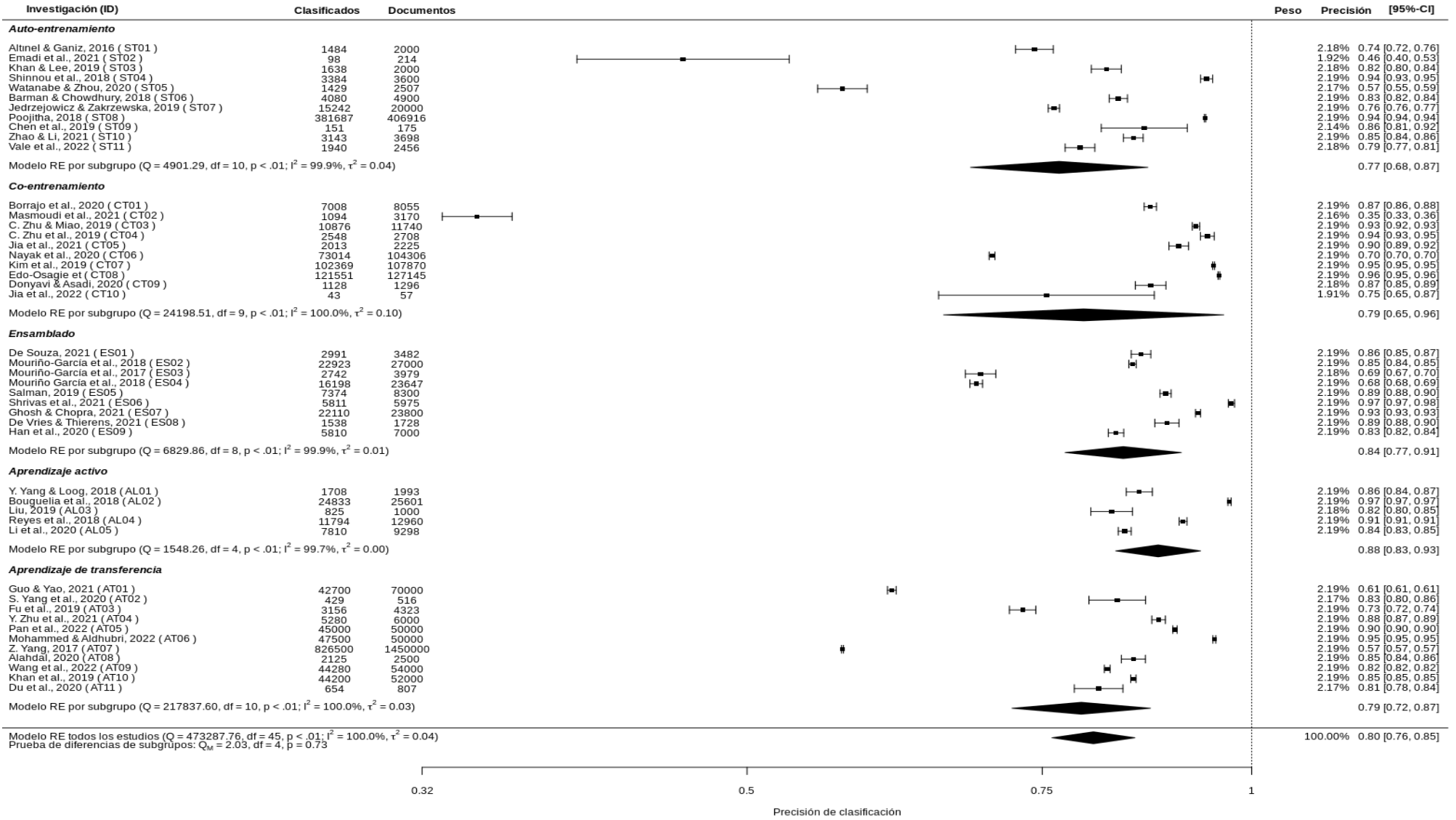
En conclusión, los modelos analizados evidencian avances significativos en transferencia de conocimiento y clasificación eficiente en dominios específicos y contextos multilingües. No obstante, persisten limitaciones en escalabilidad, adaptabilidad y experimentación, lo que sugiere la necesidad de enfoques más flexibles y robustos para superar estos desafíos.

3.6. Meta-análisis comparativo de los diferentes tipos de modelos SSL

En las secciones previas se analizaron 47 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 6 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 6. Forest plot agrupado de niveles de precisión de modelos semi-supervisados
(Tomado de [27])



Utilizando un modelo de efectos aleatorios (RE), se obtuvo un promedio general para todos los subgrupos analizados, con una precisión de 0.80 y un intervalo de confianza del 95% (CI [0.74 - 0.86]). Al evaluar el rendimiento por tipo de modelo, se observó lo siguiente en orden descendente: el aprendizaje activo obtuvo la mayor precisión (RE de 0.88, 95%CI [0.83 - 0.95]), seguido por los modelos ensamblados (RE de 0.84, 95%CI [0.75 - 0.92]). El co-entrenamiento alcanzó un RE de 0.79 (95%CI [0.62 - 1.00]), mientras que el aprendizaje por transferencia tuvo un RE de 0.79 (95%CI [0.68 - 0.89]). Finalmente, el auto-entrenamiento presentó un RE de 0.77 (95%CI [0.63 - 0.88]).

En el esquema Forest Plot se distinguen dos bloques de modelos. El primero incluye los modelos de auto-entrenamiento, co-entrenamiento y ensamblado, que se centran en trabajar con conjuntos limitados de documentos etiquetados. El segundo bloque está compuesto por los modelos de aprendizaje activo y de transferencia, caracterizados por utilizar recursos externos, como etiquetadores o bases de datos adicionales pre-entrenadas, para incrementar la cantidad de documentos etiquetados. Cada bloque incluye modelos con métricas destacadas, así como fortalezas y limitaciones específicas. Aunque el segundo bloque presenta la mejor métrica de precisión, es importante señalar que el modelo de aprendizaje activo depende de un etiquetador manual, lo que restringe la automatización del proceso de etiquetado. Además, la disponibilidad de estos etiquetadores puede estar limitada por factores como el tiempo y los recursos.

3.7. Ventajas y desventajas de los modelos SSL

En la Revisión Sistemática de Literatura (SLR) realizada efectuada para esta tesis [27], se identifican y definen las fortalezas y limitaciones más relevantes y recurrentes de los modelos de aprendizaje semi-supervisado. 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.

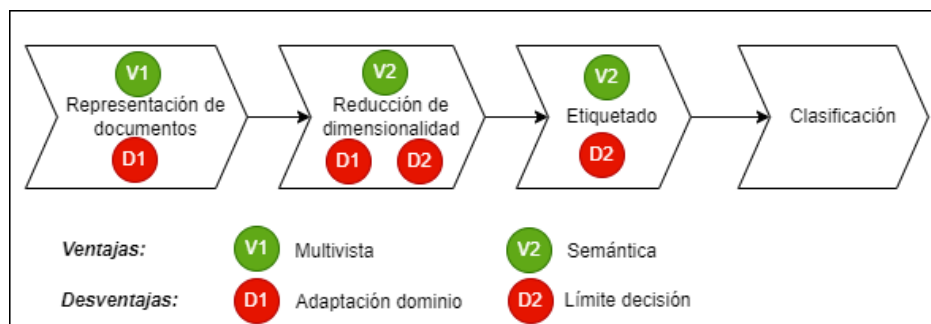


Figura 9. Ventajas y desventajas del proceso de clasificación.

En la Figura 9 se ilustra el flujo del proceso de clasificación de documentos en los modelos de aprendizaje semi-supervisado (SSL). En este esquema se ubican las fortalezas

y debilidades identificadas en cada una de las etapas correspondientes. Entre las fortalezas más destacadas se encuentran el enfoque *Multivista* (V1), que permite integrar diversas perspectivas de los datos para mejorar la clasificación; la *Semántica de características* (V2), que optimiza la comprensión del contexto y reduce la ambigüedad en la representación de datos; y el uso de *Crowdsourcing* (V3), que facilita la generación de etiquetas mediante la colaboración externa, ampliando los recursos disponibles.

Por otro lado, también se reflejan algunas limitaciones significativas. La *Adaptación de dominio* (D1) sigue siendo un desafío en escenarios donde los datos de entrenamiento y prueba provienen de contextos externos. Asimismo, el *Límite de decisión* (D2) puede afectar la precisión en casos donde un documento se ubique en una región que separa una categoría de otra, donde existe un alto margen de error en clasificación. Finalmente, el *Consenso entre clasificadores* (D3), característico de modelos ensamblados, puede introducir inconsistencias cuando los clasificadores débiles no logran acuerdos efectivos. Este análisis proporciona una visión equilibrada de las capacidades y limitaciones de los modelos SSL a lo largo del proceso de clasificación, a continuación, se detalla el desempeño asociado a las ventajas y desventajas identificadas.

- *Multivista*

El enfoque multisensorial (vista, oído, tacto) permite al cerebro aprender y adaptarse desde diversas perspectivas [95]. Inspirados en esta analogía, los modelos de aprendizaje SSL incorporan estrategias multivista, utilizando diferentes formas de representación para mejorar su entrenamiento. En contextos donde la cantidad de documentos etiquetados es limitada, estas representaciones alternativas desempeñan un papel crucial al ampliar el alcance del aprendizaje. Por ejemplo [96], un documento etiquetado puede estar disponible en varios idiomas (inglés, francés y español) o en diferentes formatos como texto, audio o video. Estas variaciones enriquecen la etapa de representación, aumentando la diversidad de documentos etiquetados disponibles para el modelo y, con ello, su capacidad de clasificación.

En este marco multivista, se han desarrollado diversos modelos diseñados para manejar múltiples representaciones de documentos. Entre las estrategias identificadas tenemos, por ejemplo, vistas basadas en títulos, contenidos y url del documento [31], vistas basadas en diferentes técnicas de representaciones como TF-IDF frente a representaciones basadas en LDA [71], y vistas que combinan categorías del documento con palabras asociadas a sentimientos positivos y negativos [72].

Además, se han identificado estudios que extienden el enfoque multivista a datos de gran escala, como audio y video. Modelos como SOMVFV (Semi-supervised One-pass Multi-View learning with variable Features and Views) [67], SSOPMV (Semi-supervised one pass multi view) [68] y SMDDRL (Semi-supervised multi-view deep discriminant representation learning) [69] han desarrollado estructuras específicas para la categorización de estos datos complejos. El uso de co-entrenamiento en estos sistemas multivista ha demostrado un incremento significativo en el rendimiento de clasificación, superando en cinco puntos el desempeño promedio de los modelos basados únicamente en auto-entrenamiento (ver Tabla 6). Este enfoque no solo potencia la precisión en la

clasificación, sino que también amplía el horizonte de aplicaciones de los modelos SSL al abordar escenarios más diversos y complejos.

- *Semántica de características*

El análisis de la semántica de las características es una etapa clave en el proceso de clasificación de documentos, particularmente durante el etiquetado (ver Figura 2). Su objetivo principal es asignar un valor semántico a los conjuntos de características del documento, permitiendo que estas puedan ser reutilizadas para clasificar documentos no etiquetados. Esto se logra al identificar características similares entre documentos etiquetados y no etiquetados, facilitando la asignación de categorías a los nuevos documentos. Según [14], la incorporación de semántica en las características ayuda a superar desafíos como la polisemia (palabras con múltiples significados) y los sinónimos, incrementando así la precisión de los modelos.

Diversos estudios en el ámbito del SSL han explorado enfoques para reforzar la semántica de las características mediante distintas técnicas. Por ejemplo, en [14] se propone el modelo HCSC (Hybrid Class Semantics Classifier), que se basa en relaciones semánticas entre términos dentro de clases, proporcionando un contexto más completo del documento y utilizando esta información para etiquetar nuevos datos. De manera similar, [17] presenta el modelo SSApolo, que emplea el método del círculo de Apolonio para identificar puntos de alta densidad en los datos. Este método agrupa puntos no etiquetados dentro de un círculo generado por un punto máximo, asignándoles la etiqueta correspondiente.

En [15] se introduce un framework con una capa de ingeniería diseñada para procesar las características del documento. Este sistema utiliza dos extractores semánticos: uno interno, que clasifica características gramaticales como sustantivos, adjetivos, verbos y adverbios, y otro externo, que emplea léxicos predefinidos de palabras relacionadas con sentimientos para asignar significado al documento. Así también, los diccionarios semánticos, como Word2Vec utilizado en [63], proporcionan una base robusta para incorporar significado a las características, permitiendo que los modelos sean más efectivos en su clasificación. Estos enfoques destacan la importancia de fortalecer la semántica como un elemento esencial para mejorar el desempeño de los modelos de aprendizaje SSL.

- *Adaptación de dominio*

El aprendizaje por transferencia permite incorporar conocimiento previamente generado al modelo en desarrollo, como diccionarios de palabras o conjuntos de datos etiquetados provenientes de distintos contextos, a los que se le denomina dominio fuente. Este conocimiento puede ser reutilizado para cumplir con objetivos específicos en un nuevo modelo, denominado dominio destino [97]. Sin embargo, la alineación entre el conocimiento existente y el dominio destino requiere de una adaptación de dominio, sin este ajuste, el conocimiento transferido podría no ser relevante, lo que afectaría negativamente el entrenamiento del modelo y comprometería su rendimiento en lugar de mejorarlo.

La adaptación de dominio se lleva a cabo, en la mayoría de los casos, durante la etapa de representación de documentos. Por ejemplo, en [62] se presenta un método que asigna pesos a las características del dominio fuente y, mediante el uso de coeficientes de correlación, identifica aquellas características que muestran mayor compatibilidad con el dominio objetivo. Así también, en [91] se aborda esta adaptación mapeando características desde el dominio fuente al dominio destino según su frecuencia en documentos pre-entrenados, con el propósito de fortalecer el proceso de entrenamiento de etiquetas.

A pesar de estos enfoques, el proceso de adaptación presenta desafíos significativos, particularmente en el análisis semántico de las características. Tanto en [62] como en [91], la falta de un análisis profundo de la semántica puede limitar la precisión, la interpretabilidad del documento y la optimización de la dimensionalidad. Esto evidencia que, para garantizar una transferencia de conocimiento efectiva, es fundamental incluir métodos que consideren la semántica como un eje central en la adaptación de dominio.

- *Límite de Decisión o Límite Lineal*

El proceso de etiquetado de documentos a menudo se basa en técnicas de agrupamiento, donde a los documentos dentro de un grupo se les asigna una categoría específica. No obstante, puede ocurrir que ciertos documentos se encuentren en los límites entre dos agrupaciones, lo que dificulta determinar su categoría con claridad. Este fenómeno, conocido como límite de decisión, introduce ambigüedad en el etiquetado, lo que puede impactar negativamente en la precisión y el rendimiento general del modelo de clasificación [17].

Diversos estudios, como los de [14], [17], [15] y [63], han implementado diferentes técnicas de etiquetado para minimizar los errores asociados al límite de decisión. Por ejemplo, en [17] se propone una metodología basada en gráficos geométricos para crear estructuras de grupos vecinales. Este enfoque alcanzó una precisión del 92,75% al clasificar documentos en tres categorías, pero el rendimiento disminuyó drásticamente al aumentar el número de categorías, registrando un 50,07% de precisión. Por su parte, [15] aborda este problema utilizando un esquema de agrupación en dos niveles. En el primer nivel, los documentos se agrupan según las categorías gramaticales de sus características (adjetivos, adverbios, verbos y sustantivos). En el segundo nivel, dentro de cada agrupación, se ordenan las características por sentimientos identificados. Este doble nivel de análisis busca reducir las incertidumbres propias del límite de decisión en ambas etapas del proceso.

3.8. Conclusiones del capítulo

Este capítulo ha explorado una revisión del contexto científico y tecnológico de los modelos de aprendizaje semi-supervisado (SSL) aplicados a la clasificación de documentos, evaluando su rendimiento, ventajas y limitaciones, bajo las siguientes variables: el dominio, el tipo de documento, el número de clases, la cantidad de datos

etiquetados y sus niveles de precisión. Se han analizado diferentes modelos, incluyendo el auto-entrenamiento, co-entrenamiento, ensamblados, aprendizaje activo y aprendizaje por transferencia. Además, se ha realizado un meta-análisis comparativo, proporcionando una visión integral del estado del arte en la clasificación de documentos mediante SSL.

Los resultados revelan dos enfoques principales en los modelos analizados: el primero corresponde a aquellos que optimizan su entrenamiento con un conjunto de datos etiquetados limitado (autoentrenamiento, coentrenamiento y ensamblados), mientras que el segundo incluye modelos que dependen de recursos externos, como bases de datos pre-entrenadas, para fortalecer su aprendizaje (aprendizaje activo y aprendizaje por transferencia). Los hallazgos indican que, dentro del primer grupo, los modelos de co-entrenamiento y ensamblado presentan las mejores métricas de precisión. En el segundo grupo, el aprendizaje activo logra el mayor nivel de precisión, aunque su dependencia de un etiquetador manual limita la automatización del proceso de etiquetado, lo que puede verse afectado por restricciones de tiempo y disponibilidad de recursos.

Además, se ha identificado que la adaptación de dominio y los límites de decisión representan desafíos críticos que impactan significativamente en la precisión de los modelos. Estos hallazgos resaltan la importancia de seleccionar el enfoque adecuado en función de las características del conjunto de datos y los requerimientos específicos de la tarea de clasificación. Finalmente, los resultados obtenidos en este capítulo proporcionan una base fundamental para el desarrollo del modelo propuesto en la presente tesis.

Capítulo 4

4. Explorando documentos científicos por áreas de investigación

El propósito del Capítulo 4 es diseñar un modelo semi-supervisado que permita automatizar la clasificación de los documentos científicos generados por la Universidad Técnica de Cotopaxi (UTC) según las áreas de investigación abordadas. Para ello, se elabora un conjunto de datos a partir del repositorio institucional de documentos científicos, el cual es alimentado por el personal académico de la universidad. El primer paso en la preparación de los datos consistió en definir las etiquetas de clasificación para los documentos recopilados en el repositorio institucional. La UTC cuenta con cinco facultades, cada una con áreas de investigación específicas definidas según las necesidades de su entorno académico, social y productivo. Siguiendo esta estructura, las etiquetas de clasificación se definieron en dos niveles. El primero corresponde a las facultades de ciencias: (C1) Agropecuarias, (C2) Recursos Naturales, (C3) Ingeniería y Ciencias Aplicadas, (C4) Administrativas y Económicas, y (C5) Sociales, Arte y Educación. Este nivel agrupa los documentos según su origen institucional. El segundo nivel se basa en las áreas de investigación específicas de cada facultad, permitiendo categorizar los temas abordados en función de sus principales líneas de investigación:

- *C1. Agropecuarias:* (SC1) Desarrollo y seguridad alimentaria, y (SC2) Salud animal.
- *C2. Recursos Naturales:* (SC3) Análisis, conservación y aprovechamiento de la biodiversidad local, y (SC4) Planificación y gestión de turismo sostenible.
- *C3. Ingeniería y Ciencias Aplicadas:* (SC5) Energías renovables y alternativas, eficiencia energética y protección ambiental; (SC6) Procesos industriales; y (SC7) Tecnologías de información y comunicación.
- *C4. Administrativas y Económicas:* (SC8) Administración y economía para el desarrollo humano y social, y (SC9) Gestión de calidad y seguridad laboral.
- *C5. Sociales, Arte y Educación:* (SC10) Educación, comunicación y diseño gráfico para el desarrollo humano y social; y (SC11) Cultura, patrimonio y saberes ancestral.

Este enfoque proporciona una organización jerárquica clara y sistemática de los documentos, optimizando su análisis y facilitando su aplicación en modelos de clasificación. El siguiente paso consistió en recuperar la metadata de los documentos científicos centralizados en el repositorio institucional, identificando publicaciones

registradas desde julio de 2018. Durante este proceso, se priorizó la extracción de información de documentos publicados en revistas científicas y aquellos presentados en conferencias académicas, considerando que estos representan un aporte significativo al desarrollo y la difusión del conocimiento en sus respectivas áreas. Para garantizar la calidad del procesamiento de texto, se seleccionaron como campos clave el título, el resumen y las palabras clave de cada documento, ya que estos contienen información esencial para el análisis semántico y la clasificación automática. Estos elementos proporcionan una descripción precisa del contenido y permiten identificar patrones relevantes que facilitan la asignación de etiquetas y la implementación de modelos.

4.1. Introducción

La producción científica es un indicador importante para evaluar la actividad investigativa de una universidad. No obstante, la creciente cantidad de información digital y la centralización de documentos en repositorios digitales han generado la necesidad de organizar y clasificar un volumen significativo de documentos [98]. En el ámbito académico, clasificar correctamente estos documentos resulta esencial para mejorar el acceso a información relevante, promoviendo así el intercambio de conocimientos [99]. Automatizar este proceso de clasificación no solo facilitaría la búsqueda y recuperación de información para los investigadores, sino que también permitiría identificar patrones en la producción científica [100], contribuyendo a la toma de decisiones en las áreas de investigación de las organizaciones [101].

Para lograr una automatización efectiva en la clasificación de documentos, es necesario contar con un conjunto de documentos etiquetados que respalde el entrenamiento del modelo. Cuanto mayor sea el número de documentos etiquetados disponibles, mayor será la efectividad de la clasificación [102]. Sin embargo, la disponibilidad de estos documentos etiquetados es limitada debido a restricciones de recursos y tiempo [61]. En este contexto, las estrategias de co-entrenamiento y transferencia de aprendizaje resultan efectivas para abordar este problema, ya que permiten optimizar la generación de etiquetas mediante el autoetiquetado y el uso de conjuntos de datos preentrenados [103].

Ambos modelos minimizan la dependencia de grandes conjuntos de datos etiquetados. En el caso del modelo de co-entrenamiento, la combinación de múltiples vistas con diversas perspectivas y criterios de decisión contribuye a mitigar la influencia de la limitada disponibilidad de datos etiquetados, reduciendo el ruido, el sesgo y la varianza en el modelo final. Esto mejora el proceso de etiquetado incluso con un número limitado de etiquetas [80]. Por otro lado, los modelos de transferencia aprovechan el aprendizaje derivado de conjuntos de datos previamente etiquetados en modelos pre-entrenados. Entre las arquitecturas más destacadas se encuentran BERT (Bidirectional Encoder Representations from Transformers) y BART (Bidirectional Auto-Regressive Transformers), que mejoran constantemente en volumen de datos pre-entrenados y velocidad de procesamiento mediante el uso de transformers. Estos factores contribuyen a reducir la dependencia de datos etiquetados [104].

En los últimos años, ha habido avances significativos en el desarrollo de modelos de co-entrenamiento y transferencia. Según el análisis de revisión de literatura realizado para esta tesis, sobre los modelos de aprendizaje semi-supervisados [27], entre los modelos más destacados, considerando la distribución de resultados en contextos diversos, con menor dispersión y un desempeño promedio superior en términos de precisión para la clasificación de documentos, se encuentran los modelos de co-entrenamiento (0,79) y los modelos de transferencia (0,79). En este estudio, estos modelos se aplicaron a conjuntos de datos compuestos principalmente por documentos y textos de reseñas de usuarios, con un promedio de cinco clases a clasificar.

La clasificación de documentos mediante modelos de co-entrenamiento o de transferencia, de forma independiente, es un tema recurrente en el campo de la investigación. Sin embargo, son escasos los estudios que exploran experimentos donde ambos modelos se combinen para la clasificación de documentos digitales [31]. Diversas investigaciones han abordado estos modelos de forma individual. Por ejemplo, en [69], se desarrolla un modelo de co-entrenamiento basado en una arquitectura MDDRL (Multi-view Deep Discriminant Representation Learning) con dos vistas. Este modelo realiza un procesamiento de datos en el que se representa documentos de páginas web de noticias mediante técnicas de extracción de ortogonalidad (orthogonality constraint). La estructura admite múltiples vistas, cada una con un tratamiento específico y compartido para su entrenamiento, predicción y clasificación. Estas vistas son co-entrenadas para establecer un proceso de etiquetado mediante redes neuronales.

En [70], se propone un modelo de co-entrenamiento para procesar documentos de reseñas de productos utilizando dos vistas: una de resolución fina y otra de resolución gruesa. La primera se basa en representaciones de características con bajo nivel de granularidad, como palabras individuales, mientras que la segunda utiliza representaciones con mayor granularidad, como conjuntos de palabras o frases completas. Ambas vistas son entrenadas y para integrar las distintas resoluciones, la alineación de instancias y la predicción de etiquetas utiliza mecanismos de atención. Aunque este modelo aprovecha diversas representaciones de documentos, tiende a sobreajustarse debido a la abundancia de elementos en la resolución fina, lo que resulta en una pérdida de rendimiento.

En cuanto a los modelos de clasificación de documentos que emplean aprendizaje por transferencia, [88] utiliza una estructura combinada con modelos generativos (DGMs) mediante el decodificador NX-VAE (autoencoder variacional) para extraer conocimiento del modelo pre-entrenado BERTW y clasificar documentos de noticias en cuatro categorías. Otro estudio es el de [89] en este modelo se construye una ontología para organizar las categorías de clasificación en una estructura jerárquica. Los documentos recopilados de Google Scholar se procesan, y el texto se extrae usando n-gramas. Este conjunto de características es entrenado y etiquetado mediante la técnica de clasificación de cero disparos (zero-shot), que utiliza el transformer pre-entrenado BART para etiquetar documentos de acuerdo con las categorías definidas por la ontología.

Se ha presentado varios estudios de clasificación de documentos en entornos de aprendizaje por co-entrenamiento y de transferencia, no obstante, escasas son las investigaciones enfocadas en la clasificación de artículos científicos con modelos de co-

entrenamiento utilizando fuentes de datos pre-entrenadas [31]. En el análisis de modelos por transferencia se identifica el uso de la clasificación de texto por pipeline [105], este método de clasificación multiclase recibe un conjunto de etiquetas que son sometidas al modelo pre-entrenado de BART para clasificar un texto predeterminado. Las limitantes de esta técnica radican en la sensibilidad al ruido de las categorías y el texto a clasificar, si las descripciones son ambiguas se pierde precisión en la clasificación.

Por esta razón en el presente estudio se busca diseñar un modelo que mejore los niveles de rendimiento de clasificación de documentos científicos, a través de un esquema de co-entrenamiento con dos vistas que extraen conocimiento de modelos pre-entrenados como BERT a través de estructuras Transformers. Para su entrenamiento en la primera vista se utiliza los textos de los títulos de los documentos científicos y en la segunda vista el texto de los resúmenes de los documentos. Para evaluar el rendimiento del modelo, la propuesta es comparada con el rendimiento del clasificador de texto por pipeline, que utiliza un modelo por transferencia que extrae conocimiento pre-entrenado de BART a través de un esquema de Transformers. Se espera expandir el conocimiento de la clasificación de documentos científicos ante los desafíos de mejorar la eficiencia del entrenamiento de los modelos de co-entrenamiento, en términos de su capacidad para lidiar con conjuntos de datos grandes, documentos complejos, heterogéneos y desequilibrados, así como reducir el ruido y los errores en los datos etiquetados.

4.2. Estructura del modelo COTRA

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 [31]. Los documentos etiquetados pueden ser tratados de forma específica o compartida entre las vistas con el fin de reforzar el proceso de entrenamiento [69]. La evaluación de la predicción, se convierte en un proceso más complejo por la diversidad de clasificadores existentes y las estrategias de decisión [74]. En la Figura 2 se presenta el esquema de trabajo de las etapas del proceso de aprendizaje de un modelo de co-entrenamiento.

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 [27]. 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 (ver Figura 10).

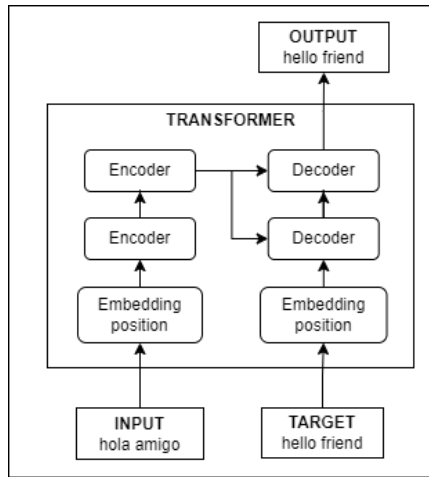


Figura 10. Estructura de transformer.

En esta sección se describe la estructura del modelo de co-entrenamiento (COTRA) propuesto, el modelo 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 científicos 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 Figura 11 se puede apreciar un resumen del modelo.

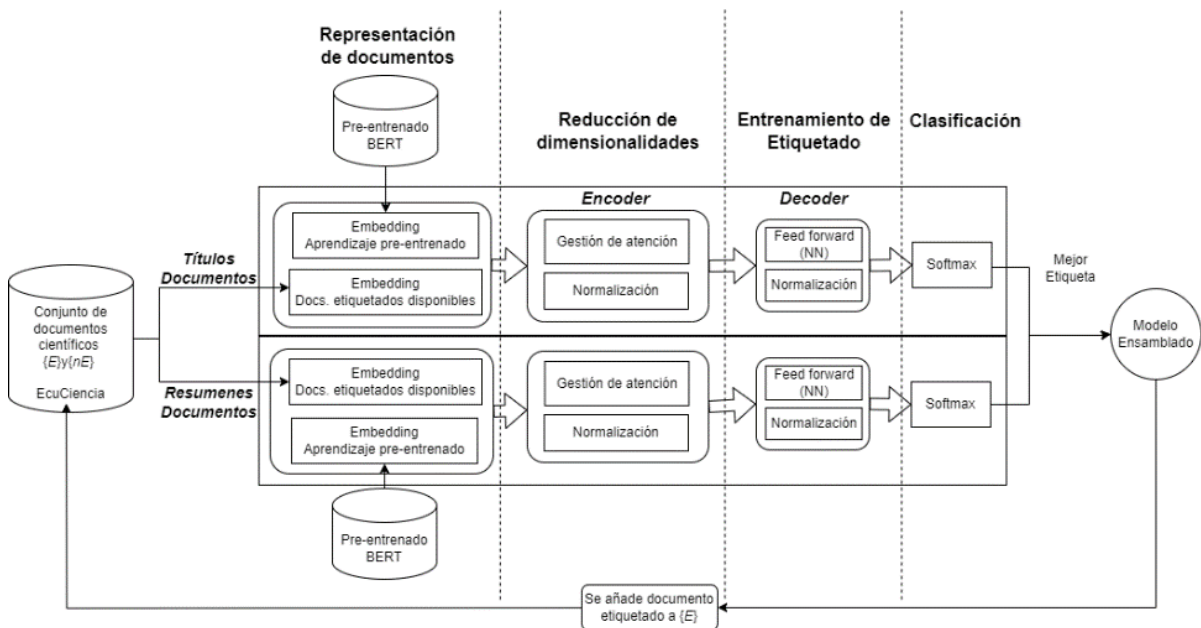


Figura 11. Estructura del modelo COTRA.

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.- La estructura de BERT permite la tokenización bidireccional de los títulos y resúmenes de los documentos científicos, para su representación el transformer utiliza embeddings que considerando su bidireccionalidad generan un mecanismo de atención que establece niveles de semántica entre tokens lo que permite tener un modelo focalizado en el relevancia del significado del texto.

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.

Generalmente los procesamientos de texto son secuenciales para este modelo el procesamiento de texto es paralelo, esto ha dado lugar a la paralelización del proceso de entrenamiento pudiendo ser realizado por GPUs, lo que permite el aprovechamiento de recursos y la aceleración del proceso.

Como entrada COTRA recibe: las dos vistas (título y resúmenes) $\{V\}=\{(V_i)|i=1,2\}$, el conjunto $\{Esd\}=\{(d_i)|i=1,\dots,n\}$ que contiene documentos etiquetados con pre-entrenamiento de BERT (SD), el conjunto $\{Etd^{(v1)}\}=\{(d_j)|j=1,\dots,n\}$ que contiene la vista 1 con los títulos de los documentos etiquetados del TD, el conjunto $\{Etd^{(v2)}\}=\{(d_j)|j=1,\dots,n\}$ que contiene la vista 2 con los resúmenes de los documentos etiquetados del TD, el conjunto $\{nEtd^{(v1)}\}=\{(d_k)|k=n+1,\dots,n+m\}$ que contiene documentos no etiquetados de la vista 1, el conjunto $\{nEtd^{(v2)}\}=\{(d_l)|l=n+1,\dots,n+m\}$ que contiene documentos no etiquetados de la vista 2, el conjunto $\{Cl\}=\{(c_k)|k=1,\dots,n\}$ 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 $\{Esd\}=\{(d_i)|i=1, \dots, n\}$;

Documentos (títulos) etiquetados TD $\{Etd^{(v1)}\}=\{(d_j)|j=1, \dots, n\}$;

Documentos (resúmenes) etiquetados TD $\{Etd^{(v2)}\}=\{(d_j)|j=1, \dots, n\}$;

Documentos no etiquetados TD $\{nEtd^{(v1)}\}=\{(d_j)|j=n+1, \dots, n+m\}$;

Documentos no etiquetados TD $\{nEtd^{(v2)}\}=\{(d_j)|j=n+1, \dots, n+m\}$;

Clases de distribución $\{Cl\}=\{(c_k)|k=1, \dots, n\}$;

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 $V2$;

$\{dr\}=\{\text{embeddings}\}$;

$\{C\}=\{NN\}$;

1: **repeat**

2: Representación de documentos $dr^{(v1)}$ y $dr^{(v2)}$ en cada vista para los documentos de SD $\rightarrow Esd^{(v)}$, TD $\rightarrow Etd^{(v)}$ 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 $\{E\}$

4.3. Caso de estudio

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 (ecuciencia.utc.edu.ec) que recopila documentos científicos de los investigadores de la UTC. En la presente sección se evalúa la eficiencia de clasificación del modelo propuesto con métricas de rendimiento de precisión y un conjunto de datos formado por los títulos y resúmenes de los documentos científicos recopilados de la plataforma EcuCiencia.

En la Tabla 7 se proporciona información sobre el conjunto de datos EcuCiencia, posee 898 documentos científicos recopilados del repositorio. Para apreciar el comportamiento del modelo propuesto, el 10% del conjunto de datos (90docs.) 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 7. Características y parámetros del conjunto de datos

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

En la Tabla 8 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 8. 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
C4	SC7 Tecnologías de información y comunicación.
	SC8 Administración y economía para el desarrollo humano y social
C5	SC9 Gestión de la calidad y seguridad laboral
	SC10 Educación, comunicación y diseño gráfico para el desarrollo humano y social
	SC11 Cultura, patrimonio y saberes ancestrales

La Tabla 9 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), así como con el modelo DGMs [88] que combina modelos generativos y aprendizaje por transferencia para extraer conocimiento del modelo pre-entrenado BERTW.

Tabla 9. 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
DGMs	Combinado	Pre-entrenado / Títulos&Resúmenes	Embeddings	NN
COTRA	Co- entrenamiento	Pre-entrenado / Títulos&Resúmenes	Embeddings	NN

En la Figura 12 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.

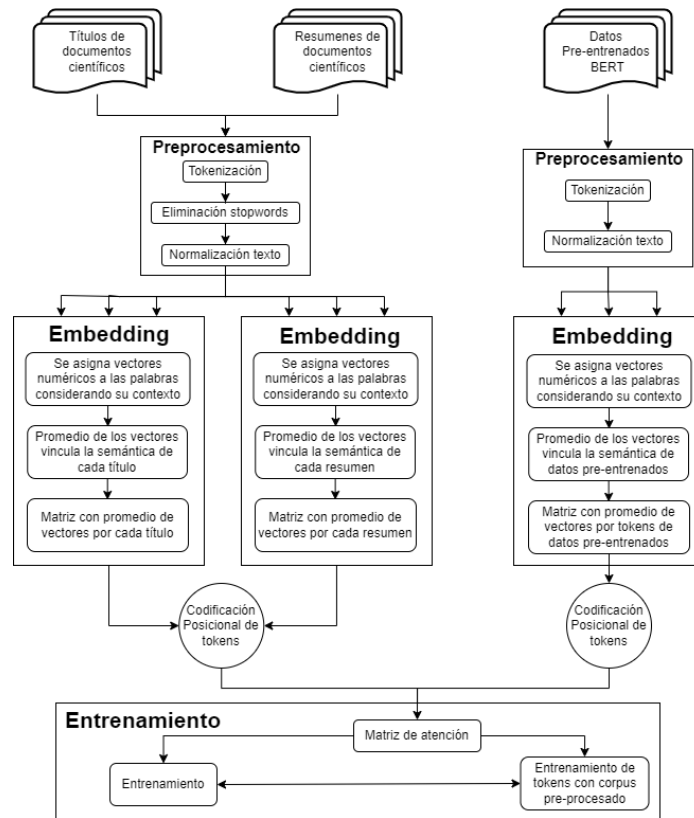


Figura 12. Preprocesamiento y representación de documentos del modelo.

La Tabla 10 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. Finalmente, los modelos DGMs emplean dos espacios latentes, z_1 y z_2 , cada uno con 768 dimensiones, para la representación de características.

Tabla 10. Configuración de parámetros de los algoritmos.

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

PIP2	Modelo	<i>Parámetros:</i> task=classification; model="bart-large-mnli". <i>Hiperparámetros:</i> padding='max_length'; max_length= 150; truncation=True; return_tensors='pt'.
COPIP	Modelo	<i>Parámetros:</i> dropout=0.5; nn.Linear(768, [5/11]); task=classification; model="bart-large-mnli". <i>Hiperparámetros:</i> max_length= 150; truncation=True; return_tensors='pt'; batch_size=16; Epochs=7; LR = 5e-5.
DGMs	Modelo	<i>Parámetros:</i> dropout=0.5; classifier (768, [5/11]); z1_dim=768; z2_dim=768. <i>Hiperparámetros:</i> padding='max_length'; max_length= 200; truncation=True; batch_size=4; Epochs=50; LR = 2e-5; vocabulary_size=1e-5; tie_embedding=true.
COTRA	Repr. Docs. y Modelo	<i>Parámetros:</i> dropout=0.5; nn.Linear(768, [5/11]); nn.LogSoftmax(dim=1). <i>Hiperparámetros:</i> 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 10. 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.

4.4. Evaluación y resultados

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 11 se presenta los valores de precisión de los modelos de clasificación V1, V2, PIP1, PIP2, COPIP, DGMs 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 0,68 (10%) a 0,73 (30%) para cinco clases (C), y de 0,35 (10%) a 0,64 (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 0,48 y 0,69 para SC.

Con respecto al modelo PIP1, entrenado con los títulos de los documentos en una estructura BART, presenta una precisión que varía entre 0,71 y 0,78 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, mientras que DGMs presenta la segunda mejor precisión, con un rango de 0,63 a 0,82 para C y de 0,52 a 0,73 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 0,77 y 0,87 para C y entre 0,56 y 0,79 para SC.

Tabla 11. Valores de precisión de los modelos

Modelo	C (5 clases)			SC (11 clases)		
	10%	20%	30%	10%	20%	30%
V1	0,68	0,7	0,73	0,35	0,50	0,64
V2	0,71	0,75	0,78	0,48	0,65	0,69
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
DGMs	0,63	0,71	0,82	0,52	0,62	0,73
COTRA	0,77	0,83	0,87	0,56	0,70	0,79

La evaluación del rendimiento de COTRA se presenta mediante una comparación con varios modelos individuales tradicionales (Figura 13) y modelos combinados (Figura 14), 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 de los modelos DGMs y V2. En contraste, la curva correspondiente a los modelos PIP presenta la menor precisión en comparación con el resto de modelos.

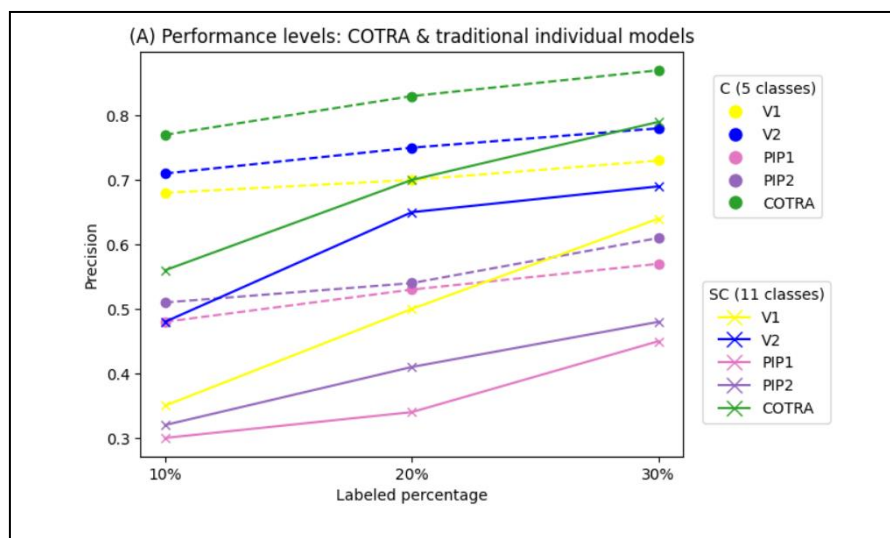


Figura 13. Niveles de rendimiento con modelos tradicionales individuales.

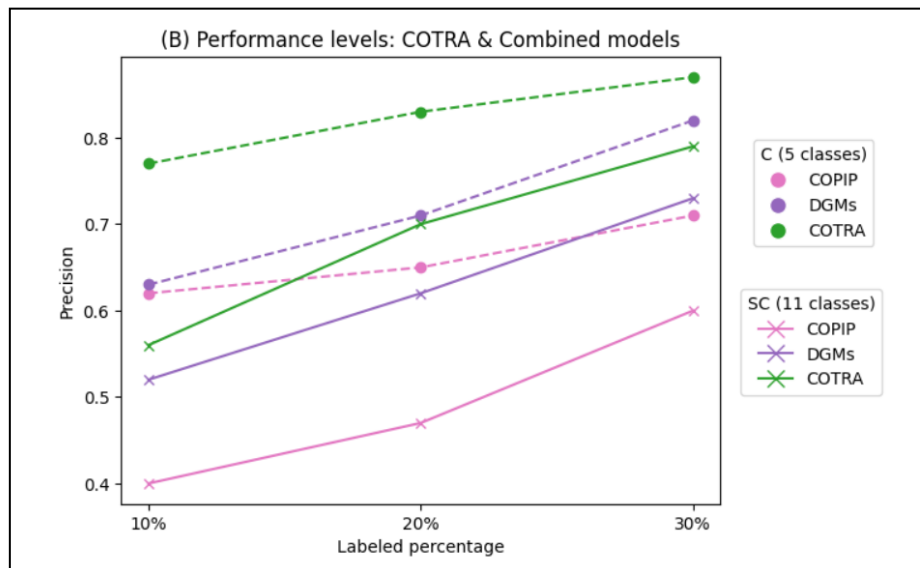


Figura 14. Niveles de rendimiento con modelos combinados.

Para validar estadísticamente las diferencias en la precisión de los modelos evaluados, se realizó un análisis de varianza (ANOVA) que permite determinar si las variaciones observadas en los resultados son estadísticamente significativas. En este estudio se aplicaron dos enfoques complementarios: el ANOVA de dos vías y el ANOVA de medidas repetidas. El ANOVA de dos vías se utilizó para evaluar el impacto de dos factores clave en la precisión de los modelos: el porcentaje de datos etiquetados (10%, 20% y 30%) y el número de clases en la clasificación (5 y 11 clases). Este análisis permitió identificar si existe una interacción significativa entre estos factores y la precisión alcanzada por los modelos. Adicionalmente, se aplicó un ANOVA de medidas repetidas, ya que cada modelo fue evaluado en múltiples condiciones experimentales (distintos porcentajes de etiquetado), lo que permite analizar las variaciones dentro de cada modelo a lo largo de las condiciones de evaluación.

El análisis de *ANOVA de dos vías* muestra que el porcentaje de datos etiquetados (10%, 20%, 30%) tiene un efecto significativo en la precisión de los modelos ($p=0.0047$), lo que indica que a medida que se incrementa la cantidad de datos etiquetados, los modelos mejoran su desempeño. Este comportamiento es notable en el modelo COTRA, que presentó una mejora progresiva en la precisión al aumentar el porcentaje de etiquetado. En la clasificación de 5 clases, su precisión pasó de 0.77 con el 10% de datos etiquetados a 0.87 con el 30%. De manera similar, en la clasificación de 11 clases, la precisión aumentó de 0.56 a 0.79 en los mismos niveles de etiquetado.

Por otro lado, el número de clases también tuvo un impacto significativo en la precisión de los modelos ($p=0.0001$). La reducción en el número de clases mejoró la precisión de todos los modelos evaluados. En el caso de COTRA, se observó que su rendimiento fue superior cuando la clasificación se realizó en 5 clases en lugar de 11. Esto sugiere que

una mayor cantidad de categorías dificulta la tarea de clasificación, reduciendo la capacidad de los modelos para generar predicciones precisas.

En cuanto a la interacción entre ambos factores, los resultados indican que no existe un efecto significativo ($p=0.4588$). Esto implica que la mejora en la precisión de los modelos debido al incremento en el porcentaje de datos etiquetados ocurre de manera similar tanto en la clasificación de 5 clases como en la de 11 clases. En el caso de COTRA, este hallazgo confirma que el modelo se beneficia del aumento en la cantidad de datos etiquetados sin que el número de clases modifique sustancialmente este efecto.

En cuanto al *ANOVA de medidas repetidas*, se consideraron los niveles de etiquetado, los resultados obtenidos muestran un valor de $F = 31.94$, con 2 grados de libertad en el numerador y 12 en el denominador, y un p-valor menor a 0.001. Estos resultados indican que existe una diferencia estadísticamente significativa entre los valores de precisión obtenidos por el modelo COTRA en los distintos niveles de datos etiquetados (10%, 20% y 30%). Dado que el p-valor es inferior a 0.05, se rechaza la hipótesis nula, lo que sugiere que la cantidad de datos etiquetados tiene un efecto significativo sobre la precisión del modelo.

El incremento de la precisión con un mayor porcentaje de datos etiquetados sugiere que el modelo COTRA aprovecha la información adicional disponible para mejorar su rendimiento. Esto refuerza la importancia de la búsqueda de estrategias para incrementar la cantidad de datos etiquetados en tareas de clasificación. Además, la magnitud del estadístico F sugiere que el impacto del porcentaje de datos etiquetados es considerable en comparación con la variabilidad residual.

4.5. Conclusiones del capítulo

En este capítulo 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.

Capítulo 5

5. Conclusión y trabajo futuro

5.1. Conclusiones

En el presente trabajo se ha determinado que los modelos de aprendizaje semi-supervisado son una herramienta eficaz para la clasificación de documentos institucionales, destacándose por su capacidad de adaptación a diversos conjuntos de datos y entornos. Además, se ha planteado una metodología para analizar el estado del arte en esta área, agrupando los modelos según sus técnicas de clasificación y comparando su desempeño mediante un análisis basado en la síntesis de sus conjuntos de datos y el rendimiento de clasificación.

La revisión de literatura efectuada en la presente tesis, ha identificado que técnicas como la semántica de características y el aprendizaje multivista contribuyen significativamente a mejorar la precisión de los modelos de clasificación documental. Además, se ha evidenciado que factores como la cantidad de documentos etiquetados y el número de clases a categorizar tienen un impacto directo en su desempeño. No obstante, los desafíos más recurrentes detectados en los distintos tipos de modelos semi-supervisados analizados, son la adaptación de dominio y la estabilidad en los límites de decisión. Estos hallazgos han dado lugar al desarrollo de soluciones especializadas para optimizar la eficacia de los modelos en distintos entornos y obtener una clasificación más precisa y robusta.

A partir del análisis de las fortalezas y debilidades de los distintos modelos de aprendizaje semi-supervisado, así como del desempeño observado en los casos de estudio, se diseñó el modelo semi-supervisado COTRA. Este modelo integra técnicas de co-entrenamiento para reducir el margen de error en los límites de decisión y técnicas de aprendizaje por transferencia mediante embeddings, con el objetivo de mejorar la eficacia en la adaptación de dominio.

Como parte de la experimentación, se preparó un conjunto de documentos científicos de la UTC para su clasificación por áreas de estudio. Debido a que los títulos y resúmenes de estos documentos contienen datos no estructurados, la arquitectura del modelo incorpora dos vistas del conjunto de datos, cada una con características diferenciadas, lo que enriquece la representación documental y su proceso de entrenamiento. Además, la estructura de aprendizaje se fortaleció con datos pre-entrenados de los transformers BERT y BART, lo que permitió una mayor precisión en el etiquetado de documentos.

La combinación de aprendizaje multivista y transferencia de conocimiento redujo el margen de error y optimizó el entrenamiento del modelo. Asimismo, el intercambio de etiquetas en los esquemas de co-entrenamiento y aprendizaje por transferencia contribuyó a mitigar el desafío de la escasez de documentos etiquetados, asegurando una clasificación más eficiente y precisa.

Este modelo puede aplicarse en otros dominios donde predomine la información no estructurada en los conjuntos de datos, como en la clasificación de libros, informes de auditoría, manuales, documentos legales, normativos, entre otros. Asimismo, la arquitectura de COTRA representa una mejora significativa respecto a los métodos tradicionales de clasificación de documentos, los cuales, en muchos casos, dependen de enfoques supervisados y no supervisados convencionales que requieren grandes volúmenes de datos etiquetados o, en su ausencia, generan distribuciones poco representativas y con alta dispersión. Su capacidad para reducir esta dependencia y optimizar la eficiencia del entrenamiento lo convierte en una alternativa robusta para entornos con recursos limitados.

En conclusión, COTRA representa un avance en la clasificación automatizada de documentos científicos, reduciendo la dependencia de grandes volúmenes de datos etiquetados y mejorando la precisión del proceso. No obstante, futuras investigaciones deberían enfocarse en optimizar su eficiencia computacional y evaluar su desempeño en conjuntos de datos más amplios y diversos para consolidar su aplicabilidad en diferentes contextos.

Evaluación SSL.- Se ha diseñado una estructura para clasificar los modelos semi-supervisados identificados (ver Sección 2.1), lo que ha permitido identificar y analizar estudios y técnicas relevantes para la clasificación de documentos. Esta estructura, en conjunto con la recopilación y evaluación de variables del conjunto de datos, como el dominio, el tipo de documento, el número de clases, la cantidad de datos etiquetados y los niveles de precisión del modelo, permitió realizar un análisis comparativo (Sección 3). Para ello, se llevó a cabo una síntesis de datos mediante Forest Plot, basada en la precisión de clasificación de documentos reportada en cada caso de estudio. Para analizar el desempeño de cada agrupación, se empleó el modelo de efecto aleatorio (RE), basado en la precisión reportada en los estudios que conforman cada grupo. Esta metodología permitió determinar un rendimiento global de los modelos, obteniendo un valor de precisión de 0.80, con un intervalo de confianza del 95 % (CI [0.74 – 0.86]).

Así también, según lo expuesto en la Tabla 6, se observa que un aumento en la cantidad de datos etiquetados o una disminución en el número de categorías a clasificar favorecen una mayor precisión en el modelo. Durante el análisis del desempeño en la clasificación, se reconoció esta tendencia en las distintas variantes evaluadas, evidenciando que tanto el número de documentos etiquetados como la cantidad de clases influyen directamente en las métricas de precisión.

Es fundamental considerar que los desafíos en la clasificación de documentos, así como los conjuntos de datos y los recursos disponibles, varían según cada caso. No obstante, los modelos semi-supervisados han demostrado ser adaptables a distintas condiciones. En este trabajo, se identificaron los modelos más efectivos en función de la disponibilidad de recursos y fuentes externas pre-entrenadas. Se observa que, cuando se dispone de recursos suficientes y se puede contar con etiquetadores manuales, el aprendizaje activo los incorpora en su estructura, alcanzando un rendimiento óptimo de 0,89. En contraste, en ausencia de estos recursos, los modelos de ensamblado utilizan clasificadores débiles para mejorar el proceso de etiquetado sin intervención humana, logrando resultados igualmente significativos de 0,83. Respecto a los conjuntos de datos, si existen fuentes externas pre-entrenadas, los modelos de aprendizaje por transferencia pueden aprovechar esta información para el proceso de categorización, obteniendo un desempeño moderado de 0,78. En ausencia de estas fuentes, el modelo de co-entrenamiento aprovecha las diferentes vistas generadas a partir del conjunto de datos, permitiendo una clasificación eficiente con índices de precisión aceptables 0,79.

Se ha identificado que los principales desafíos que enfrentan estos modelos incluyen la adaptación de dominio y el límite de decisión, los cuales se analizan en detalle en la sección 3.7. El límite de decisión es un aspecto recurrente en todos los tipos de modelos, y varios estudios han intentado establecer límites más estables y menos propensos a errores. No obstante, muchas de estas soluciones resultan efectivas únicamente para datos lineales [14] [15] [17] [63] mientras que en el caso de datos no lineales aún persisten desafíos por resolver. En cuanto a la adaptación de dominio, esta técnica permite aprovechar el conocimiento propagado en la red para ajustarlo a un problema específico de clasificación. Sin embargo, el principal reto radica en que los dominios de origen y destino pueden presentar diferencias significativas [62] [91]. Por esta razón, es fundamental realizar una adaptación previa de los dominios con el fin de maximizar el aprovechamiento del conocimiento y mejorar la eficacia del modelo.

Límite de decisión.- Se estructura un modelo basado en un enfoque de múltiples vistas, donde se utilizan dos representaciones distintas del mismo conjunto de datos, como se puede apreciar en la Figura 11. Esta estrategia introduce redundancia en la primera fase de representación de los documentos de las etapas del esquema de auto-entrenamiento (ver Figura 2), lo que contribuye a reducir la ambigüedad en la toma de decisiones. En particular, cuando la clasificación de una de las vistas presenta incertidumbre, la otra vista proporciona información complementaria que refuerza la decisión final.

Este enfoque estructural disminuye la variabilidad en la frontera de decisión, lo que se refleja en un incremento en los niveles de precisión del modelo. Los resultados experimentales presentados en la Tabla 11 demuestran que los modelos de múltiples vistas superan en rendimiento a aquellos basados en una única vista. En particular, los modelos multivista COPIP y COTRA muestran métricas superiores a PIP1, PIP2, V1 o V2 respectivamente, destacándose especialmente en entornos con datos ruidosos o distribuciones de clases desbalanceadas.

Además, la estrategia de selección de vistas con mejor desempeño en clasificación, basada en métricas de predicción y confianza, refuerza la estabilidad y confiabilidad del sistema. Este mecanismo no solo optimiza la capacidad de generalización del modelo, sino que también mejora su resiliencia ante la incertidumbre en los datos, garantizando un desempeño más robusto y consistente en diversas condiciones experimentales.

Adaptabilidad de dominio.- Para abordar los desafíos de la adaptación de dominio en modelos semi-supervisados, la arquitectura propuesta no solo incorpora un enfoque multivista, sino que también integra técnicas de transferencia de aprendizaje. Esto permite mejorar la capacidad de generalización de los modelos al aplicarlos en nuevos contextos. Se ha integrado a la estructura un mecanismo de transferencia de conocimiento mediante la incorporación de un transformer, lo que permite que un dominio fuente (SD) aporte información relevante a un dominio destino (TD). En esta implementación, SD corresponde a un conjunto de datos multilingüe pre-entrenado BERT, que contiene aproximadamente 3.3 mil millones de palabras, mientras que TD está compuesto por 898 documentos científicos recopilados para la tarea de clasificación por área de estudio. Estos documentos han sido organizados para su entrenamiento en dos conjuntos diferenciados: uno que contiene los títulos de los artículos científicos y otro que almacena los resúmenes de los mismos.

La incorporación de la transferencia de aprendizaje en la arquitectura del modelo ofrece varias ventajas significativas. En primer lugar, permite aprovechar las representaciones latentes adquiridas en el dominio fuente para mejorar la capacidad de generalización en el dominio objetivo. Esto es particularmente útil cuando el conjunto de datos etiquetado en el dominio objetivo es limitado o costoso de obtener. Además, al utilizar un modelo basado en transformers, se optimiza el proceso de ajuste fino (fine-tuning), lo que facilita la adaptación del modelo a las características específicas de TD sin requerir un re-entrenamiento completo desde cero.

La efectividad del mecanismo de transferencia implementado en el modelo se ha evaluado mediante los niveles de precisión alcanzados en la clasificación de documentos científicos. Los resultados presentados en la Tabla 11 evidencian que el modelo COTRA supera en precisión a DGMs y COPIP. En el caso de DGMs, la transferencia de aprendizaje se realiza a través de modelos generativos basados en el decodificador NX-VAE, el cual emplea una codificación latente para representar las características del dominio fuente. Por su parte, COPIP utiliza una estrategia de transferencia mediante el modelo de abstracción de clasificación de texto pipeline zero-shot. La ventaja de COTRA radica en su capacidad para aprovechar de manera más eficiente las representaciones aprendidas del dominio fuente, logrando una mayor generalización y precisión en contextos diversos. Esto se debe a que el modelo aprovecha estructuras lingüísticas y patrones semánticos previamente adquiridos en SD, lo que permite obtener representaciones más robustas y evitar problemas de sobreajuste.

Modelo de co-entrenamiento.- Se ha logrado estructurar un modelo de co-entrenamiento integrado con transferencia de aprendizaje, diseñado para mejorar la capacidad de clasificación de documentos. La arquitectura propuesta aprovecha la colaboración entre dos vistas del conjunto de datos, lo que permite enriquecer la representación de los documentos mediante características diferenciadas. Además, la incorporación de modelos pre-entrenados, como BERT, facilita la transferencia de conocimiento desde dominios fuente hacia el dominio objetivo, optimizando el proceso de aprendizaje. Esta combinación sinérgica de co-entrenamiento y transferencia de aprendizaje fortalece la generalización del modelo, permitiendo un rendimiento más eficiente y preciso en tareas de clasificación, incluso cuando la disponibilidad de datos etiquetados es limitada.

La combinación del co-entrenamiento y el aprendizaje por transferencia ha permitido reducir el margen de error en el etiquetado de documentos, logrando un proceso de entrenamiento más eficiente. Asimismo, el mecanismo de compartición de etiquetas en los esquemas de co-entrenamiento y transferencia contribuye a mitigar las dificultades derivadas de la limitada disponibilidad de documentos etiquetados. Los resultados experimentales confirman que la eficiencia de clasificación del modelo COTRA alcanzó las mejores métricas de precisión en todos los casos, superando el entrenamiento individual de cada vista, los modelos de clasificación zero-shot y los modelos combinados (V1, V2, PIP1, PIP2, COPIP y DGMs). Para el entrenamiento del modelo, se emplearon diferentes porcentajes de documentos etiquetados, que oscilaron entre el 10 % y el 30 %, con el objetivo de analizar el comportamiento del modelo a medida que aumenta la cantidad de documentos etiquetados. En términos generales, se observó que el incremento significativo en la precisión del modelo es directamente proporcional al aumento del porcentaje de documentos etiquetados utilizados en el entrenamiento. Este resultado demuestra la importancia de ampliar la cantidad de documentos etiquetados para mejorar el rendimiento del modelo en tareas de clasificación.

5.2. Trabajo futuro

No obstante, 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.

Acrónimos o siglas

ASC	Adversarial Similarity Constraint	NLTK	Natural Language Toolkit
BART	Bidirectional Auto-Regressive Transformers	NN	Neural network
BERT	Bidirectional Encoder Representations from Transformers	NSGA-II	Non-dominated Sorting Genetic Algorithm
BoW	Bag of Words	OC	Orthogonality Constraint
C4.5	Decision tree classifier	PCA	Principal Component Analysis
CBoW	Continuous bag of words	PoS	Speech
CNN	Convolutional Neural Networks	RDT	Random decision tree
CSA	Crow search algorithm	RE	Random effect
CSW	Critical software	RESSELL	Reliable semi-supervised ensemble learning
CSWE	Cosine similarity weight Extraction	RF	Random Forest
DPC	Density Peaks Clustering	RoBERTa	Robustly Optimized BERT pre-training Approach
DTGMO-SSC	Diverse training dataset generation based on a multi-objective optimization for semi-Supervised classification	SDGMs	Semi-supervised deep generative models
DVEM	Document vector extensión model	SD-TD	Source domain - Target domain
EM	Expectation Maximization	SLLM	Self learning linear mode
FlexCon-C2	Flexible Confidence Classifier 2	SMDRL	Semi-supervised multi-view deep discriminant representation learning
FRBS	Fast radio bursts	SNN	Semantic convolution neural network
GBC	Gradient Boosting Classifier	SOM	Self organizing map
GNB	Gaussian Naive Bayes	SOMVfV	Semi-supervised One-pass Multi-View learning with variable Features and Views
GP	Genetic Programming	SSC/SCM	Semantic similarity computation / Strong correlation method)
HCSC	Hybrid Class Semantics Classifier	SSDTM	Semi-supervised model based on dynamic threshold and multiple classifiers
HMM	Hidden Markov Model	SSKMS	Semi-supervised k-means with seeds
HTF	Hidden feature transformation	SSMT	Single source Multiple target domain
IG	Information Gain	SSOPMV	Semi-supervised one pass multi view
KNN	K-nearest neighbors	ssSCL-ST	Semi-supervised learning with SCL and space transfer
LDA	Linear discriminant analysis	STDP	Self-Training with Density Peaks
LR	Logistic Regression	STDPNaN	Self-training method based on density peaks and natural neighbors
LSTM	Long short-term memory	STFW	Static threat factor weight
MCT	Multi Co-training	SVM	Support Vector Machine
MIL	Multi Instance Learning	TF-IDF	Term Frequency-Inverse Document Frequency
MLP	Multilayer perceptron	TrAdaBoost	Transfer AdaBoost
MLSMOTE	Multilabel Synthetic Minority Over-sampling Technique	VAE	Variational autoencoder
mLVQb	Batch multi-label learning vector quantization	W2V	Word2Vec
MMC	Maximum model change	WD1	Weighted disagreement 1
MMSL	Multi-model Sentiment Learning Layer	WELM	Weighted extreme learning machine
MS	Margin sampling	WM	Wikipedia Miner
NB	Naive Bayes	WMVC	Weighted multi-view clustering
NBoW	Neural Bag of Words	WSAL	Warm Start Active Learning

Referencias

- [1] G. Macgregor, “Digital repositories and discoverability: definitions and typology,” in *Discoverability in Digital Repositories*, Routledge, 2023, pp. 11–31.
- [2] U. Kuckartz, “Qualitative text analysis: A systematic approach,” *Compendium for early career researchers in mathematics education*, pp. 181–197, 2019.
- [3] G. Ginde *et al.*, “ScientoBASE: a framework and model for computing scholastic indicators of non-local influence of journals via native data acquisition algorithms,” *Scientometrics*, vol. 108, no. 3, pp. 1479–1529, 2016, doi: 10.1007/s11192-016-2006-2.
- [4] E. C. McNie, A. Parris, and D. Sarewitz, “Improving the public value of science: A typology to inform discussion, design and implementation of research,” *Res Policy*, vol. 45, no. 4, pp. 884–895, 2016, doi: 10.1016/j.respol.2016.01.004.
- [5] A. Ibáñez, “Machine Learning in Scientometrics,” Universidad Politécnica de Madrid, 2015. [Online]. Available: <http://oa.upm.es/36488/>
- [6] M. Maridueña, M. Leyva, and A. Febles, “Modelado y análisis de indicadores de ciencia y tecnología mediante mapas cognitivos difusos,” *Ciencias de la información*, vol. 47, no. 1, pp. 17–24, 2016, [Online]. Available: <http://biblat.unam.mx/es/revista/ciencias-de-la-informacion/articulo/modelado-y-analisis-de-indicadores-de-ciencia-y-tecnologia-mediante-mapas-cognitivos-difusos>
- [7] T. Jiang, J. L. Gradus, and A. J. Rosellini, “Supervised machine learning: a brief primer,” *Behav Ther*, vol. 51, no. 5, pp. 675–687, 2020.
- [8] M. T. Almuqati, F. Sidi, S. N. Mohd Rum, M. Zolkepli, and I. Ishak, “Challenges in Supervised and Unsupervised Learning: A Comprehensive Overview.,” *Int J Adv Sci Eng Inf Technol*, vol. 14, no. 4, 2024.
- [9] J. E. Van Engelen and H. H. Hoos, “A survey on semi-supervised learning,” *Mach Learn*, vol. 109, no. 2, pp. 373–440, 2020, doi: 10.1007/s10994-019-05855-6.
- [10] Y. Padmanabha, P. Viswanath, and B. Eswara, “Semi-supervised learning: a brief review,” *International Journal of Engineering & Technology*, vol. 7, no. 1.8, p. 81, 2018, doi: 10.14419/ijet.v7i1.8.9977.
- [11] J.-C. Su, Z. Cheng, and S. Maji, “A realistic evaluation of semi-supervised learning for fine-grained classification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12966–12975.
- [12] S. Shrutika and M. Prabukumar, “Semi-Supervised Techniques Based Hyperspectral Image Classification : A Survey,” *International Conference on Innovations in Power and Advanced Computing Technologies*, pp. 1–8, 2017.
- [13] A. Cevallos-Culqui, C. Pons, and G. Rodr\u00edguez, “A Co-Training Model Based in Learning Transfer for the Classification of Research Papers,” in *2024 IEEE 12th International Conference on Intelligent Systems (IS)*, 2024, pp. 1–6.

-
- [14] B. Altinel and M. C. Ganiz, "A new hybrid semi-supervised algorithm for text classification with class-based semantics," *Knowl Based Syst*, vol. 108, pp. 50–64, 2016, doi: 10.1016/j.knosys.2016.06.021.
- [15] J. Khan and Y. K. Lee, "LeSSA: A unified framework based on lexicons and semi-supervised learning approaches for textual sentiment classification," *Applied Sciences*, vol. 9, no. 24, 2019, doi: 10.3390/app9245562.
- [16] D. Barman and N. Chowdhury, "A novel semi supervised approach for text classification," *International Journal of Information Technology*, 2018, doi: 10.1007/s41870-018-0137-9.
- [17] M. Emadi, J. Tanha, M. E. Shiri, and M. H. Aghdam, "A Selection Metric for semi-supervised learning based on neighborhood construction," *Inf Process Manag*, vol. 58, no. 2, 2021, doi: 10.1016/j.ipm.2020.102444.
- [18] J. Li, Q. Zhu, Q. Wu, and D. Cheng, "An effective framework based on local cores for self-labeled semi-supervised classification," *Knowl Based Syst*, vol. 197, Jun. 2020, doi: 10.1016/j.knosys.2020.105804.
- [19] Y. Yang and M. Loog, "A benchmark and comparison of active learning for logistic regression," *Pattern Recognit*, vol. 83, pp. 401–415, 2018, doi: 10.1016/j.patcog.2018.06.004.
- [20] M. Liu, "Weak Supervision and Active Learning for Natural Language Processing," Monash University, 2019.
- [21] S. Guo and N. Yao, "Document Vector Extension for Documents Classification," *IEEE Trans Knowl Data Eng*, vol. 33, no. 8, pp. 3062–3074, 2021, doi: 10.1109/TKDE.2019.2961343.
- [22] M. Mohammed, L. Yu, A. Aldhubri, and G. R. S. Qaid, "Study on sentiment classification strategies based on the fuzzy logic with crow search algorithm," *Res Sq*, 2022.
- [23] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.
- [24] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, 2019, pp. 15–18.
- [25] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," *arXiv preprint arXiv:2006.05278*, 2020.
- [26] G. Alene, "The Role of Management Information System in Improving Organizational Performance and Effectiveness in Case of Debre Markos City Administration Revenue Authority, Ethiopia," *ICTACT Journal on Management Studies*, vol. 4, no. 1, pp. 691–697, 2018.
- [27] A. Cevallos-Culqui, C. Pons, and G. Rodriguez, "Semi-supervised learning models for document classification: A systematic review and meta-analysis," *Inteligencia Artificial*, vol. 26, no. 72, pp. 81–111, Jun. 2023, doi: 10.4114/intartif.vol26iss72pp81-111.
- [28] N. Chen, B. Ribeiro, C. Tang, and A. Chen, "Multi-label learning vector quantization for semi-supervised classification," *Intelligent Data Analysis*, vol. 23, no. 4, pp. 839–853, 2019, doi: 10.3233/IDA-184195.

-
- [29] L. Borrajo, A. Seara Vieira, and E. L. Iglesias, "An HMM-based synthetic view generator to improve the efficiency of ensemble systems," *Log J IGPL*, vol. 28, no. 1, pp. 4–18, 2020, doi: 10.1093/jigpal/jzz067.
- [30] I. Khong, N. A. Yusuf, A. Nuriman, and A. B. Yadila, "Exploring the impact of data quality on decision-making processes in information intensive organizations," *APTISI Transactions on Management*, vol. 7, no. 3, pp. 253–260, 2023.
- [31] A. Masmoudi, H. Bellaaj, K. Drira, and M. Jmaiel, "A co-training-based approach for the hierarchical multi-label classification of research papers," *Expert Syst*, vol. 38, no. 4, pp. 1–19, 2021, doi: 10.1111/exsy.12613.
- [32] A. Søgaard, *Semi-supervised learning and domain adaptation in natural language processing*. Springer Nature, 2022.
- [33] M. Das, X. Chen, X. Yuan, and L. Zhang, "Federated semi-supervised domain adaptation via knowledge transfer," *arXiv preprint arXiv:2207.10727*, 2022.
- [34] M. Alemi, A. Bosaghzadeh, and F. Dornaika, "Graph-Based Semi-Supervised Learning with Bipartite Graph for Large-Scale Data and Prediction of Unseen Data," 2024.
- [35] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," *arXiv preprint arXiv:2101.06329*, 2021.
- [36] R. Liu *et al.*, "Recent Advances in Hierarchical Multi-label Text Classification: A Survey," *arXiv preprint arXiv:2307.16265*, 2023.
- [37] T. Fredriksson, J. Bosch, H. H. Olsson, and D. I. Mattos, "Machine learning algorithms for labeling: Where and how they are used?," in *2022 IEEE International Systems Conference (SysCon)*, 2022, pp. 1–8.
- [38] A. Mahadevan and M. Mathioudakis, "Cost-Effective Retraining of Machine Learning Models," *arXiv preprint arXiv:2310.04216*, 2023.
- [39] M. H. Tanveer, Z. Fatima, S. Zardari, and D. Guerra-Zubiaga, "An In-Depth Analysis of Domain Adaptation in Computer and Robotic Vision," *Applied Sciences*, vol. 13, no. 23, p. 12823, 2023.
- [40] J. Serrano-Pérez and L. E. Sucar, "Semi-Supervised Hierarchical Multi-Label Classifier Based on Local Information," *arXiv preprint arXiv:2405.00184*, 2024.
- [41] J. M. Duarte and L. Berton, "A review of semi-supervised learning for text classification," *Artif Intell Rev*, vol. 56, no. 9, pp. 9401–9469, 2023, doi: 10.1007/s10462-023-10393-8.
- [42] D. Ahfock and G. J. McLachlan, "Semi-supervised learning of classifiers from a statistical perspective: A brief review," *Econom Stat*, vol. 26, pp. 124–138, 2023.
- [43] Y. Fan, A. Kukleva, D. Dai, and B. Schiele, "Revisiting consistency regularization for semi-supervised learning," *Int J Comput Vis*, vol. 131, no. 3, pp. 626–643, 2023.
- [44] A. Murtadha *et al.*, "Rank-Aware Negative Training for Semi-Supervised Text Classification," 2023. [Online]. Available: <https://arxiv.org/abs/2306.07621>
- [45] D. Garigliotti, "Semi-supervised learning for word sense disambiguation," *arXiv preprint arXiv:1908.09641*, 2019.

-
- [46] H. Zou and C. Caragea, “JointMatch: A Unified Approach for Diverse and Collaborative Pseudo-Labeling to Semi-Supervised Text Classification,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 7290–7301. doi: 10.18653/v1/2023.emnlp-main.451.
- [47] T. Kim, I. Hwang, G.-C. Kang, W.-S. Choi, H. Kim, and B.-T. Zhang, “Label propagation adaptive resonance theory for semi-supervised continuous learning,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4012–4016.
- [48] J. Haddock *et al.*, “Semi-supervised Nonnegative Matrix Factorization for Document Classification,” 2022. [Online]. Available: <https://arxiv.org/abs/2203.03551>
- [49] A. Shukla, G. S. Cheema, and S. Anand, “Semi-supervised clustering with neural networks,” in *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, 2020, pp. 152–161.
- [50] H. Chen, W. Han, and S. Poria, “SAT: Improving Semi-Supervised Text Classification with Simple Instance-Adaptive Self-Training,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 6141–6146. doi: 10.18653/v1/2022.findings-emnlp.456.
- [51] A. Kumagai, T. Iwata, and Y. Fujiwara, “Semi-supervised anomaly detection on attributed graphs,” in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8.
- [52] H.-M. Xu, L. Liu, and E. Abbasnejad, “Progressive class semantic matching for semi-supervised text classification,” *arXiv preprint arXiv:2205.10189*, 2022.
- [53] H. Xu, L. Liu, and E. Abbasnejad, “Progressive Class Semantic Matching for Semi-supervised Text Classification,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 3003–3013. doi: 10.18653/v1/2022.naacl-main.219.
- [54] P. Karisani and N. Karisani, “Semi-supervised text classification via self-pretraining,” in *Proceedings of the 14th ACM international conference on web search and data mining*, 2021, pp. 40–48.
- [55] C. Caragea, F. Bulgarov, and R. Mihalcea, “Co-training for topic classification of scholarly data,” in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2357–2366.
- [56] M. Hasnain, I. Ghani, S. R. Jeong, and A. Ali, “Ensemble Learning Models for Classification and Selection of Web Services: A Review.,” *Computer Systems Science & Engineering*, vol. 40, no. 1, 2022.
- [57] B. Miller, F. Linder, and W. R. Mebane Jr, “Active learning approaches for labeling text: review and assessment of the performance of active learning approaches,” *Political Analysis*, vol. 28, no. 4, pp. 532–551, 2020.

-
- [58] M. Li, H. Zhou, J. Hou, P. Wang, and E. Gao, "Is cross-linguistic advert flaw detection in Wikipedia feasible? A multilingual-BERT-based transfer learning approach," *Knowl Based Syst*, vol. 252, p. 109330, 2022.
- [59] I. Ameer, N. Bölücü, M. H. F. Siddiqui, B. Can, G. Sidorov, and A. Gelbukh, "Multi-label emotion classification in texts using transfer learning," *Expert Syst Appl*, vol. 213, p. 118534, 2023, doi: <https://doi.org/10.1016/j.eswa.2022.118534>.
- [60] M. Sao and H.-J. Lim, "MIroBERTa: Mental Illness Text Classification With Transfer Learning on Subreddits," *IEEE Access*, vol. 12, pp. 197454–197466, 2024, doi: [10.1109/ACCESS.2024.3522465](https://doi.org/10.1109/ACCESS.2024.3522465).
- [61] K. Watanabe and Y. Zhou, "Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches," *Soc Sci Comput Rev*, pp. 1–21, 2020, doi: [10.1177/0894439320907027](https://doi.org/10.1177/0894439320907027).
- [62] H. Shinnou, K. Komiya, and M. Sasaki, "Domain Adaptation for Document Classification by Alternately Using Semi-supervised Learning and Feature Weighted Learning," pp. 1–23, 2018, doi: https://doi.org/10.1007/978-981-10-8438-6_17.
- [63] J. Jedrzejowicz and M. Zakrzewska, *Text classification using LDA-W2V hybrid algorithm*, vol. 142. Springer Singapore, 2019. doi: [10.1007/978-981-13-8311-3_20](https://doi.org/10.1007/978-981-13-8311-3_20).
- [64] B. Poojitha, "Machine learning for text categorization: Experiments using clustering and classification," Kansas University, 2018.
- [65] S. Zhao and J. Li, "A semi-supervised self-training method based on density peaks and natural neighbors," *J Ambient Intell Humaniz Comput*, vol. 12, no. 2, pp. 2939–2953, Feb. 2021, doi: [10.1007/s12652-020-02451-8](https://doi.org/10.1007/s12652-020-02451-8).
- [66] K. M. O. Vale, A. C. Gorgonio, F. Da Luz E Gorgonio, and A. M. De Paula Canuto, "An Efficient Approach to Select Instances in Self-Training and Co-Training Semi-Supervised Methods," *IEEE Access*, vol. 10, pp. 7254–7276, 2022, doi: [10.1109/ACCESS.2021.3138682](https://doi.org/10.1109/ACCESS.2021.3138682).
- [67] C. Zhu and D. Miao, *Semi-supervised One-Pass Multi-view Learning with Variable Features and Views*, vol. 50, no. 1. Springer US, 2019. doi: [10.1007/s11063-019-10037-5](https://doi.org/10.1007/s11063-019-10037-5).
- [68] C. Zhu, Z. Wang, R. Zhou, L. Wei, X. Zhang, and Y. Ding, "Semi-supervised one-pass multi-view learning," *Neural Comput Appl*, vol. 31, no. 11, pp. 8117–8134, 2019, doi: [10.1007/s00521-018-3654-3](https://doi.org/10.1007/s00521-018-3654-3).
- [69] X. Jia *et al.*, "Semi-supervised multi-view deep discriminant representation learning," *IEEE Trans Pattern Anal Mach Intell*, vol. 43, no. 7, pp. 2496–2509, 2021, doi: [10.1109/TPAMI.2020.2973634](https://doi.org/10.1109/TPAMI.2020.2973634).
- [70] G. Nayak, R. Ghosh, X. Jia, V. Mithal, and V. Kumar, "Semi-supervised Classification using Attention-based Regularization on Coarse-resolution Data," *Proceedings of the 2020 SIAM International Conference on Data Mining, SDM 2020*, pp. 253–261, 2020, doi: [10.1137/1.9781611976236.29](https://doi.org/10.1137/1.9781611976236.29).
- [71] D. Kim, D. Seo, S. Cho, and P. Kang, "Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec," *Inf Sci (N Y)*, vol. 477, pp. 15–29, 2019, doi: [10.1016/j.ins.2018.10.006](https://doi.org/10.1016/j.ins.2018.10.006).

-
- [72] O. Edo-Osagie, G. Smith, I. Lake, O. Edeghere, and B. De La Iglesia, "Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance," *PLoS One*, vol. 14, no. 7, pp. 1–29, 2019, doi: 10.1371/journal.pone.0210689.
- [73] Z. Donyavi and S. Asadi, "Diverse training dataset generation based on a multi-objective optimization for semi-Supervised classification," *Pattern Recognit*, vol. 108, Dec. 2020, doi: 10.1016/j.patcog.2020.107543.
- [74] W. Jia, X. Liu, Y. Wang, W. Pedrycz, and J. Zhou, "Semisupervised Learning via Axiomatic Fuzzy Set Theory and SVM," *IEEE Trans Cybern*, vol. 52, no. 6, pp. 4661–4674, Jun. 2022, doi: 10.1109/TCYB.2020.3032707.
- [75] E. De Souza, "Intelligent Document Validation Intelligent Document Validation using Natural Language Processing and Computer Vision," Coimbra, 2021.
- [76] M. Mouriño García, R. Pérez Rodríguez, L. Anido Rifón, and M. Vilares Ferro, "Wikipedia-based hybrid document representation for textual news classification," *Soft comput*, vol. 22, no. 18, pp. 6047–6065, 2018, doi: 10.1007/s00500-018-3101-5.
- [77] M. A. Mouriño García, R. Pérez Rodríguez, and L. E. Anido Rifón, "A bag of concepts approach for biomedical document classification using Wikipedia knowledge: Spanish-English cross-language case study," *Methods Inf Med*, vol. 56, no. 5, pp. 370–376, 2017, doi: 10.3414/ME17-01-0028.
- [78] M. Mouriño García, R. Pérez Rodríguez, and L. Anido Rifón, "Leveraging Wikipedia knowledge to classify multilingual biomedical documents," *Artif Intell Med*, vol. 88, pp. 37–57, 2018, doi: 10.1016/j.artmed.2018.04.007.
- [79] H. M. Salman, "Text Classification Based on Weighted Extreme Learning Machine," *Ibn AL- Haitham Journal For Pure and Applied Science*, vol. 32, no. 1, p. 203, 2019, doi: 10.30526/32.1.1978.
- [80] A. K. Shrivastava, A. K. Dewangan, S. M. Ghosh, and D. Singh, "Development of proposed ensemble model for spam e-mail classification," *Information Technology and Control*, vol. 50, no. 3, pp. 411–423, 2021, doi: 10.5755/j01.itc.50.3.27349.
- [81] S. Ghosh and A. Chopra, "Using Transformer Based Ensemble Learning to Classify Scientific Articles," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12705 LNAI, pp. 106–113, 2021, doi: 10.1007/978-3-030-75015-2_11.
- [82] S. de Vries and D. Thierens, "A reliable ensemble based approach to semi-supervised learning," *Knowl Based Syst*, vol. 215, Mar. 2021, doi: 10.1016/j.knosys.2021.106738.
- [83] Y. Han, Y. Liu, and Z. Jin, "Sentiment analysis via semi-supervised learning: a model based on dynamic threshold and multi-classifiers," *Neural Comput Appl*, vol. 32, no. 9, pp. 5117–5129, May 2020, doi: 10.1007/s00521-018-3958-3.
- [84] M. R. Bouguelia, S. Nowaczyk, K. C. Santosh, and A. Verikas, "Agreeing to disagree: active learning with noisy labels without crowdsourcing," *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 8, pp. 1307–1319, 2018, doi: 10.1007/s13042-017-0645-0.

-
- [85] O. Reyes, A. H. Altalhi, and S. Ventura, "Statistical comparisons of active learning strategies over multiple datasets," *Knowl Based Syst*, vol. 145, pp. 274–288, 2018, doi: 10.1016/j.knosys.2018.01.033.
- [86] S. Yang, R. Wei, J. Guo, and H. Tan, "Chinese semantic document classification based on strategies of semantic similarity computation and correlation analysis," *Journal of Web Semantics*, vol. 63, p. 100578, 2020, doi: 10.1016/j.websem.2020.100578.
- [87] W. Fu, B. Xue, X. Gao, and M. Zhang, "Genetic Programming based Transfer Learning for Document Classification with Self-taught and Ensemble Learning," *2019 IEEE Congress on Evolutionary Computation, CEC 2019 - Proceedings*, pp. 2260–2267, 2019, doi: 10.1109/CEC.2019.8790318.
- [88] Y. Zhu, E. Shareghi, Y. Li, R. Reichart, and A. Korhonen, "Combining deep generative models and multi-lingual pretraining for semi-supervised document classification," *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, no. 1, pp. 894–908, 2021, doi: 10.18653/v1/2021.eacl-main.76.
- [89] Z. Pan, P. Soong, S. Rafatirad, Z. Pan, P. Soong, and S. Rafatirad, "Ontology-Driven Scientific Literature Classification using Clustering and Self-Supervised Learning Ontology-Driven Scientific Literature Classification using Clustering and Self-Supervised Learning," *Easy Chair*, 2022.
- [90] Z. Yang, "Incorporating Structural Bias into Neural Networks for Natural Language Processing," Carnegie Mellon University, 2019. [Online]. Available: <http://www.cs.cmu.edu/~zichaoy/proposal.pdf>
- [91] S. Alahdal, "Diary mining: predicting emotion from activities, people and places," Cardiff University, 2020.
- [92] D. Wang *et al.*, "Cross-Lingual Knowledge Transferring by Structural Correspondence and Space Transfer," *IEEE Trans Cybern*, vol. 52, no. 7, pp. 6555–6566, Jul. 2022, doi: 10.1109/TCYB.2021.3051005.
- [93] F. H. Khan, U. Qamar, and S. Bashir, "Enhanced cross-domain sentiment classification utilizing a multi-source transfer learning approach," Jul. 01, 2019, *Springer Verlag*. doi: 10.1007/s00500-018-3187-9.
- [94] X. Du, Z. Zhou, B. Yin, and G. Xiao, "Cross-project bug type prediction based on transfer learning," *Software Quality Journal*, vol. 28, no. 1, pp. 39–57, Mar. 2020, doi: 10.1007/s11219-019-09467-0.
- [95] Z. Okray *et al.*, "Multisensory learning binds neurons into a cross-modal memory engram," *Nature*, vol. 617, no. 7962, pp. 777–784, 2023, doi: 10.1038/s41586-023-06013-8.
- [96] X. Chen *et al.*, "PaLI: A Jointly-Scaled Multilingual Language-Image Model," 2023. [Online]. Available: <https://arxiv.org/abs/2209.06794>
- [97] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, "Transfer learning: a friendly introduction," *J Big Data*, vol. 9, no. 1, p. 102, 2022, doi: 10.1186/s40537-022-00652-w.
- [98] A. Q. Wynne Harlen OBE, *The Teaching of Science in Primary Schools*, Sixth. England: Routledge, 2014.

-
- [99] N. Pulido and A. Mata, *El documento digital: aspectos para garantizar su integridad por la ciudadanía*, Astro data. Potosí, México, 2022. [Online]. Available: <https://www.researchgate.net/publication/365839063>
- [100] A. Bhavani and B. Santhosh, "A Review of State Art of Text Classification Algorithms," *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, no. Iccmc, pp. 1484–1490, 2021, doi: 10.1109/ICCMC51019.2021.9418262.
- [101] S. Franko, "Multiclass analysis of automatic text classification techniques," Galatasaray University, 2018.
- [102] V. Prabhu, S. Khare, D. Kartik, and J. Hoffman, "SENTRY: Selective Entropy Optimization via Committee Consistency for Unsupervised Domain Adaptation," *Computer Vision Foundation*, 2021, [Online]. Available: <https://github.com/virajprabhu/SENTRY>.
- [103] M. Mirończuk and J. Protasiewicz, "A recent overview of the state-of-the-art elements of text classification," *Expert Syst Appl*, vol. 106, pp. 36–54, 2018, doi: 10.1016/j.eswa.2018.03.058.
- [104] A. Mustar, S. Lamprier, B. Piwowarski, and B. Piwowarski Using BERT, "Using BERT and BART for Query Suggestion," Samatan, París, 2020. [Online]. Available: <https://hal.sorbonne-universite.fr/hal-02989015>
- [105] W. Alhoshan, A. Ferrari, and L. Zhao, "Zero-shot learning for requirements classification: An exploratory study," *Inf Softw Technol*, vol. 159, p. 107202, Jul. 2023, doi: 10.1016/j.infsof.2023.107202.