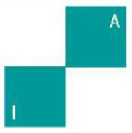
Inteligencia Artificial, 26(72) (2023), 81-111
doi: 10.4114/intartif.vol26iss72pp81-111



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Semi-supervised learning models for document classification: A systematic review and meta-analysis

Alex Cevallos-Culqui[1, A], Claudia Pons[1, 2, 3, B], Gustavo Rodríguez[4, C]

[1]LIFIA - Facultad de Informática. Universidad Nacional de La Plata, La Plata, Argentina.
[2]CAETI - Facultad de Tecnología Informática. UAI., Buenos Aires, Argentina.
[3]CIC - Comisión de Investigaciones Científicas de Buenos Aires, Buenos Aires, Argentina.
[4]Departamento de Tecnologías de Información y Comunicación, UTC, Latacunga, Ecuador.

[A]alex.cevallosc@info.unlp.edu.ar
[B]claudia.pons@uai.edu.ar
[C]gustavo.rodriguez@utc.edu.ec

Abstract The proliferation of digital documents in the internet has given rise to the search for information patterns that allow for the categorization of organizational documents to generate knowledge in a institution. One of the Artificial Intelligence techniques for this purpose is text classification, which for its application uses labels (categorized documents) with supervised (with labels) or unsupervised (without labels) training models. Both traditional models with their advantages and disadvantages have been consolidated into semi-supervised models to extract the best qualities of each one, however, the labeling process involves resources that need to be optimized to improve the classification accuracy. An analysis of the types of semi-supervised models would show the strengths of their training and how the structure of each of them affects the accuracy of their classification. The present study proposes a structure of semi-supervised model in document classification, in order to analyze the qualities of each one in their categorization process, it through a systematic literature review (SLR) that analyzes the performance of the studies to conduct a meta-analysis. Further, the study search strategy was defined by the PICOC method (Population, Intervention, Comparison, Outcome, Context), supported by two research questions and delimited in a search chain that allowed the collection of 332 research studies. These papers were filtered using the PRISMA method and the determination of exclusion criteria, in total 46 papers have been selected for the present study.From this SLR, an organizational structure has been obtained for semi-supervised models and a scheme for the classification process. In addition, the advantages and disadvantages of different learning types have been analyzed, evaluating their classification performance in each type of learning through a meta-analysis. This has determined that the models that present the best levels of performance are active learning model (0.88) and ensemble learning model (0.84).

Resumen La proliferación de documentos digitales en la red ha dado lugar a la búsqueda de patrones de información que permitan la categorización de documentos organizacionales para generar conocimiento en una determinada institución. Una de las técnicas de la Inteligencia Artificial para este efecto es la clasificación de texto, la cual para su aplicación emplea etiquetas (documentos categorizados) con modelos de entrenamiento supervisados (con etiquetas) o no-supervisados (sin etiquetas). Ambos modelos tradicionales con sus ventajas y desventajas, se han visto cohesionados en los modelos semi-supervisados que extraen las mejores cualidades de cada uno, sin embargo, el proceso de etiquetado implica recursos que buscan ser optimizados para mejorar la precisión de clasificación. Un análisis de los tipos de modelos semi-supervisados mostraría las fortalezas de su entrenamiento y la forma en que la estructura de cada uno de ellos incide en la precisión de su clasificación.

ISSN: 1137-3601 (print), 1988-3064 (on-line)

En el presente estudio se propone una estructura de los tipos de modelos semi-supervisados en la clasificación de documentos, para de esta manera analizar las cualidades de cada uno de ellos en su proceso de categorización, esto a través de una SLR (Revisión de literatura sistemática) que analiza el rendimiento de los estudios para efectuar un meta-análisis. La estrategia de búsqueda de estudios ha sido definida con el método PICOC (Población, Intervención, Comparación, Salidas, Contexto), el cual, apoyado en dos preguntas de investigación, define una cadena de búsqueda que ha permitido recopilar 332 investigaciones, filtradas con el método de la declaración PRISMA y la determinación de criterios de exclusión, seleccionando así 46 investigaciones para el estudio. De la SLR se ha obtenido una estructura de organización para los modelos semi-supervisados y un esquema del proceso de clasificación. También, se ha analizado las ventajas y desventajas de los diferentes tipos de aprendizaje, evaluando su desempeño de clasificación en cada tipo de aprendizaje a través de un meta-análisis. Se determina que los modelos que presentan los mejores niveles de rendimiento son el aprendizaje activo (0.88) y ensamblado (0.84).

Keywords: Text classification, Document classification, Semi-supervised models, systematic review.
Palabras Clave: Clasificación de texto, clasificación documentos, modelos semi-supervisado, revisión sistemática.

1. Introducción

El entorno digital día tras día acumula una gran cantidad de información en forma de documentos, la acentuada necesidad del intercambio de datos digitalizados a nivel corporativo y público ha dado lugar al almacenamiento de datos estructurados, no estructurados y semiestructurados. El 80 % de esta información recopilada corresponde a datos no estructurados y pertenece a una determinada entidad ya sea ésta una persona, lugar, animal o cosa [1]. En este tipo de estructuras existen patrones de información ocultos que poseen valioso conocimiento para la institución y su toma de decisiones [2]. Esto ha motivado a que la investigación científica por medio del análisis de texto de documentos, busque las mejores técnicas para identificar estas relaciones existentes entre conjuntos de datos y entidades, para así establecer una adecuada representación del conocimiento [3].

El nivel de precisión de la clasificación de documentos a través del procesamiento de texto, depende en gran medida de su forma de entrenamiento y estructura [1]. Un adecuado entrenamiento se lleva a cabo con la generación de aprendizaje a través de documentos etiquetados, pero no siempre se dispone de documentos ya categorizados, disponerlos se ha convertido en una tarea que requiere tiempo y expertos en el dominio, lo cual implica recursos y costos [4]. Es así como el aprendizaje semi-supervisado toma protagonismo en la clasificación de documentos por su flexibilidad al momento de incrementar el conjunto de documentos categorizados, esto se lo realiza con el auto-etiquetado de nuevos documentos que son sometidos a un entrenamiento a partir de una muestra representativa ya clasificada [5].

Sin embargo, el auto-etiquetado es un proceso propenso a errores, principalmente cuando el conjunto de elementos etiquetados inicial es escaso [6]. A menor etiquetas el entrenamiento del modelo puede adquirir mayor entropía; mientras que si el conjunto incrementa su etiquetado correctamente la entropía del modelo es menor, mejorando la clasificación de los documentos [4]. En la búsqueda de eficiencia de este proceso se plantean diferentes modelos de aprendizaje semi-supervisado con distintas técnicas en cada una de sus etapas, que han sido diseñados para solventar necesidades en diferentes dominios [7].

En varios estudios de clasificación semi-supervisada en distintos dominios concluyen que es necesario determinar bajo que condiciones el entrenamiento de datos no etiquetados puede mejorar la precisión de la clasificación [8][4] [9] [10]. Es por esta razón, que en el presente estudio se busca analizar aquellos modelos semi-supervisados utilizados en el entorno de la clasificación de documentos, considerando su nivel de rendimiento, ventajas y desventajas en sus etapas de procesamiento de texto.

Los procesos de clasificación semi-supervisados son una temática de investigación de interés en revistas de inteligencia artificial y minería de texto, así es como se han propuesto varios modelos semi-supervisados de clasificación de documentos, sin embargo, escazas son las revisiones sistemáticas existentes que hayan recopilado y analizado esta variedad. Entre estas limitadas revisiones tenemos el caso de [9] que proporciona una conceptualización rápida de los tipos de aprendizaje, definiendo la clasificación semi-supervisada y la agrupación(clustering) semi-supervisada. Por otro lado, en [10] se hace una revisión de los modelos de clasificación semi-supervisada en relación al análisis de imágenes. En el estudio [7] se hace una revisión en referencia a una recolección de los últimos métodos de aprendizaje semi-supervisados. Sin embargo, estas revisiones no focalizan el análisis en la clasificación de documentos y tampoco consideran en su

investigación una comparativa del desempeño de los diferentes modelos.

La presente revisión de literatura busca expandir el conocimiento de las revisiones existentes brindando un análisis y comparación de modelos semi-supervisados focalizados en la tarea de clasificación de documentos. Las contribuciones de la presente revisión radican en: Incluir un significativo número de investigaciones (332) que han sido analizados para la interpretación de la temática; Analizar investigaciones más actualizadas incluyendo estudios entre 2017 y 2022; Elaborar un meta-análisis y revisión sistemática en base a un protocolo de revisión establecido; Y, categorizar las investigaciones identificadas acorde al tipo de modelo de aprendizaje semi-supervisado.

La revisión de literatura sistemática (SLR) y de meta-análisis es elaborada considerando las recomendaciones de los siguientes estudios: la guía para la realización de literatura de revisiones sistemáticas en ingeniería de software [11] y las preferencias para la presentación de informes en revisiones sistemáticas y meta-análisis de la declaración PRISMA [12]. En la presente SLR se busca estudios primarios utilizando una estrategia de búsqueda para que responda las siguientes preguntas de investigación: (P1) ¿Cuáles son los diferentes modelos de aprendizaje semi-supervisados utilizados para clasificar documentos?; (P2) ¿Cuáles son las principales ventajas y desventajas que abordan los modelos de aprendizaje semi-supervisados en la clasificación de documentos?

Considerando las preguntas planteadas, se estructura una cadena de búsqueda que ha permitido la extracción de información y el análisis de cada punto, para esto se utiliza el método de PICOC (population, intervention, comparison, outcome, context), criterio planteado por [13], con este método se ha estructurado la siguiente cadena de búsqueda: Población (documents classification); Intervención AND (semi supervised learning); Comparación AND (self-training OR co-training OR ensembled OR (active learning) OR (transfer learning)); Salida AND (accuracy OR comparison).

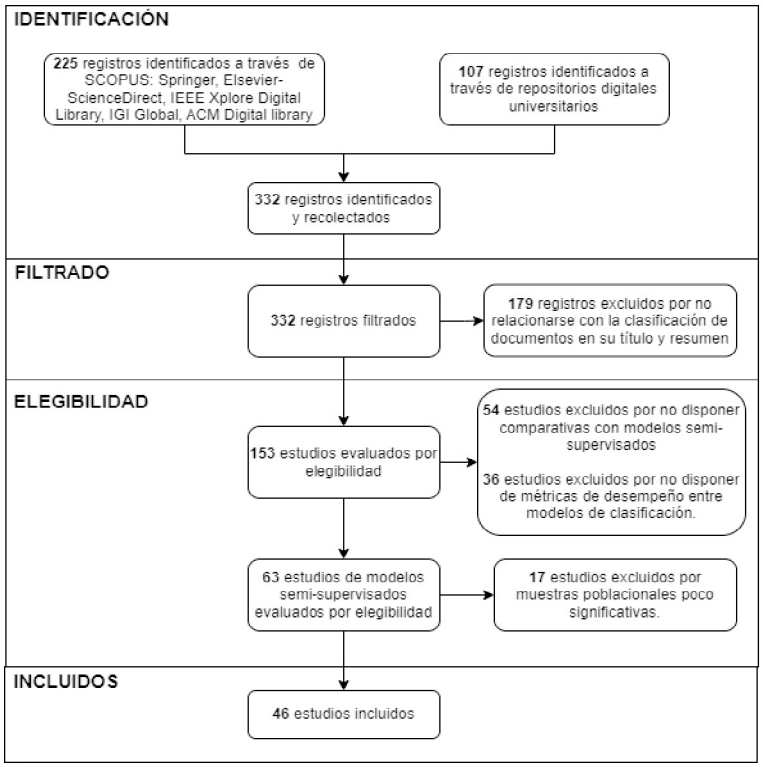


Figura 1: Diagrama de flujo PRISMA

Con la cadena de búsqueda definida, ésta ha sido aplicada en la base de datos bibliográfica SCOPUS, así es como se ha recopilado investigaciones para la SLR de los siguientes repositorios: (1) IEEE Xplore Digital Library, (2) Springer, (3) Elsevier, (4) IGI Global y (5) ACM Digital Library. Como resultado se ha obtenido 332 artículos publicados en revistas, congresos y repositorios, de los cuales luego de un

proceso de filtrado por criterios de inclusión y exclusión (ver Figura 1), se han elegido 153 para ser analizados por contenidos y finalmente incluir 46 investigaciones en el estudio del presente SLR.

Los estudios incluidos fueron agrupados de acuerdo al tipo de aprendizaje semi-supervisado del modelo con el fin de realizar comparaciones entre las diferentes técnicas. Los análisis estadísticos han sido conducidos utilizando la herramienta estadística de software R-Studio, empleando la gráfica de meta-análisis de forestplot, para analizar e interpretar resultados. El desempeño de clasificación de documentos de cada estudio incluido ha sido evaluado considerando la precisión del modelo, esta precisión es obtenida relacionando los verdaderos positivos (TP) y los falsos positivos (FP) de la siguiente manera: TP/(TP+FP).

La estructura del estudio tiene la siguiente organización: En la sección 2, se responde a la pregunta de investigación P1, se identifican los estudios y tipos de modelos semi-supervisados utilizados para la clasificación de documentos, se ha analizado cada estudio determinando sus fortalezas, debilidades y niveles de precisión de clasificación para compararlos. En la sección 3, se da respuesta a la pregunta de investigación P2, se describe las ventajas y desventajas identificadas en el contexto general del proceso de clasificación. Finalmente, en la sección 4 se emiten las conclusiones y trabajo futuro del presente estudio.

2. Modelos semi-supervisados para clasificación de documentos

En respuesta a la pregunta de investigación P1. En la presente SLR se han encontrado diferentes modelos de clasificación de documentos semi-supervisados, considerando la diversidad de estrategias de entrenamiento o etiquetado de los modelos semi-supervisados se plantea una estructura de agrupación para el conjunto de modelos. Varios tipos de entrenamiento en clasificación de documentos semi-supervisados han sido identificados en estudios como [1] [2], en cada investigación cambia la forma de clasificar los modelos, sin embargo considerando el análisis de los estudios seleccionados en torno a clasificación de documentos se expresa la distribución con una estructura de agrupación que ayuda a organizar y comparar los modelos recopilados en cinco categorías: auto-entrenamiento, co-entrenamiento, ensamblado, aprendizaje activo y de transferencia (ver Figura 2).

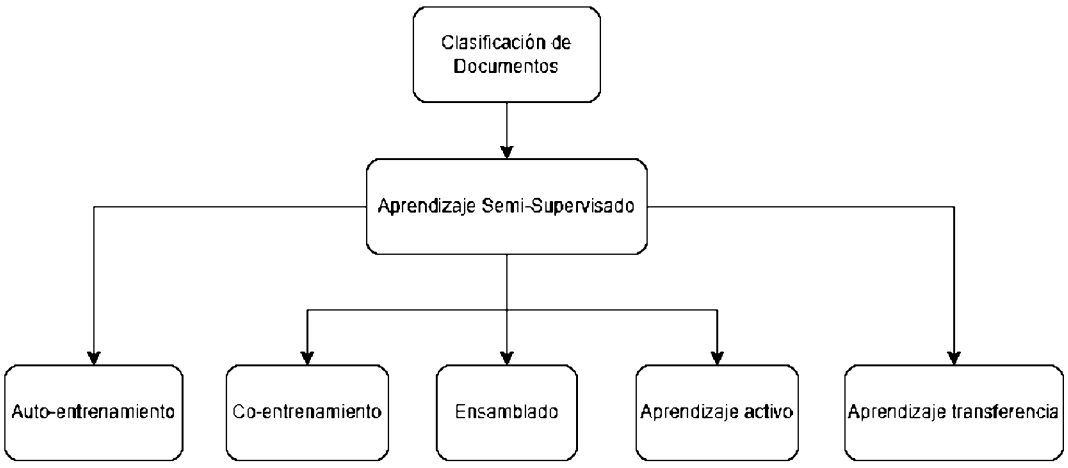


Figura 2: Tipos de entrenamiento en modelos semi-supervisados

En la Figura 3 de acuerdo a los cinco tipos de entrenamiento definidos se presenta un resumen de la cantidad de estudios recopilados por año en el periodo de tiempo establecido para esta SLR. Considerando que el modelo de auto-entrenamiento es la base del resto de modelos, se puede apreciar que sus estudios han sido relativamente constantes en el rango de tiempo; el aprendizaje activo y de co-entrenamiento inician con un elevado interés de investigación que se va reduciendo en el transcurso del tiempo; de forma opuesta el aprendizaje por transferencia y ensamblado acentúan su análisis y estudio en los últimos años.

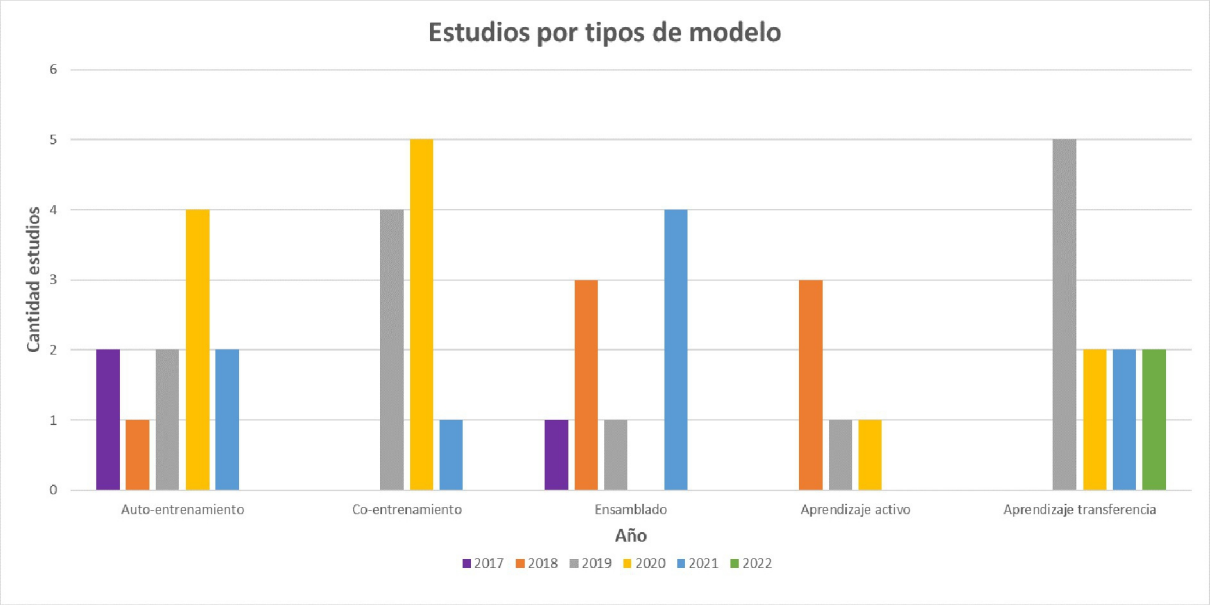


Figura 3: Estudios distribuidos por año y tipo de modelo

Así también, en la Tabla 1 se presenta agrupaciones de los estudios incluidos, en los que se considera el dominio de los conjuntos de datos que se han sometido a experimentación en cada estudio, así se determina la frecuencia de los dominios más utilizados en la presente SLR. Por otro lado, se puede apreciar por cada modelo semi-supervisado los conjuntos de datos y los dominios que han sido utilizados en sus experimentaciones, así se ha podido conocer en que dominio ha sido más usado determinado modelo.

Tabla 1: Distribución de los dominios por dominio y tipo de aprendizaje semi-supervisado

Dominio	Auto-entrenamiento	Co-entrenamiento	Ensamblado	Aprendizaje activo	Aprendizaje transferencia
Conjunto de datos UCI(18)	9	8	1	4	3
Documentos institucionales(6)	1	1	3	0	2
Documentos médicos(3)	0	0	1	1	1
Redes sociales(2)	0	1	0	0	1
Reseñas de productos(3)	1	0	1	0	4
Wiki(2)	0	0	2	0	0
Spam(1)	0	0	1	0	0
TOTAL	11	10	9	5	11

Con estos datos se determina que para la clasificación de documentos con modelos semi-supervisados, los conjuntos de datos más utilizados son datos recopilados del repositorio de machine learning UCI, en segunda instancia están los documentos institucionales pertenecientes a industrias privadas y entidades públicas, finalmente el dominio menos utilizado es el de clasificación de spam en mensajería. Por otro lado, los modelos de aprendizaje más utilizados son el de co-entrenamiento y transferencia probablemente por un aprendizaje que puede implicar dos o más clasificadores que ayudan a mejorar la precisión, a diferencia

de los modelos de auto-entrenamiento y aprendizaje activo que están siendo olvidados por su sencillez y complejidad respectivamente.

Proceso de clasificación

Los distintos modelos semi-supervisados poseen un conjunto de etapas con algunas actividades para el efecto de su proceso de clasificación de documentos. Se identifica que estos tipos de modelos tienen una estructura base en común, con etapas semejantes en su proceso. En la Figura 4 se presenta el esquema de trabajo de las etapas del proceso de un modelo semi-supervisado.

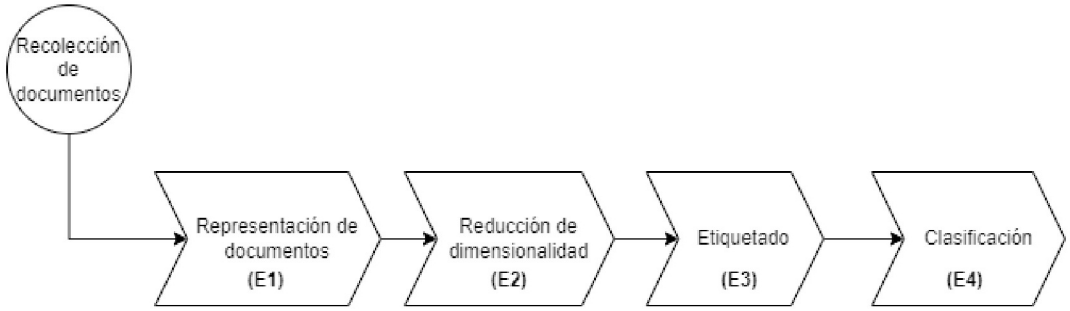


Figura 4: Proceso de clasificación de un modelo semi-supervisado. Adaptado de [7]

La estructura general de un proceso de clasificación de documentos por aprendizaje semi-supervisado posee los elementos presentados en la Figura 4. El proceso da inicio con la recolección de datos, la identificación de fuentes de información acorde al dominio en estudio, puede darse en repositorios, redes sociales, bases de datos corporativas, internet y otros. Focalizado el conjunto de documentos se analiza cuál es el porcentaje de datos etiquetados y no etiquetados para pre-procesarlos de manera segmentada.

E1: Se efectúa como parte del pre-procesamiento y consiste en analizar el corpus del texto de cada documento para seleccionar una representación significativa de cada una de sus palabras o características.

E2: En la mayoría de los casos estos conjuntos de características son de gran cantidad de variables, a los cuales se aplica técnicas de reducción de dimensionalidad para disponer únicamente de las características más significativas del documento con una mejor representación de datos y baja redundancia.

E3: Diferentes técnicas de aprendizaje semi-supervisado pueden ser utilizadas para el entrenamiento del modelo de clasificación con los datos etiquetados existentes, así el clasificador predice una seudo-etiqueta, categoría o clase para cada nuevo documento.

E4: Ya con suficientes datos etiquetados y seudo-etiquetados, el desempeño del clasificador puede ser evaluado en medidas como: índices, recall, f-measure, accurancy [5].

En los distintos modelos semi-supervisados de esta SLR, se identifican varias técnicas para cada etapa del proceso de clasificación, cada técnica posee características propias en determinadas circunstancias más eficientes o deficientes de acuerdo a los tipos de documentos a procesar. A continuación, se describe las técnicas más representativas del entorno de cada etapa:

Representación de documentos (E1).- En esta etapa se identifican técnicas que construyen características (una o más palabras) para determinar su ocurrencia en el corpus del documento acorde a una clase o categoría, en el caso de BoW [8] la característica corresponde a una palabra representada por un valor o peso; mientras que en N-gram la característica corresponde a dos o más palabras en un valor; W2V [14] que es un método de incrustación de palabras recopila conjuntos con dos o más características pero representadas en un vector de valores de N-dimensiones que pueden ser comparados con palabras de similar significado y medir su distancia. También se identifica que, para este proceso de agrupación de palabras del corpus, se utiliza la técnica de PoS [15] que, a más de identificar las palabras más relevantes, anexa una etiqueta diferenciando a la palabra de un sustantivo, pronombre, adjetivo, verbo, adverbio o preposición. En [16] se utiliza SOM una técnica no supervisada para la representación de documentos a través de la agrupación de las características.

Reducción de dimensionalidad (E2).- Su implementación es necesaria ya que muchas de las veces una representación de datos implica una alta dimensionalidad que dificulta el procesamiento de identificación de contenidos, lo que puede dar lugar a un modelo sobre ajustado. Se identifican varias técnicas

de reducción de dimensionalidad, en las investigaciones [4], [14] y [17] utilizan LDA que es un método de búsqueda de características que maximiza la variabilidad de dos o más clases de objetos. Por otro lado, se identifica la propuesta [8] llevada a cabo con la técnica de IG, la cual permite determinar el nivel de relevancia de cada una de las características en cada clase, así se establece un ranking de características que permite determinar la probabilidad de predicción que tiene cada una de ellas, priorizando el uso de las de mayor rango. Así también, es el caso de TF-IDF, la cuál es una de las técnicas más populares para asignar pesos a las características a través de la frecuencia de las palabras en el documento y por medio de los pesos reducir aquellas palabras comunes en el corpus [15].

Etiquetado (E3).- En la mayoría de investigaciones se dispone de diferentes porcentajes mínimos de documentos ya etiquetados que son la base del auto-entrenamiento para que el resto de conjunto de datos no etiquetados adquieran una seudo-etiqueta y se fortalezca el entrenamiento del modelo. Cada estudio posee un porcentaje de documentos etiquetados y plantean algoritmos afianzados o una extensión de los mismos, acoplados a los métodos propuestos en su estructura para seudo-etiquetar el resto de documentos. Estos algoritmos son iterativos y no iterativos, los más consolidados son EM [18], Kohonen [16], Cross Validation [17] y DPC [19]; mientras que HCSC [8], Newsmap [4], LDA-W2V [14] y MMSL [15] son algoritmos diseñados por los autores con el fin de generar auto-entrenamiento.

Clasificación (E4).- Una vez que el modelo se encuentra entrenado se procede a la clasificación de documentos de prueba para evaluar su desempeño y precisión, en esta etapa se puede hacer uso de clasificadores supervisados o no supervisados ya que se dispone de un conjunto de documentos etiquetados y seudo-etiquetados con un modelo entrenado. Algunos estudios como [8] [15] [19] [16] y [20] hacen uso del clasificador supervisado SVM, otro clasificador supervisado de uso frecuente es NB identificado en [18] [4] y [17], así también en esta etapa se identifica el uso de aprendizaje profundo con algoritmos de redes neuronales, NN y KNN en [21] [22] [23] y [24].

Finalmente, es importante mencionar que para evaluar el desempeño del rendimiento de las etapas de los modelos semi-supervisados en los estudios recopilados. Se determinan varios factores que influyen en el rendimiento del clasificador, entre estos factores tenemos: el número de documentos, la cantidad de clases o categorías en las cuales se quiere clasificar, el porcentaje de documentos etiquetados, el porcentaje de documentos sin etiquetar o el número de documentos sometidos a prueba.

2.1. Auto-entrenamiento

A este tipo de modelo semi-supervisado también se lo conoce como self-labeling o self-learning, en primera instancia se lo puede considerar como un clasificador supervisado que entrena con un mínimo de documentos etiquetados para pseudo-etiquetar un conjunto de documentos no etiquetados, posteriormente el modelo adquiere su propiedad de semi-supervisado al re-entrenar iterativamente con los documentos etiquetados y pseudoetiquetados hasta que el conjunto de documentos no etiquetados sea vacío [7]. El detalle de su aplicación según [25] involucra la estructura presentada en el Algoritmo 1.

Algoritmo 1: Modelo de aprendizaje por auto-entrenamiento

Entrada: Documentos etiquetados {E} y Documentos no etiquetados {nE}

Se dispone de {E}={$(\mathrm{d}_i)|i = 1, .., n$}$y\{nE\} = \{(d_j)|j = n + 1, .., n + m\}$;
repeat
  Entrena de forma semi-supervisada el clasificador C basado en {E};
  Se aplica C en los documentos de {nE} para predecir una seudo-etiqueta;
  Retira de {nE} los documentos con mejor grado de confianza y genera el subconjunto con documentos pseudoetiquetados sE={ds1, ds2,.., dsn};
  Agrega el subconjunto sE = {(d, f (d))—d $\in sE\}alconjuntoE$;
until $\{nE\} = 0$;
Salida: Documentos etiquetados {E}

En la Tabla 2 se listan los estudios con auto-entrenamiento identificados en la SLR, en sus columnas se anexa las diferentes técnicas utilizadas en cada una de las etapas del proceso de clasificación del modelo planteado, así también se identifica las ventajas y desventajas de la estructura del modelo de auto-entrenamiento propuesto en cada investigación. Entre los beneficios más destacados de estos estudios, se

encuentran modelos [18] [4] [16] focalizados en buscar un eficiente procesamiento con la menor cantidad de datos etiquetados posible. Se identifica también estructuras que, para tener un etiquetado con menos errores, en su representación de documentos y reducción de dimensionalidad sus características son representadas considerando el significado de sus palabras [8] [14] o también reducen el uso de parámetros para la distribución de características [23]. En el etiquetado se destaca las técnicas de generación de métricas y grados de confianza [19][15][24] para evaluar la clasificación del documento, existe mayoritariamente un tratamiento manual del etiquetado, pero existen modelos con tendencia a automatizarlo [21] [24].

Si bien es cierto en el listado de estudios se identifican algunos beneficios en el proceso de clasificación de documentos, sin embargo, es importante mencionar que estas propuestas tienen aún sus limitaciones. Por ejemplo, mientras menos parámetros se considera para la distribución de características [18] mayor es el margen de error de agrupación y la dificultad de evaluar el significado de las palabras en características [16] [14] [23], dando lugar a la replicación de características en distintas agrupaciones [17]. Así también la automatización iterativa del etiquetado aún no es eficiente en definir rangos adecuados de confianza y seleccionar las instancias con mejor grado de confianza [21] [24].

La tendencia que se identifica en este tipo de modelos de auto-entrenamiento es la sencillez de uso de su estructura y la automatización de su procesamiento de etiquetado para que sea un proceso iterativo, esto los convierte mayoritariamente en modelos con esquemas sin mucha carga de procesamiento. En cuanto a su rendimiento, los modelos con mejor precisión [18] en 94 % y [17] en 93.8 % son los que disponen de mayor cantidad de documentos etiquetados para su entrenamiento, 17 % y 70 %, respectivamente; mientras que los de más baja precisión [19] en 50.7 % y [4] en 57 % son aquellos con menor cantidad de etiquetados 10 % y 0 %. El resto de modelos se encuentran en un rango de precisión promedio al grupo con un número de clases que van de 2 a 20 y con una cantidad de etiquetados del 5 % al 10 %.

Tabla 2: Listado de estudios y técnicas usadas en las etapas del modelo de auto-entrenamiento

Id	Autor	E1	E2	E3	E4	Dataset	Docs.	Clases	E	NE	Prueba	P(%)	Ventajas	Desventajas
ST01	Altınel & Ganiz, 2016 [8]	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 [19]	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 [18]	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 [4]	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 [14]	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 [17]	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 [21]	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 [23]	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 [24]	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.

2.2. Co-entrenamiento

La propuesta del modelo de co-entrenamiento es una extensión del auto-entrenamiento, que permite entrenar a dos o más clasificadores a partir de una base de documentos etiquetados para seudo-etiquetar no etiquetados, el fin es compartir las seudo-etiquetas entre los clasificadores buscando mejorar la precisión de la predicción [26]. Para cada clasificador se busca enfoques distintos de las características de los documentos etiquetados denominados vistas, mientras menos correlacionadas estén las características en las vistas mejor será la predicción, por esta razón a este modelo se lo conoce como modelo multivista y genera entrenamiento a través de una red de aprendizaje [27]. En el Algoritmo 2 se presenta la estructura del modelo:

Algoritmo 2: Modelo de aprendizaje por co-entrenamiento

Entrada: Documentos etiquetados $\{E\} = \{(d_i)|i = 1, .., n\}$
Documentos no etiquetados $\{nE\} = \{(d_j)|j = n + 1, .., n + m\}$
Para cada documento se establece dos o más vistas $d_i = [d_i^{(V1)}, .., d_i^{(Vn)}]$
Cada vista $\{V_i|i = 1, .., n\}$ recopila diferentes atributos $\{V\} = \{(a_j)|j = 1, .., n\}$;
Donde $V_1 \cap V_2 .. \cap V_n = 0$;
Así;
Instancias de entrenamiento etiquetadas: $E^{(V1)} = d_i^{(V1)}, .., d_n^{(V1)} .. E^{(Vn)} = d_i^{(Vn)}, ..., d_n^{(Vn)}$;
Instancias de entrenamiento no etiquetadas: $nE^{(V1)} = d_{n+1}^{(V1)}, .., d_{n+m}^{(V1)} . nE^{(Vn)} = d_{n+1}^{(Vn)}, .., d_{n+m}^{(Vn)}$;
Donde;
$E = E^{(V1)} \cup E^{(V2)} .. \cup E^{(Vn)}$
$nE = nE^{(V1)} \cup nE^{(V2)} .. \cup E^{(Vn)}$
a=1;

repeat
  Entrena clasificadores $C^{(1)} .. C^{(n)}$ para cada vista de documentos etiquetados $E^{(V1)} .. E^{(Vn)}$;
  Clasifica documentos de las vistas no etiquetadas $nE^{(V1)} .. nE^{(Vn)}$ usando $C^{(1)} .. C^{(n)}$;
  for *cada* $C^{(k)}$ *con* $k = 1$ *hasta* n **do**
    if $k \neq a$ **then**
      Los documentos con mejor predicción clasificados por $C^{(k)}$ son añadidos al conjunto de etiquetados $E^{(Va)}$;
    end
  end
  a++;
  Se retira de $nE^{(V1)} .. nE^{(Vn)}$ los documentos clasificados con mejor grado de confianza;
until $\{nE\} = 0$;
Salida: Documentos etiquetados $\{E\}$

En la Tabla 3 se lista cada uno de los estudios con co-entrenamiento identificados en la revisión, anexando las diferentes técnicas utilizadas en cada una de las etapas del modelo planteado para conseguir una clasificación de documentos, la mayoría de los estudios de co-training utilizan como base de su estructura el auto-entrenamiento. El factor extra radica en su definición de vistas que son un aporte destinado para mejorar la precisión de sus modelos. Se identifican varias estrategias para definir vistas, tal es el caso de [26] que define dos técnicas diferentes de efectuar la representación de documentos y con cada una elabora una vista para retroalimentar su modelo con las mejores precisiones. También tenemos a [28] que plantea para su entrenamiento cinco vistas considerando documentos en cinco idiomas diferentes con el mismo contenido. En [29] las vistas son estructuradas considerando las partes de la url de un documento: base, texto de la imagen de la url y el destino de la url, así se arma conjuntos de datos en las tres dimensiones. O el estudio de [30] que recopila conjuntos de documentos multimedia y prepara vistas de entrenamiento diferentes con el video, audio y texto del documento.

Entre las fortalezas de estos modelos se identifica estructuras que no tienen límite de generación de vistas [31] o que aprovechan el conjunto de no etiquetados para entrenarlos en una vista independiente [30]. Además, se determina diseños para el análisis de documentos de alta escala (video, audio) [28], así como

también documentos con actualización en tiempo real (big data) [30]. Esta diversidad ha dado lugar a técnicas que analicen las características específicas y compartidas en cada vista para evitar la redundancia en el procesamiento [29], o técnicas como la teoría de atención [31] e índices de etiquetado difusos [32]. En cuanto a los inconvenientes de los modelos de co-entrenamiento tenemos que algunos de los esquemas tienen un límite en la cantidad de vistas [27] [30], si bien es cierto esto puede deteriorar la precisión de clasificación; el exceso de vistas en cambio puede incrementar el uso de recursos y deteriorar el rendimiento del proceso [30] [31] [33] [34]. Para la distribución de características, la agrupación sigue siendo una de las técnicas más usadas, sus limitantes identificadas radican en la falta de métricas para medir la diversidad de los datos y los conflictos existentes cuando los datos son desequilibrados [35]. Otro de los problemas con el planteamiento de vistas es la automatización de la asignación de las características específicas y compartidas [29], si se brinda mayor importancia a una de otra se pueden perder características [28] o generar redundancia de ellas [33]. En cuanto a la evaluación de la predicción, se convierte en un proceso más complejo por la diversidad de clasificadores existentes y las estrategias de decisión [32].

Las métricas de precisión del rendimiento de los modelos de co-entrenamiento en general son destacadas, así tenemos estudios como [28] con 93 %, [30] con 94 %, [29] con 96 %, [33] con 95 %, y [34] con 96 %, estas investigaciones manejan diferentes planteamientos y números (entre 2 y 5) de vistas, considerables cantidades de clases (entre 2 y 10) y reducidos conjuntos de etiquetados (entre 2 % y 20 %), sin embargo, las estructuras de sus algoritmos de etiquetado son eficientes en la predicción de clasificación. En el otro extremo de los índices de rendimiento se identifica el estudio [27] con 34,5 % cuya métrica reduce el rendimiento promedio de los modelos, varios factores influyen en su métrica, el uso del límite de vistas (2), la clasificación de documentos multietiqueta, un algoritmo MLSMOTE de etiquetado poco eficiente en el análisis semántico de características y la evaluación de sus predicciones.

Tabla 3: Listado de estudios y técnicas usadas en las etapas del modelo de co-entrenamiento

Id	Autor	Vistas	E1	E2	E3	E4	Dataset	Docs.	Clases	E	NE	Test	P(%)	Ventajas	Desventajas
CT01	Borrajo et al., 2020 [26]	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 [27]	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 [28]	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 [30]	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 [29]	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 [31]	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 [33]	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 [34]	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 [35]	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 [32]	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.

(*) Estudios que buscan una clasificación multi-etiqueta

2.3. Ensamblado

Este tipo de modelos también se los conoce como modelos en conjunto y plantean una estructura de entrenamiento con múltiples clasificadores, se combina los resultados de sus predicciones para acorde a un análisis seleccionar el más idóneo del conjunto [36]. Teóricamente se asemeja mucho a la conceptualización del co-entrenamiento, pero la principal diferencia radica en su conjunto de datos para entrenamiento, co-training busca generar conjuntos diferentes de datos denominados vistas para entrenar sus modelos considerando las diferentes cualidades de sus datos; mientras que en los ensamblados trabajan su entrenamiento con el mismo conjunto de datos en diferentes modelos o denominados también clasificadores débiles, no separa los datos en vistas. Varias técnicas permiten la aplicación de este tipo de aprendizaje, entre las primarias tenemos a Bagging, Boosting y entre las secundarias AdaBoost, generalización de apilados, combinación de expertos, basados en voto [5]. En el Algoritmo 3 se presenta la estructura del modelo:

Algoritmo 3: Modelo de aprendizaje por ensamblado

Entrada: Documentos etiquetados $\{E\} = \{(d_i)|i = 1, .., n\}$
Documentos no etiquetados $\{nE\} = \{(d_j)|j = n + 1, .., n + m\}$
Para {E} se establece dos o más modelos de clasificación $C_1..C_n$
Así, por cada clasificador
Instancias de entrenamiento etiquetadas: $E^{(M1)}, E^{(M2)}..E^{(Mn)}$
Instancias de entrenamiento no etiquetadas: $nE^{(M1)}, nE^{(M2)}..nE^{(Mn)}$

repeat
- Entrena cada clasificador $C^{(1)}..C^{(n)}$ para documentos etiquetados $E^{(M1)}..E^{(Mn)}$;
- Clasifica $nE^{(M1)}..nE^{(Mn)}$ con los diferentes clasificadores $C_1..C_n$;
- Documentos con mejor predicción por $C_1..C_n$ se añaden al conjunto de etiquetados $E^{(M1)}..E^{(Mn)}$;
- Se retira de $nE^{(M1)}..nE^{(Mn)}$ los documentos clasificados con mejor grado de confianza;

until $\{nE\} = 0$;
Se Elige el $E^{(M1)}..E^{(Mn)}$ que ha registrado el mejor grado de confianza;
Salida: Documentos etiquetados $\{E\}$

En la Tabla 4 se presentan investigaciones con modelos ensamblados, se identifican estructuras que pueden disponer por cada clasificador débil un proceso diferente de clasificación de documentos con distintas técnicas [37] [38] [39], al final se realiza un consenso para la selección de la mejor predicción. Estos modelos brindan apertura al reforzamiento de su entrenamiento a través de la incorporación de conjuntos de documentos externos ya pre-entrenados [38], así como también el entrenamiento multilingüe que aprovecha el etiquetado de documentos en otros lenguajes para incorporarlo al lenguaje del documento en estudio [40] [41]. Se determina también un trabajo en la semántica y secuencia de las características del documento con el uso de la teoría de transformer [37] [38] [40], logrando conseguir eficientes representaciones de documentos con reducida dimensionalidad. En cuanto a la predicción y consenso del etiquetado considerando la variedad de clasificadores débiles, se encuentran estudios que buscan la automatización de este proceso ajustando el consenso de predicción [42] [43] [44] [22]. Ante las propuestas de estos estudios, uno de las mayores limitantes que se identifica es el alto costo computacional necesario para el procesamiento de las robustas estructuras [37] [38] [40] [42] [43] [39] [22]. Así también se aprecia la pérdida de características cuando el modelo incorpora pre-entrenamiento externo o de otro lenguaje [41]. Si bien es cierto plantean consensos eficientes para la evaluación de predicciones, este proceso está abierto a mejorar su automatización agregando mejores técnicas en la toma de decisión [44] .

Tabla 4: Listado de estudios y técnicas usadas en las etapas del modelo de ensamblado

Id	Autor	CsD	E1	E2	E3	E4	Dataset	Docs.	Clases	E	NE	Test	P(%)	Ventajas	Desventajas
ES01	De Souza, 2021 [37]	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 [38]	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 [40]	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 [41]	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 [42]	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 multiclasificador.	Las características resultantes son de alta dimensionalidad, la diversidad de clasificadores débiles genera un alto costo computacional.
ES06	Shrivas et al., 2021 [43]	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 entramiento 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 [39]	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 [44]	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 [22]	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

Las métricas de rendimiento de clasificación de estos modelos determinan una vez más que el número de documentos etiquetados es importante para su precisión, los modelos ensamblados más eficientes son [43] 97.25 %, [39] 92.9 %, [44] 88.69 % y efectivamente son los modelos que mayor número de etiquetados disponen, con 30 %, 47 % y 65 % respectivamente del total de documentos. Otros factores que influyen en la eficiencia de estos modelos son su cantidad (5) de clasificadores débiles [42] [43] [39], su entrenamiento por capas (cross-validation) [42], la apertura a documentos pre-entrenados [39] y su eficiencia en el manejo del consenso para la predicción [42] [43]. Mientras que las características de los menos eficientes [40] y [41] recaen en menor cantidad de clasificadores débiles (2), menor cantidad de etiquetados y modelos que en su estructura consideran conjuntos de documentos con diversos lenguajes para su entrenamiento que deterioran la precisión.

2.4. Aprendizaje activo

Los modelos de aprendizaje activos son denominados también modelos asistidos, este tipo de aprendizaje es usado para reducir el esfuerzo y costo de etiquetado e incrementar el desempeño de precisión del modelo [45] [46] . Se lo conoce como asistido porque permite la interacción de un etiquetador que generalmente puede ser un experto en el dominio [47] y con la identificación de importantes puntos de datos o documentos ejemplo no etiquetados, el etiquetador sugiere etiquetas que sirvan de patrones para un entrenamiento eficiente, reduciendo así, el error de etiquetado [36]. En el Algoritmo 4 se presenta los pasos genéricos del procedimiento:

Algoritmo 4: Modelo de aprendizaje por aprendizaje activo

Entrada: Documentos no etiquetados $\{nE\} = \{(d_j)|j = n+1, .., n+m\}$
Etiquetadores $\mathrm{Et}_1..Et_n$;

repeat
- $Et_1..Et_n$ selecciona de $\{nE\}$ los documentos más relevantes y los etiqueta en $\{E\} = (d_i)|i = 1, , n$;
- Se entrena el clasificador C basado en los documentos etiquetados $\{E\}$;
- Se aplica C en los documentos de $\{nE\}$ para predecir una seudo-etiqueta;
- Retira de $\{nE\}$ los documentos con mejor grado de confianza;
- Genera el subconjunto con documentos pseudoetiquetados $sE = \{ds_1, ds_2, .., ds_n\}$;
- Agrega el subconjunto $sE = \{(d, f(d))|d \in sE\}$ al conjunto E;

until $\{nE\} = 0$;
Salida: Documentos etiquetados $\{E\}$

En la Tabla 5 se presenta los modelos semi-supervisados con un aprendizaje activo, la fortaleza de estos modelos se encuentra en su etiquetador [47] o etiquetado activo [20] que son los encargados de seleccionar las instancias con mayor influencia o importancia [47] para etiquetarlas, así se dispone de un conjunto de etiquetados más significativo que mejoran la eficiencia de la predicción de clasificación [47] [45]. Generalmente los conjuntos de datos son de un número considerable, de acuerdo a la cantidad y tipo de documentos, el etiquetador selecciona los documentos más esenciales para que sirvan de entrenamiento en el proceso de etiquetado del resto de documentos del conjunto.

Si bien es cierto el etiquetador permite categorizar adecuadamente documentos clave con un mínimo margen de error, hay que considerar que mayoritariamente este proceso es manual y depende de recursos externos [47]; se trata de automatizar este proceso sin embargo las limitaciones identificadas se encuentran en la dificultad de identificar los documentos más relevantes del conjunto [48] y los considerables márgenes de error en el etiquetado activo automático [20].

Tabla 5: Listado de estudios y técnicas usadas en las etapas del modelo de aprendizaje activo

Id	Autor	E1	E2	E3	E4	Dataset	Docs.	Clases	E	NE	Test	P(%)	Ventajas	Desventajas
AL01	Y. Yang & Loog, 2018 [46]	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 [47]	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 [48]	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-entramiento y documentos multilingües.	En el etiquetado automático existe dificultad para identificar las características más significativas.
AL04	Reyes et al., 2018 [45]	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 [20]	s/reg	SSKMS	STDP	SVM	UCI/USPS	9298	10	93 *(1%)*	9112 *(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.

Se aprecia que el nivel de rendimiento de los modelos de aprendizaje activo varía acorde a su estrategia de etiquetado ya sea manual o automático, el modelo de mejor rendimiento [47] con 97,2 % tiene un apoyo en 50 puntos de control para validar la categoría de documentos clave, lo que ayuda a disponer de un conjunto de entrenamiento más eficiente. En [48] con 82,5 %, el rendimiento se reduce, el modelo no dispone de puntos de control para etiquetado, posee un algoritmo automatizado que tiene limitantes en la evaluación de la semántica de las características de los documentos. Sin embargo, los valores de desempeño de este tipo de aprendizaje activo son eficientes considerando su bajo número de documentos etiquetados [20], su alto número de clases [47] [45] [20] y la fortaleza de un etiquetador que en ciertos casos es un experto en el dominio [46] [47].

2.5. Aprendizaje de transferencia

Este tipo de aprendizaje ha sido de uso reiterativo en la presente SLR para clasificación de documentos. El método propone un entrenamiento para clasificadores a partir del aprendizaje con un conjunto de datos de un dominio fuente (SD), el conocimiento generado puede ser transferido como aprendizaje para la clasificación de un conjunto de datos en otro dominio destino (TD) relacionado [49]. Junto a esta estructura aparece el concepto de pre-entrenamiento que son modelos de aprendizaje recopilados y entrenados en experiencias previas vinculantes a la necesidad de mejorar el etiquetado y clasificación del conjunto de datos en estudio, entre ejemplos de este tipo de modelos de pre-entrenamiento tenemos a Word2Vec, ELMO, BERT entre otros [7]. En el Algoritmo 5 presentamos los pasos genéricos del modelo:

Algoritmo 5: Modelo de aprendizaje por transferencia

Entrada:
Documentos etiquetados con pre-entrenamiento de un SD $\{Esd\} = \{(d_i)|i = n+1,,n+m\}$;
Documentos etiquetados del TD $\{Etd\} = \{(d_j)|j = n+1,,n+m\}$;
Documentos no etiquetados del TD $\{nEtd\} = \{(d_k)|k = n+1,,n+m\}$;

repeat
- Entrena el clasificador C basado en los documentos etiquetados {Etd};
- Con aprendizaje de transferencia se refuerza el entrenamiento de C con documentos {Esd};
- Se aplica C en los documentos de {nEtd} para predecir una seudo-etiqueta;
- Retira de {nE} los documentos con mejor grado de confianza;
- Genera el subconjunto con documentos pseudoetiquetados $sE = \{ds_1, ds_2, .., ds_n\}$;
- Agrega el subconjunto $sE = \{(d, f(d))|d \in sE\}$ al conjunto E;

until $\{nEtd\} = 0$;
Salida: Documentos etiquetados $\{Etd\}$

En la Tabla 6 se presenta el listado de estudios identificados como modelos de transferencia de aprendizaje, con cada una de sus técnicas utilizadas en el proceso de clasificación, para la transferencia de aprendizaje de los documentos pre-entrenados disponibles se identifican varias técnicas, la más frecuente es la de SD (Source domain) y TD (Target domain) [50] y [49] que se especializa en la transferencia de conocimiento de un modelo fuente a otro objetivo. En los estudios [51] y [52] identificamos al método generativo VAE que es capaz de capturar complejas distribuciones de datos con eficientes representaciones latentes vía pre-entrenamiento. Así también, tenemos a DVEM, SNN, BART y CSA en [53] [54] [55] y [56] respectivamente, que son modelos que se desempeñan en entornos de pre-entrenamiento y son los encargados de capturar el conocimiento generado por diccionarios externos para transferir y acoplar el aprendizaje a la clasificación de documentos.

En esta variedad de técnicas de transferencia de aprendizaje se identifica modelos con flexibilidad de adaptación, que permiten: configurar varias entradas SD para una salida TD o una entrada SD para varias salidas TD [57], aprovechar la riqueza de aprendizaje de un lenguaje para transferirlo a otro [51] [58], extraer conocimiento de diccionarios o palabras semilla [52] [49]. Todo esto acompañado de procedimientos focalizados en la adecuada distribución de las características [59], el análisis semántico con teorías de atención [54] [55] [52] o la experimentación de modelos entrenados únicamente con documentos externos pre-entrenados sin ninguna etiqueta en el domino destino [50].

Uno de los mayores problemas de los modelos de transferencia de aprendizaje se encuentra en la adaptación de dominio, para conseguir adaptación es importante validar la relación existente entre los documentos pre-entrenados (SD) y los documentos por etiquetar (TD), pues si no existe el vínculo adecuado el entrenamiento del modelo en estudio no se podría efectuar adaptación o sería ineficiente [51] [58] [57]. Otra limitante identificada es la dificultad para evaluar la predicción de clasificación, como en ciertos casos no se dispone de documentos etiquetados del conjunto de datos en estudio no se puede realizar una eficiente evaluación con documentos pre-entrenados que son de un conjunto externo [50] [59] .

El rendimiento de este tipo de modelos es aceptable, considerando que no se dispone de documentos etiquetados en su conjunto de datos propio, se aprecia que en algunos casos ya no se diferencian documentos etiquetados y no etiquetados, la información registrada corresponde a la cantidad de documentos pre-entrenados que dispone el experimento para conseguir la transferencia de aprendizaje. En los estudios [54] [51] [55][53] y [49] de mejor desempeño se puede apreciar que los principales factores que influyen en su rendimiento son el alto número de documentos pre-entrenados y la reducida cantidad de clases. Por otro lado, en las investigaciones [56] [50] y [52] con menor rendimiento, se aprecia que independientemente del número de documentos pre-entrenados, son los estudios que manejan mayor número de clases.

Tabla 6: Listado de estudios y técnicas usadas en las etapas del modelo de aprendizaje de transferencia

Id	Autor	E1	E2	E3	E4	Dataset	Docs.	Clases	Pre-E	Test	P(%)	Ventajas	Desventajas
AT01	Guo & Yao, 2021 [56]	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 [54]	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 [50]	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 [51]	s/n	BoW	VAE	SDGMs	UCI/ Multilengua	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 [55]	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 [53]	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 [52]	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 [49]	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 [58]	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 perdida 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 [57]	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 [59]	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.

2.6. Meta-análisis comparativo de los modelos semi-supervisados

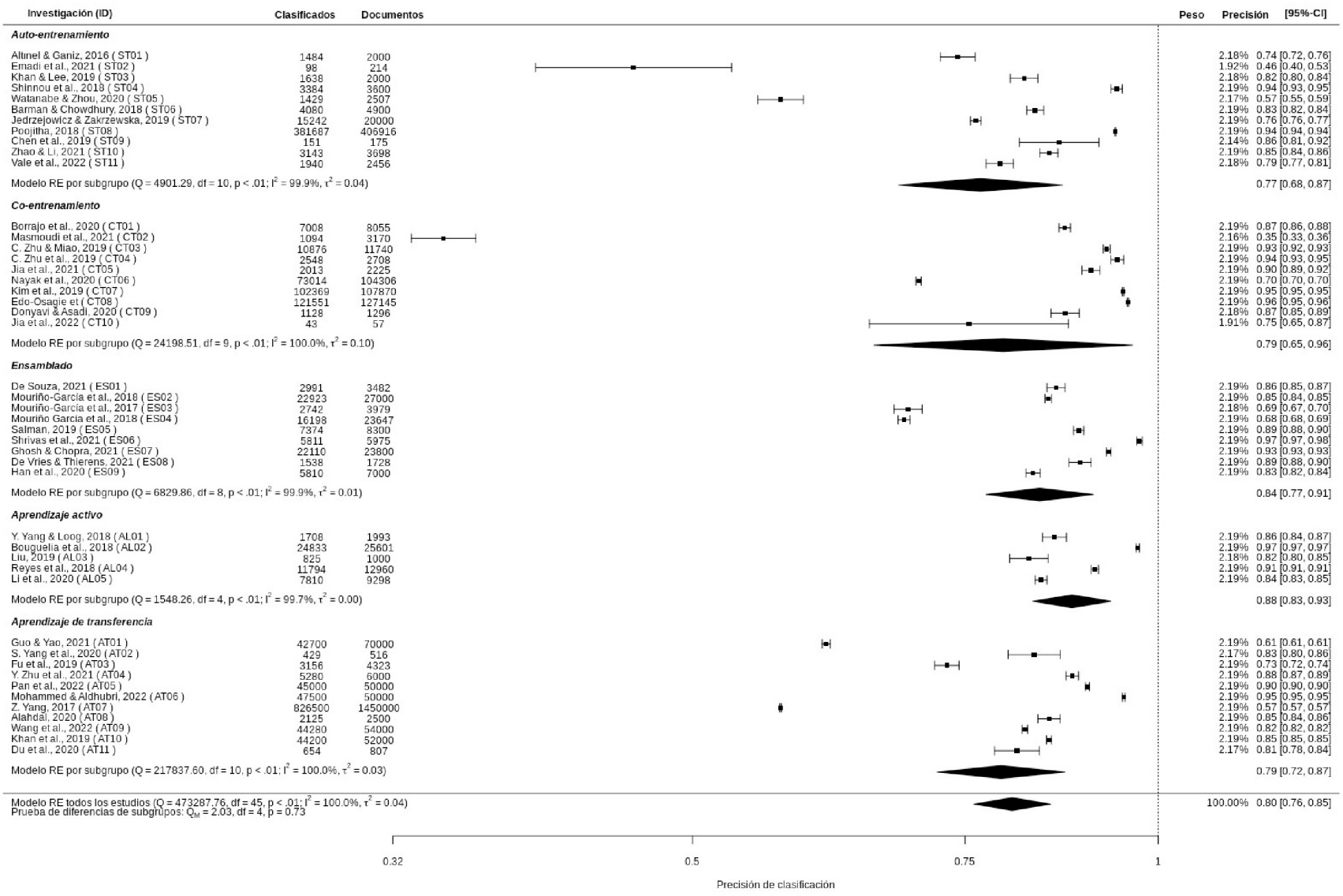
Los 47 estudios identificados en la revisión han sido agrupados de acuerdo a su tipo de modelo de aprendizaje semi-supervisado y se ha considerado el desempeño de precisión de clasificación de documentos del modelo para la estimación de un meta-análisis. Es importante mencionar que se ha hecho hincapié en identificar aquellos tipos de modelos focalizados en la clasificación de documentos. En este ámbito se determina 11 investigaciones con modelos de auto-entrenamiento (ST), 10 de co-entrenamiento (CT), 9 modelos ensamblados (ES), 5 de aprendizaje activo (AL) y 11 con aprendizaje de transferencia (AT).

En la Tabla 7 a través de un ForestPlot, se puede apreciar una comparativa de los distintos modelos identificados en la SLR, todos los estudios identificados se agrupan en 5 tipos de modelos de aprendizaje, de los cuales se adjunta, número de documentos clasificados correctamente, número total de documentos sometidos a clasificar, porcentaje de peso asignado al estudio acorde al ajuste del modelo en base a los resultados de precisión de sus observaciones, porcentaje del nivel de precisión en la clasificación de documentos, el rango de intervalo de confianza al 95 % y el logaritmo de la relación de riesgo que es diseñado considerando las medidas de salida de un grupo individual con variables dicotómicas en formato de proporción logarítmica transformada (PLN).

Aplicando el modelo de efectos aleatorios (Random Effect - RE) a los subgrupos identificados se obtiene un valor promedio para toda su población el cual se ha fijado en 0.80 con un intervalo de confianza de (95 %CI [0.74 - 0.86]). Mientras que el rendimiento individual por tipo de aprendizaje en orden descendente ha sido el siguiente: Modelo por Aprendizaje activo con una precisión en RE del 0.88 en un intervalo de confianza de (95 %CI [0.83 - 0.95]); en una segunda instancia tenemos al modelo Ensamblado con una precisión en RE del 0.84 en un intervalo de confianza de (95 %CI [0.75 - 0.92]); posteriormente tenemos al modelo de Co-entrenamiento con una precisión en RE del 0.79 en un intervalo de confianza de (95 %CI [0.62 - 1.00]); luego se encuentra el modelo de Aprendizaje por transferencia con una precisión en RE del 0.79 en un intervalo de confianza de (95 %CI [0.68 - 0.89]) y en última instancia se encuentra el modelo de Auto-entrenamiento con una precisión en RE del 0.77 en un intervalo de confianza de (95 %CI [0.63 - 0.88]).

En el listado de tipos de aprendizaje semi-supervisado empleados para la clasificación de documentos se aprecia dos agrupaciones acordes a su estructura de funcionamiento, 3 de ellos (auto-entrenamiento, co-entrenamiento y ensamblado) trabajan con reducidos conjuntos de documentos etiquetados; dos de ellos (aprendizaje activo y transferencia) en cambio disponen de etiquetadores o fuentes externas que aportan a incrementar el número de documentos etiquetados con un mayor empleo de recursos. El modelo de mejor rendimiento en el primer grupo es el aprendizaje ensamblado (0.84) y en el segundo grupo el modelo de aprendizaje activo (0.88), considerando la disparidad importante en la disposición de fuentes externas por parte del aprendizaje activo la diferencia es de 4 puntos lo cual es una muestra de eficiencia de los modelos en los dos ámbitos.

Tabla 7: Forest plot agrupado de niveles de precisión de modelos semi-supervisados



3. Ventajas y desventajas de los modelos semi-supervisados

Para responder a la pregunta de investigación 2, se ha considerado el análisis realizado en la sección anterior de los estudios incluidos en la SLR para identificar las principales ventajas y desventajas de los modelos semi-supervisados en las diferentes etapas del proceso de clasificación de texto en su contexto general. Los estudios han sido agrupados de acuerdo a las temáticas (ventaja o desventaja) en común determinadas, en la Figura 5 se presenta el proceso de clasificación de documentos, ubicando las fortalezas y debilidades que se han identificado en sus etapas.

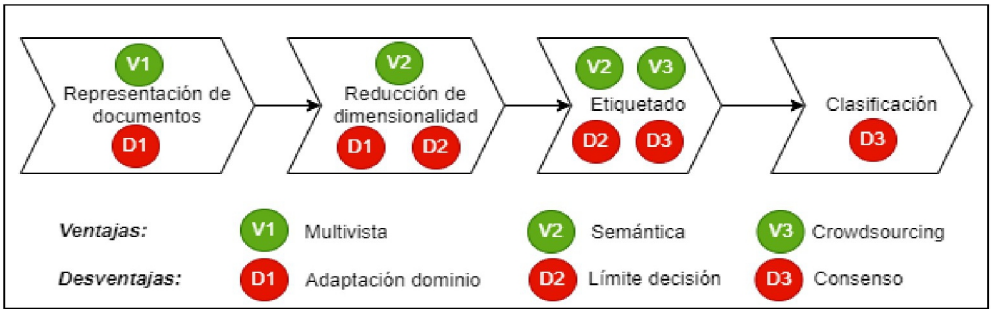


Figura 5: Ventajas y desventajas del proceso de clasificación

Ventajas

V1. Multivista.- Genera una instancia diferente del conjunto de datos perteneciente a un documento a la que se le denomina vista, cada instancia plantea su propia representación de características del documento desde distintos puntos de vista [28]. Esta tarea se efectúa en la etapa de representación de documentos.

V2. Semántica de características.- Establece un valor significativo para los términos de un documento, valor necesario para disponer de una mejor interpretabilidad del texto y poder asignar su categoría de clasificación de forma más precisa [8]. Esta tarea se puede llevar a cabo en la etapa de reducción de dimensionalidad o etiquetado.

V3. Crowdsourcing.- Define un oráculo (experto en dominio) para etiquetar documentos ruidosos o de difícil clasificación, se los etiqueta de forma activa para reducir el error de categorización y mejorar la entropía del modelo [47]. Esta tarea se ejecuta en la etapa de etiquetado.

Desventajas

D1. Adaptación de dominio.- Cuando existe un conocimiento entrenado previo, este conocimiento puede ser reutilizado en un nuevo modelo de categorización de documentos con un dominio similar, a través de una adaptación de dominio, la adaptación no tiene un proceso definido y puede desorientar el entrenamiento del modelo [50]. Esta actividad se la puede llevar a cabo en la etapa de representación de documentos, reducción de dimensionalidades o etiquetado.

D2. Límite de decisión.- Los documentos de texto acorde a los valores de sus vectores de características son ubicados en un espacio vectorial divido en clases o categorías. Un documento que se ubique en una región que separa una categoría de otra, posee un alto margen de error en clasificación [19]. Esta tarea se lleva a cabo en la etapa de etiquetado.

D3. Consenso entre clasificadores.- La combinación de vistas o modelos para mejorar el entrenamiento de etiquetado implica el uso de varios clasificadores, que necesitan cumplir con propiedades propias y compartidas para consensuar una predicción en conjunto [29]. Esta actividad se lleva a cabo en la etapa de etiquetado o clasificación.

3.1. Multivista

Los seres humanos perciben el mundo exterior a través de múltiples sentidos (vista, oído, tacto), es decir, la información que recepta el cerebro es multivista y puede aprender desde diferentes puntos de vista [29]. Con esta analogía los modelos de aprendizaje semi-supervisados buscan reforzar su entrenamiento usando diferentes formas de representación o vistas de los documentos. La escasez de documentos etiquetados es un factor común en el entrenamiento de este tipo de modelos, sin embargo, existen determinadas circunstancias en las que los documentos etiquetados pueden tener diferentes instancias de representatividad, ya sea por idioma, formato, descripción u otros. Por ejemplo, el mismo documento etiquetado puede estar disponible en inglés, francés y español; o se puede disponer de un documento en los formatos de audio, texto o video. Esta variedad de alternativas ayuda a fortalecer la etapa de representación de documentos, pudiendo incrementar el conjunto de documentos etiquetados que facilitan la categorización y desempeño del modelo.

En este entorno multivista se proponen varios modelos y frameworks estructurados para soportar distintas representaciones de documentos, entre ellos HMM en [26], MLSMOTE en [27], MCT en [33], EM en [34], estos modelos extraen de sus conjuntos de documentos distintas representaciones que alimentan su entrenamiento. Su estructura plantea en promedio dos vistas con subconjuntos de los documentos a entrenar generando nuevas instancias del documento con elementos tales como: vista con el tituloresumen del documento y vista con las citas del documento [27]; vista con representación de datos TF-IDF y vista con representación de datos LDA [33]; vista con las categorías del documento y vista con palabras de sentimiento positivasnegativas del documento [34]. Se identifica también estudios, en los que la técnica multivista no solo trata datos de pequeña escala (documentos) sino también datos de gran escala (audio, video) con frecuencia actualizada (streaming). Modelos como SOMVFV en [28], SSOPMV en [30], SMDDRL en [29] proponen estructuras para la categorización de este tipo de datos, el co-entrenamiento de esta técnica multivista mejora el desempeño de clasificación en cinco puntos al rendimiento promedio de los modelos de auto-entrenamiento (Ver Tabla 7).

3.2. Semántica de características

El análisis de la semántica de las características es una tarea que se puede llevar a cabo en la etapa de etiquetado del proceso de clasificación de documentos (ver Figura 4). Su propósito es determinar valor significativo en los conjuntos de características del documento, estas características valoradas en documentos etiquetados pueden identificarse en otros documentos no etiquetados, así se puede reutilizar el significado de las características para orientar la categoría del nuevo documento sin etiquetar. Según [8] agregar semántica a los conjuntos de características ayuda a mitigar los problemas que genera la polisemia (palabras con significados múltiples) y sinónimos en la clasificación de documentos, permitiendo mejorar el desempeño de la precisión de los modelos.

En los modelos de aprendizaje semi-supervisados se identifica varios estudios focalizados en fortalecer la semántica de las características usando diferentes técnicas. Así es el caso de [8] que plantea el modelo semántico HCSC basado en clases que extraen las relaciones de significación entre términos, lo cual brinda un mejor contexto del documento, utilizando esta semántica para etiquetar nuevos documentos. Con la misma perspectiva, [19] plantea el modelo SSApolo que utiliza como base el método de circulo de apollonius, considerando puntos (documentos) máximos se realiza un agrupamiento de picos de densidad, por cada punto máximo se forma un círculo de apollonius, así los puntos no etiquetados en el circulo toman la etiqueta del punto máximo. En [15] se estructura un framework con una capa de ingeniería específica para procesar las características del documento, aquí existen dos extractores de semántica uno interno y otro externo, el primero identifica las características como adjetivo, adverbio, verbo y sustantivo; mientras que el segundo, consulta con léxicos de palabras de sentimientos definidas para asignar una semántica al documento. Los léxicos de palabras definidas son un soporte para fortalecer la semántica de los conjuntos de características en un modelo, por esta razón se preparan diccionarios de palabras acorde al dominio de estudio, en [14] se utiliza Word2Vec que es un diccionario de palabras que facilita la incorporación de significado a las características en estudio.

3.3. Crowdsourcing

En la etapa de etiquetado, existe dificultad al momento de asignar una categoría a aquellos documentos que se encuentran en el límite de decisión (posición difusa de pertenencia a varias categorías), la mayoría de los modelos toman una decisión aleatoria para esta clasificación, descuidando la consistencia del modelo. Sin embargo, los modelos de aprendizaje activo, dan apertura al uso de técnicas de crowdsourcing en distintos puntos del modelo, así, por medio de la colaboración de expertos se puede mejorar la decisión de clasificación para aquellos documentos en límite de decisión y fortalecer del desempeño del modelo de clasificación.

Esta técnica se aplica al modelo como un agente externo que implica costos de recursos humanos y tiempo, con resultados cuantiosos de su colaboración. A fin de reducir costos se ha buscado estrategias para automatizar este proceso, así es el caso de [47] que define medidas informativas para las instancias de los documentos, para generar un entrenamiento más estable con las instancias de mejor métrica y más distantes al límite de decisión. Por otro lado, también tenemos [48] que establece estrategias de aprendizaje para que por medio de la heurística se pueda adaptar al entrenamiento conceptos de diccionarios complementarios.

3.4. Adaptación de dominio

Con el aprendizaje de transferencia se puede agregar al modelo en estudio nuevo conocimiento generado previamente, este conocimiento se puede disponer de diccionarios de palabras o conjuntos de datos ya etiquetados (Source Domain SD) en diferentes dominios que pueden ser reutilizados de acuerdo a los propósitos de un nuevo modelo de trabajo (Target Domain TD). Sin embargo, no siempre el conocimiento existente se alinea al dominio de trabajo actual, por esta razón es necesario una adaptación de dominio para que este conocimiento sea de utilidad en el modelo de trabajo, de no existir su adecuada adaptación, el entrenamiento del modelo se puede volver difuso y su desempeño se puede deteriorar antes que mejorar.

La adaptación de dominio puede llevarse a cabo en distintas etapas del proceso de clasificación de documentos, en [18] se plantea una técnica en la etapa de representación de documentos, se asigna pesos a las características del SD y por coeficiente de correlación se evalúa las características con mejor adaptabilidad al TD. En [49] la adaptación se la realiza en la representación de documentos del SD, se evalúa la concurrencia de las características en TD y acorde a su disponibilidad en los documentos pre-entrenados estos son mapeados para el entrenamiento de su etiqueta. En este proceso de adaptación de conocimiento, el análisis de la semántica de las características es una tarea compleja, en el caso de [18] y [49] no se la efectúa, perdiendo precisión de la información, interpretabilidad del documento y dimensionalidad.

3.5. Límite de decisión o límite linear

El etiquetado de documentos sin etiquetar generalmente se lo realiza en base a técnicas de clustering, a los documentos pertenecientes al grupo se les asigna su categoría, sin embargo, pueden existir documentos cuya representatividad se encuentra en los bordes de una agrupación que está junto a otra y su etiqueta se torna difusa, a este suceso se lo denomina límite de decisión y afecta el etiquetado óptimo, factor vinculante al buen desempeño de la clasificación [19].

Los estudios [8], [19], [15] y [14] utilizan diferentes técnicas de etiquetado para sus documentos, todas ellas ajustan su estructura para mitigar el margen de error del límite de decisión. En [19] se plantea la construcción de estructuras de grupos vecinales basados en gráficos geométricos obteniendo un rendimiento del 92,75 % de precisión con tres categorías, sin embargo, el desempeño del modelo con un mayor número de categorías se deteriora llegando hasta un 50.07 % de precisión. Por otro lado [15] genera dos niveles de agrupaciones para buscar reducir el margen de decisión, en un primer nivel genera subconjuntos para agrupar adjetivos, adverbios, verbos y sustantivos; en un segundo nivel dentro de cada subconjunto se agrupa características por sentimiento identificado, así es como, el paradigma de margen de decisión se determina en los dos niveles.

3.6. Consenso entre clasificadores

Las técnicas de co-entrenamiento por vistas y ensamblado de modelos implica el uso de múltiples clasificadores, para los cuales es importante cumplir con propiedades de complementariedad y consenso para su estructura general de multiclasificador, estas propiedades son relevantes para evitar la redundancia de los conjuntos de datos en el modelo, permiten discernir de mejor manera la información propia o específica (complementariedad) de cada clasificador frente a la información compartida (consenso) del modelo para un mejor entrenamiento.

En un modelo multiclasificador existe dificultad en mantener estas propiedades, ya que se puede retirar conjuntos de características redundantes entre clasificadores sin considerar que esta información específica puede ser relevante para determinado clasificador. Y si se conserva todas las características de cada clasificador el modelo adquiere una representación pesada para su entrenamiento [29]. La pérdida de características específicas no solo se puede dar por la búsqueda de eficiencia, sino también como el caso de [28] existen circunstancias en las que el clasificador pierde complementariedad porque el emisor pierde información por razones naturales o técnicas, como por ejemplo la indisponibilidad del servicio de una cámara.

4. Conclusiones y trabajo futuro

Los diferentes estudios que se han incorporado en la presente SLR han dado muestra de la variedad e importancia que tienen los modelos de aprendizaje semi-supervisados en la clasificación de documentos institucionales. En este artículo se ha propuesto una estructura de los tipos de modelos identificados, la cual ha permitido analizar el estado del arte de la investigación y conocer nuevos estudios y técnicas vinculantes a la clasificación de documentos (Sección 3). Con esta estructura se ha logrado agrupar estudios, acorde a sus diferentes técnicas usadas en el proceso de clasificación, así se puede diferenciar el desempeño de cada tipo de modelo. Con estos insumos se ha logrado efectuar un análisis comparativo de estos tipos de modelos (Sección 3.7), para su efecto se ha utilizado la síntesis de datos de Forest Plot considerando el nivel de precisión de clasificación de documentos de cada estudio. Para medir el rendimiento de cada agrupación se ha empleado el modelo de efecto randómico (RE) que toma como base la precisión de cada estudio del grupo, así se puede obtener también un rendimiento general de todos los modelos, el cual corresponde a un 0.80 del valor de desempeño de precisión, con un intervalo de confianza de (95 % CI [0.74 0.86]).

Entre los hallazgos más importantes que se han identificado en la estructura de ambientes de aprendizaje semi-supervisados se encuentra la semántica de características descrito en el apartado 3.2, varias investigaciones tales como [8] [19] [15] [14] centran la estructura de su modelo en torno a esta propiedad, ya que se constata que la determinación del significado de las palabras del documento mejora el rendimiento de categorización, es por esta razón que investigadores buscan ampliar y afinar esta cualidad en el proceso de clasificación. Así tenemos también a las particularidades de multivista y crowdsourcing descritas en el apartado 3.1 y 3.3 respectivamente, estas propiedades de los modelos semi-supervisados más allá de la demanda de costos de recursos, mejoran las métricas de rendimiento en la predicción de etiquetas como es el caso de [34] [43] [47]. Con respecto a las técnicas de mayor singularidad que se han identificado en la estructura de ambientes de aprendizaje semi-supervisados están, los algoritmos genéticos [50] , la lógica difusa [35] y las redes neuronales con arquitectura transformer [37] [34][39].

En relación al rendimiento de clasificación de los modelos se ha identificado un factor en común en todos sus tipos, el número de documentos etiquetados y el número de clases a categorizar son elementos incidentes en las métricas de precisión, en la Tabla 7 se puede apreciar que, a mayor cantidad de etiquetados o menor cantidad de clases, mejor es su nivel de precisión. No obstante, es importante considerar que las problemáticas de clasificación de documentos, los conjuntos de datos y recursos disponibles son diversos, sin embargo, los distintos tipos de modelos semi-supervisado han presentado flexibilidad ante los medios existentes. Por ejemplo, en el caso de existir recursos como son los etiquetadores, el aprendizaje activo los acopla a su estructura obteniendo buenos desempeños (0.89); y en el caso de no existir recursos los modelos de ensamblado se apoyan en clasificadores débiles para mejorar su etiquetado sin la necesidad de un etiquetador consiguiendo también índices representativos (0.83). Por otro lado considerando los conjuntos de datos, si existen fuentes externas de documentos ya clasificados y que puedan

ser reutilizados el modelo de aprendizaje de transferencia aprovecha esta información para su procesos de clasificación con resultados modestos (0.78); y en el caso de no disponer de fuentes externas tenemos al modelo de co-entrenamiento que aprovecha las diferentes vistas que se pueden extraer del conjunto de datos para su proceso para su categorización, con índices aceptables (0.79).

Es así como se ha determinado que los principales desafíos a los que se enfrenta este tipo de modelos son, la adaptación de dominio, el límite de decisión y el consenso entre clasificadores, cada uno puntualizado en la sección 3.4, 3.5 y 3.6 respectivamente. El límite de decisión es un factor recurrente en todos los tipos de modelos, varios estudios buscan límites más estables y menos propensos a errores, sin embargo, muchas de las propuestas son adecuadas para datos lineales [8], [19], [15] y [14] mientras que para los datos no lineales aún existe tarea pendiente. El consenso entre clasificadores es un problema que se ha identificado en técnicas donde se combinan varias predicciones de distintos clasificadores, si bien es cierto estas técnicas mejoran los niveles de precisión del modelo la contraparte es su redundancia y robustez [28] [29], por tal razón se busca eficiencia para este procedimiento. En cuanto a la adaptación de dominio, es una técnica que aprovecha el conocimiento propagado en la red para adaptarlo a un problema de clasificación en específico, sin embargo, el problema radica en que los dominios origen y destino pueden ser dispares [18] y [49], por esta razón es importante una adaptación previa de dominios para aprovechar de mejor manera el conocimiento y plantear un modelo eficiente.

Con el efecto de la presente SLR se ha identificado como trabajo futuro, el planteamiento de un método que gestione de forma eficiente el límite de decisión con datos no lineales; en el caso de consenso de clasificadores se podría buscar métricas para identificar los niveles de redundancia y buscar el consenso más eficiente; para la adaptación de dominio sería posible establecer un procedimiento para transferir de forma eficiente conocimiento de un lenguaje a otro. O también se puede establecer un modelo que focalice su desempeño en una representación de documentos cuyas características tengan una adecuada gestión del significado de las palabras controlando sinónimos y polisemia, así también buscar medir en el proceso de etiquetado los documentos más idóneos para reducir el margen de error al momento de generar nuevas etiquetas. En igual medida, se podría preparar un modelo que acorde a una problemática de clasificación de documentos planteada y la disponibilidad de sus recursos, busque el tipo de modelo semi-supervisado más eficiente para disponer de la mejor precisión en la categorización de su conjunto de documentos.

Apéndice A. Lista de acrónimos

AdaBoost	Adaptive Boosting	**N-gram**	Graph Representation Model
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
CI	Confidence intervals	**Pre-E**	Pre-entrenados
CNN	Convolutional Neural Networks	**P**	Precisión
CSA	Crow search algorithm	**RDT**	Random decision tree
CSA	Crow search algorithm	**RE**	Random effect
CsD	Clasificador débil	**RESSELL**	Reliable semi-supervised ensemble learning
CSW	Critical software	**RF**	Random Forest
CSWE	Cosine similarity weight Extraction	**RoBERTa**	Robustly Optimized BERT pretraining Approach
DPC	Density Peaks Clustering	**SCL**	Structural correspondence learning

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
E	Documentos Etiquetados	**SLLM**	Self learning linear mode
NE	Documentos No etiquetados	**SMDDRL**	Semi-supervised multi-view deep discriminant representation learning
ELMO	Embeddings from Language Model	**SNN**	Semantic convolution neural network
EM	Expectation Maximization	**SOM**	Self organizing map
FlexCon-C2	Flexible Confidence Classifier 2	**SOMVFV**	Semi-supervised One-pass Multi-View learning with variable Features and Views
FRBS	Fast radio bursts	**SSApolo**	Semi-supervised self-training method based on Apollonius
GBC	Gradient Boosting Classifier	**SSC/SCM**	Semantic similarity computation / Strong correlation method)
GNB	Gaussian Naive Bayes	**SSDTM**	Semi-supervised model based on dynamic threshold and multiple classifiers
GP	Genetic Programming	**SSKMS**	Semi-supervised k-means with seeds
HCSC	Hybrid Class Semantics Classifier	**SSMT**	Single source Multiple target domain
HMM	Hidden Markov Model	**SSOPMV**	Semi-supervised one pass multi view
HTF	Hidden feature transformation	**ssSCL-ST**	Semi-supervised learning with SCL and space transfer
IG	Information Gain	**STDP**	Self-Training with Density Peaks
KNN	K-nearest neighbors	**STDPNaN**	Self-training method based on density peaks and natural neighbors
LDA	Linear discriminant analysis	**STFW**	Static threat factor weight
LR	Logistic Regression	**SVM**	Support Vector Machine
LSTM	Long short-term memory	**TF-IDF**	Term Frequency-Inverse Document Frequency
MCT	Multi Co-training	**TrAdaBoost**	Transfer AdaBoost
MIL	Multi Instance Learning	**UCI**	University of California, Irvine
MLP	Multilayer perceptron	**VAE**	Variational autoencoder
MLSMOTE	Multilabel Synthetic Minority Over-sampling Technique	**W2V**	Word2Vec
mLVQb	Batch multi-label learning vector quantization	**WD1**	Weighted disagreement 1
MMC	Maximum model change	**WELM**	Weighted extreme learning machine
MMSL	Multi-model Sentiment Learning Layer	**WM**	Wikipedia Miner
MS	Margin sampling	**WMVC**	Weighted multi-view clustering
NB	Naive Bayes	**WSAL**	Warm Start Active Learning
NBoW	Neural Bag of Words		

Referencias

[1] Muthuraman Thangaraj and Muthusamy Sivakami. Text classification techniques: A literature review. *Interdisciplinary journal of information, knowledge, and management*, 13:117, 2018.

[2] A Bhavani and B Santhosh Kumar. A review of state art of text classification algorithms. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1484–1490. IEEE, 2021.

[3] Semuel Franko. Multiclass analysis of automatic text classification techniques. Master's thesis, Fen Bilimleri Enstitüsü, 2018.

[4] Kohei Watanabe and Yuan Zhou. Theory-driven analysis of large corpora: Semisupervised topic classification of the un speeches. *Social Science Computer Review*, 40(2):346–366, 2022.

[5] Protasiewicz Mironczuk. Mironczuk mm, protasiewicz j. *A recent overview of the state-of-the-art elements of text classification, Expert Systems With Applications*, 106:36–54, 2018.

[6] Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8558–8567, 2021.

[7] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020.

[8] Berna Altınel and Murat Can Ganiz. A new hybrid semi-supervised algorithm for text classification with class-based semantics. *Knowledge-Based Systems*, 108:50–64, 2016.

[9] YCAP Reddy, P Viswanath, and B Eswara Reddy. Semi-supervised learning: A brief review. *Int. J. Eng. Technol*, 7(1.8):81, 2018.

[10] Shrutika S Sawant and M Prabukumar. Semi-supervised techniques based hyper-spectral image classification: a survey. *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*, pages 1–8, 2017.

[11] Staffs Keele et al. Guidelines for performing systematic literature reviews in software engineering, 2007.

[12] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and the PRISMA Group*. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, 151(4):264–269, 2009.

[13] Mark Petticrew and Helen Roberts. *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons, 2008.

[14] Joanna Jedrzejowicz and Magdalena Zakrzewska. Text classification using lda w2v hybrid algorithm. In *Intelligent Decision Technologies 2019: Proceedings of the 11th KES International Conference on Intelligent Decision Technologies*, pages 227–237. Springer, 2020.

[15] Jawad Khan and Young-Koo Lee. Lessa: A unified framework based on lexicons and semi-supervised learning approaches for textual sentiment classification. *Applied Sciences*, 9(24):5562, 2019.

[16] Debaditya Barman and Nirmalya Chowdhury. A novel semi supervised approach for text classification. *International Journal of Information Technology*, 12:1147–1157, 2020.

[17] Poojitha Bikki. Machine learning for text categorization: experiments using clustering and classification. 2018.

[18] Hiroyuki Shinnou, Kanako Komiya, and Minoru Sasaki. Domain adaptation for document classification by alternately using semi-supervised learning and feature weighted learning. In *Computational Linguistics: 15th International Conference of the Pacific Association for Computational Linguistics, PACLING 2017, Yangon, Myanmar, August 16–18, 2017, Revised Selected Papers 15*, pages 205–216. Springer, 2018.

[19] Mona Emadi, Jafar Tanha, Mohammad Ebrahim Shiri, and Mehdi Hosseinzadeh Aghdam. A selection metric for semi-supervised learning based on neighborhood construction. *Information Processing & Management*, 58(2):102444, 2021.

[20] Junnan Li, Qingsheng Zhu, Quanwang Wu, and Dongdong Cheng. An effective framework based on local cores for self-labeled semi-supervised classification. *Knowledge-Based Systems*, 197:105804, 2020.

[21] Ning Chen, Bernardete Ribeiro, Chaosheng Tang, and An Chen. Multi-label learning vector quantization for semi-supervised classification. *Intelligent Data Analysis*, 23(4):839–853, 2019.

[22] Yue Han, Yuhong Liu, and Zhigang Jin. Sentiment analysis via semi-supervised learning: a model based on dynamic threshold and multi-classifiers. *Neural Computing and Applications*, 32:5117–5129, 2020.

[23] Suwen Zhao and Junnan Li. A semi-supervised self-training method based on density peaks and natural neighbors. *Journal of Ambient Intelligence and Humanized Computing*, 12:2939–2953, 2021.

[24] Karliane Medeiros Ovidio Vale, Arthur Costa Gorgônio, E Gorgônio Flavius Da Luz, and Anne Magály De Paula Canuto. An efficient approach to select instances in self-training and co-training semi-supervised methods. *IEEE Access*, 10:7254–7276, 2021.

[25] Massih-Reza Amini, Vasilii Feofanov, Loic Pauletto, Emilie Devijver, and Yury Maximov. Self-training: A survey. *arXiv preprint arXiv:2202.12040*, 2022.

[26] L Borrajo, A Seara Vieira, and Eva Lorenzo Iglesias. An hmm-based synthetic view generator to improve the efficiency of ensemble systems. *Logic Journal of the IGPL*, 28(1):4–18, 2020.

[27] Abir Masmoudi, Hatem Bellaaj, Khalil Drira, and Mohamed Jmaiel. A co-training-based approach for the hierarchical multi-label classification of research papers. *Expert Systems*, 38(4):e12613, 2021.

[28] Changming Zhu and Duoqian Miao. Semi-supervised one-pass multi-view learning with variable features and views. *Neural Processing Letters*, 50:189–226, 2019.

[29] Xiaodong Jia, Xiao-Yuan Jing, Xiaoke Zhu, Songcan Chen, Bo Du, Ziyun Cai, Zhenyu He, and Dong Yue. Semi-supervised multi-view deep discriminant representation learning. *IEEE transactions on pattern analysis and machine intelligence*, 43(7):2496–2509, 2020.

[30] Changming Zhu, Zhe Wang, Rigui Zhou, Lai Wei, Xiafen Zhang, and Yi Ding. Semi-supervised one-pass multi-view learning. *Neural Computing and Applications*, 31:8117–8134, 2019.

[31] Guruprasad Nayak, Rahul Ghosh, Xiaowei Jia, Varun Mithafi, and Vipin Kumar. Semi-supervised classification using attention-based regularization on coarse-resolution data. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 253–261. SIAM, 2020.

[32] Wenjuan Jia, Xiaodong Liu, Yuangang Wang, Witold Pedrycz, and Juxiang Zhou. Semisupervised learning via axiomatic fuzzy set theory and svm. *IEEE transactions on cybernetics*, 52(6):4661–4674, 2020.

[33] Donghwa Kim, Deokseong Seo, Suhyoun Cho, and Pilsung Kang. Multi co training for document classification using various document representations: Tf idf, lda, and doc2vec. *Information Sciences*, 477:15–29, 2019.

[34] Oduwa Edo-Osagie, Gillian Smith, Iain Lake, Obaghe Edeghere, and Beatriz De La Iglesia. Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance. *PloS one*, 14(7), 2019.

[35] Zahra Donyavi and Shahrokh Asadi. Diverse training dataset generation based on a multi–objective optimization for semi–supervised classification. *Pattern Recognition*, 108:107543, 2020.

[36] Majigsuren Enkhsaikhan. Geological knowledge graph construction from mineral exploration text. 2021.

[37] Eduardo de Souza Pais. Intelligent document validation intelligent document validation using natural language processing and computer vision. Master's thesis, 2021.

[38] Marcos Antonio Mouriño-García, Roberto Perez-Rodriguez, Luis Anido-Rifon, and Manuel Vilares-Ferro. Wikipedia-based hybrid document representation for textual news classification. *Soft Computing*, 22:6047–6065, 2018.

[39] Sohom Ghosh and Ankush Chopra. Using transformer based ensemble learning to classify scientific articles. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2021 Workshops, WSPA, MLMEIN, SDPRA, DARAI, and AI4EPT, Delhi, India, May 11, 2021 Proceedings 25*, pages 106–113. Springer, 2021.

[40] Marcos A Mouriño-García, Roberto Pérez-Rodríguez, and Luis E Anido-Rifón. A bag of concepts approach for biomedical document classification using wikipedia knowledge. *Methods of Information in Medicine*, 56(05):370–376, 2017.

[41] Marcos Antonio Mouriño García, Roberto Pérez Rodríguez, and Luis Anido Rifón. Leveraging wikipedia knowledge to classify multilingual biomedical documents. *Artificial intelligence in medicine*, 88:37–57, 2018.

[42] Hayder Mahmood Salman. Text classification based on weighted extreme learning machine. *Ibn AL-Haitham Journal For Pure and Applied Science*, 32(1):197–204, 2019.

[43] Akhilesh Kumar Shrivas, Amit Kumar Dewangan, SM Ghosh, and Devendra Singh. Development of proposed ensemble model for spam e-mail classification. *Information Technology and Control*, 50(3), 2021.

[44] Sjoerd de Vries and Dirk Thierens. A reliable ensemble based approach to semi-supervised learning. *Knowledge-Based Systems*, 215:106738, 2021.

[45] Oscar Reyes, Abdulrahman H Altalhi, and Sebastián Ventura. Statistical comparisons of active learning strategies over multiple datasets. *Knowledge-Based Systems*, 145:274–288, 2018.

[46] Yazhou Yang and Marco Loog. A benchmark and comparison of active learning for logistic regression. *Pattern Recognition*, 83:401–415, 2018.

[47] Mohamed-Rafik Bouguelia, Slawomir Nowaczyk, KC Santosh, and Antanas Verikas. Agreeing to disagree: Active learning with noisy labels without crowdsourcing. *International Journal of Machine Learning and Cybernetics*, 9:1307–1319, 2018.

[48] Ming Liu. *Weak Supervision and Active Learning for Natural Language Processing.* PhD thesis, Monash University, 2019.

[49] Shahd Alahdal. *Diary mining: predicting emotion from activities, people and places.* PhD thesis, Cardiff University, 2020.

[50] Wenlong Fu, Bing Xue, Xiaoying Gao, and Mengjie Zhang. Genetic programming based transfer learning for document classification with self-taught and ensemble learning. In *2019 IEEE Congress on Evolutionary Computation (CEC)*, pages 2260–2267. IEEE, 2019.

[51] Yi Zhu, Ehsan Shareghi, Yingzhen Li, Roi Reichart, and Anna Korhonen. Combining deep generative models and multi-lingual pretraining for semi-supervised document classification. *arXiv preprint arXiv:2101.10717*, 2021.

[52] Zichao Yang. Incorporating structural bias into neural networks for natural language processing, 2019.

[53] Mazen Mohammed, Lasheng Yu, Ali Aldhubri, and Gamil RS Qaid. Study on sentiment classification strategies based on the fuzzy logic with crow search algorithm. 2022.

[54] Shuo Yang, Ran Wei, Jingzhi Guo, and Hengliang Tan. Chinese semantic document classification based on strategies of semantic similarity computation and correlation analysis. *Journal of Web Semantics*, 63:100578, 2020.

[55] Zhengtong Pan, Patrick Soong, and Setareh Rafatirad. Ontology-driven scientific literature classification using clustering and self-supervised learning. In *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2022*, pages 133–155. Springer, 2022.

[56] Shun Guo and Nianmin Yao. Document vector extension for documents classification. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3062–3074, 2019.

[57] Farhan Hassan Khan, Usman Qamar, and Saba Bashir. Enhanced cross-domain sentiment classification utilizing a multi-source transfer learning approach. *Soft Computing*, 23:5431–5442, 2019.

[58] Deqing Wang, Junjie Wu, Jingyuan Yang, Baoyu Jing, Wenjie Zhang, Xiaonan He, and Hui Zhang. Cross-lingual knowledge transferring by structural correspondence and space transfer. *IEEE Transactions on Cybernetics*, 52(7):6555–6566, 2021.

[59] Xiaoting Du, Zenghui Zhou, Beibei Yin, and Guanping Xiao. Cross-project bug type prediction based on transfer learning. *Software Quality Journal*, 28:39–57, 2020.