ANÁLISIS DE LAS ESTRATEGIAS DE ETIQUETADO EN CONJUNTOS DE DATOS PARA CLASIFICACIÓN SUPERVISADA Y SEMI-SUPERVISADA: UNA REVISIÓN SISTEMÁTICA

Omam Sebastian Escalona Espinosa[1], Dra. Maria del Pilar Ortiz Vilchis[2],
Dr. Aldo Ramirez Arellano[1], Dra. Elena Fabiola Ruiz Ledemsa[3]

[1] Sección de Estudios de Posgrado e Investigación-Unidad Profesional Interdisciplinaria de Ingeniería y Ciencias Sociales y Administrativas. Instituto Politécnico Nacional,

[2] Sección de Estudios de Posgrado e Investigación-Escuela Superior Medicina. Instituto Politécnico Nacional,

[3] Sección de Estudios de Posgrado e Investigación-Escuela Superior de Cómputo. Instituto Politécnico Nacional
oescalone1500@alumno.ipn.mx, mportizv@ipn.mx, aramirezar@ipn.mx, eruizl@ipn.mx

Referencia de este artículo [1].

RESUMEN

La construcción de los conjuntos de datos etiquetados a utilizar en diferentes investigaciones es un aspecto importante, aunque poco reportado en algunos estudios de clasificación. La siguiente revisión sistemática analiza algunas de las estrategias de etiquetado empleadas en estudios publicados que comprenden de 2021 a 2026. Siguiendo el marco PICO y la metodología PRISMA, para este trabajo se buscó en las bases de datos IEEE Xplore, Springer Link y Scopus, identificando en total 7881 registros de los cuales 370 cumplieron los criterios de elegibilidad. Se seleccionó una muestra de 73 estudios. Los resultados encontrados muestran que las estrategias más frecuentes son: pseudo-labeling, weak supervision, active learning y etiquetado asistido por LLMs . Sin embargo, la mayoría de los estudios no describen completamente el proceso de etiquetado, limitando la reproducibilidad de los mimos. De una muestra de 20 artículos solo 9 estudios compararon explícitamente los diferentes enfoques. Esta revisión proporciona una visión estructurada parcial de las prácticas actuales de etiquetado e identifica posibles brechas que requieren atención en futuras investigaciones.

ABSTRACT

The construction of labeled datasets for use in different research projects is an important, though often underreported, aspect of classification studies. This systematic review analyzes some of the labeling strategies employed in published studies from 2021 to 2026. Following the PICO framework and the PRISMA methodology, the IEEE Xplore, Springer Link, and Scopus databases were searched, identifying a total of 7,881 records, of which 370 met the eligibility criteria. A sample of 73 studies was selected. The results show that the most frequent strategies are pseudo-labeling, weak supervision, active learning, and LLM-assisted labeling. However, most studies do not fully describe the labeling process, limiting their reproducibility. Of a sample of 20 articles, only 9 studies explicitly compared the different approaches. This review provides a partial, structured overview of current labeling practices and identifies potential gaps that require attention in future research.

Keywords: *Data annotation, Data labeling, Classification, Supervised learning, Semi-supervised learning, Annotated datasets, Dataset construction*
Palabras clave: *Anotación de datos, Etiquetado de datos, Clasificación, Aprendizaje supervisado, Aprendizaje semisupervisado, Conjuntos de datos anotados, Construcción de conjuntos de datos*

1. Introducción

Tanto el aprendizaje supervisado y semi-supervisado han demostrado alto rendimiento en tareas de clasificación automática de textos, imágenes y audio [2]. Sin embargo, su desempeño depende de la calidad de los conjuntos de datos etiquetados empleados [3].

Su construcción implica: Procesos costosos de selección de anotadores, medición de acuerdo inter-anotador y resolución de desacuerdos, estos aspectos se reportan de manera insuficiente en la literatura [3]. por lo que la falta de transparencia limita la reproducibilidad, además de dificultar la comparación entre enfoques y oculta posibles sesgos que podrían ser introducidos durante la anotación [3,4].

El presente trabajo de revisión tiene como objetivo analizar los métodos de etiquetado empleados en estudios de clasificación automática tanto supervisada como semi-supervisada publicados entre 2021 y 2026, identificando las prácticas empleadas en los estudios y posibles brechas en la construcción de los conjuntos de datos. Se empleo la técnica PICO [5] y la metodología PRISMA, se realizó una revisión sistemática en IEEE Xplore, Springer Link y Scopus. Se aplicaron criterios de inclusión y exclusión, y se analizó una muestra representativa de 73 artículos. Los hallazgos permitirán tener un panorama del estado actual del reporte de procesos de etiquetado, identificar las estrategias utilizadas entre ellas pseudo-labeling, weak-supervision, active learning y LLM-assisted labeling [6] y señalar áreas que requieren atención en futuras investigaciones.

2. Metodología

2.1 Formulación de la pregunta (PICO)

Para la formulación de la pregunta se utilizó la técnica PICO:

Tabla 1. Técnica pico.

Componente	Descripción
P (Problema)	Estudios de clasificación automática en diferentes tipos de datos
I (Intervención)	Estrategias de anotación y etiquetado utilizadas en los conjuntos de datos
C (Comparación)	Comparación entre diferentes enfoques de etiquetado
O (Outcome)	Identificación de las prácticas de etiquetado y su nivel de reporte en los estudios

Pregunta de investigación: ¿Cómo se construyen y describen las etiquetas en los conjuntos de datos utilizados en tareas de clasificación automática supervisada y semi-supervisada?

Objetivo: Analizar los métodos de etiquetado empleados en estudios de clasificación automática supervisada y semi-supervisada, con el fin de identificar las prácticas utilizadas en la construcción de sus conjuntos de datos.

2.2 Estrategia de búsqueda

Tabla 2. Consultas utilizadas en las bases de datos

Base de datos	Ecuación de búsqueda	Resultados iniciales	Tras filtros
IEEE Xplore	("All Metadata":"classification") AND ("All Metadata":"supervised learning" OR "All Metadata":"semi-supervised learning") AND ("All Metadata":"annotation" OR "All Metadata":"labeling" OR "All Metadata":"ground truth" OR "All Metadata":"pseudo-label" OR "All Metadata":"weak supervision" OR "All Metadata":"self-training")	6159	365
Springer Link	("text classification" OR "document classification") AND ("supervised learning" OR "semi-supervised learning") AND ("data annotation" OR "data labeling" OR "ground truth")	1472	282
Scopus	TITLE-ABS-KEY(("text classification" OR "document classification") AND ("supervised learning" OR "semi-supervised learning") AND ("annotation" OR "labeling" OR "ground truth")) AND PUBYEAR > 2020 AND PUBYEAR < 2027 AND (LIMIT-TO(DOCTYPE,"ar")) AND (LIMIT-TO(LANGUAGE,"English")) AND (LIMIT-TO(SRCTYPE,"j")) AND (LIMIT-TO(OA,"all"))	250	26

2.3 Criterios de selección

2.3.1 Criterios de inclusión:

- Artículos científicos publicados en revistas
- Publicados entre 2021 y 2026
- Escritos en idioma inglés
- De acceso abierto (Open Access)
- Que aborden tareas de clasificación automática
- Que empleen aprendizaje supervisado o semi-supervisado
- Que utilicen datos etiquetados o hagan uso de etiquetas (manuales o generadas automáticamente)

Se consideraron estudios de etiquetado manual con estrategias automáticas o semi-supervisadas, incluyendo técnicas como pseudo-labeling, self-training y weak supervision, debido a su relevancia en la generación y asignación de etiquetas en aprendizaje automático.

2.3.2 Criterios de exclusión:

- No aborden tareas de clasificación
- No utilicen etiquetas o datos anotados
- No empleen aprendizaje supervisado o semi-supervisado
- Sean editoriales, libros o tesis
- No presenten resultados experimentales
- No tengan acceso al texto completo
- No describan el proceso de construcción u anotación de etiquetas

3. Resultados

3.1 Identificación y selección de estudios

En la tabla 3 se puede observar el resumen numérico donde se obtuvo una muestra representativa de 73 del total de 370 estudios elegibles, se seleccionó una muestra representativa para el análisis detallado 20, priorizando aquellos que:

- Comparan explícitamente diferentes enfoques de etiquetado,
- Describen detalladamente el proceso de anotación, o
- Provienen de las diferentes bases de datos para garantizar diversidad.

Tabla 3. Resumen numérico por base de datos

Etapa	IEEE	Springer	Scopus	Total
Identificados	6159	1472	250	7881
Eliminados (inclusión)	5794	1190	224	7205
Examinados	**365**	**282**	**26**	**673**
Excluidos (título/abstract)	0	204	0	204
Recuperados (texto completo)	**365**	**78**	**26**	**469**
Excluidos (texto completo)	74	5	20	99
Elegibles	**282**	**78**	**6**	**370**
Incluidos (muestra)	**58**	**14**	**1**	**73**

Para la creación del diagrama de la figura 1 se consideraron como criterios de inclusión los artículos de revista de acceso abierto en inglés de 2021 a 2026. los registros examinados fueron aquellos que superaron los filtros automáticos de las bases de datos con las ecuaciones de búsqueda definidas. Los criterios de exclusión 1-5 son aquellos donde no se abordan clasificación, no usan etiquetas, no emplean aprendizaje supervisado y semi-supervisado, libros o tesis que no presentan resultados experimentales. Entre los criterios de exclusión 6-7 se excluyeron los que no describen el proceso de construcción de etiquetas o solo usan data sets pre-etiquetados sin describir su origen.
Finalmente, los estudios elegibles son el total de artículos que cumplen con todos los criterios de inclusión y ninguno de exclusión.

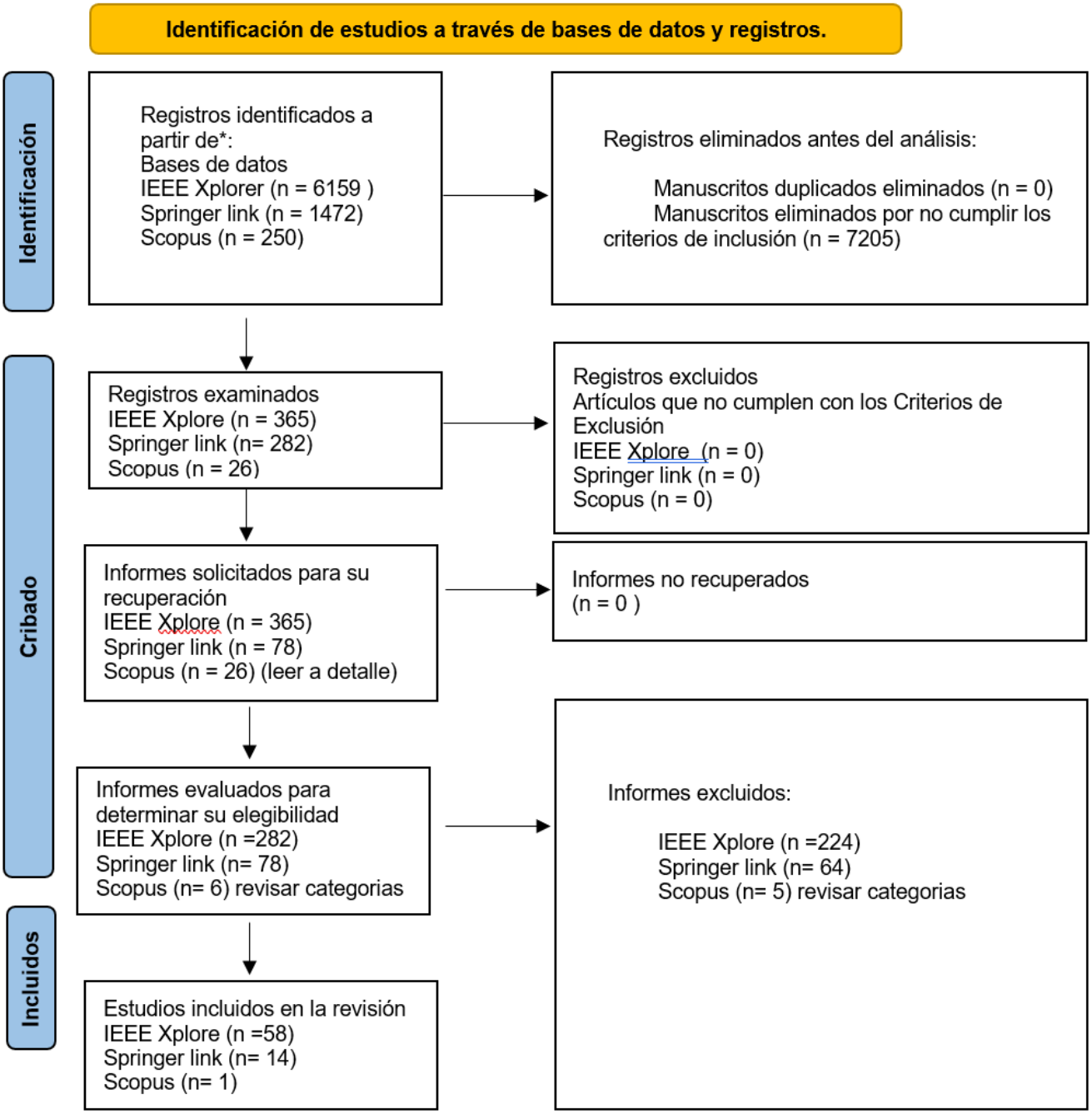

***Figura** 1. Diagrama de flujo adaptación de PRISMA*

3.2 Artículos incluidos (muestra de 20)

A continuación, en la tabla 4, se presentan los 20 artículos más representativos de la muestra, seleccionados por su relevancia en la comparación de enfoques de etiquetado además por la descripción detallada del proceso de anotación:

Tabla 4. Artículos representativos

#	Título abreviado	Primer autor	Base de datos
1	Pseudo-Labeling with LLMs for Multi-Label Emotion Classification	Malik [7]	IEEE
2	Transductive Text Classification with LLM-Assisted Labeling	Oliveira [8]	IEEE
3	Toward Integrating ChatGPT into Satellite Image Annotation	Beck [9]	IEEE
4	Weakly Supervised Deep Learning for Arabic Tweet Sentiment with Snorkel	Alotaibi [10]	IEEE

5	A Noise-Resilient Auto-Labeling Framework with Transition Matrix	Lee [11]	IEEE
6	Semi-Supervised Data Labeling for Elevator Safety	Rajibul Islam [12]	IEEE
7	Guiding Labelling Effort for Efficient Learning	Yamada [13]	IEEE
8	Contrastive Meta-Learner for Automatic Text Labeling	Cooper [14]	IEEE
9	Semi-Automatic Annotation of 3D Radar and Camera	Agrawal [15]	IEEE
10	Point-Supervised Facial Expression Spotting	Deng [16]	IEEE
11	Generating Labeled Training Datasets for NIDS	Ishibashi [17]	IEEE
12	Multiple Weak Supervision for Short Text Classification	Li-Ming Chen [18]	Springer
13	Prompt Selection Matters: Enhancing Text Annotations with LLMs	Abraham [19]	Springer
14	Comparing Active Learning Uncertainty Measures	Gonsior [20]	Springer
15	Ontology-driven Weak Supervision for Clinical Entity Classification	Jason Fries [21]	Springer
16	Multi-task Weak Supervision for Abnormality Detection	Sabri Eyuboglu [22]	Springer
17	Hybrid Sentiment Analysis with Active Learning on Greek Social Media	Kyriakos Skoularikis [23]	Springer
18	A comparative analysis of active learning for biomedical text mining	Usman Naseem [24]	Scopus
19	Semi-supervised self-training for COVID-19 misinformation detection	Siri Frisli [25]	Springer
20	Weakly Supervised Image Segmentation for Detecting Defects	Younghwan Lee [26]	IEEE

3.3 Estudios que comparan enfoques de etiquetado (componente C del PICO)

En la tabla 5 se muestra los 9 estudios que comparan explícitamente diferentes estrategias de etiquetado:

Tabla 5. Estudios con diferentes estrategias de etiquetado

#	Título	Publicación	Enfoque comparado
1	Hybrid sentiment analysis with domain-specific lexicons and active learning on Greek social media [23]	Soc Network Analysis	Lexicones vs Active Learning
2	Prompt selection matters: enhancing text annotations for social sciences with LLMs [19]	J Computational Soc Sci	Diferentes prompts para anotación
3	Comparing and Improving Active Learning Uncertainty Measures for Transformer Models [20]	Information Systems Frontiers	Diferentes medidas de incertidumbre
4	Ontology-driven weak supervision for clinical entity classification in EHRs [21]	Nature Communications	Weak supervision vs supervisado tradicional
5	Multi-task weak supervision enables anatomically-resolved abnormality detection [22]	Nature Communications	Weak supervision multi-tarea
6	Toward Integrating ChatGPT into Satellite Image Annotation Workflows [9]	IEEE JSTARS	Humano vs ChatGPT
7	A Real-Life Evaluation of Supervised and Semi-Supervised ML for Indoor Occupancy [12]	IEEE Access	Supervisado vs semi-supervisado
8	Comparing Active Learning for Satellite Imagery [26]	IEEE JSTARS	Diferentes estrategias de active learning
9	Active Learning for Imbalanced Classification [24]	IEEE Access	Diferentes estrategias de AL

3.4 Categorización de los estudios

A continuación, en la tabla 6, se presenta la categorización de los 9 artículos principales de la muestra según: tipo de aprendizaje, tipo de dataset y tipo de dato.

Tabla 6. Categorización según tipo

Artículo	Objetivo	Sup	SS	DS. No STD	DS STD	Tipo dato
Malik (LLM labeling)	LLM labeling	**X**	**X**	**X**		Texto
Oliveira (LLM labeling)	LLM labeling	**X**	**X**		**X**	Texto
Alotaibi (weak supervision)	Weak supervision		**X**	**X**		Texto
Lee (auto-labeling)	Auto-labeling	**X**		**X**		Imagen
Rajibul (semi-sup labeling)	Semi-sup labeling		**X**	**X**		Señal
Yamada (guiar etiquetado)	Guiar etiquetado		**X**		**X**	Imagen
Cooper (auto-labeling)	Auto-labeling	**X**			**X**	Texto
Agrawal (semi-auto annotation)	Semi-auto annotation		**X**	**X**		Imagen
Deng (point supervision)	Point supervision		**X**	**X**		Imagen

Legenda: STD. = Estandar; SS. = Semi-supervisado; Sup. = Supervisado; DS. = Data Set.

4. Discusión

Como se describió anteriormente en la presente revisión se identificó un total de 370 artículos que cumplían con los criterios de elegibilidad, de los cuales se seleccionó una muestra representativa de estudios para el análisis detallado donde se hizo un análisis extenso de 20 muestras. Los hallazgos preliminares indican que las estrategias de etiquetado más reportadas en la literatura que comprende de 2021 a 2026 incluye:

- Pseudo-labeling y self-training presentes en artículos como P-PseudoLabel, STDPboost.
- Weak supervision asistida por ontologías o heurísticas ejemplo Ontology-driven weak supervision, Snorkel
- Active learning para selección de muestras informativas como en Comparing Active Learning Uncertainty Measures
- Etiquetado asistido por grandes modelos de lenguaje (LLMs) ejemplo de ello Pseudo-Labeling with LLMs, Toward Integrating ChatGPT

Se observó que los estudios que utilizan data sets personalizados tienden a describir con mayor detalle el proceso de etiquetado en comparación con aquellos que emplean data sets estándar, los cuales frecuentemente omiten información sobre la construcción original de las etiquetas. Esta falta de reporte sistemático del proceso de anotación representa una limitación importante para la reproducibilidad y la comparación entre estudios.

En cuanto a la comparación de enfoques, se identificaron 9 estudios que comparan explícitamente diferentes estrategias de etiquetado, destacando aquellos que comparan anotación humana vs. asistida por LLM, y diferentes métodos de active learning. Estos estudios son valiosos para responder a la pregunta de investigación.

4.1 Limitaciones

Esta revisión presenta las siguientes limitaciones:

- Se limitó a artículos de acceso abierto publicados en inglés entre 2021-2026, lo que pudo haber excluido literatura relevante en otros idiomas o en fuentes no indexadas.
- La muestra representativa de 20 de 73 artículos (de 370 elegibles) pudo haber introducido un sesgo de selección se priorizaron criterios objetivos para mitigar lo más posible.
- La fase de extracción de datos se basó únicamente en la información reportada en los artículos, sin acceso a materiales suplementarios en algunos casos.

5. Conclusiones preliminares

En conclusión, la revisión sistemática permitió identificar, aunque limitada por la muestra de artículos las principales estrategias de etiquetado utilizadas en estudios de clasificación supervisada y semi-supervisada en los últimos 5 años. Se concluye preliminarmente que existe una tendencia creciente hacia el uso de estrategias semiautomáticas de etiquetado, en particular weak supervision y LLM-assisted labeling, que reducen la dependencia de anotación manual exhaustiva. La mayoría de los estudios no reportan de manera completa el proceso de construcción de etiquetas cuando no se usan data sets estándares esto dificulta la evaluación de la calidad de los data sets y la reproducibilidad de los resultados. Entre los estudios que comparan explícitamente diferentes enfoques de etiquetado los resultados son escasos, lo que representa una oportunidad para futuras investigaciones. Finalmente, como se infiere los data sets personalizados tienden a estar mejor documentados en términos de proceso de anotación que los data sets estándar.

5.1 Futuros trabajos

- Se recomienda a la comunidad científica estandarizar el reporte del proceso de construcción de data sets, incluyendo por mencionar algunos aspectos número de anotadores, métricas de acuerdo inter-anotador, estrategias de resolución de desacuerdos, y origen de los datos.
- Fomentar la comparación explícita de diferentes estrategias de etiquetado dentro de un mismo estudio.
- Promover el uso de técnicas semi-automáticas (weak supervision, active learning, LLM-assisted labeling y las que surjan) en contextos con recursos limitados para anotación.

6. Referencias

[1] Omam Sebastian Escalona Espinosa, Maria del Pilar Ortiz Vilchis, Aldo Ramirez Arellano, Elena Fabiola Ruiz Ledemsa **(julio – agosto, 2026)** *Análisis de las estrategias de etiquetado en conjuntos de datos para clasificación supervisada y semi-supervisada: una revisión sistemática. Boletín UPIITA. año 21, (115) 2026. ISSN 2007-6150.*

[2] LeCun, Y., Bengio, Y., & Hinton, G. ***(2015).*** *Deep learning. *Nature*, 521(7553), 436-444.*

[3] Northcutt, C. G., Athalye, A., & Mueller, J. ***(2021).*** *Pervasive label errors in test sets destabilize machine learning benchmarks.* *arXiv preprint arXiv:2103.14749*.

[4] Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., & Crawford, K. ***(2021).*** *Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92.*

[5] Schardt, C., Adams, M. B., Owens, T., Keitz, S., & Fontelo, P. ***(2007).*** *Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics and Decision Making*, 7(1), 16.*

[6] Bommasani, R., et al. ***(2021).*** *On the opportunities and risks of foundation models.* *arXiv preprint arXiv:2108.07258*.

[7] Malik, U., et al. ***(2024).*** *"Pseudo-Labeling With Large Language Models for Multi-Label Emotion Classification of French Tweets." IEEE Access, 12, 15902-15916.*

[8] Oliveira, V. V. d., et al. ***(2026).*** *"Transductive Text Classification With Concept Bipartite Graphs and LLM-Assisted Labeling." IEEE Access, 14, 63698-63715.*

[9] Beck, J., et al. ***(2025).*** *"Toward Integrating ChatGPT Into Satellite Image Annotation Workflows: A Comparison of Label Quality and Costs of Human and Automated Annotators." IEEE JSTARS, 18, 4366-4381.*

[10] Alotaibi, A., et al. ***(2025).*** *"Weakly Supervised Deep Learning for Arabic Tweet Sentiment Analysis on Education Reforms: Leveraging Pre-Trained Models and LLMs With Snorkel." IEEE Access, 13, 30523-30542.*

[11] Lee, W., & Hur, Y. ***(2025).*** *"A Noise-Resilient Auto-Labeling Framework With Transition Matrix." IEEE Access, 13, 185790-185801.*

[12] Rajibul Islam, M., et al. ***(2025).*** *"A Novel Approach to Elevator Safety: Semi-Supervised Data Labeling for Function Classification and Uneven Rope Tension Detection Using FBG Sensors." IEEE Access, 13, 191673-191688.*

[13]Yamada, T., et al. ***(2023).*** *"Guiding Labelling Effort for Efficient Learning With Georeferenced Images." IEEE TPAMI, 45(1), 593-607.*

[14] Cooper, R., et al. ***(2024).*** *"Contrastive Meta-Learner for Automatic Text Labeling and Semantic Textual Similarity." IEEE Access, 12, 166792-166799.*

[15] Agrawal, S., et al. ***(2024).*** *"Semi-Automatic Annotation of 3D Radar and Camera for Smart Infrastructure-Based Perception." IEEE Access, 12, 34325-34341.*

[16] Deng, Y., et al. ***(2026).*** *"Point-Supervised Facial Expression Spotting With Gaussian-Based Instance-Adaptive Intensity Modeling." IEEE TBIOM, 8(3), 378-391.*

[17]Ishibashi, R., et al. ***(2022).*** *"Generating Labeled Training Datasets Towards Unified Network Intrusion Detection Systems." IEEE Access, 10, 53972-53986.*

[18] Chen, L.-M., et al. ***(2022).*** *"Multiple weak supervision for short text classification". Applied Intelligence, 52, 9101-9116.*

[19] Abraham, L., Arnal, C., & Marie, A. ***(2025).*** *Prompt selection matters: enhancing text annotations for social sciences with large language models. Journal of Computational Social Science, 8:73.*

[20] Gonsior, J., Falkenberg, C., Magino, S., Reusch, A., Hartmann, C., Thiele, M., & Lehner, W. ***(2024).*** *Comparing and Improving Active Learning Uncertainty Measures for Transformer Models by Discarding Outliers. Information Systems Frontier.*

[21] Fries, J. A., Steinberg, E., Khattar, S., Fleming, S. L., Posada, J., Callahan, A., & Shah, N. H. ***(2021).*** *Ontology-driven weak supervision for clinical entity classification in electronic health records. Nature Communications, 12:2017.*

[22] Eyuboglu, S., Angus, G., Patel, B. N., Pareek, A., Davidzon, G., Long, J., Dunnmon, J., & Lungren, M. P. ***(2021).*** *Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body FDG-PET/CT. Nature Communications, 12:1880.*

[23] Skoularikis, K., Savvas, I., Garani, G., & Kakarontzas, G. ***(2026).*** *Hybrid sentiment analysis approach with domain-specific lexicons with active learning on Greek social media texts. Social Network Analysis and Mining, 16:35*

[24] Naseem, U., et al. ***(2021).*** *"A comparative analysis of active learning for biomedical text mining". Applied System Innovation, 4(1), 23*

[25] Frisli, S. ***(2025).*** *Semi-supervised self-training for COVID-19 misinformation detection: analyzing Twitter data and alternative news media on Norwegian Twitter. Journal of Computational Social Science, 8:39.*

[26] Lee, Y., & Kim, S. B. **(2024).** *"Weakly Supervised Image Segmentation for Detecting Defects From Scanning Electron Microscopy Images in Semiconductor." IEEE Access, 12, 184896-184908*.