

Scientia et PRAXIS

Vol. 06. No.11. Ene-Jun (2026): 1-23

<https://doi.org/10.55965/setp.6.11.a1>

eISSN: 2954-4041

Hacia una innovación de proceso de evaluación docente sostenible: Revisión Sistemática de Aplicaciones de IA, Ciencia de Datos y PLN en Educación Superior

Toward a Sustainable Innovation in the Faculty Evaluation Process: A Systematic Review of AI, Data Science, and NLP Applications in Higher Education

Cristian Ulises Barenca-Sotelo. ORCID [0009-0000-5091-1900](https://orcid.org/0009-0000-5091-1900)

Universidad de Guadalajara, Centro Universitario de Ciencias Económico Administrativas,
Zapopan, Jalisco, México

e-mail: cristian.barenca1897@alumnos.udg.mx

Ma. Del Rocío Maciel-Arellano. ORCID [0000-0002-5548-2073](https://orcid.org/0000-0002-5548-2073)

Universidad de Guadalajara, Centro Universitario de Ciencias Económico Administrativas,
Zapopan, Jalisco, México

e-mail: ma.maciel@academicos.udg.mx

Víctor Manuel Larios-Rosillo. ORCID [0000-0002-2899-724X](https://orcid.org/0000-0002-2899-724X)

Universidad de Guadalajara, Centro Universitario de Ciencias Económico Administrativas,
Zapopan, Jalisco, México

e-mail: victor.larios@academicos.udg.mx

Palabras Clave: evaluación docente, educación superior, inteligencia artificial, procesamiento de lenguaje natural, revisión sistemática.

Keywords: faculty evaluation, higher education, artificial intelligence, natural language processing, systematic review.

Recibido: 12-Nov-2025; **Aceptado:** 22-Feb-2026

RESUMEN

Contexto. La evaluación docente en universidades públicas multicampus depende aún de promedios y encuestas cerradas, limitadas para captar la complejidad del desempeño académico. Ante ello, este estudio presenta una revisión sistemática (2019–2024) sobre el uso de la Inteligencia Artificial (IA) y el Procesamiento de Lenguaje Natural (PLN) en procesos de evaluación docente universitaria, siguiendo como innovaciones de proceso enunciadas en el Manual de Oslo, alineadas con los Objetivos de Desarrollo Sostenibles (ODS) 4 y 9.

Problema. Los sistemas tradicionales en instituciones de alta matrícula presentan baja precisión analítica, escaso aprovechamiento de datos cualitativos y retroalimentación tardía. Ante esto, se plantea la pregunta de investigación: ¿qué técnicas y algoritmos reporta la literatura para el análisis integrado de datos cuantitativos y cualitativos en la evaluación docente universitaria?

Objetivo. Analizar críticamente el uso de la ciencia de datos y la IA —especialmente el PLN— para mejorar la retroalimentación docente, superando las limitaciones de los enfoques tradicionales y promover procesos evaluativos oportunos, profundos y escalables.

Metodología. Se siguió el protocolo PRISMA mediante búsquedas con operadores booleanos en Scopus, Web of Science, IEEE Xplore y ERIC. Tras aplicar criterios de inclusión y un proceso de revisión por pares en dos fases, se analizaron 17 estudios publicados entre 2019 y 2024.

Hallazgos. Las técnicas como el análisis de sentimientos, el modelado de tópicos (LDA) y modelos de lenguaje de gran escala (LLM) —notablemente DistilBERT, con precisiones cercanas al 93%— superan consistentemente a los métodos tradicionales en la gestión de grandes volúmenes de información.

Originalidad. Radica en la integración de literatura dispersa sobre IA y PLN en educación superior bajo un marco de innovación de procesos, articulando rigor técnico, empírico y ético.

Conclusiones y Limitaciones. Se concluye que los modelos avanzados de IA y PLN poseen un alto potencial para transformar la evaluación en redes universitarias latinoamericanas al permitir retroalimentación personalizada. No obstante, persisten desafíos en la interpretabilidad de modelos, mitigación de sesgos, escasez de datos etiquetados en español y resistencia institucional, lo que abre relevantes líneas para futuras investigaciones y desarrollos aplicados.

ABSTRACT

Context. Teaching evaluation in public multicampus universities still relies on averages and closed-ended surveys, which are limited in capturing the complexity of academic performance. In response, this study presents a systematic review (2019–2024) on the use of Artificial Intelligence (AI) and Natural Language Processing (NLP) in university teaching evaluation processes, conceptualized as process innovations from Oslo Manual and aligned with Sustainable Development Goals (SDGs) 4 and 9.

Problem. Traditional systems in high-enrollment institutions exhibit low analytical precision, limited use of qualitative data, and delayed feedback. Accordingly, the following research question is posed: which techniques and algorithms does the literature report for the integrated analysis of quantitative and qualitative data in university teaching evaluation?

Purpose. To critically analyze the use of data science and AI—particularly NLP—to enhance teaching feedback, overcoming the limitations of traditional approaches and promoting timely, in-depth, and scalable evaluation processes.

Methodology. The PRISMA protocol was followed through searches using Boolean operators in Scopus, Web of Science, IEEE Xplore, and ERIC. After applying inclusion criteria and a two-phase peer-review process, 17 studies published between 2019 and 2024 were analyzed.

Findings. Techniques such as sentiment analysis, topic modeling (LDA), and large language models (LLMs)—notably DistilBERT, with accuracy levels close to 93%—consistently outperform traditional methods in managing large volumes of information.

Originality. The study’s originality lies in integrating dispersed literature on AI and NLP in higher education within a coherent process-innovation framework, combining technical, empirical, and ethical rigor.

Conclusions and Limitations. Advanced AI and NLP models show high potential to transform evaluation in Latin American university networks by enabling personalized feedback. However, challenges remain regarding model interpretability, bias mitigation, the scarcity of labeled data in Spanish, and institutional resistance to change, opening relevant avenues for future research and applied developments.

1. INTRODUCCIÓN

En las instituciones de educación superior de América Latina, particularmente en universidades públicas de gran escala organizadas en redes universitarias y con estructuras multicampus, la evaluación del desempeño docente constituye un proceso estratégico para el aseguramiento de la calidad educativa. No obstante, dichos procesos continúan apoyándose mayoritariamente en modelos tradicionales basados en promedios numéricos y cuestionarios cerrados, los cuales resultan insuficientes para capturar la complejidad de la práctica docente y para aprovechar de manera sistemática la información cualitativa generada por el estudiantado. En este contexto, universidades con altos volúmenes de matrícula y una amplia diversidad disciplinaria —como aquellas que conforman la Red Universitaria de la Universidad de Guadalajara (**UDG**)— enfrentan desafíos significativos asociados al procesamiento masivo de datos, la oportunidad de la retroalimentación y la toma de decisiones informadas.

Centros universitarios como el de Ciencias Económico Administrativas (**CUCEA**) ilustran estas tensiones, al operar sistemas de evaluación que deben atender millones de registros por ciclo académico, situación que refleja problemáticas comunes a numerosas instituciones latinoamericanas y pone de manifiesto la necesidad de transitar hacia modelos de evaluación docente apoyados en innovaciones de proceso basadas en ciencia de datos, Inteligencia Artificial y Procesamiento de Lenguaje Natural.

La evaluación docente en el ámbito universitario ha dependido históricamente de métodos tradicionales que priorizan promedios simples y encuestas de preguntas cerradas, si bien estos enfoques permiten generar métricas básicas, resultan insuficientes para capturar la complejidad del desempeño académico y ofrecer retroalimentación significativa y oportuna.

Asimismo, en la **UDG**, y particularmente en el **CUCEA**, estas limitaciones se evidencian en el procesamiento de más de 2'354,631 registros por ciclo, lo que retrasa la toma de decisiones pedagógicas y conduce a análisis superficiales que no reflejan el desempeño real del profesorado.

Ahora bien, frente a estas deficiencias, la **IA**, Ciencia de Datos y el **PLN** emergen como enfoques innovadores de procesos organizacionales, según el Manual de Oslo, ya que estas tecnologías permiten automatizar el análisis de grandes volúmenes de datos y transformar comentarios cualitativos de estudiantes en información útil mediante técnicas como análisis de sentimientos y modelos de lenguaje (**LLM**), proporcionando retroalimentación casi inmediata y

personalizada. Así, la implementación de **IA** y **PLN** no solo representa una mejora tecnológica, sino también un cambio organizacional que optimiza la gestión de la evaluación docente y la toma de decisiones institucionales. Además, esta innovación se vincula directamente con los **ODS**, especialmente con el **ODS 4** (Educación de Calidad), al elevar la precisión y utilidad de la retroalimentación docente y fomentar prácticas educativas más efectivas, y con el **ODS 9** (Industria, Innovación e Infraestructura), al fortalecer la infraestructura digital y las capacidades institucionales para adoptar tecnologías sostenibles y escalables en educación superior.

Por otro lado, el presente artículo realiza una revisión sistemática de investigaciones publicadas entre 2019 y 2024, con el objetivo de analizar críticamente el uso de ciencia de datos e **IA**, con énfasis en **PLN**, aplicado a la retroalimentación docente personalizada.

Por lo anterior, para guiar este análisis se plantean tres preguntas de investigación: (1) ¿Qué técnicas y algoritmos se reportan para el análisis de datos cuantitativos y cualitativos en la evaluación docente?; (2) ¿Cómo se validan estos modelos y qué nivel de efectividad muestran frente a los métodos tradicionales?; y (3) ¿Cuáles son los beneficios percibidos y los principales desafíos técnicos, éticos e institucionales asociados con su implementación? Las respuestas a estas preguntas se presentan en la **sección 5.1 Síntesis de respuestas** a las preguntas de investigación de este documento. Asimismo, esta revisión sistemática ofrece una síntesis integral que permite mapear las soluciones tecnológicas actuales y evaluar su impacto para modernizar la evaluación docente y promover la mejora continua en el **CUCEA**, constituyendo un aporte relevante a la innovación educativa sostenible.

2. CONTEXTUALIZACIÓN

A continuación, se presenta la contextualización del uso de la **IA** en la evaluación docente universitaria, enmarcando la problemática de los sistemas tradicionales y la transición global hacia soluciones basadas en Ciencia de Datos y **PLN**.

2.1. Mundial

A nivel mundial, el panorama de la educación superior se ha caracterizado históricamente por sistemas de evaluación docente dominados por métodos tradicionales que priorizan promedios

simples y encuestas de preguntas cerradas, estableciendo que estos enfoques resultan insuficientes para capturar la complejidad del desempeño y proporcionar retroalimentación oportuna mientras que, por otro lado, se observa una rápida adopción de la **IA** y el Aprendizaje Automático (**ML**) en el ecosistema educativo para el análisis de sentimientos y la evaluación del desempeño docente (Okoye et al., 2022, 2023).

Ahora bien, las fuentes revelan una tendencia clara hacia la superioridad de los modelos avanzados basados en **PLN** y Aprendizaje Profundo (**DL**) demostrando que técnicas **LLMs** como **BERT** y **DistilBERT**, el análisis de sentimientos y el modelado de tópicos (**LDA**) superan ampliamente a los métodos tradicionales en velocidad, precisión y profundidad analítica para procesar grandes volúmenes de retroalimentación textual (Shuqin & Raga, 2024).

2.2. Internacional

En el ámbito internacional, las instituciones de educación superior funcionan como instituciones productoras de conocimiento debido a que las tendencias globales de evaluación giran en torno a políticas convergentes como el crecimiento de la matrícula, la descentralización administrativa y una mayor competencia por recursos (Martínez-Téllez & Celaya-Lozano, 2023).

En cuanto a la adopción de **IA**, la literatura internacional demuestra un interés creciente, con estudios empíricos en diversas regiones:

- China y Australia han desarrollado modelos híbridos avanzados (como **BERT-BiLSTM-Attention** y **CNN-BiLSTM**) para clasificar emociones complejas en comentarios estudiantiles con alta precisión hasta, 0.9664 **F1-score** (Shuqin & Raga, 2024).
- En Europa y Estados Unidos (**E.U.A.**), se han empleado métodos como **VADER** y análisis léxico para clasificar emociones y habilidades cognitivas (Taxonomía de Bloom), buscando una comprensión más rica del contexto pedagógico (Resuello et al., 2024).
- América Latina, se reporta una actividad emergente liderada por instituciones en Colombia y México, como la Universidad del Valle y la Pontificia Universidad Javeriana, las cuales han comenzado a adoptar **LLMs** y técnicas de **PLN** preentrenados para español —como **Py-sentimiento**—, logrando precisiones del 93% en el análisis de comentarios (Peña-Torres, 2024). No obstante, persiste una brecha investigativa regional marcada por la escasez de datos etiquetados en el idioma local y la fragmentación de los estudios y (Okoye et al., 2022).

2.3. País o localidad.

En México, la evaluación de la enseñanza ha evolucionado mediante herramientas tecnológicas que permiten medir con mayor precisión el desempeño docente y la satisfacción estudiantil (Arriaga-Cárdenas et al., 2022). Se han procesado 471,968 comentarios abiertos en español mediante minería de texto, empleando modelos como **EPDM** para la extracción de sentimientos y valencias emocionales, junto con algoritmos de aprendizaje automático como K vecinos más cercanos (**KNN**) y Máquinas de Vector de Soporte (**SVM**) para predecir recomendaciones docentes basadas en género, emociones y factores pedagógicos integrando métricas estadísticas avanzadas como **MANCOVA**, **ANCOVA**, Kruskal-Wallis, precisión, sensibilidad y curvas (**ROC**), además de modelos de **PLN** que van desde **SVM** hasta **DistilBERT** con precisiones del 93% en español (Okoye et al., 2022, 2023).

Por otro lado, en el **CUCEA**, los sistemas tradicionales de evaluación presentan limitaciones críticas, ya que el procesamiento manual de más de 2'354,631 de registros por ciclo escolar retrasa la entrega de retroalimentación en formato **PDF**, lo que dificulta la toma de decisiones pedagógicas oportunas y evidencia la necesidad de implementar estas soluciones tecnológicas (Santamaría-Velasco, 2024). Por lo anterior, la modernización de los procesos de evaluación requiere no solo herramientas tecnológicas, sino también estrategias para balancear comentarios, evitar sesgos y conformar equipos interdisciplinarios con conocimientos en **IA**, análisis de datos y pedagogía (Vera-Noriega, 2023). Así, el **CUCEA** puede pasar de un análisis superficial a la generación de retroalimentación accionable y personalizada, mejorando la enseñanza y la experiencia estudiantil.

3. REVISIÓN DE LA LITERATURA

La presente sección expone de manera integrada el estado del arte y la base teórica que sustentaron el uso de modelos de Inteligencia Artificial (**IA**) y Procesamiento de Lenguaje Natural (**PLN**) en la evaluación docente universitaria, así como las hipótesis derivadas de la revisión. Para ello se realizó un análisis sistemático de estudios empíricos publicados entre 2019 y 2024, periodo considerado clave para capturar los avances recientes en aprendizaje profundo, modelos transformadores y sistemas de análisis automatizado de texto aplicados a la retroalimentación

educativa, por lo que siguiendo el protocolo **PRISMA**, se revisaron 17 estudios seleccionados de **Scopus**, **Web of Science**, **IEEE Xplore** y **ERIC** donde en la **Tabla 1** se muestra sistematizadamente la relación entre los estudios revisados y su procedencia.

Tabla 1. Caracterización del corpus de estudio

Estudio	País	Metodología / Modelo IA	Hallazgo Principal
Pramod et al. (2022)	India	Análisis de sentimientos y Word Cloud en texto y Random Forest como modelo predictivo, destacando modelos de conjunto.	Tras comparar diversos algoritmos (como Naïve Bayes, SVM y Regresión Logística), el modelo RandomForestClassifier fue seleccionado como el más efectivo para predecir la eficacia docente.
Hu et al. (2020)	China	Análisis de sentimientos con SVM (kernels lineal, polinómico, RBF); preprocesamiento con word2vec, K-means para clustering y TFIDF para ponderar características.	El modelo SVM utilizando un núcleo de base radial (RBF) es el método más eficaz y estable para clasificar sentimientos en textos de evaluación docente F1 90.27%
Esparza et al. (2018)	México	Análisis de sentimiento con SVM usando kernels lineal, RBF y polinomial, aprovechando su alto rendimiento en clasificación.	El algoritmo de Máquinas de Vectores de Soporte (SVM) con un kernel lineal obtuvo los mejores resultados, logrando una exactitud equilibrada (balanced accuracy) de 0.8149
Peña-Torres et al. (2024)	Colombia	Análisis de sentimiento con NLP usando LLMs preentrenados:Py-sentimiento (Robertuito/RobERTa), versión personalizada de VADER y modelo DistilBERT .	El modelo basado en Distilbert alcanzó una métrica del 93% de precisión en la clasificación de la polaridad de los comentarios estudiantiles.
Nikolić et al. (2020)	Serbia	Sistema ABSA con SVM lineal para aspectos y sentimientos, Bag-of-Words con TFIDF, diccionarios y reglas basadas en expresiones regulares.	Se logró un F-measure promedio de 0.90. De manera específica, el sentimiento negativo se detectó con un F-measure de 0.94, mientras que el positivo alcanzó 0.83.
Okoye et al. (2022)	México	Se aplicó el modelo EPDM de minería de texto para extraer sentimientos y valencias emocionales de los estudiantes.	Las métricas de rendimiento del modelo alcanzaron valores óptimos de 1.00 (100%) en precisión, exhaustividad (recall), especificidad, exactitud (accuracy) y puntuación F1 , con una tasa de error de 0.00
H. Shuqin & R. C. Raga, (2024)	China	BERT para representaciones contextuales profundas. BiLSTM para captar dependencias bidireccionales. Atención para enfocar la información relevante.	Los modelos superaron en rendimiento a otros enfoques de aprendizaje profundo existentes en el análisis de sentimientos de evaluaciones de cursos. • Exactitud (Accuracy): 0.9608. • Precisión (Precision): 0.9589. • Valor F1 (F1-score) : 0.9664.
Wu et al. (2023)	EEUU	Análisis de sentimiento con VADER .	Al añadir el análisis de sentimiento a la evaluación, la correlación de Pearson con las calificaciones reales de los estudiantes aumentó a 0.16.

Fargues et al. (2023)	Noruega	Se aplicó NLP usando análisis léxico: VADER para sentimientos, NRC para emociones generales y un diccionario personalizado para seis emociones educativas.	Se utilizó una biblioteca de PLN para clasificar el texto en una escala de polaridad: positivo (0.05 a 1), negativo (-0.05 a -1) y neutral (-0.05 a 0.05)
Peng et al. (2022)	China	Modelo de análisis de sentimientos que combina CNN , BLSTM y atención, usando Softmax para clasificar emociones complejas y optimizado con descenso de gradiente mini-batch.	El modelo propuesto obtuvo un F1 de 0.845 y un mínimo de 0.748 en los mecanismos de atención.
Shaik et al. (2022)	Australia	IA y PNL : preprocesamiento, TextBlob para sentimiento, Word2Vec para vectores y un modelo Bi-LSTM para clasificar comentarios en 19 subdominios de Biggs.	En estudios relacionados citados, se reportó un F1-score de 86.13% para la identificación de categorías de aspectos y un 82.10% para su clasificación de sentimiento.
Pu et al. (2021)	China	Modelo que combina Word2Vec y GloVe para representar palabras, usa CNN y BiLSTM para extraer características, y optimiza un ensemble con el algoritmo MOGWO .	Se obtuvo un F1 superior al 91% (específicamente 0.9184), superando a otros siete modelos de comparación como SVM , RNN y CNN tradicionales.
Resuello et al. (2024)	Filipinas	Se aplicaron técnicas de minería de texto, NLP , VADER y LDA .	Se alcanzó un grado de acuerdo del 70.14% entre VADER y TextBlob en los comentarios negativos
Okoye et al. (2023)	México	Se aplicó el modelo EPDM de minería de texto para extraer sentimientos y valencias emocionales de los estudiantes.	para cuantificar datos textuales, detectando que la confianza (~25%) fue la emoción predominante en las evaluaciones a instructores, mientras que la preocupación (~20%) destacó en las recomendaciones entre estudiantes.
Das et al. (2022)	India	Regresión Logística, XGBoost y Árboles de Decisión.	El valor más alto alcanzado fue un F1-score de 0.728 mediante la combinación de Regresión Logística y TF-IDF .
Yuan & Hernandez, (2023)	Filipinas	Bi-LSTM .	Las métricas clave obtenidas por este modelo son: • Exactitud de validación: Alcanzó un 90.77%,. • Exactitud de prueba: Logró un 84.33%,.
Ponmalar et al. (2024)	India	PLN y BERT .	Se evaluaron las respuestas descriptivas en educación en línea con una precisión promedio del 90.6% (en un rango del 87% al 95%)

Fuente: Elaboración propia basada en la caracterización del corpus de estudio.

Ahora bien, el análisis bibliométrico mostró un predominio de investigaciones orientadas a métricas técnicas (**precisión** y **F1-score**), con un desarrollo acelerado en Asia y Norteamérica,

mientras que los estudios en contextos hispanohablantes permanecieron limitados, fragmentados y con escasez de corpus etiquetados en español. Por otro lado, la exclusión de artículos en idiomas distintos al inglés y español constituyó un sesgo recurrente, y resaltó la limitada diversidad lingüística de la literatura disponible.

Históricamente, la evaluación docente se fundamentó en instrumentos tradicionales basados en encuestas cerradas y promedios simples, enfoques que resultaron insuficientes para capturar la complejidad del desempeño y la diversidad de percepciones estudiantiles (Shaik et al., 2022). En contraste, el estado del arte mostró que las técnicas modernas de **PLN** superaron estos métodos en precisión, velocidad y profundidad analítica, ya que modelos clásicos como Máquinas de Vector de Soporte (**SVM**) se aplicaron en estudios Mexicanos para clasificar comentarios en español, aunque su desempeño varió según el kernel utilizado (Okoye et al., 2022).

Sin embargo, los métodos basados en aprendizaje profundo demostraron superioridad: por ejemplo, el modelo híbrido representaciones contextuales bidireccionales con redes de memoria a largo y corto plazo y mecanismos de atención (**BERT-BiLSTM-Attention**) en China alcanzó la métrica **F1-score** de 0.9664, y el uso de las Redes de Memoria a Largo y Corto Plazo Bidireccionales (**Bi-LSTM**) en Filipinas superó el 90% de precisión en tendencias emocionales, donde particularmente relevantes fueron los Modelos de Lenguaje de Gran Escala **LLMs** adaptados al español.

Asimismo, en un estudio comparativo, la versión destilada y multilingüe de **BERT multilingual-DistilBERT** logró un 93% de precisión, superando ampliamente a enfoques léxicos como el Diccionario y Razonador de Sentimientos Consciente de la Valencia (**VADER**) de 0.79, lo que evidenció su capacidad predictiva y comprensión contextual avanzada (Shuqin & Raga, 2024).

Por otro lado, la literatura también destacó herramientas de mayor profundidad analítica, por ejemplo, el modelado de tópicos mediante la Asignación Latente de Dirichlet (**LDA**) se utilizó para vincular temas subyacentes en grandes volúmenes de texto cualitativo, mientras que el Análisis de Sentimiento Basado en Aspectos (**ABSA**) permitió detectar simultáneamente el aspecto educativo mencionado y la polaridad asociada, ofreciendo un entendimiento más fino de la percepción estudiantil (Nikolić et al., 2020).

Otros trabajos clasificaron comentarios según la **Taxonomía de Bloom**, ampliando el valor pedagógico del análisis cualitativo (Fargues et al., 2023). Asimismo, modelos predictivos integrados en sistemas de soporte a la decisión, como el Modelo de Descubrimiento de Parámetros Emocionales más Aprendizaje Automático (**EPDM+ML**), mostraron capacidad para anticipar niveles de satisfacción estudiantil y apoyar procesos de mejora institucional (Okoye et al., 2022). A pesar de sus ventajas, el estado del arte señaló desafíos relevantes, una baja interpretabilidad de los modelos de aprendizaje profundo, la presencia de sesgos algorítmicos (por ejemplo, de género), la resistencia institucional a la adopción tecnológica y la ausencia de validaciones longitudinales que evaluaran el impacto sostenido de estas herramientas (**LDA**, **ABSA** y **EPDM+ML**). Además, se subrayó también la necesidad de complementar los modelos con validación humana para fortalecer su legitimidad y utilidad educativa (Peña-Torres, 2024).

Por lo tanto, a partir del análisis sistemático de la literatura, se identifican patrones de rendimiento que permiten formular los supuestos de esta investigación. En la **Tabla 1** se sistematiza los **17 estudios que conforman el corpus de esta revisión**, detallando el modelo utilizado y su precisión, lo que sirve de base empírica para las **hipótesis** planteadas a continuación:

H1: “Basado en los resultados de Shuqin & Raga (2024) y la comparativa de modelos multilingües analizada en la **Tabla 1**, se postula que los modelos de aprendizaje profundo basados en transformadores (como **DistilBERT**) adaptados al español, ofrecen una precisión analítica superior al 90% en la clasificación de sentimientos, superando significativamente la capacidad de los métodos léxicos tradicionales como **VADER** (0.79).”

H2: “A partir de la capacidad demostrada por las técnicas de **ABSA** y **LDA** para desglosar dimensiones cualitativas (Nikolić et al., 2020), se asume que su integración permite identificar no solo la polaridad, sino los aspectos específicos del proceso educativo, generando una retroalimentación más detallada y accionable que los promedios numéricos simples.”

H3: “Considerando que los modelos de **IA / PLN** reducen los tiempos de procesamiento en grandes volúmenes de texto (Okoye et al., 2022), se hipotetiza que su implementación en el contexto del **CUCEA** permitirá procesar de forma automatizada los más de 2’354,631 de registros por ciclo, transformando datos masivos en decisiones oportunas para la mejora docente.”

Donde para el análisis de las hipótesis planteadas, se establecieron criterios analíticos de comparación basados exclusivamente en la revisión sistemática reportada y el análisis comparativos de la literatura. En el caso de la **H1**, la comparación se sustenta en la precisión analítica reportada y en el desempeño relativo de los modelos de aprendizaje profundo basados en transformadores frente a métodos léxicos tradicionales. Por el contrario, en la **H2** se analiza a partir del nivel de granularidad del análisis, considerando la capacidad de distintas técnicas para identificar dimensiones cualitativas y aspectos específicos más allá de la polaridad global. Finalmente, la **H3** se fundamenta en la capacidad de procesamiento y escalabilidad de los modelos de **IA** y **PLN**, evaluando su potencial para automatizar el análisis de grandes volúmenes de texto y apoyar la toma de decisiones oportunas en contextos institucionales, sin introducir nuevas variables ni mediciones propias.

4. METODOLOGÍA

Esta revisión sistemática se diseñó siguiendo un enfoque riguroso y transparente, con el objetivo de identificar, seleccionar, analizar e interpretar estudios relevantes sobre el uso de técnicas de **IA**, particularmente aquellas basadas en **NLP**, para la retroalimentación docente en el contexto de la educación superior, por lo tanto, a continuación, se describen los métodos empleados para garantizar la validez y reproducibilidad del proceso.

4.1 Estrategia de búsqueda

Se definió una estrategia de búsqueda estructurada mediante operadores Booleanos y de proximidad, con el fin de capturar de manera precisa los estudios más relevantes, por lo que las cadenas de búsqueda fueron diseñadas para cubrir cuatro dimensiones clave: Evaluación docente, técnicas de **PLN**, métodos de **IA** y contexto universitario. Un ejemplo de cadena utilizada es:

("teacher evaluation" OR "faculty evaluation" OR "teaching feedback") AND ("natural language processing" OR "text mining" OR "sentiment analysis") AND ("machine learning" OR "artificial intelligence") AND ("higher education" OR "university").

Estas búsquedas se llevaron a cabo en bases de datos académicas como: **Scopus**, **Web of Science**, **IEEE Xplore** y **ERIC**, considerando un rango temporal de 2019 a 2024, con el objetivo

de capturar tanto los desarrollos más recientes como estudios relevantes que sirvan como antecedentes clave en el área.

4.2. Criterios

Por otro lado, se establecieron criterios claros de inclusión y exclusión para asegurar la pertinencia y calidad de los estudios seleccionados: Criterios de inclusión se incluirán estudios empíricos (cuantitativos, cualitativos o mixtos) que describan el desarrollo, implementación o evaluación de modelos o sistemas de **IA** aplicados específicamente a la retroalimentación docente en contextos universitarios y criterios de exclusión: Se excluirán artículos que no estén escritos en inglés o español, estudios enfocados en educación primaria o secundaria, artículos de opinión, revisiones teóricas o sin datos empíricos, así como modelos aplicados a otros ámbitos que no involucren directamente la evaluación docente universitaria.

4.3. Fases

El proceso de selección se realizó en dos fases sucesivas, primero, se examinaron los títulos y resúmenes de los estudios recuperados para filtrar aquellos que cumplen con los criterios de elegibilidad. En la segunda fase, se analizó el texto completo de los estudios preseleccionados realizándose de forma independiente por dos revisores, donde cualquier discrepancia fue resuelta mediante consenso o la intervención de un tercer revisor. Además se documentó el proceso completo utilizando el diagrama de flujo Preferred Reporting Items for Systematic Reviews and Meta-Analyses (**PRISMA** por sus siglas en inglés) ver **Figura 1**.

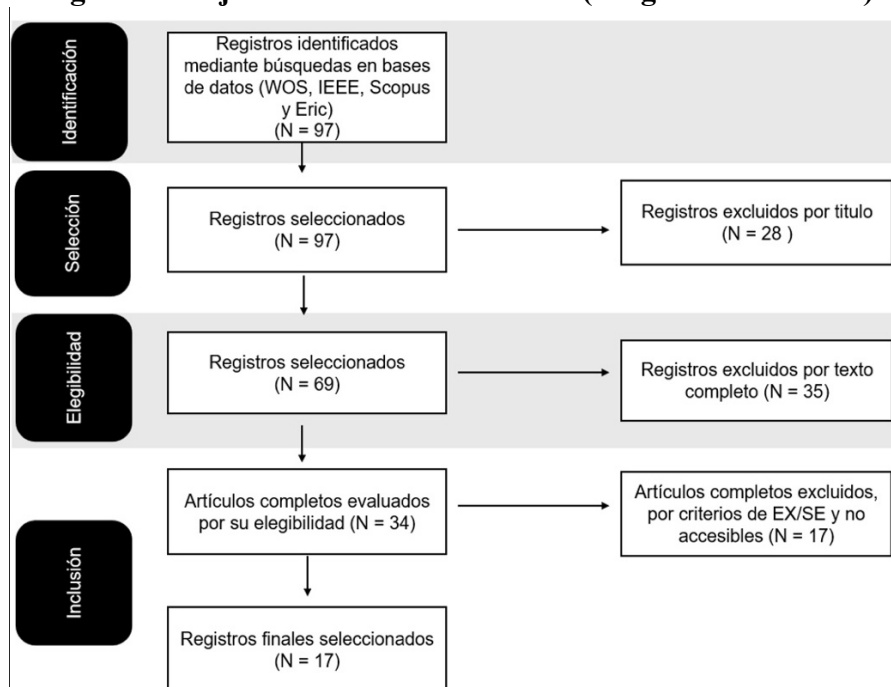
Para la extracción de datos se diseñó una hoja de recolección estandarizada que incluye las siguientes variables: Autor(es), año de publicación, país de origen del estudio, objetivo principal, técnicas de **IA** utilizadas, tipo de datos analizados (cuantitativos y/o cualitativos), métricas de validación empleadas (tanto técnicas como de percepción), resultados principales, limitaciones reportadas y desafíos identificados. Asimismo, el análisis de los datos fue de carácter narrativo y temático, guiado por las preguntas de investigación previamente planteadas, siendo los hallazgos agrupados en torno a las siguientes categorías: Técnicas y algoritmos utilizados, métodos de validación y efectividad, beneficios percibidos, y barreras técnicas, éticas e institucionales para la

implementación de estas tecnologías en entornos universitarios, presentándose en la **Tabla 2** de la siguiente sección.

5. RESULTADOS

El proceso de selección se desarrolló bajo los lineamientos del diagrama **PRISMA**, partiendo de 97 registros identificados en **Web of Science**, **IEEE Xplore**, **Scopus** y **ERIC**. Así mismo para la fase de cribado o selección fue realizada por dos revisores de manera independiente, donde las discrepancias surgidas durante este proceso se resolvieron mediante la discusión técnica y el consenso mutuo, garantizando así la rigurosidad en la aplicación de los criterios. No se aplicó el estadístico Kappa de Cohen debido a que el diseño metodológico priorizó la validación cualitativa por consenso sobre la medición de la concordancia aleatoria, asegurando que cada inclusión fuera validada por la totalidad del equipo investigador. Tras excluir 28 registros por título y 35 por criterios de inclusión en texto completo, se evaluaron 34 artículos finales. De estos, 17 fueron descartados por incumplimiento de criterios específicos o inaccesibilidad, resultando en 17 estudios incluidos. La **Figura 1** detalla este flujo de decisiones mediante el diagrama de flujo **PRISMA**.

Figura 1. Flujo de selección de estudios (Diagrama PRISMA)



Fuente: Elaboración propia basada en el proceso de selección conforme a los lineamientos del diagrama PRISMA.

Ahora bien, con el fin de organizar y presentar de manera clara la información recopilada en esta revisión, se ha elaborado una tabla, ver **Tabla 2**, la cual resumen distintos aspectos clave de los estudios analizados permitiendo visualizar de forma estructurada los datos más relevantes como: País de origen, objetivo del estudio, técnicas de **IA** utilizadas, tipo de datos de entrada, métricas de validación, resultados principales, limitaciones reportadas y desafíos identificados relevantes para esta revisión.

Tabla 2. Principales características de los estudios revisados

Estudio	País	Objetivo del estudio	Técnicas de IA utilizadas	Tipo de datos de entrada
Pramod et al. (2022)	India	Predecir la efectividad docente mediante machine learning.	Análisis de sentimientos y Word Cloud en texto y Random Forest como modelo predictivo, destacando modelos de conjunto.	Análisis de datos cuantitativos y cualitativos de cuestionarios estudiantiles.
Hu et al. (2020)	China	Proponer un método de análisis de sentimientos aplicando machine learning.	Análisis de sentimientos con SVM (kernels lineal, polinómico, RBF); preprocesamiento con word2vec, K-means para clustering y TFIDF para ponderar características.	Datos cualitativos: 581 evaluaciones cortas en chino, 566 usadas como corpus (466 entrenamiento, 100 prueba), con textos de 25 a 150 palabras.
Esparza et al. (2018)	México	Clasificar comentarios reales en español.	Análisis de sentimiento con SVM usando kernels lineal, RBF y polinomial, aprovechando su alto rendimiento en clasificación.	Datos cualitativos: 1040 comentarios en español de estudiantes de ingeniería, evaluando 21 profesores; 99 características identificadas tras limpieza.
Peña-Torres et al. (2024)	Colombia	Mejorar la práctica docente.	Análisis de sentimiento con NLP usando LLMs preentrenados: Py-sentimiento (Robertuito/RoBERTa), versión personalizada de VADER y modelo DistilBERT .	Enfoque mixto: cuantitativo (medidas de rendimiento LLM) y cualitativo (caso de estudio, prototipo). Datos: 365 comentarios en español.
Nikolić et al. (2020)	Serbia	Analizar sentimientos en texto libre de encuestas estudiantiles en serbio.	Sistema ABSA con SVM lineal para aspectos y sentimientos, Bag-of-Words con TFIDF ,	Datos cualitativos en serbio: Presencial (2,472 reseñas, 2,798 oraciones) y C2 online

Estudio	País	Objetivo del estudio	Técnicas de IA utilizadas	Tipo de datos de entrada
			diccionarios y reglas basadas en expresiones regulares.	(3,863 reseñas y 6,896 oraciones).
Okoye et al. (2022)	México	Evaluar la evaluación de del conocimiento adquirido por los estudiantes.	Se aplicó el modelo EPDM de minería de texto para extraer sentimientos y valencias emocionales de los estudiantes.	Cualitativamente se cuantificaron 471,968 comentarios abiertos en español.
Shuqin & Raga (2024)	China	Proponer el modelo BERTBiLSTM-ATTENTION (BBA) para extraer sentimientos de evaluaciones de cursos.	BERT para representaciones contextuales profundas. BiLSTM para captar dependencias bidireccionales. Atención para enfocar la información relevante.	Se analizaron 9944 comentarios reales de estudiantes de cursos de informática en la plataforma MOOC de una universidad China.
Wu et al (2023)	EEUU	Mejorar resultados de aprendizaje y rediseñar el curso AME4163 .	Análisis de sentimiento con VADER .	Textos no estructurados de estudiantes en forma de take-aways (TA) y oraciones de aprendizaje (LS) estructuradas como tripletas (Experiencia, Aprendizaje y Valor).
Fargues et al. (2023)	Noruega	Analizar la retroalimentación abierta de estudiantes.	Se aplicó NLP usando análisis léxico: VADER para sentimientos, NRC para emociones generales y un diccionario personalizado para seis emociones educativas.	Los datos de entrada son cualitativos, provenientes de retroalimentación abierta de estudiantes.
Peng et al. (2022)	China	Desarrollar un modelo de análisis de sentimientos aplicado a las evaluaciones docentes.	Modelo de análisis de sentimientos que combina CNN , BLSTM y atención, usando Softmax para clasificar emociones complejas y optimizado con descenso de gradiente mini-batch.	Datos cuantitativos (precisión, pérdida, recall, F1-score) y cualitativos (comentarios de estudiantes con carga emocional).
Shaik et al. (2022)	Australia	Desarrollar una aplicación que permita a una institución educativa analizar comentarios textuales para evaluar y mejorar	IA y PNL : preprocesamiento, TextBlob para sentimiento, Word2Vec para vectores y un modelo Bi-LSTM	El estudio analizó 5,585 comentarios cualitativos y cuantitativos de estudiantes (texto y calificaciones de 1 a 5)

Estudio	País	Objetivo del estudio	Técnicas de IA utilizadas	Tipo de datos de entrada
		sus sistemas y servicios.	para clasificar comentarios en 19 subdominios de Biggs.	del semestre 2 de 2021.
Pu et al. (2021)	China	Proponer un modelo de aprendizaje profundo en conjunto para analizar sentimientos en evaluaciones de cursos online.	Modelo que combina Word2Vec y GloVe para representar palabras, usa CNN y BiLSTM para extraer características, y optimiza un ensemble con el algoritmo MOGWO .	El conjunto de datos preprocesado incluye 20,240 entradas con etiquetas emocionales.
Resuello et al. (2024)	Filipinas	Utilizar algoritmos de minería de texto para analizar comentarios estudiantiles sobre desempeño docente.	Se aplicaron técnicas de minería de texto, NLP , VADER y LDA .	Se analizaron principalmente comentarios cualitativos de estudiantes, con un total de 2,312 comentarios recopilados entre 2020 y 2023.
Okoye et al. (2023)	México	El estudio analizó la evaluación de la enseñanza por parte de los estudiantes (SET) para identificar los factores clave que afectan el desempeño y la enseñanza en la Tecnología.	Se aplicó el modelo EPDM de minería de texto para extraer sentimientos y valencias emocionales de los estudiantes.	Enfoque mixto (cualitativo y cuantitativo): Cualitativamente se cuantificaron 471,968 comentarios abiertos en español.
Das et al. (2022)	India	Mejorar la enseñanza usando ML .	Regresión Logística, XGBoost y Árboles de Decisión.	El estudio se enfoca en la Retroalimentación Cualitativa (QLF).
Yuan & Hernandez (2023)	Filipinas	Analizar la tendencia emocional de los estudiantes.	Bi-LSTM .	Textos de evaluación docente de estudiantes obtenidos mediante evaluación en línea.
Ponmalar et al. (2024)	India	Proponer el AI-Enhanced Context-Aware Semantic Grading Network (AI-CASGN).	PLN y BERT .	Evaluación de respuestas descriptivas (datos cualitativos, texto), que requieren una visión más profunda de la comprensión, la capacidad de razonamiento y el pensamiento analítico de los estudiantes.

Fuente: Elaboración propia basada en las principales características de los estudios revisados

Por otro lado, respecto a la primera interrogante planteada en la introducción, la literatura científica revisada evidencia una transición tecnológica clara, donde el **PLN** ha evolucionado desde el análisis de sentimientos basado en diccionarios léxicos y modelos de aprendizaje automático tradicional, como el modelo **SVM**, hacia arquitecturas de aprendizaje profundo altamente sofisticadas. En este sentido, el uso predominante de modelos como las redes de memoria a largo y corto plazo bidireccionales y los transformadores tipo **BERT** ha permitido no solo clasificar polaridades, sino también identificar temas subyacentes mediante el modelado de tópicos con la técnica de Asignación Latente de Dirichlet (**LDA**) y categorizar el nivel de pensamiento crítico a través de la Taxonomía de Bloom.

En lo que respecta a la segunda pregunta de investigación, los datos indican que los modelos avanzados como DistilBERT logran una efectividad superior a los métodos tradicionales, alcanzando precisiones cercanas al 93% en contraste con el 79% que suelen reportar herramientas léxicas como **VADER**. Esta superioridad técnica se valida mediante métricas estandarizadas como el **F1-score** y el área bajo la curva, permitiendo además una eficiencia operativa notable al reducir los tiempos de respuesta a escasos 3.2 segundos.

En cuanto a la tercera pregunta de investigación, sobre los beneficios y desafíos, la investigación revela que la automatización analítica permite generar retroalimentación personalizada y escalable para grandes volúmenes de datos, como los más de 2'354,631 de registros generados por ciclo en el **CUCEA**. No obstante, este potencial se enfrenta a retos éticos y técnicos significativos, tales como la baja interpretabilidad de los modelos de "caja negra", la presencia de sesgos algorítmicos de género y la resistencia cultural al cambio tecnológico dentro del sector educativo. Atendiendo a estas implicaciones.

6. DISCUSIÓN

Desde una perspectiva institucional, los resultados evidencian consistencia entre los planteamientos teóricos formulados y los patrones identificados en la literatura especializada. En este sentido, las hipótesis (**H1**, **H2** y **H3**) encuentran respaldo conceptual en la evidencia sintetizada, al observarse convergencias significativas en los estudios analizados. Dicho respaldo se fundamenta en análisis comparativo y revisión sistemática, y no en validación experimental directa ni en datos primarios generados por la investigación. Además es importante señalar, que la

principal contribución del artículo es de carácter metodológico-organizacional, al proponer un marco replicable de evaluación docente basado en **IA/PLN** como innovación de proceso, más que el desarrollo de un algoritmo específico. Casos como los centros universitarios de la Red Universitaria de la Universidad de Guadalajara —entre ellos el Centro Universitario de Ciencias Económico-Administrativas el cual se presenta como un contexto institucional ilustrativo, y no como un estudio empírico aplicado. — ilustran desafíos comunes a numerosas instituciones latinoamericanas, donde el procesamiento de millones de registros por periodo limita la oportunidad y utilidad de la retroalimentación.

En este contexto, la literatura revisada evidencia que la integración de **IA**, ciencia de datos y **PLN** constituye una innovación de proceso organizacional, en los términos establecidos por el Manual de Oslo, al automatizar el análisis de información cualitativa a gran escala y mejorar la eficiencia, calidad y trazabilidad de los procesos evaluativos. Modelos como **AI-CASGN**, capaces de generar retroalimentación en segundos, y arquitecturas basadas en **DistilBERT**, que alcanzan altos niveles de precisión en español, demuestran la superioridad de estos enfoques frente a métodos tradicionales y su viabilidad para entornos universitarios latinoamericanos.

Asimismo, el uso de modelos híbridos como **BERT-BiLSTM-ATTENTION** y técnicas de modelado de tópicos permite capturar tanto la semántica como las sutilezas emocionales presentes en los comentarios estudiantiles, favoreciendo una comprensión más profunda de la experiencia educativa y fortaleciendo los mecanismos de mejora continua del desempeño docente.

Desde la perspectiva del desarrollo sostenible, los hallazgos de este estudio contribuyen directamente al fortalecimiento del **ODS 4**, al proponer mecanismos innovadores que mejoran la calidad de la educación superior mediante procesos de evaluación docente más justos, oportunos y formativos. De igual manera, se alinean con el **ODS 9** al promover el uso estratégico de infraestructura digital y tecnologías avanzadas para optimizar procesos organizacionales en instituciones educativas.

La adopción de modelos de **IA** y **PLN** no solo incrementa la eficiencia operativa de los sistemas de evaluación, sino que también favorece una gestión del conocimiento más sostenible, al transformar grandes volúmenes de datos en información útil para la toma de decisiones, el diseño de políticas académicas y la mejora institucional en universidades latinoamericanas.

A diferencia de revisiones previas centradas exclusivamente en métricas técnicas o en aplicaciones aisladas de **IA** en educación, este estudio integra de manera sistemática la evaluación docente universitaria, la Inteligencia Artificial y el Procesamiento de Lenguaje Natural desde un enfoque de innovación de proceso organizacional, alineado con el Manual de Oslo y los Objetivos de Desarrollo Sostenible. Esta aproximación permite articular rigor metodológico, análisis técnico e implicaciones institucionales, ofreciendo un marco analítico replicable para redes universitarias y universidades públicas de América Latina.

6.1 Implicaciones Teóricas (Scientia)

La presente investigación realiza una contribución significativa al estado del conocimiento al proponer un enfoque interdisciplinario que trasciende la visión puramente técnica de la Inteligencia Artificial (**IA**), esto debido a que se diferencia de revisiones previas que abordan el análisis de sentimientos o la minería de datos de forma aislada. Asimismo, este estudio integra el uso de la **IA** y el Procesamiento de Lenguaje Natural (**PLN**) dentro de un marco que vincula la valuación docente con la innovación de proceso, bajo las directrices del Manual de Oslo, y los Objetivos de Desarrollo Sostenible (**ODS 4 y 9**).

Desde una perspectiva teórica, los hallazgos subrayan dos vacíos críticos en la literatura de habla hispana:

- **Contextualización Lingüística y Cultural:** Se identifica que, aunque modelos multilingües como **DistilBERT** alcanzan precisiones del 93%, persiste una escasez de corpus etiquetados en español que capturen los matices educativos regionales. Esta investigación aporta una base teórica para la transición de modelos léxicos hacia arquitecturas de aprendizaje profundo adaptadas.
- **Validez de Constructo y Medición Longitudinal:** Se plantea la necesidad de marcos conceptuales que no solo midan la polaridad (positivo/negativo), sino que integren dimensiones pedagógicas complejas como la Taxonomía de Bloom, ya que la literatura actual carece de evidencia longitudinal, por lo que este estudio fundamenta teóricamente cómo la retroalimentación automatizada debe evaluarse en función de la mejora sostenida de las estrategias docentes y la satisfacción estudiantil a largo plazo.

6.2 Implicaciones Prácticas

En términos de aplicación inmediata, este artículo trasciende la teoría al proponer un diseño de plataforma concreto para el contexto del Centro Universitario de Ciencias Económico Administrativas (CUCEA), ya que esta propuesta aborda directamente las limitaciones críticas de baja precisión, análisis superficial y retroalimentación tardía de los sistemas actuales a través de las siguientes dimensiones:

- **Eficiencia Operativa y Escalabilidad:** La integración de Modelos de Lenguaje de Gran Escala (LLMs) permite el procesamiento ágil de grandes volúmenes de datos, específicamente los más de 2'354,631 registros generados por ciclo en el CUCEA, garantizando una capacidad de respuesta que los métodos manuales o léxicos tradicionales no pueden alcanzar.
- **Profundidad Analítica (Accionabilidad):** Mediante la implementación de técnicas de Análisis de Sentimiento Basado en Aspectos (ABSA) y Asignación Latente de Dirichlet (LDA), la plataforma transforma comentarios abiertos en datos estructurados por dimensiones (metodología, evaluación, infraestructura). Esto permite a los gestores académicos y docentes recibir retroalimentación predictiva y en tiempo real a través de Sistemas de Soporte a la Decisión (DSS).
- **Ética y Gobernanza de Datos:** La implementación práctica exige un diseño centrado en la transparencia algorítmica, por lo que se proponen mecanismos específicos para mitigar sesgos de género y asegurar la privacidad de los datos, elementos clave para vencer la resistencia institucional y fomentar la adopción activa de la tecnología por parte del profesorado.

7. CONCLUSIÓN

El presente artículo de **revisión sistemática (2019–2024)** confirma la **insuficiencia de los modelos tradicionales de evaluación docente**, caracterizados por el uso de promedios simples y análisis manuales que generan una retroalimentación superficial, tardía y altamente subjetiva. Esta situación limita la toma de decisiones informadas para la mejora continua de la docencia y resulta especialmente crítica en contextos universitarios que enfrentan procesos de expansión, diversificación de modalidades educativas y mayores exigencias de calidad académica.

La evidencia analizada demuestra que la **incorporación de técnicas de Inteligencia Artificial**, particularmente aquellas orientadas al análisis automatizado de datos cualitativos y cuantitativos,

ofrece una **oportunidad estratégica para transformar los sistemas de evaluación docente**. Estos enfoques permiten identificar patrones complejos, generar retroalimentación oportuna y personalizada, y fortalecer una cultura de evaluación formativa centrada en el desarrollo profesional del docente, más allá del carácter meramente administrativo o sancionador.

Desde una el punto de vista prospectivo, los hallazgos de esta revisión abren la posibilidad de **desarrollar estudios longitudinales** que permitan analizar la evolución del desempeño docente a lo largo del tiempo, evaluar el impacto sostenido de la retroalimentación inteligente y medir su influencia en variables clave como la calidad del aprendizaje, la innovación pedagógica y la satisfacción estudiantil. Este tipo de estudios resultan fundamentales para transitar de diagnósticos estáticos hacia modelos dinámicos de mejora continua basados en evidencia.

Asimismo, los resultados sugieren un **alto potencial de replicabilidad del modelo en universidades latinoamericanas**, dado que muchas instituciones comparten problemáticas estructurales similares: limitaciones tecnológicas, sobrecarga administrativa y sistemas de evaluación poco contextualizados. La adaptación de soluciones basadas en **IA**, considerando marcos normativos, realidades institucionales y enfoques de justicia educativa propios de la región, puede contribuir significativamente a la modernización de la gestión académica y al fortalecimiento de la calidad educativa en América Latina.

En síntesis, esta revisión no solo evidencia la necesidad de superar los modelos tradicionales de evaluación docente, sino que también **proyecta una línea clara de investigación y acción institucional**, orientada a la implementación de sistemas inteligentes, éticos y contextualizados, capaces de responder a los desafíos presentes y futuros de la educación superior.

Finalmente, como trabajo futuro, se propone la validación aplicada del enfoque presentado a través de la implementación un estudio empírico posterior en un contexto educativo real.

8. REFERENCIAS

- Arriaga-Cárdenas, O. G., Lara-Magaña, P. del C., & Pasciuta-Marco, P. D. (2022). Los centros públicos de investigación, como eje central de la innovación y la educación en México. *Revista científica Scientia et PRAXIS*, 2(04), 66–81. <https://doi.org/10.55965/setp.2.coed.a3>
- Das, S., Roy, S., Bose, R., Acharjya, P. P., & Mondal, H. (2022). Analysis of student sentiment dynamics to evaluate teachers' performance in online courses using machine learning. In *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 668–673). IEEE. <https://doi.org/10.1109/ICAAIC53929.2022.9792958>

- Esparza, G. G., de-Luna, A., Zezzatti, A. O., Hernandez, A., Ponce, J., Álvarez, M., Cossio, E., & de Jesus Nava, J. (2018). A sentiment analysis model to analyze students reviews of teacher performance using support vector machines. In *14th International Conference Distributed Computing and Artificial Intelligence* (Vol. 14, pp. 157–164). Springer Nature. https://doi.org/10.1007/978-3-319-62410-5_19
- Fargues, M., Kadry, S., Lawal, I. A., Yassine, S., & Rauf, H. T. (2023). Automated analysis of open-ended students' feedback using sentiment, emotion, and cognition classifications. *Applied Sciences*, 13(4), 2061. <https://doi.org/10.3390/app13042061>
- Hu, X., Yang, Y., Wu, X., & Li, Y. (2020). Text analysis of teaching evaluation based on machine learning. In *Proceedings of the 10th International Workshop on Computer Science and Engineering (WCSE 2020)*. WCSE. <https://doi.org/10.18178/wcse.2020.06.004>
- Martínez-Téllez, M. Á., & Celaya-Lozano, A. (2023). Evaluación institucional de contenidos y proyectos de investigación en un CPI: Caso de estudio (2013-2022). *Scientia et PRAXIS*, 3(05), 51–77. <https://doi.org/10.55965/setp.3.coed1.a3>
- Nikolić, N., Grljević, O., & Kovačević, A. (2020). Aspect-based sentiment analysis of reviews in the domain of higher education. *The Electronic Library*, 38(1), 44–64. <https://doi.org/10.1108/EL-06-2019-0140>
- Okoye, K., Arrona-Palacios, A., Camacho-Zuñiga, C., Achem, J. A. G., Escamilla, J., & Hosseini, S. (2022). Towards teaching analytics: A contextual model for analysis of students' evaluation of teaching through text mining and machine learning classification. *Education and Information Technologies*, 27(3), 3891–3933. <https://doi.org/10.1007/s10639-021-10751-5>
- Okoye, K., Daruich, S. D. N., de La O, J. F. E., Castano, R., Escamilla, J., & Hosseini, S. (2023). A text mining and statistical approach for assessment of pedagogical impact of students' evaluation of teaching and learning outcome in education. *IEEE Access*, 11, 9577–9596. <https://doi.org/10.1109/ACCESS.2023.3239779>
- Peng, H., Zhang, Z., & Liu, H. (2022). A sentiment analysis method for teaching evaluation texts using attention mechanism combined with CNN-BLSTM model. *Scientific Programming*, 2022, 1–9. <https://doi.org/10.1155/2022/8496151>
- Peña-Torres, J. A. (2024). Towards an improved of teaching practice using sentiment analysis in student evaluation. *Ingeniería y Competitividad*, 26(2). <https://doi.org/10.25100/iyc.v26i2.13759>
- Ponmalar, S. P., V. S., Srithar, V., Morankar, G., Jebaseelan, S. D. S., & Harikrishnan, R. (2024). *AI-driven CASGN framework for evaluating descriptive answers in online education*. In *2024 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSSES)* (pp. 1–8). IEEE. <https://doi.org/10.1109/ICSSES63760.2024.10910723>
- Pramod, D., Bharathi, S. V., & Raman, R. (2022). Faculty effectiveness prediction using machine learning and text analytics. In *2022 IEEE Technology and Engineering Management Conference (TEMSCON EUROPE)* (pp. 40–47). IEEE. <https://doi.org/10.1109/TEMSCONEUROPE54743.2022.9801997>
- Pu, X., Yan, G., Yu, C., Mi, X., & Yu, C. (2021). Sentiment analysis of online course evaluation based on a new ensemble deep learning mode: Evidence from Chinese. *Applied Sciences*, 11(23), 11313. <https://doi.org/10.3390/app112311313>

- Resuello, J. L., Feliscuzo, L. S., & Pantaleon, C. B. (2024). Text mining student comments for teaching performance evaluation using VADER and Latent Dirichlet Allocation algorithm. In *2024 International Conference Automatics and Informatics (ICAI)* (pp. 260–265). IEEE. <https://doi.org/10.1109/ICAI63388.2024.10851580>
- Santamaria-Velasco, C. A., Montiel-Méndez, O. J. M.-M., & Montañez-Moya, G. S. (2024). Liderazgo transformacional y emprendimiento en estudiantes: Una vía hacia el desarrollo educativo sostenible. *Scientia et PRAXIS*, 4(08), 90–119. <https://doi.org/10.55965/setp.4.08.uady.a4>
- Shaik, T., Tao, X., Dann, C., Quadrelli, C., Li, Y., & O'Neill, S. (2022). Educational decision support system adopting sentiment analysis on student feedback. In *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (pp. 377–383). IEEE. <https://doi.org/10.1109/WI-IAT55865.2022.00062>
- Shuqin, H., & Raga, R. C. (2024). A deep learning model for student sentiment analysis on course reviews. *IEEE Access*, 12(2169–3536), 136747–136758. <https://doi.org/10.1109/ACCESS.2024.3463793>
- Vera-Noriega, J. Á., Borbón-Morales, C. G., Mejía-Trejo, J., & Durazo-Salas, F. F. (2023). Relación y comparación entre las variables de clima organizacional, satisfacción y calidad de vida en una empresa de ventas por teléfono en el noroeste de México. *Scientia et PRAXIS*, 3(05), 83–109. <https://doi.org/10.55965/setp.3.05.a4>
- Wu, Y., Ming, Z., Allen, J. K., & Mistree, F. (2023). *Evaluation of students' learning through reflection on doing based on sentiment analysis*. *Journal of Mechanical Design*, 145(3). <https://doi.org/10.1115/1.4056495>
- Yuan, H., & Hernandez, A. A. (2023). Sentiment analysis of student evaluation of teaching based on Bi-LSTM algorithm. In *2023 IEEE 15th International Conference on Advanced Infocomm Technology (ICAIT)* (pp. 417–422). IEEE. <https://doi.org/10.1109/ICAIT59485.2023.10367267>



This is an open access article distributed under the terms of the CC BY-NC license (<https://creativecommons.org/licenses/by-nc/4.0/>)