EDUCACIÓN, CREATIVIDAD E INTELIGENCIA ARTIFICIAL: NUEVOS HORIZONTES PARA EL APRENDIZAJE. ACTAS DEL VIII CONGRESO INTERNACIONAL SOBRE APRENDIZAJE, INNOVACIÓN Y COOPERACIÓN, CINAIC 2025

María Luisa Sein-Echaluce Lacleta, Ángel Fidalgo Blanco y Francisco José García Peñalvo (coords.)

1º Edición. Zaragoza, 2025

Edita: Servicio de Publicaciones. Universidad de Zaragoza.

Servicio de Publicaciones Universidad Zaragoza

EBOOK ISBN 978-84-10169-60-9

DOI 10.26754/uz.978-84-10169-60-9

Referencia a esta obra:

Sein-Echaluce Lacleta, M.L., Fidalgo Blanco, A. & García-Peñalvo, F.J. (coords.) (2025). *Educación, Creatividad e Inteligencia Artificial: nuevos horizontes para el Aprendizaje. Actas del VIII Congreso Internacional sobre Aprendizaje, Innovación y Cooperación. CINAIC 2025 (11-13 de Junio de 2025, Madrid, España)*. Zaragoza. Servicio de Publicaciones Universidad de Zaragoza. DOI 10.26754/uz.978-84-10169-60-9

Análisis de reseñas de usuarios sobre apps educativas con IA: una aproximación basada en sentimiento y modelado temático

User Review Analysis of AI-Based Educational Apps: A Sentiment and Topic Modeling Approach

Cristina Ceballos-Hernández, Juan Luis Blanco-Guzmán
cceballos@us.es, jbguzman@us.es

Departamento Economía Financiera y Dirección
de Operaciones
Universidad de Sevilla
Sevilla, España

Resumen- El trabajo analiza la percepción de los usuarios sobre aplicaciones educativas basadas en inteligencia artificial a partir de un corpus de más de 840.000 reseñas extraídas de Google Play y App Store. Utilizando técnicas avanzadas de procesamiento de lenguaje natural, se aplicaron modelos de análisis de sentimiento y de modelado temático. Los resultados muestran una valoración mayoritariamente positiva, con alta coherencia entre la puntuación otorgada y el sentimiento textual. Se identifican temas clave como la utilidad, facilidad de uso y personalización del aprendizaje. La metodología, sostenible y automatizable, es aplicable a otros contextos educativos y tecnológicos, y se propone como herramienta para mejorar el diseño y la eficacia de las apps educativas..

Palabras clave: app educativas, Inteligencia Artificial, reseñas, análisis de sentimiento.

Abstract- This study analyzes user perceptions of artificial intelligence-based educational applications using a corpus of over 840,000 reviews extracted from Google Play and the App Store. Advanced natural language processing techniques were applied, including sentiment analysis and topic modeling. The results indicate a predominantly positive evaluation, with strong alignment between user ratings and textual sentiment. Key themes identified include usefulness, ease of use, and personalized learning. The methodology—scalable, sustainable, and automatable—is transferable to other educational and technological contexts and is proposed as a valuable tool for improving the design and effectiveness of educational apps.

Keywords: educational apps, artificial intelligence, reviews, sentiment analysis.

1. INTRODUCCIÓN

Las aplicaciones móviles educativas son herramientas de software diseñadas para el aprendizaje, el desarrollo de habilidades o la entrega de contenido educativo a través de teléfonos inteligentes o tabletas. Estas aplicaciones han experimentado un crecimiento significativo y han revolucionado la forma en que estudiantes, profesores y aprendices interactúan con el contenido educativo, convirtiéndose en uno de los principales canales para la diseminación de recursos de aprendizaje, especialmente durante la pandemia de COVID-19 (Ganguly y Dasari, 2024).

Muchas de estas aplicaciones integran ahora Inteligencia Artificial (IA) para ofrecer experiencias más adaptativas y personalizadas y para mejorar la automatización e interactividad. Su uso impacta además positivamente en la participación y el rendimiento académico (Baba et al., 2024), mejora la accesibilidad y la adaptabilidad del contenido educativo (Yuen y Schlote, 2024) y aumenta la motivación de los estudiantes (Krishnan y Zaini, 2025).

Es crucial abordar los problemas de usabilidad y centrarse en mejorar la calidad del contenido y las funciones de IA para aumentar la retención de usuarios. Algunos estudios utilizan el recuento de "me gusta" en las reseñas como una métrica para priorizar los factores de la experiencia del usuario, lo que podría estar relacionado con el rendimiento percibido (Arambepola et al., 2024). Las calificaciones en sí mismas son una forma directa de evaluación de la percepción del usuario (Mondal et al., 2022). El número de descargas de una aplicación y el volumen de reseñas pueden indicar el compromiso del usuario. Sin embargo, un gran volumen de descargas no siempre se traduce en calificaciones más altas (Ganguly y Dasari, 2024).

La evaluación del rendimiento de las aplicaciones educativas móviles se realiza principalmente a través del análisis de las reseñas de los usuarios, donde se examina el sentimiento, los problemas reportados y los aspectos valorados. El sentimiento expresado en las reseñas (positivo, negativo o neutral) es un indicador clave de la satisfacción general del usuario con el rendimiento de la aplicación y la usabilidad (Mondal et al., 2022). Estas reseñas ayudan a los desarrolladores a comprender los requisitos de los usuarios, identificar problemas y a mejorar sus productos, especialmente el análisis de reseñas negativas (Zahoor & Bawany, 2023).

Este trabajo contribuye al estado del arte combinando análisis de sentimiento con modelado temático sobre un conjunto de reseñas de usuarios de gran tamaño y sobre un amplio número de apps educativas basadas en inteligencia artificial, proporcionando una visión detallada y actualizada de la experiencia desde la perspectiva del usuario.

2. Contexto y Descripción

A. Objetivo

El objetivo del trabajo es analizar las percepciones de los usuarios sobre aplicaciones educativas basadas en inteligencia artificial mediante técnicas de análisis de sentimiento y modelado temático, validando la consistencia del modelo.

B. Contexto

Para la obtención del corpus, se realizó un proceso de extracción automatizada de datos (web scraping) en febrero de 2025 desde las tiendas de aplicaciones Google Play Store y Apple App Store, empleando las librerías google_play_scraper y app_store_scraper en Python. La selección incluyó las 27 aplicaciones educativas basadas en inteligencia artificial más populares en ambos markets, según su número de descargas y puntuación media, asegurando así la representatividad del análisis respecto a las herramientas más utilizadas por los usuarios. El conjunto final estuvo compuesto por 842.248 reseñas de usuarios, escritas en múltiples idiomas.

C. Metodología

Es estudio emplea una combinación de metodologías cualitativas y cuantitativas de análisis de texto, apoyadas por herramientas avanzadas de procesamiento de lenguaje natural. El análisis de sentimiento se realizó utilizando el modelo RoBERTa (Liu et al., 2019), una arquitectura de tipo Transformer entrenada para tareas de clasificación emocional. Para extraer términos relevantes, se utilizó la métrica TF-IDF, común en análisis léxico (Salton & Buckley, 1988), y para la detección de temas se aplicó el modelo LDA (Blei et al., 2003) Todo el procesamiento se desarrolló en entorno Python con soporte de GPU en Google Colab Pro, utilizando las librerías nltk, transformers, sentence-transformers, scikit-learn, matplotlib y bertopic.

3. Resultados

Los resultados obtenidos permiten caracterizar de forma precisa la percepción de los usuarios hacia las aplicaciones educativas con inteligencia artificial.

Tabla 1. Puntuación media otorgada en cada Market

Market	*Media Score Usuarios*
Apple App Store	4.020024
Google Play Store	4.148620

El análisis de la distribución de puntuaciones mostró una clara tendencia hacia las valoraciones positivas, con una media general de 4,13 sobre 5 (Tabla 1) y una mayor concentración de reseñas en los extremos (1 y 5), lo que indica polarización en las experiencias de uso (Figura 1).

En relación con el primero de los objetivos, mediante el modelo de análisis de sentimientos basado en RoBERTa se clasificaron las reseñas en tres categorías: positiva, neutral y negativa. La mayor parte del corpus fue catalogada como positiva (57%), seguida de reseñas neutras (30%) y negativas (13%) (Figura 2). Esta distribución evidencia una percepción mayoritariamente favorable, aunque se identificaron también áreas de mejora.

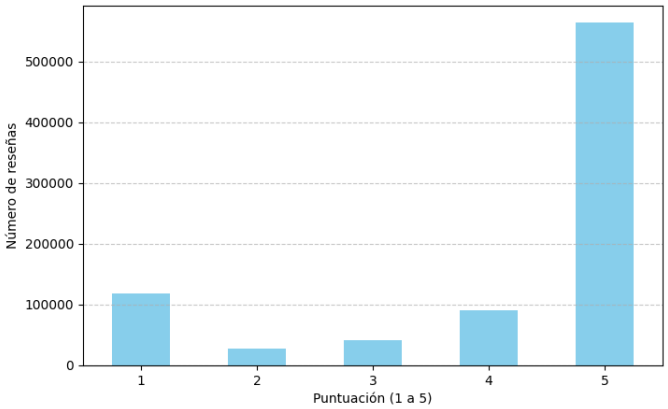

Figura 1. Distribución de puntuaciones

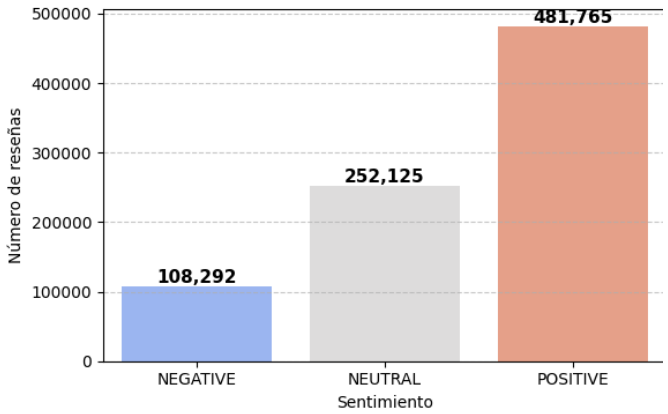

Figura 2. Distribución de sentimientos en las reseñas

Para validar la consistencia del modelo se analizó la relación entre la puntuación otorgada por el usuario y la confianza del modelo de análisis de sentimiento mediante el coeficiente de correlación de Pearson. Aunque la correlación obtenida fue baja ($r = 0{,}1365$; $p < 0{,}001$), el resultado fue estadísticamente significativo y sugiere una leve tendencia: cuanto mayor es la puntuación, mayor es la seguridad del modelo al clasificar el sentimiento del comentario (Tabla 2).

Tabla 2. Confianza del modelo según Score

Score	*Media_confianza*	*Desv_estándar*	*N reseñas*
1	0.7220	0.1539	118659
2	0.6843	0.1496	27530
3	0.6866	0.1503	41557
4	0.7311	0.1588	90478
5	0.7693	0.1615	563958

Este patrón se reforzó visualmente mediante un diagrama de caja, que mostró que las reseñas con puntuaciones de 4 y 5 presentan niveles de confianza más altos y menor variabilidad (Figura 3). Por el contrario, las puntuaciones bajas mostraron mayor dispersión, lo que podría estar relacionado con expresiones más ambiguas, sarcásticas o emocionales que dificultan la clasificación automática.

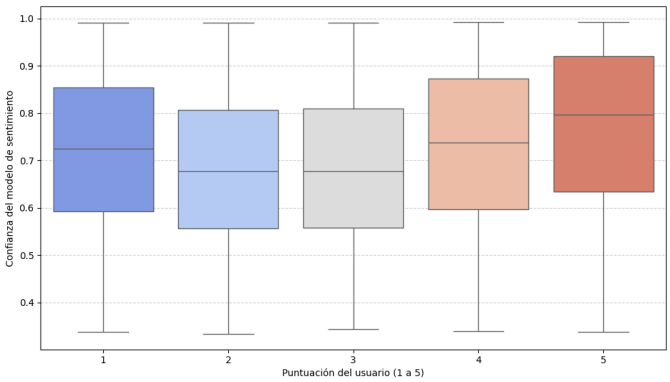

Figura 3. Confianza del modelo por puntuación del usuario

Para analizar la relación entre la puntuación numérica otorgada por el usuario y el sentimiento detectado en su reseña, se codificaron las categorías emocionales en una escala numérica (-1 para negativo, 0 para neutro y 1 para positivo) y se calculó el coeficiente de correlación de Pearson. El resultado obtenido ($r = 0{,}5601$; $p < 0{,}001$) indica una correlación positiva de intensidad moderada-alta, estadísticamente significativa. Este hallazgo respalda la existencia de una alineación consistente entre la valoración emocional expresada en el texto y la puntuación cuantitativa, reforzando la validez del modelo de análisis semántico aplicado.

Se calculó la puntuación media asignada por los usuarios dentro de cada categoría de sentimiento detectado (Tabla 3). Los resultados muestran una clara alineación entre ambos elementos: las reseñas clasificadas como positivas presentan una puntuación media significativamente más alta (4,70) mientras que las negativas se sitúan en el extremo opuesto (2,20). Las reseñas neutras presentan una puntuación intermedia (3,87), lo que respalda la coherencia general entre la valoración textual y la puntuación cuantitativa. Esta correspondencia valida la capacidad del modelo utilizado para captar con precisión el tono emocional de los comentarios. En consecuencia, se demuestra su utilidad para convertir información cualitativa no estructurada en datos analizables, incluso en muestras masivas como la aquí utilizada.

Tabla 3. Relación entre la puntuación del usuario y el sentimiento expresado en las reseñas

Sentiment_label	*Puntuación media*	*Desviación estándar*	*Núm. reseñas*
NEGATIVE	2.20	1.58	108292
NEUTRAL	3.87	1.57	252125
POSITIVE	4.70	0.78	481765

En lo que respecta al análisis de la dimensión semántica de las reseñas, segundo de nuestros objetivos, se comienza con la generación de una nube de palabras a partir del contenido textual de las reseñas, con el objetivo de identificar visualmente los términos más recurrentes utilizados por los usuarios. Esta representación gráfica permite observar rápidamente las temáticas predominantes y el lenguaje común empleado en los comentarios, aportando una primera aproximación a los intereses, valoraciones o problemáticas más mencionadas (Figura 4). Para analizar las percepciones desde un punto de vista semántico en profundidad, se analizó el corpus desde dos perspectivas complementarias en: la extracción de palabras clave y el modelado temático.

Figura 4. Nube de palabras de las reseñas

Para identificar los términos claves del discurso de los usuarios según la polaridad emocional, se aplicó el algoritmo TF-IDF (Term Frequency – Inverse Document Frequency). La Tabla 4 muestra los términos con mayor puntuación TF-IDF en los grupos de reseñas positivas y negativas. En las reseñas positivas destacan expresiones como nice, help, amazing, learn o excellent, todas ellas asociadas a experiencias satisfactorias, gratitud, eficacia pedagógica y facilidad de uso. Por el contrario, en las reseñas negativas aparecen con frecuencia términos como app, dont, cant, like o doesnt, que reflejan problemas técnicos, insatisfacción con el funcionamiento o limitaciones en el acceso.

Tabla 4. Análisis TF-IDF de las reseñas

Reviews Positivas		*Reviews Negativas*	
word	score	word	score
nice	4.08804754	app	2.01410354
help	3.9800589	dont	2.90400895
amazing	3.97221573	cant	3.12883177
much	3.96831709	like	3.21111788
learning	3.96056511	doesnt	3.2908565
recommend	3.93389686	even	3.35197532
learn	3.92640619	bad	3.36041419
use	3.87194627	use	3.44665544
excellent	3.78729339	get	3.54104226
helpful	3.75534179	useful	3.55381383

El modelado temático se aplicó para identificar automáticamente los principales contenidos abordados por los usuarios en sus reseñas. Se empleó el enfoque Latent Dirichlet Allocation (LDA), que permitió agrupar palabras que coaparecen frecuentemente en contextos similares, generando cinco temas principales representados por sus diez términos más relevantes. Entre los temas detectados (Tabla 5) se encuentran la valoración positiva de la utilidad y resolución de problemas (Tema 1), las limitaciones del acceso gratuito o funciones premium (Tema 2), la apreciación emocional y experiencia satisfactoria (Tema 3), los problemas técnicos y funcionamiento de la aplicación (Tema 4) y facilidad de uso y experiencia de aprendizaje positiva (Tema 5). Este análisis permitió extraer patrones latentes relevantes para comprender en mayor profundidad las experiencias y valoraciones expresadas por los usuarios.

Tabla 5. Agrupación por temas de las reseñas LDA

Tema 1	good, application, lot, helps, app, recommend, helped, solve, help, understand
Tema 2	app, free, im, pay, like, premium, learn, use, dont, using
Tema 3	app, love, best, excellent, helpful, amazing, useful, thank, great, study
Tema 4	app, like, answer, dont, work, time, use, doesnt, keyboard, answers
Tema 5	great, easy, use, app, learn, learning, really, way, fun, brainlyDebe contener el impacto, forma de evaluar dicho impacto y resultados.

4. CONCLUSIONES

Los resultados permiten responder al objetivo general del estudio: caracterizar las percepciones de los usuarios sobre las apps educativas con inteligencia artificial mediante análisis de sentimiento y modelado temático, y validar la consistencia del modelo. En relación con el primero de los objetivos, el análisis de sentimiento clasificó el 57 % de las 842.248 reseñas como positivas, el 30 % como neutras y el 13 % como negativas. Este alto porcentaje reseñas clasificadas como positivas y una puntuación media de 4,13 sobre 5 reflejan una experiencia de usuario satisfactoria. Para validar la consistencia del modelo, se calculó la correlación entre sentimiento textual y puntuación. El coeficiente de Pearson ($r = 0{,}5601$; $p < 0{,}001$) mostró una asociación positiva y significativa. Esto permite concluir que existe una alineación consistente entre la valoración emocional expresada en el texto y la puntuación cuantitativa, lo que refuerza la validez del modelo de análisis semántico aplicado. Además, el modelo RoBERTa presentó mayor confianza en reseñas con puntuaciones altas, lo que sugiere que los comentarios positivos son más claros desde el punto de vista emocional. En lo que respecta al análisis de la dimensión semántica de las reseñas, se emplearon dos técnicas complementarias: TF-IDF para identificar términos clave y LDA para detectar patrones temáticos. En cuanto a los términos clave en las reseñas positivas destacaron “nice” “help” “amazing” “learning” que sugieren experiencias satisfactorias, gratitud y eficacia. En las negativas, “app”, “dont”, “cant”, “like” que reflejan insatisfacción con el funcionamiento de la aplicación. Con el modelo LDA se agruparon las reseñas en cinco temas principales: (1) utilidad y resolución de problemas; (2) críticas al acceso limitado o funciones premium; (3) satisfacción emocional; (4) fallos técnicos; y (5) facilidad de uso y aprendizaje personalizado. Estos resultados no solo aportan evidencias empíricas sobre la experiencia del usuario con este tipo de apps, sino que también validan la metodología utilizada, al mostrar una fuerte coherencia entre puntuación, sentimiento y temas. En conjunto, se demuestra que el enfoque propuesto permite convertir información cualitativa masiva en conocimiento útil para el diseño y la mejora de aplicaciones educativas basadas en IA.

La metodología utilizada (basada en técnicas de minería de texto, análisis de sentimiento y modelado temático con aprendizaje automático) ha demostrado ser sostenible, automatizable y aplicable a gran escala. En trabajos futuros se analizará en profundidad las reseñas negativas para identificar patrones de insatisfacción, así como estudiar los desafíos éticos asociados al uso de inteligencia artificial en la educación, especialmente en términos de equidad, privacidad y transparencia. Este enfoque contribuye a la mejora continua de las tecnologías educativas, al ofrecer una herramienta robusta para captar y analizar la opinión del usuario a gran escala.

REFERENCIAS

Arambepola, N., Munasinghe, L. and Warnajith, N. (2024). Factors Influencing Mobile App User Experience: An Analysis of Education App User Reviews, 2024 4th International Conference on Advanced Research in Computing (ICARC), Belihuloya, Sri Lanka, 2024, pp. 223-228, doi: 10.1109/ICARC61713.2024.10499727.

Baba, K., El Faddouli, N.-E., & Cheimanoff, N. (2024). Mobile-Optimized AI-Driven Personalized Learning: A Case Study at Mohammed VI Polytechnic University. International Journal of Interactive Mobile Technologies, 18(4), pp. 81–96. https://doi.org/10.3991/ijim.v18i04.46547

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(Jan), 993-1022.

Ganguly, R. and Dasari, N. (2024). Comparative Study Of AI-Driven Ed-Tech Applications: Insights From Google Play Store Data. International Conference on TVET Excellence & Development (ICTeD), Melaka, Malaysia, 2024, pp. 237-243, doi: 10.1109/ICTeD62334.2024.10844660

Krishnan, V. & Zaini, H. (2025). A Systematic Literature Review on Artificial Intelligence in English Language Education. International Journal of Research and Innovation in Social Science, 9 (3), pp. 17-27. https://doi.org/10.47772/ijriss.2025.903sedu0002

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach (No. arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Mondal, A. S., Zhu, Y., Bhagat, K. K., & Giacaman, N. (2022). Analysing user reviews of interactive educational apps: a sentiment analysis approach. Interactive Learning Environments, 32(1), 355–372. https://doi.org/10.1080/10494820.2022.2086578

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

Yuen, C. L., & Schlote, N. (2024). Learner Experiences of Mobile Apps and Artificial Intelligence to Support Additional Language Learning in Education. Journal of Educational Technology Systems, 52(4), pp. 507-525. https://doi.org/10.1177/00472395241238693

Zahoor, K., & Bawany, N. Z. (2023). Explainable artificial intelligence approach towards classifying educational android app reviews using deep learning. Interactive Learning Environments, 32(9), 5227–5252. https://doi.org/10.1080/10494820.2023.22127