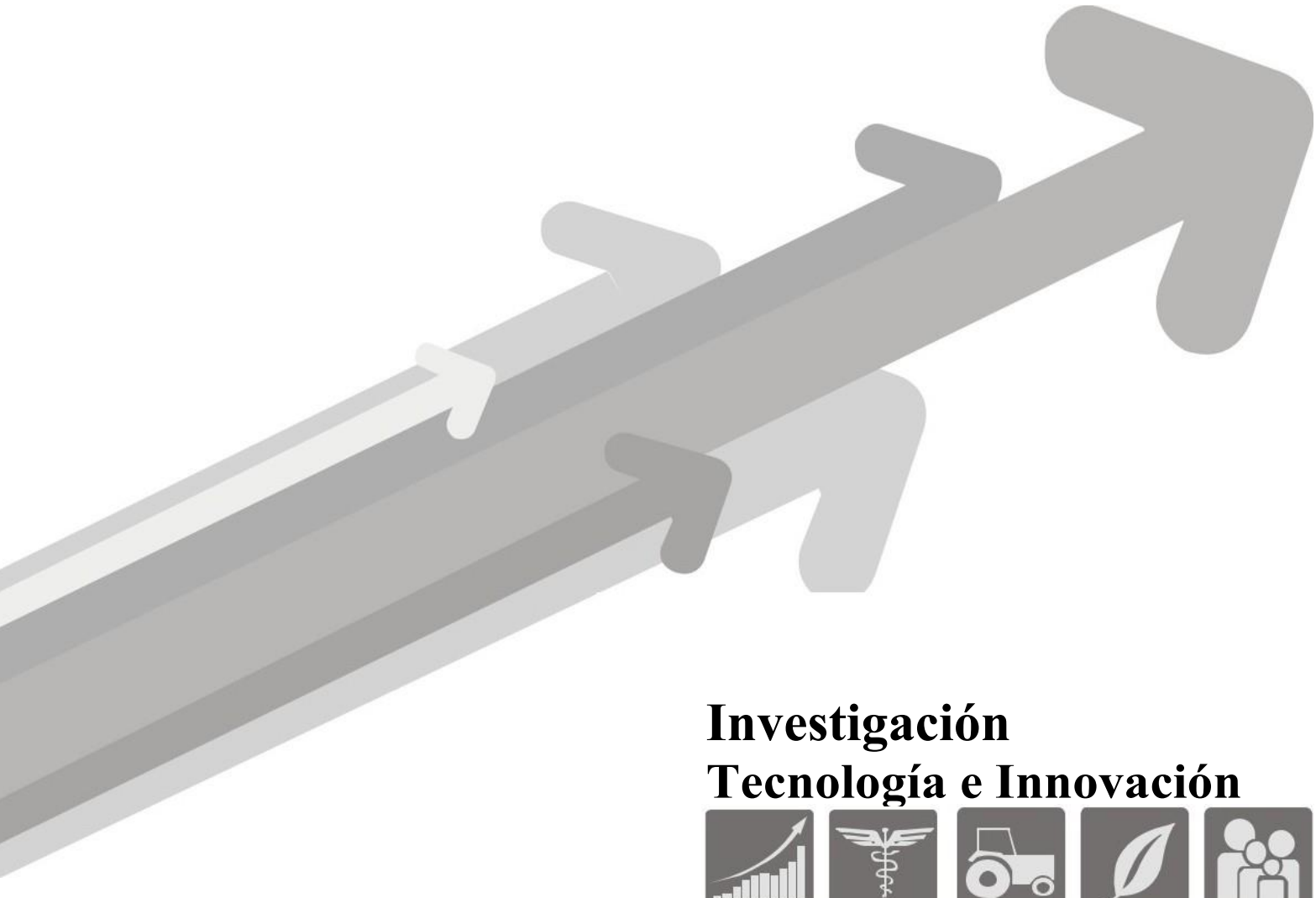
INVESTIGACIÓN, TECNOLOGÍA E INNOVACIÓN Vol. 17 No. 24 NOVIEMBRE - 2025 (94-105)
ISSN: 1390-5147 / **e-ISSN:** 2661-6548
DOI: https://doi.org/10.53591/iti.v17i24.2639

Inteligencia Artificial en el Proceso de Evaluación por Pares en la Educación Superior

Artificial Intelligence in the Peer Assessment Process in Higher Education

Anthony Holguin-Parraga
Maricela Pinargote-Ortega



Investigación
Tecnología e Innovación

INVESTIGACIÓN, TECNOLOGÍA E INNOVACIÓN Vol. 17 No. 24 NOVIEMBRE - 2025 (94-105)
ISSN: 1390-5147 / **e-ISSN:** 2661-6548
DOI: https://doi.org/10.53591/iti.v17i24.2639

Inteligencia Artificial en el Proceso de Evaluación por Pares en la Educación Superior

Artificial Intelligence in the Peer Assessment Process in Higher Education

Anthony Holguin-Parraga[1], y Maricela Pinargote-Ortega[2]

Como citar: Holguin-Párraga, A., Pinargote-Ortega, M. (2025). Inteligencia Artificial en el Proceso de Evaluación por Pares en la Educación Superior. *Investigación, Tecnología e Innovación*. 17(24), 94-105. DOI: https://doi.org/10.53591/iti.v17i24.2639

RESUMEN

Contexto: En los últimos años, la inteligencia artificial ha contribuido en la automatización de procesos de evaluación por pares en la educación superior. Sin embargo, persisten desafíos en la calificación de las evaluaciones cualitativas en la Universidad Técnica de Manabí del Ecuador. **Objetivo:** Crear un modelo de análisis de sentimientos basado en aspectos de retroalimentación por pares en idioma español. **Método:** Se aplicó una metodología experimental, mediante las fases del proceso estándar interindustrial para el aprendizaje automático. Se empleó técnicas de tokenización contextual, etiquetado BIO, reconocimiento de entidades nombradas, y balanceo de clases a través de aumento de datos. Se utilizó representaciones de codificadores bidireccionales a partir de transformadores para la comprensión contextual y una red neuronal convolucional para la detección de patrones. **Resultados:** El modelo alcanzó una precisión de 93,61%, y una puntuación F1 de 93,62%, con una pérdida de 16,77%, y una perplejidad de 1,18. **Conclusiones:** El modelo computacional se postula como una herramienta útil para analizar retroalimentaciones en procesos de evaluación por pares en el ámbito de la educación superior. Como trabajo futuro, se planea ampliar el conjunto de datos, e implementar otros algoritmos de análisis de sentimientos para mejorar la identificación de aspectos y la predicción de sentimiento por cada aspecto, consolidando su aplicabilidad en diversos contextos educativos en el idioma español.

Palabras clave: análisis de sentimientos basado en aspecto, educación superior, evaluación por pares, inteligencia artificial, procesamiento de lenguaje natural.

ABSTRACT

Context: In recent years, artificial intelligence has contributed to the automation of peer assessment processes in higher education. However, challenges remain in grading qualitative assessments at the Technical University of Manabí in Ecuador. **Objective:** To create a sentiment analysis model based on aspects of peer feedback in the Spanish language. **Method:** An experimental methodology, utilizing the phases of the cross industry standard process for machine learning. Techniques such as contextual tokenization, BIO tagging, named entity recognition, and class balancing through data augmentation were employed. Bidirectional encoder representations from transformers were used for contextual understanding, along with a convolutional neural network for pattern detection. **Results:** The model achieved an accuracy of 93.61%, an F1 score of 93.62%, a loss of 16.77%, and a perplexity of 1.18. **Conclusions:** The computational model is proposed as a useful tool for analyzing feedback in peer assessment processes within higher education. Future work aims to expand the data set and

[1]Máster en Ciencias de Datos y Máquinas de Aprendizaje, Universidad Técnica de Manabí, Ecuador. Correo electrónico: aholguin0694@utm.edu.ec
[2]Doctora en Computación Avanzada, Energía y Plasmas, Universidad Técnica de Manabí, Ecuador. Correo electrónico: maricela.pinargote@utm.edu.ec

ISSN: 1390-5147 / e-ISSN: 2661-6548
DOI: https://doi.org/10.53591/iti.v17i24.2639

implement other sentiment analysis algorithms to enhance the identification of aspects and the prediction of sentiment for each aspect, solidifying its applicability in various educational contexts in the Spanish language.

Keywords: aspect-based sentiment analysis, artificial intelligence, higher education, natural language processing, peer feedback.

Fecha de recepción: Septiembre 19, 2025.

Fecha de aceptación: Noviembre 12, 2025.

Introducción

En los últimos años, el uso de la inteligencia artificial ha causado una gran transformación en varios escenarios de la educación superior, en los que se destaca el análisis de sentimientos, que es una técnica NLP (Natural Language Processing) que permite identificar el sentimiento de las opiniones, emociones y valoraciones subjetivas de textos. Esta técnica se ha implementado en varios contextos académicos para la automatización de procesos como la retroalimentación en entornos virtuales y revisión por pares (Cárdenas, 2023; Ulises et al., 2021).

La evaluación por pares es una estrategia pedagógica que consiste en el análisis crítico y el aprendizaje colaborativo, y entre sus limitaciones están la subjetividad, el tiempo que conlleva hacerlo, y la inconsistencia en la calidad de la retroalimentación. Diversos trabajos han demostrado que el uso de la inteligencia artificial puede automatizar tareas esenciales en la educación superior como la clasificación de sentimientos, verificación de autenticidad y contexto de las retroalimentaciones, dando como resultado una mejora de la eficiencia y la calidad de la retroalimentación (Alqahtani et al., 2023; Pinargote-Ortega et al., 2023).

Sin embargo, las técnicas tradicionales de análisis de sentimientos, como Bag of Words, Word Embenddings, tienen las limitaciones al momento de captar el contexto semántico (Kasri et al., 2022). En consecuencia, a estas restricciones, se han desarrollado modelos más avanzados como BERT (Bidirectional Encoder Representations from Transformers), que trabajan con representaciones contextuales más precisas del lenguaje. Se ha evidenciado en trabajos recientes que la combinación de modelos BERT con CNN (Convolutional Neural Network), se obtiene como resultado un incremento importante de la precisión del análisis de sentimientos por aspectos (Martínez-Seis et al., 2022; Montañez Castelo et al., 2024).

En este contexto, se propone la creación de un modelo computacional aplicando la combinación de BETO (variante de BERT en español) y una arquitectura CNN, con la finalidad de realizar análisis de sentimientos orientado a aspectos (ABSA), de retroalimentación por pares en el ámbito educativo. El procedimiento se estructuró a partir de trabajos recientes de ABSA, con técnicas como la tokenización contextual, la extracción automática de aspectos con modelos NER (Named Entity Recognition), y la clasificación supervisada con Transformers (Gul et al., 2025; Wang & Xu, 2025).

El modelo híbrido BETO-CNN se desarrolló empleando las fases de la metodología CRISP-ML (Cross Industry Standard Process For Machine Learning), alcanzando buenos resultados en el análisis de sentimientos de retroalimentación por pares en la Universidad Técnica de Manabí del Ecuador. El modelo servirá como una herramienta adaptable a diferentes entornos educativos, contribuyendo a mejorar la eficiencia y calidad de los procesos de evaluación.

El artículo está organizado de la siguiente manera: la Sección 1 presenta la introducción y literatura relevante, la Sección 2 describe la metodología empleada, la Sección 3 presenta los hallazgos

ISSN: 1390-5147 / e-ISSN: 2661-6548
DOI: https://doi.org/10.53591/iti.v17i24.2639

obtenidos, la Sección 4 detalla el análisis con otros estudios, y finalmente, la Sección 5 exterioriza las conclusiones del estudio.

Trabajo relacionado

En los últimos años, las investigaciones sobre evaluación por pares han ido en aumento en el ámbito de la educación superior. En la Universidad de Istanbul Kültür y la de Qatar se analizó el efecto del aprendizaje a distancia durante el brote de COVID-19, aplicaron el modelo Long Short-Term Memory (LSTM) y Word2Vec con más de 160 000 tweets en turco para clasificar sentimientos, obtuvieron una precisión de 75,9% (Sadigov et al., 2024).

En la Universidad de Mansoura y la de King Faisal, se desarrolló un sistema para automatizar la evaluación de exámenes universitarios, que utilizó técnicas de NLP y Logistic Regression (LR), en conjunto con CountVectorizer, se alcanzó una precisión promedio de 98,5% (Ahmed & Sorour, 2024).

En la Universidad de McMaster de Canadá, aplicaron un pipeline de machine learning para evaluar comentarios narrativos basados en el entorno laboral utilizando vectorización de n-Grams y Lineal Regression, logrando una precisión del 87% en la identificación de residentes en riesgo de bajo rendimiento (Yilmaz et al., 2022).

En la Universidad de Omán, realizaron un estudio de análisis de un conjunto de datos de 6 514 estudiantes de un histórico de diez años con el objetivo de predecir resultados relacionados con la aprobación académica, utilizaron algoritmos como Logit Boost, Vote, Bagging y Decisión Trees (DT), siendo este último el que tuvo mejor desempeño con una precisión de 82,4% (Al-Alawi et al., 2023).

En la Universidad de Utara Malaysia desarrollaron un algoritmo con DT y LR con un conjunto de datos de 7706 registros estudiantiles en inglés de 2018 a 2019, siendo el DT, el que dio mejor rendimiento con una precisión de 89,49% utilizando un Split de 80 y 20 (Roslan et al., 2025).

En la Universidad de National Textile y la de King Saud, desarrollaron un framework con ABSA en una arquitectura Multi-Task Learning que integra paraphrasing, el cual fue probado en un conjunto de datos de 11.000 evaluaciones de cursos de inglés, el algoritmo que tuvo mejor rendimiento fue SP-ML-BERT con 96,3% en exactitud y 98,79% en precisión (Gul et al., 2025).

En la Universidad de Dian Nuswantoro propusieron un método para ABSA utilizando Embeddings de Sentence-BERT, Bayesian Search Clustering y mecanismos de atención Bigbird, el conjunto de datos utilizado corresponde a reseñas hoteleras, logrando una precisión del 99%, que destaca el potencial de arquitecturas basadas en Transformers en el análisis de sentimiento (Marutho et al., 2024).

En la Universidad de Yang-En y la de East Chine Jiaotong experimentaron con un modelo Hybrid Graph Neural Network que logró mejorar la comprensión del contexto de tipos de dependencias y las relaciones inter-aspectos, destacándose de los modelos tradicionales en varios conjuntos de datos estándar, logrando una precisión de 87% con el conjunto de datos de restaurants con el modelo HGNN-BERT (Zhao et al., 2024).

En la Universidad de Diponegoro y la de Muria Kudus desarrollaron un modelo de ABSA, utilizando el modelo IndoBert para evaluar retroalimentación estudiantil en idioma indonesio, el conjunto de datos contenía 10 000 reseñas estudiantiles, se realizaron ajustes a IndoBERT para poder usar matices lingüísticos específicos del idioma, el modelo obtuvo una precisión de 97,9% y un F1-score de 97,4% (Jazuli et al., 2024).

ISSN: 1390-5147 / e-ISSN: 2661-6548
DOI: https://doi.org/10.53591/iti.v17i24.2639

En la Universidad Tecnológica de la Habana "José Antonio Echeverría" y la de Castilla-La Mancha trabajaron en el desarrollo de un modelo basado en aprendizaje profundo para mejorar ABSA en español, aplicando modelos BERT para la extracción de aspectos y RoBERTa para la clasificación de sentimientos, el conjunto de datos entrenado proviene de reseñas de restaurants en español SemEval-2016 Task 5, alcanzando una puntuación de F1-score de 84,8% y una precisión de 86,6% (Montañez Castelo et al., 2024).

En la Universidad de Zhengzhou, realizaron un sistema de soporte de decisiones denominado Aspect2Labels (A2L), que utiliza multi-layer topic modeling, zero-shot learning y técnicas avanzadas de clasificación de sentimientos para anotar y analizar retroalimentación estudiantil; el modelo tuvo un rendimiento del 97% con Support Vector Machines y 93% con Artificial Neural Networks (Hussain et al., 2022).

En el Instituto Tecnológico de Monterrey en México analizaron un conjunto de datos de opiniones sobre actitudes de la planta docente hacia la tecnología educativa y las tendencias pedagógicas, aplicaron métodos estadísticos y análisis de sentimientos basado en TextBlob, como resultado obtuvieron que el 84,1% de los profesores opinaban sentimientos positivos, revelando diferencias en la apertura a la innovación, insights útiles en la aplicación de estrategias de desarrollo docente (Mahrishi et al., 2025).

En la Universidad Metropolitana de Hong Kong, aplicaron la estrategia de aumento de datos utilizando grandes modelos de lenguaje como GPT-3.5, mejorando la clasificación de sentimientos basada en aspectos usando BERT, generando nuevos términos de aspectos, obteniendo como resultado una precisión del 87,23% con el modelo BERT-MATP en combinación con el conjunto de datos restaurant, y con el conjunto de datos Laptop obtuvo 81,50% (Wang & Xu, 2025).

En la Universidad de Correos y Telecomunicaciones de Xi'an en China, propusieron DCASAM, que combina DBiLSTM y Graph Convolutional Networks densamente conectadas, el modelo se evaluó en un conjunto de datos standars como restaurant14, laptop14, y Twitter, logrando como resultado un 80,56% en precisión y del 77% en F1-Score (Jiang et al., 2024).

Materiales y métodos

Se utilizó un enfoque cuantitativo-computacional, siguiendo las fases del modelo CRISP-ML, adaptado de Studer (2021). A continuación, se describe las fases:

I fase. Comprensión del problema y recolección de datos

Los datos fueron recolectados previamente mediante un prototipo de evaluación entre pares implementado en la Universidad Técnica de Manabí, donde los estudiantes proporcionaron retroalimentación en español sobre las tareas de sus compañeros (Pinargote-Ortega et al., 2023). Se utilizaron datos obtenidos de estudiantes de diversas carreras, incluyendo la carrera de ingeniería de sistemas informáticos en modalidades presencial, virtual sincrónica y asincrónica; tecnologías de la información en modalidad presencial; y pedagogía de la química en modalidad virtual asincrónica.

Se tuvo como objetivo desarrollar un componente de análisis de sentimientos basado en aspectos de retroalimentación entre pares, utilizando modelos de lenguaje avanzados entrenados en español.

II fase. Preparación de los datos

Se aplicó un preprocesamiento orientado a modelos Transformers para análisis de sentimientos en español. No se eliminaron palabras funcionales, ni se realizó lematización, dado que los modelos como BETO utilizan el contexto completo de las frases para construir representaciones semánticas

INVESTIGACIÓN, TECNOLOGÍA E INNOVACIÓN Vol. 17 No. 24 NOVIEMBRE - 2025 (94-105)
ISSN: 1390-5147 / **e-ISSN:** 2661-6548
DOI: https://doi.org/10.53591/iti.v17i24.2639

(Montañez Castelo et al., 2024).

Se efectuaron los siguientes pasos:

- Normalización básica: se hizo corrección ortográfica, y eliminación de espacios innecesarios.
- Tokenización contextualizada: se aplicó BertTokenizerFast, que está incluido en BETO.
- Codificación: el texto se transformó en vectores numéricos por medio de embeddings contextuales como entrada al modelo.
- Extracción de aspectos: se utilizó el modelo BETO fine-tuned como token-classifier (NER).
- Clasificación de sentimiento por aspecto: se aplicó un clasificador supervisado basado en Transformers, entrenado con anotaciones previas.
- Validación manual parcial: se realizó una verificación para garantizar la confiabilidad del etiquetado.

III fase. Entrenamiento del modelo

Se realizó un modelo híbrido, con BETO para la comprensión contextual de los aspectos y CNN para la detección de patrones y clasificación. Enfoque que ha dado excelentes resultados en otros estudios de análisis de sentimiento basado en aspectos al combinar la comprensión profunda de los Transformers y la capacidad de detección de patrones (Jazuli et al., 2024; Martínez-Seis et al., 2022).

IV fase. Evaluación del modelo

El corpus se dividió en conjuntos de entrenamiento y validación mediante un muestreo estratificado para preservar la distribución de clases. Para validar el rendimiento del modelo, se emplearon métricas estándar utilizadas en tareas de clasificación multiclase (Jazuli et al., 2024).

- Accuracy: se evaluaron todas las clases.

$$Accuracy = \frac{(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP})} \quad (1)$$

- Precision y Recall: se realizaron por clase, evaluando exactitud y cobertura.

$$Precision = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})} \quad (2)$$

$$Recall = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \quad (3)$$

- F1-score: se valoró el equilibrio entre precisión y recuperación.

$$F1 = \frac{2 \text{ x Precision x Recall}}{\text{Precision+Recall}} \quad (4)$$

- Log Loss: se calculó el ajuste probabilístico del modelo, penalizando las predicciones incorrectas con mayor incertidumbre; valores más bajos indican una mejor calibración del modelo.

$$Log\ Loss = -\frac{1}{\mathrm{N}}\sum_{i=1}^{N}\sum_{c=1}^{M} y_{i,c}\log(p_{i,c}) \quad (5)$$

- Perplexity: se midió la incertidumbre del modelo, donde una menor perplejidad indica predicciones más confiables y consistentes.

$$Perplexiy = e^{-\frac{1}{N}\sum_{i=1}^{N}\log(p_i)} \quad (6)$$

INVESTIGACIÓN, TECNOLOGÍA E INNOVACIÓN Vol. 17 No. 24 NOVIEMBRE - 2025 (94-105)
ISSN: 1390-5147 / **e-ISSN:** 2661-6548
DOI: https://doi.org/10.53591/iti.v17i24.2639

Resultados

En esta sección se presentan las configuraciones del experimento, el procesamiento del conjunto de datos, la configuración de hiperparámetros y la arquitectura del modelo.

Comprensión del problema

Se utilizó un conjunto de datos de evaluación entre pares, compuesto por 10 118 instancias, provenientes del prototipo publicado en Pinargote-Ortega et al. (2023). Luego, se aplicó desbalanceo de clases para mejorar la generalización del modelo, y se implementó una estrategia de aumento de datos basada en traducción automática, que implicó la traducción de los comentarios originales en español a un idioma intermedio, posteriormente al inglés y de vuelta a la lengua española, lo que permitió dispersar la semántica de los enunciados y producir oraciones nuevas semánticamente equivalentes, pero con una diversidad sintáctica mayor.

En la **Tabla 1.** se detalla la distribución de las clases de sentimiento en el conjunto de datos tras aplicar aumento de datos mediante traducción inversa, alcanzando un total de 14 963 instancias.

Tabla 1. Conjunto de datos de retroalimentación entre pares

Instancias del prototipo	Instancias tras aplicación de aumento de datos	Sentimiento
6 838	6 550	Positivo
1 224	5 107	Neutral
2 056	3 306	Negativo
10 118	14 963	Total

Fuente: Autores

Preparación de datos

El conjunto de datos se preparó utilizando la librería spaCy y el modelo BETO. Se normalizaron los comentarios mediante la conversión a minúsculas, y la eliminación de las tildes. Luego, se etiquetaron los aspectos de los comentarios aplicando el modelo es_core_new_md de spaCy, y se implementó una función heurística para filtrar los artículos asociados a los aspectos. Una vez obtenido el conjunto de datos con tokens y etiquetas BIO (Beginning, Inside, Outside), se guardó en formato JSON para ser utilizado en el entrenamiento del modelo.

Modelado

Se diseñó una arquitectura híbrida, con la capacidad contextual de BETO para representar los comentarios en español, y con la eficiencia de CNN para identificar patrones discriminativos de sentimiento, localizados en torno a los aspectos detectados.

Bajo esa perspectiva, el modelo se desarrolló con dos capas: una capa para la extracción de aspectos mediante NER, y otra capa para la clasificación de sentimiento basado en aspectos, utilizando el modelo BETO y CNN. Para realizar la detección automática de aspectos en comentarios en español con NER, se utilizó fine-tuning del modelo BETO. Además, se usó BertForTokenClassification para tokenizar la salida del modelo y realizar las predicciones a nivel de token.

Una vez que se detectó los aspectos mediante NER, se procedió a la clasificación del sentimiento asociado a cada aspecto utilizando una arquitectura híbrida basada en representaciones contextuales de lenguaje y redes convolucionales. Con ese enfoque, se reutilizó BETO para generar embeddings generalizado, sin fine-tuning. El conjunto de datos compuesto por comentarios con su respectivo

ISSN: 1390-5147 / e-ISSN: 2661-6548
DOI: https://doi.org/10.53591/iti.v17i24.2639

sentimiento fue entrenado, junto con su aspecto correspondiente a la representación semántica del comentario, por lo cual la red neuronal convolucional se entrenó para identificar los patrones discriminativos del sentimiento.

La configuración de los hiperparámetros del modelo se efectuó con el optimizador AdamW, con cross-entropy como función objetivo. La primera capa convolucional fue de una dimensión (Conv1D), con 128 filtros, el tamaño del kernel fue de 3. En la siguiente capa se utilizó una función ReLu, con el objetivo de que no sea lineal, y aumentar una mayor representación del modelo.

En los primeros experimentos no se usó dropout, donde se obtuvo resultados de accuracy por debajo de 87%, lo que conllevó a realizar el aumento de datos para balancear el conjunto de datos, y se agregó un dropout de 30% después de la primera capa convolucional, y antes de la capa densa final, esto ayudó a conseguir un accuracy del 93%, reduciendo el sobreajuste.

Además, se aplicó un max-pooling adaptativo con el objetivo de extraer las características más relevantes, incluso en sentencias largas. En la última capa, se configuró un Dense Layer Linear de 128 unidades, con una salida de 3 neuronas, cada una correspondiente a las clases positivo, neutral y negativo. Esta capa no utiliza una función de activación explícita, porque la función de pérdida CrossEntropyLoss usa softmax implícitamente para obtener una distribución de probabilidad sobre las clases.

Evaluación

Los resultados se visualizaron mediante gráficas comparativas de precisión y pérdida durante el entrenamiento y la validación, que muestran una curva de mejora progresiva y consistente en el rendimiento del modelo, alcanzando una precisión del 94,23% en el entrenamiento y del 93,61% en la validación en la época 30. Asimismo, la pérdida disminuye de manera sostenida, estabilizándose en 15,28% en el entrenamiento y en validación en 16,71%, como se muestra en la **Figura 1**.

Se observa que en las épocas 9 y 13 se presentan pequeñas variaciones en el rendimiento durante el entrenamiento y la validación. No obstante, a partir de la época 15, el modelo se estabiliza, logrando la convergencia en la época 30, con diferencias mínimas de 0,01 en accuracy y en Loss. Estos resultados evidencian que el modelo tiene un excelente equilibrio entre el ajuste de los datos y la capacidad de generalización **Tabla 2** y **Figura 1**. Además, la efectividad del modelo para tareas de clasificación multiclase se corrobora por lo alcanzado en F1-score (93,62%), Log Loss (16,77%) y perplexity (1,18) (**Tabla 3**).

Para validar la eficacia del modelo, se realizaron pruebas con frases que contienen varios aspectos. El modelo descompone la frase por aspectos, luego emplea el modelo NER para identificar y abstraer los aspectos. De esta forma, se obtiene el input aspecto [SEP] comentario, el cual es procesado por el modelo CNN + BETO, para predecir el sentimiento asociado a cada aspecto (**Tabla 4**).

INVESTIGACIÓN, TECNOLOGÍA E INNOVACIÓN Vol. 17 No. 24 NOVIEMBRE - 2025 (94-105)
ISSN: 1390-5147 / **e-ISSN:** 2661-6548
DOI: https://doi.org/10.53591/iti.v17i24.2639

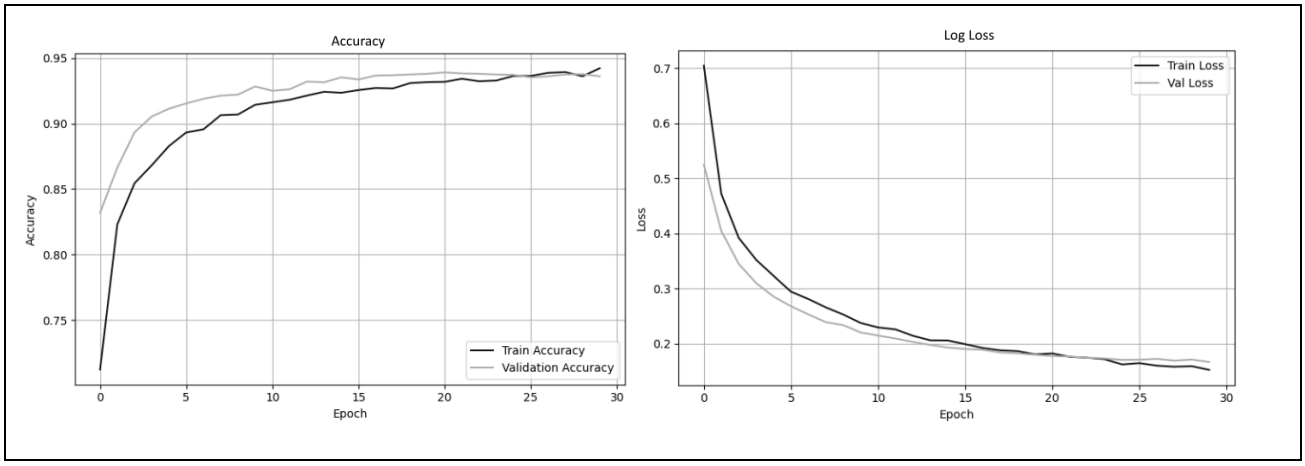


Figura 1. Curvas de precisión y pérdida del entrenamiento y validación**.**

Fuente: Autores

Tabla 2. Entramiento y validación del modelo

Época	Accuracy Entrenamiento	Accuracy Validación	Train Loss	Val Loss
1	0,7122	0,8319	0,7044	0,5249
5	0,8829	0,9114	0,3232	0,2856
10	0,9144	0,9284	0,2378	0,2206
15	0,9236	0,9354	0,2061	0,1933
20	0,9317	0,9381	0,1809	0,1802
25	0,9364	0,9372	0,1626	0,1709
28	0,9394	0,9375	0,1585	0,1695
29	0,9362	0,9378	0,1595	0,1714
30	**0,9423**	**0,9361**	**0,1528**	**0,1671**

Fuente: Autores

Tabla 3. Resumen de los resultados del mejor modelo

Métrica	Resultado
Accuracy	93,61%
F1-score	93,62%
Precision	93,64%
Recall	93,61%
Log Loss	16,77%
Perplexity	1,18%

Fuente: Autores

INVESTIGACIÓN, TECNOLOGÍA E INNOVACIÓN Vol. 17 No. 24 NOVIEMBRE - 2025 (94-105)
ISSN: 1390-5147 / **e-ISSN:** 2661-6548
DOI: https://doi.org/10.53591/iti.v17i24.2639

Tabla 4. Ejemplo de detección de aspecto y predicción de sentimiento

Comentario	Aspecto detectado	Sentimiento predicho	Negativo	Neutral	Positivo
Los actores están bien especificados y son necesarios en el sistema. El actor sistema está mal	actores	Positivo	0,0027	0,1572	**0,8401**
	sistema	Positivo	0,0027	0,3011	**0,6961**
	actor sistema	Negativo	**0,7864**	0,0915	0,1220
La estructura es parcialmente adecuada, el caso de uso no está comunicado con ningún actor, pero los actores involucrados si cumplen la función	estructura	Neutral	0,0004	**0,9970**	0,0026
	caso de uso	Negativo	**0,6838**	0,0386	0,2775
	actores involucrados	Positivo	0,1104	0,0520	**0,8376**

Fuente: Autores

Tabla 5. Comparación del estudio con otras investigaciones

Estudio	Precisión	Comparación con este trabajo
Este trabajo	**93,61%**	---
(Montañez Castelo et al., 2024)	86,60%	Inferior
(Wang & Xu, 2025)	81,50%	Inferior
(Yilmaz et al., 2022)	87,00%	Inferior
(Al-Alawi et al., 2023)	82,40%	Inferior
(Roslan et al., 2025)	89,49%	Inferior
(Zhao et al., 2024)	87,00%	Inferior
(Mahrishi et al., 2025)	84,10%	Inferior
(Jiang et al., 2024)	86,70%	Inferior
(Sadigov et al., 2024)	75,90%	Inferior

Fuente: Autores

Discusión

El modelo de análisis de sentimientos basado en aspectos de retroalimentación entre pares, obtuvo una precisión del 93,61%. Este resultado es consistente con otros estudios que han superado el umbral del 80%, como el de Montañez Castelo et al. (2024) que lograron una precisión del 86,6%, Wang & Xu (2025) que alcanzaron el 81,50%, Yilmaz et al. (2022) que obtuvieron el 87%, Al-Alawi et al. (2023) que consiguieron el 82,4%, Roslan et al. (2025) que obtuvieron el 89,49%, Zhao et al. (2024) que alcanzaron el 87%, Mahrishi et al. (2025) que concretaron el 84,1%, y Jiang et al. (2024) que adquirieron el 80,56%. En contraste con el estudio de Sadigov et al. (2024) que consiguieron una precisión del 75,9% (**Tabla 5**).

El modelo se destaca por su adaptabilidad lingüística al español, lo que lo convierte en una herramienta aplicable a contextos educativos en países hispanohablantes. Similar con el estudio de Montañez Castelo et al. gul (2024) que aplicaron modelos preentrenados en español para mejorar el análisis de sentimiento en este idioma. Sin embargo, se diferencia de otros modelos desarrollados en

ISSN: 1390-5147 / e-ISSN: 2661-6548
DOI: https://doi.org/10.53591/iti.v17i24.2639

otros idiomas, como el inglés (Gul et al., 2025), el turco (Sadigov et al., 2024) y el indonesio (Jazuli et al., 2024).

Además, el modelo presenta una arquitectura híbrida, que integra BETO para la comprensión contextual en español y CNN para la detección de patrones discriminativos. Esta combinación es comparable a la utilizada por Montañez Castelo et al. (2024), que utilizaron una combinación de BERT y RoBERTa, pero no incorporaron CNN para la detección de patrones. En contraste con otros estudios, como el de Sadigov et al. (2024) que emplearon LSTM con Word2Vec, una técnica efectiva para secuencias temporales, pero menos precisa en la captura de contexto bidireccional. Asimismo, Yilmaz et al. (2022) aplicaron la regresión lineal y n-grams con menos costos computacional, pero carece de la profundidad contextual.

En la recolección de datos mediante la evaluación entre pares se aplicaron principios éticos y el modelo pedagógico constructivista. Los estudiantes proporcionaron retroalimentaciones utilizando rúbricas con criterios personalizados para cada tarea, bajo la guía del docente, lo que fomentó el pensamiento crítico y promovió una calificación justa. Este enfoque es similar al planteado por Atarés Huerta et al. (2021), quienes destacan la importancia de establecer criterios claros en la evaluación por pares para reducir la subjetividad. Asimismo, Rodríguez-Espinosa et al. (2016) enfatizan la necesidad de garantizar la equidad en la evaluación, subrayando que la percepción del estudiantado respecto a la justicia y utilidad de la retroalimentación es fundamental en el ámbito pedagógico.

El modelo representa un avance significativo en la eficiencia y calidad de la retroalimentación por pares en la educación superior. No obstante, su impacto académico y social podría maximizarse si su implementación se extiende a otras asignaturas, carreras e instituciones educativas, manteniendo el marco ético y la estrategia pedagógica. Este enfoque es respaldado por Atarés Huerta et al. (2021), quienes afirman que la evaluación y el acuerdo de calificaciones entre estudiantes contribuyen a su formación académica, así como por Longarela-Ares & Rodríguez-Padín (2023), que relacionan el aprendizaje colaborativo y práctico con la evaluación entre pares, resaltando los beneficios de la interacción en este proceso.

Finalmente, es relevante resaltar que el modelo es aplicable en contextos educativos de idioma español, debido a la implementación del modelo BETO y a la capacidad de detección de patrones de CNN para identificar sentimientos basado en aspectos de comentarios sintácticamente variados. En contraste con Jazuli et al. (2024) que usaron IndoBERT con matices lingüísticos en idioma indonesio sin incorporar CNN.

Conclusiones

El modelo híbrido BETO-CNN logró una precisión de 93,61% en el conjunto de validación, con un Log Loss de 16,77% y una Perplexity de 1,18. Los resultados muestran un alto nivel de efectividad y confiabilidad en la clasificación de sentimientos multiclase de positivo, neutral y negativo en retroalimentación por pares.

La aplicación de traducción inversa mediante aumento de datos redujo el desbalance de clases y mejoró la generalización del modelo, ampliando la diversidad sintáctica sin alterar la semántica de las retroalimentaciones, lo que contribuyó a superar el umbral inicial de precisión del 87% en experimentos preliminares a más del 93%.

La incorporación de max-pooling adaptativo y de dropout ayudó a reducir el sobreajuste, en consecuencia, se reflejó en la estabilización de las curvas de precisión y pérdida durante el entrenamiento y la validación del modelo, y favoreció la identificación de características relevantes, incluso en secuencias largas.

ISSN: 1390-5147 / **e-ISSN:** 2661-6548
DOI: https://doi.org/10.53591/iti.v17i24.2639

El modelo tiene la capacidad de identificar aspectos y predecir el sentimiento asociado a cada aspecto, lo que lo presenta como una herramienta útil para analizar evaluaciones cualitativas en procesos de evaluación por pares en el ámbito de la educación superior. No obstante, este estudio presenta limitaciones en el tamaño del conjunto de datos, en comparación con corpus más amplios utilizados en otras investigaciones. En este sentido, en futuros trabajos se ampliará el corpus de datos y se explorarán otros algoritmos de ABSA para mejorar la detección de aspectos y predicción de sentimiento, fortaleciendo así la aplicabilidad del modelo en entornos educativos en español.

Referencias bibliográficas

Ahmed, H. M. M., & Sorour, S. E. (2024). Classification-driven intelligent system for automated evaluation of higher education exam paper quality. *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-12555-9

Al-Alawi, L., Al Shaqsi, J., Tarhini, A., & Al-Busaidi, A. S. (2023). Using machine learning to predict factors affecting academic performance: the case of college students on academic probation. *Education and Information Technologies*, *28*, 12407-12432. https://doi.org/10.1007/s10639-023-11700-0

Alqahtani, T., Badreldin, H. A., Alrashed, M., Alshaya, A. I., Alghamdi, S. S., bin Saleh, K., Alowais, S. A., Alshaya, O. A., Rahman, I., Al Yami, M. S., & Albekairy, A. M. (2023). The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy*, *19*(8), 1236-1242. https://doi.org/10.1016/J.SAPHARM.2023.05.016

Atarés Huerta, L., Antonio Llorens Molina, J., & Marin-Garcia, J. A. (2021). La evaluación por pares en Educación Superior. *Educación Química*, *1*, 32. https://doi.org/10.22201/fq.18708404e.2021.1.75905

Cárdenas, J. (2023). Inteligencia artificial, investigación y revisión por pares: escenarios futuros y estrategias de acción Artificial intelligence, research and peer-review: future scenarios and action strategies. *Revista Española de Sociología (RES) / Spanish Journal of Sociology*, *4*, 1-15. https://doi.org/10.22325/fes/res.2023

Gul, S., Asif, M., Fazal-E-Amin, Saleem, K., & Imran, M. (2025). Advancing Aspect-Based Sentiment Analysis in Course Evaluation: A Multi-Task Learning Framework with Selective Paraphrasing. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3527367

Hussain, S., Ayoub, M., Jilani, G., Yu, Y., Khan, A., Wahid, J. A., Butt, M. F. A., Yang, G., Moller, D. P. F., & Weiyan, H. (2022). Aspect2Labels: A novelistic decision support system for higher educational institutions by using multi-layer topic modelling approach. *Expert Systems with Applications*, *209*. https://doi.org/10.1016/j.eswa.2022.118119

Jazuli, A., Widowati, & Kusumaningrum, R. (2024). Optimizing Aspect-Based Sentiment Analysis Using BERT for Comprehensive Analysis of Indonesian Student Feedback. *Applied Sciences*, *15*(1), 172. https://doi.org/10.3390/app15010172

Jiang, X., Ren, B., Wu, Q., Wang, W., & Li, H. (2024). *DCASAM: advancing aspect-based sentiment analysis through a deep context-aware sentiment analysis model*. https://doi.org/https://doi.org/10.1007/s40747-024-01570-5

Kasri, M., Birjali, M., Nabil, M., Beni-Hssane, A., El-Ansari, A., & El Fissaoui, M. (2022). Refining Word Embeddings with Sentiment Information for Sentiment Analysis. *Journal of ICT Standardization*, *10*(3). https://doi.org/10.13052/jicts2245-800X.1031

Longarela-Ares, Á. M., & Rodríguez-Padín, R. (2023). Aprendizaje colaborativo, learning-by-doing y evaluación entre pares en educación superior. *EDUCA. Revista Internacional para la calidad educativa*, *3*(2), 275-298. https://doi.org/10.55040/EDUCA.V3I2.66

Mahrishi, M., Rodríguez Medellín, C. E., Yau Flores, M. del R., Guzmán Brito, M. P., & Abbas, A. (2025).

ISSN: 1390-5147 / **e-ISSN:** 2661-6548
DOI: https://doi.org/10.53591/iti.v17i24.2639

Opinion mining of professors' sentiments towards the adoption of technology, the latest teaching trends, and their effectiveness. *Social Sciences and Humanities Open*, *11*. https://doi.org/10.1016/j.ssaho.2024.101249

Martínez-Seis, B. C., Pichardo-Lagunas, O., Miranda, S., Perez-Cazares, I. J., & Rodriguez-González, J. A. (2022). Deep Learning Approach for Aspect-Based Sentiment Analysis of Restaurants Reviews in Spanish. *Computacion y Sistemas*, *26*, 899-908. https://doi.org/10.13053/CyS-26-2-4258

Marutho, D., Muljono, Rustad, S., & Purwanto. (2024). Optimizing aspect-based sentiment analysis using sentence embedding transformer, bayesian search clustering, and sparse attention mechanism. *Journal of Open Innovation: Technology, Market, and Complexity*, *10*(1), 100211. https://doi.org/10.1016/J.JOITMC.2024.100211

Montañez Castelo, P., Simón-Cuevas, A., Olivas, J. A., & Romero, F. P. (2024). Enhancing Spanish Aspect-Based Sentiment Analysis Through Deep Learning Approach. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *14335 LNCS*, 215-224. https://doi.org/10.1007/978-3-031-49552-6_19

Pinargote-Ortega, M., Bowen-Mendoza, L., Meza, J., & Ventura, S. (2023). Peer Feedback Sentiment Analysis Prototype[Prototipo de análisis de sentimiento de retroalimentación textual entre pares]. *Volume 2023, Issue E61, Pages 322 - 337*, *2023*(E61), 322-337.

Rodríguez-Espinosa, H., Fernando Restrepo-Betancur Gloria Cristina Luna-Cabrera, L., Fernando Restrepo-Betancur, L., & Cristina Luna-Cabrera, G. (2016). Percepción del estudiantado sobre la evaluación del aprendizaje en la educación superior. *Revista Electrónica Educare, ISSN-e 1409-4258, Vol. 20, Nº. 3, 2016*, *20*(3), 18. https://doi.org/10.15359/ree.20-3.18

Roslan, N., Jamil, J. M., Shaharanee, I. N. M., & Alawi, S. J. S. (2025). Prediction of Student Dropout in Malaysian's Private Higher Education Institute using Data Mining Application. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *45*, 168-176. https://doi.org/10.37934/araset.45.2.168176

Sadigov, R., Yıldırım, E., Kocaçınar, B., Patlar Akbulut, F., & Catal, C. (2024). Deep learning-based user experience evaluation in distance learning. *Cluster Computing*, *27*, 443-455. https://doi.org/10.1007/s10586-022-03918-3

Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K. R. (2021). Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. *Machine Learning and Knowledge Extraction 2021, Vol. 3, Pages 392-413*, *3*(2), 392-413. https://doi.org/10.3390/MAKE3020020

Ulises, J., Salinas, S., Carlos, J., & Diaz, T. (2021). Análisis de Sentimientos en los Mensajes Recibidos en el Entorno Virtual de Aprendizaje de la Modalidad Abierta y a Distancia de la UTPL. *Revista Ibérica de Sistemas e Tecnologias de Informação*, 98-113.

Wang, W., & Xu, L. (2025). *Aspect-based sentiment classification with BERT and AI feedback*. https://doi.org/10.1109/TBDATA.2025.3536934

Yilmaz, Y., Nunez, A. J., Ariaeinejad, A., Lee, M., Sherbino, J., & Chan, T. M. (2022). Harnessing Natural Language Processing to Support Decisions Around Workplace-Based Assessment: Machine Learning Study of Competency-Based Medical Education. *JMIR Medical Education*, *8*. https://doi.org/10.2196/30537

Zhao, H., Cui, C., & Wu, C. (2024). Hybrid Graph Neural Network-Based Aspect-Level Sentiment Classification. *Electronics*, *13*(16), 3263. https://doi.org/10.3390/electronics13163263