Artículo de Investigación Original

# Arquitecturas neuronales para la clasificación de sentimientos: una evaluación empírica de LSTM, BERT y CNN usando PyTorch

## Neural architectures for sentiment classification: an empirical evaluation of LSTM, BERT, and CNN using PyTorch

Gabriela Belén Uquillas Trujillo [1][0009-0009-2478-0711], Rebeca Mariana Moposita Lasso [2][0009-0004-9181-1008].

[1] Escuela Superior Politécnica de Chimborazo – Facultad de Ingeniería Informática y Electrónica – Escuela de Posgrado – Ecuador.
[2] Escuela Superior Politécnica de Chimborazo – Facultad de Ingeniería Informática y Electrónica – Tecnología de la Información – Ecuador.

gabriela.uquillas@espoch.edu.ec, rebeca.moposita@espoch.edu.ec.

**CITA EN APA:**
Uquillas Trujillo, G. B., & Moposita Lasso, R. M. (2025). Arquitecturas neuronales para la clasificación de sentimientos: una evaluación empírica de LSTM, BERT y CNN usando PyTorch . *Technology Rain Journal*, *4*(2). https://doi.org/10.55204/trj.v4i2.e100

**Recibido:** 14 de Mayo-2025
**Aceptado:** 23 de Junio-2025
**Publicado:** 15 de Octubre-2025

Technology Rain Journal
**ISSN:** 2953-464X



**Resumen**.
El análisis de sentimientos en español es una tarea fundamental debido al desarrollo de contenido en Internet La presente investigación estudia el desempeño de tres arquitecturas neuronales, CNN, Bi-LSTM y BERT para la clasificación de sentimientos en español, utilizando el conjunto de datos de reseñas cinematográficas IMDB. Siguiendo la metodología de Investigación en Ciencia del Diseño, se implementaron modelos en PyTorch con parámetros equivalentes para garantizar una comparación equitativa. Los resultados evidencian la superioridad de BERT con una precisión del 87.92%, F1-score de 87.89% y AUC-ROC de 0.947, seguido por Bi-LSTM, 86.29% y CNN, 85.64%. BERT destaca en la identificación de sentimientos negativos, mientras que CNN muestra un rendimiento más equilibrado entre clases. No obstante, BERT demanda recursos computacionales mayores, con un tiempo de entrenamiento de 2h14m versus 1m22s de Bi-LSTM. Este estudio proporciona evidencia empírica para la selección de arquitecturas neuronales en aplicaciones de análisis de sentimientos en español, considerando el balance entre precisión y eficiencia computaciona

**Palabras Clave:** Análisis de sentimientos, Procesamiento del lenguaje natural (PLN), Redes neuronales profundas, PyTorch.

**Abstract**:
Sentiment analysis is an essential task due to the great development of content on the Internet. This research examines the performance of three neural architectures, CNN, Bi-LSTM, and BERT, for sentiment classification in the Spanish language, using movie reviews from the IMDB dataset. Following the Design Science Research methodology, optimized models were implemented in PyTorch with equivalent parameters to ensure a fair comparison. The results demonstrate the superiority of BERT with an accuracy of 87.92%, F1-score of 87.89%, and AUC-ROC of 0.947, followed by Bi-LSTM, 86.29% and CNN, 85.64%. BERT excels in identifying negative sentiments, while CNN shows a more balanced performance across classes. However, BERT demands significantly greater computational resources, with a training time of 2h14m versus 1m22s for Bi-LSTM. This study provides empirical evidence for selecting neural architectures in Spanish sentiment analysis applications, considering the balance between accuracy and computational efficiency.

**Keywords:** Sentiment analysis, Natural language processing, Deep neural networks, PyTorch.

## 1. INTRODUCIÓN

El análisis de sentimientos ha sumado importancia dentro del procesamiento del lenguaje natural hoy en día debido a su amplia aplicación en varios tipos de industrias (Liu, 2020). Como ejemplo podemos enlistar a las empresas que brindan servicios, ya que han encontrado una manera

efectiva de entender la percepción que tienen los consumidores sobre sus marcas. Los datos se encuentran principalmente en redes sociales y se pueden monitorear incluso en tiempo real para validar los comentarios en línea y obtener valiosa información acerca del nivel de satisfacción de los clientes y sus expectativas, reseñas e incluso poder anticipar su comportamiento, según Fu et al. en 2022.

Otra de las industrias que se ha beneficiado del análisis de sentimientos es la del entretenimiento, cuyo objetivo se ha basado en realizar una interpretación de la reacción y aceptación del público ante diferentes tipos de contenido, e incluso anticipar tendencias y preferencias (Poria et al., 2023). Por lo cual, esta investigación se enfoca en el estudio de sentimientos en el idioma español, lo que plantea un desafío debido a la complejidad de nuestro idioma, dado que hay elementos como el sarcasmo, la ironía, o las ambigüedades presentes en el contexto de las reseñas cinematográficas que no pueden ser identificados de forma natural por una máquina o un algoritmo. Además, es fundamental considerar las diversas variaciones regionales y culturales que permitan entender e interpretar correctamente una opinión (Zhang et al., 2022).

Ante esta problemática, hemos decidido realizar un estudio comparativo entre diferentes arquitecturas de redes neuronales para la clasificación de sentimientos, estas son: las redes neuronales convolucionales (CNN), Bi-LSTM y BERT. Nuestro objetivo es establecer una sólida base teórica que permita entender las diferencias fundamentales entre estas arquitecturas, evaluar cuán apropiadas son para la clasificación de sentimientos, implementarlas, compararlas de manera sistemática, y finalmente interpretar los resultados (González-Carvajal & Garrido-Merchán, 2021). En cuanto a los inicios del estudio de la temática de análisis de sentimientos identificamos principalmente un precursor, que es Kim, quien en su estudio realizado en 2014 hizo una adaptación de las redes neuronales convolucionales (CNN) para el procesamiento de texto. Como resultado, logró demostrar la efectividad para identificar características locales en los textos. Años después aparecieron las redes LSTM (Long Short-Term Memory), que fueron originalmente desarrolladas por Hochreiter y Schmidhuber (1997), su particularidad fue que se comprobó que son más poderosas al capturar relaciones contextuales de largo alcance en secuencias textuales, es decir, pueden interpretar palabra incluso si se encuentran separadas dentro de la oración. Sin embargo, Devlin et al. (2019) revolucionó el campo del procesamiento del lenguaje natural con la introducción de BERT (Bidirectional Encoder Representations from Transformers). Este es un modelo basado en la arquitectura de transformadores que tiene la capacidad de generar representaciones contextuales bidireccionales.

Dentro de este contexto, decidimos implementar nuestro proyecto de investigación utilizando el entorno de trabajo de Pytorch, ya que permite evaluar y comparar las arquitecturas de redes neuronales con flexibilidad y optimización de rendimiento en el código y sus modificaciones

cuando sea necesario (Taylor & Kriegeskorte, 2023). Además, es un entorno ampliamente adoptado en la comunidad científica por su fácil reproducción en un experimento, según indica Alahmari et al., 2020.

La evaluación empírica de redes neuronales en tareas de clasificación de sentimientos no es simplemente útil en la actualidad, sino que se ha vuelto una necesidad. Esto se debe a que nos permite obtener datos concretos sobre el rendimiento real de diferentes modelos, lo cual es fundamental para tomar decisiones informadas al momento de seleccionar una arquitectura para aplicaciones del mundo real, según lo expuesto por Minaee et al., (2021). Particular que toma relevancia cuando estamos trabajando con datos en el idioma español, ya que se requiere de evaluaciones sistemáticas y rigurosas que permitan comparar el rendimiento y eficiencia de distintos enfoques, según Moreno-Ortiz y García-Gámez (2023).

La literatura ha sido revisada e identificamos que existe un vacío en cuanto a comparaciones sistemáticas que consideren tanto la precisión de los modelos como su eficiencia computacional, la factibilidad práctica y su aplicabilidad real pese a que las arquitecturas LSTM, BERT y CNN han demostrado individualmente su efectividad (Prottasha et al., 2024). Asimismo, se ha investigado que, múltiples estudios relacionados con el análisis de sentimientos se han llevado a cabo en el idioma inglés, lo que percibe una brecha en nuestro entendimiento sobre cómo se desempeñan estas arquitecturas neuronales cuando trabajan con textos en español, específicamente en el contexto de reseñas cinematográficas, expuesto por Angel et al., 2021.

El análisis de sentimientos ya ha sido estudiado previamente utilizando arquitecturas neuronales, los avances han sido remarcables en la última década. Por ejemplo, Kim (2014) estableció un punto de inflexión mediante su investigación "Convolutional Neural Networks for Sentence Classification", en donde demostró que las redes convolucionales no solamente se adaptan para la clasificación de imágenes, sino que, pueden alcanzar resultados competitivos y eficaces en la clasificación de texto. Posteriormente, Huang et al. (2021), aplicaron las redes LSTM al análisis de sentimientos, demostrando la superioridad de las LSTM bidireccionales en la captura de relaciones contextuales emocionales. Una vez establecido un punto de partida, Ning et al. (2020) realizaron un estudio comparativo entre CNN y LSTM para la clasificación de sentimientos multilingüe, cuyo resultado fue superior en textos extensos para las redes LSTM, mientras que CNN presentó un mejor rendimiento computacional. Dentro del mismo contexto, Zhou et al. (2020) realizaron una comparación sistemática entre estas arquitecturas en diversas tareas de NLP, incluyendo la clasificación de sentimientos. Con respecto al idioma español, destacó el estudio de Gutiérrez-Fandiño et al. (2022), en donde se introdujo un modelo BERT preentrenado para español llamado MarIA. En el presente estudio se utilizará la variante BETO, que también se originó como respuesta a la falta de modelos preentrenados para el idioma español.

Con el paso del tiempo se han desarrollado múltiples teorías en las cuales se ha basado el origen de la clasificación de sentimientos utilizando las redes neuronales. Como primer enfoque se tiene la Teoría del Procesamiento Distribuido Paralelo (PDP) que fue propuesta por McClelland y Rumelhart en 1986, indica que el procesamiento de la información en los sistemas neuronales se realiza mediante la activación simultánea de múltiples unidades interconectadas, este es un concepto fundamental para las redes neuronales modernas. La siguiente teoría corresponde a la Memoria de Trabajo, en la que Baddeley (2023) investigó un modelo que explica el proceso para mantener la información de manera temporal, que sirve como base para comprender el funcionamiento de las redes LSTM. Adicionalmente, Vaswani et al. (2017) estudiaron los mecanismos de atención, que son la base teórica para las arquitecturas basadas en transformadores, cuya ventaja es permitir que los modelos se concentren en partes específicas del texto de entrada. Finalmente, en el 2021 Bengio et al., propusieron la Teoría de Representaciones Distribuidas en donde se introduce el concepto de *word embeddings*, en donde se representan las palabras en los modelos neuronales para capturar relaciones semánticas de forma más compleja y detallada.

Una vez estudiada esta base teórica y con el fin de llevar a cabo nuestro análisis comparativo, se seleccionó en conjunto de datos de reseñas cinematográficas en el idioma español disponible en un conjunto de datos IMDB en la plataforma Kaggle, que contiene críticas en español sobre películas de temas variados, y que están clasificadas según su polaridad como positivas o negativas. Este escenario nos ha parecido ideal para evaluar la efectividad de las diversas arquitecturas neuronales seleccionadas para tareas de clasificación de sentimientos (Shaukat et al., 2020). El español, en cambio, plantea ciertos retos específicos para el procesamiento automático, principalmente debido a la complejidad de su morfología, la flexibilidad en el orden de las palabras y el uso habitual del modo subjuntivo, aspectos que ya fueron señalados por Henríquez et al. (2016).

Con el fin de entender cómo se comportan en el idioma español las tres arquitecturas neuronales seleccionadas nos hemos planteado la pregunta: ¿Las arquitecturas neuronales CNN, LSTM, y BERT tienen un desempeño diferente en la clasificación de sentimientos cuando se implementan utilizando PyTorch?, para responder, nos planteamos los siguientes pasos: revisión de la literatura existente, aplicación de una metodología acorde al tema de investigación, seguida de la obtención de resultados, la discusión y conclusiones.

## 2. METODOLOGÍA O MATERIALES Y MÉTODOS

La metodología que nos planteamos en este estudio tiene como objetivo realizar una evaluación y análisis comparativo de tres arquitecturas neuronales para la clasificación de sentimientos. En este sentido, adoptamos un enfoque metodológico dual, es decir, combinamos la metodología de Investigación en Ciencia del Diseño (DSR) y la metodología PRISMA (Preferred

Reporting Items for Systematic Reviews and Meta-Analyses). La primera se encarga del desarrollo de la solución, que, según lo estudiado por Apiola et al., se enfoca en diseñar soluciones prácticas para problemas reales, Mientras que, la segunda garantiza la calidad de la revisión sistemática de literatura, la transparencia de las revisiones y metaanálisis, según Schjerven, 2024.

Dentro de este contexto, es esencial indicar que la metodología DSR consta de tres fases que corresponden a la relevancia, el diseño y el rigor. La etapa de relevancia identifica el problema y el contexto de la solución, el diseño, se centra en crear prototipos innovadores para resolver el problema, en este caso corresponde a nuestros modelos de redes neuronales. Finalmente, la etapa de diseño, que se encarga de validar la solución planteada. (Reyes et al., 2023). En la **Fig. 1** hemos detallado el entorno, diseño y base de conocimiento que sustenta nuestro estudio.

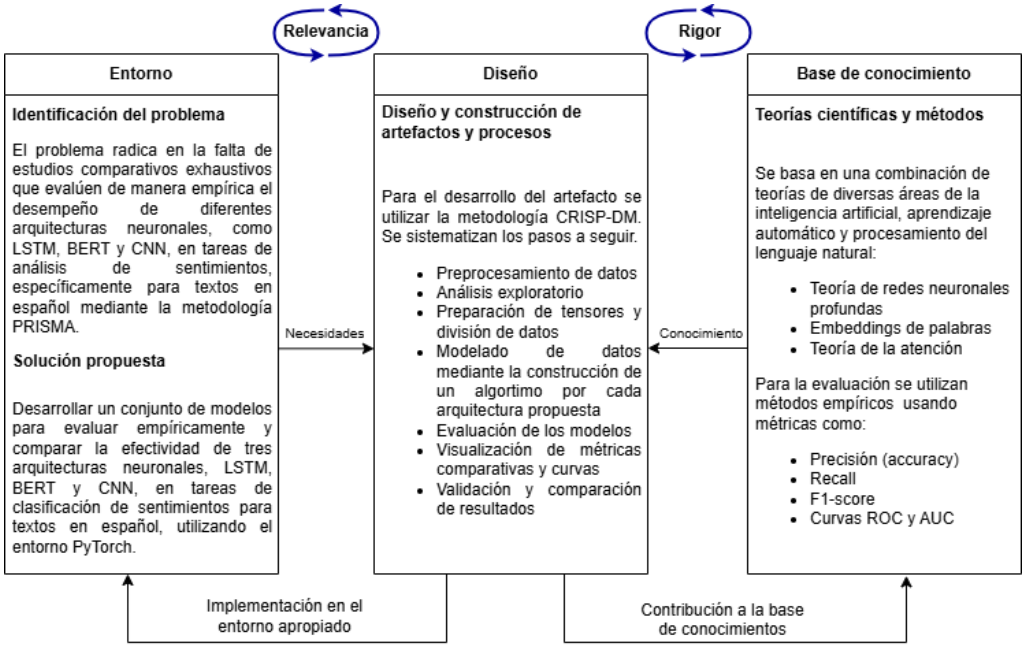


**Fig. 1. Metodología de investigación en ciencia del diseño aplicada al estudio.**

## 1.1. Ciclo de la relevancia

Dentro del ciclo de la relevancia identificamos y describimos el problema, para lo cual, realizamos una revisión exhaustiva de la literatura científica especializada en el campo del aprendizaje profundo, específicamente enfocada en la clasificación de sentimientos. Las consultas de las fuentes las realizamos en bases de datos indexadas de alto impacto como Web of Science, IEEE Xplore, Springer y Scopus. Por lo cual, garantizamos una base sólida para el desarrollo de un modelo por cada arquitectura de red neuronal, LSTM, BERT y CNN para la clasificación de sentimientos (Bellar et al., 2024). A fin de detallar el ciclo de la relevancia creamos la **Fig. 2**, donde se describe las causas, problema y consecuencias de nuestra investigación.

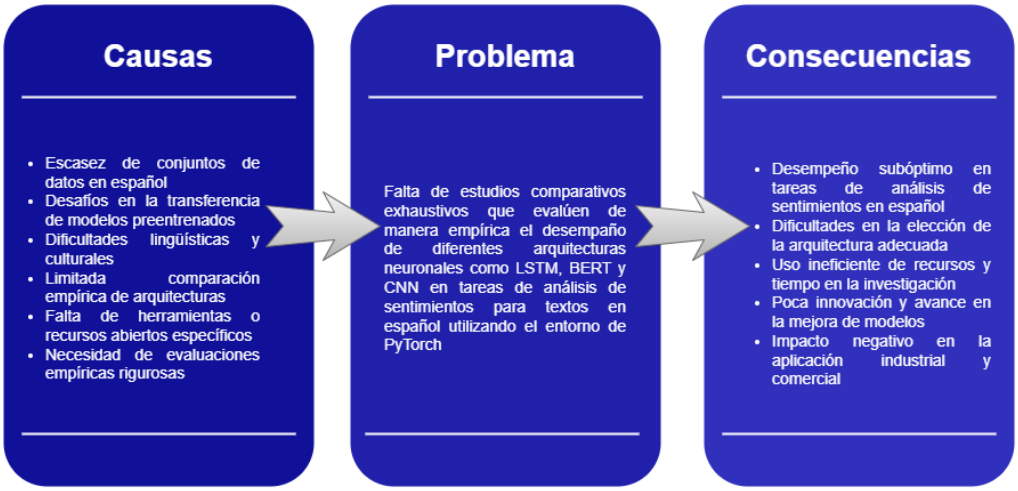


**Fig. 2. Descripción de causas, problema y consecuencias de la investigación.**

La selección de artículos realizamos basándonos en criterios como la relevancia pertinente con el tema planteado, estudios previos en donde se utilice el entorno de PyTorch, el uso de metodologías empíricas para el procesamiento de lenguaje natural, análisis de sentimientos en el idioma español y publicaciones recientes de fuentes confiables. Como resultado, obtuvimos 35 artículos que se encuentran descritos en la **Fig. 3**.

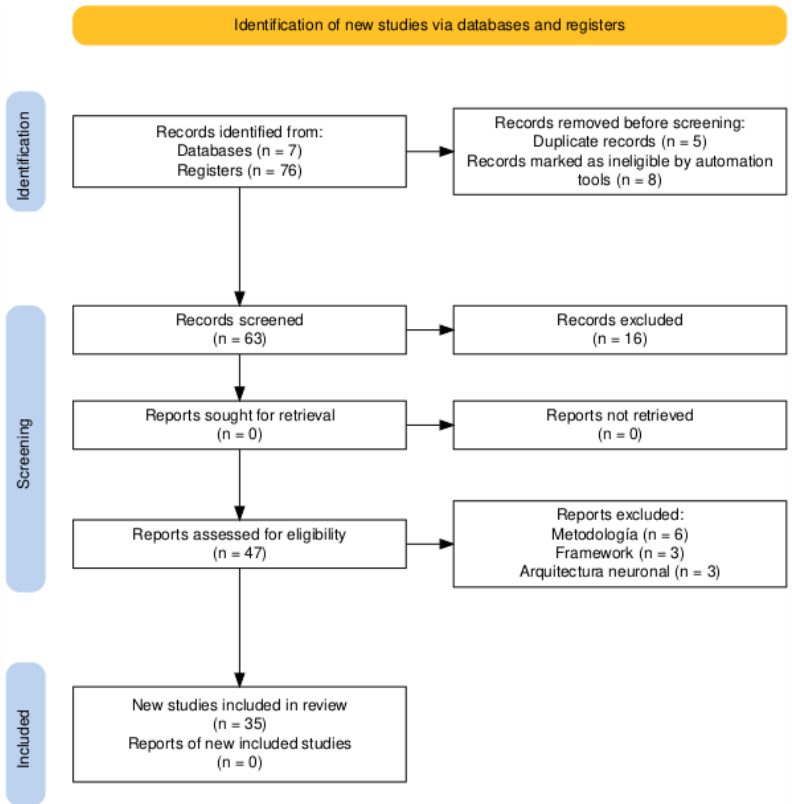


**Fig. 3. Diagrama de flujo de la metodología PRISMA para la revisión literaria.**

## 1.2. Ciclo del diseño

Para el desarrollo del trabajo investigativo nos basamos en un conjunto de datos correspondiente a reseñas cinematográficas en el idioma español, que se puede encontrar bajo el nombre de IMDB en la plataforma Kaggle (Fernández, 2021). Mediante esta investigación

planteamos realizar una evaluación empírica de arquitecturas neuronales para el análisis de sentimientos y así analizar su rendimiento para la tarea de la clasificación de sentimientos en el entorno PyTorch. En la **Fig. 4** se propone la metodología para nuestra investigación que comprende las fases de obtención de datos, preprocesamiento, modelamiento y evaluación de redes neuronales.

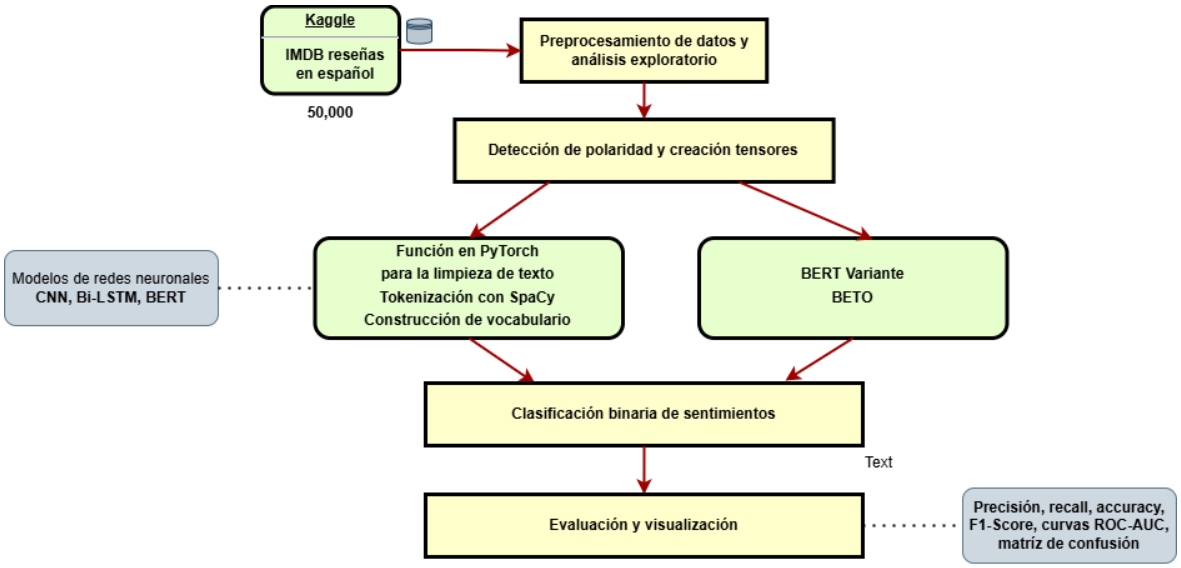


**Fig. 4. Metodología propuesta para el estudio.**

Una vez analizado el conjunto de datos obtenido de Kaggle, verificamos que contiene 50,000 reseñas de películas en inglés y español. Seleccionamos solamente aquellas que están en español, realizamos un análisis exploratorio y validamos que existe un ligero desbalance de clases (24,133 negativas y 24,050 positivas). Por lo cual, se graficó un histograma a fin de estudiar principalmente la variabilidad en la longitud de las reseñas. En la **Fig. 5** se detalla la distribución de las reseñas cinematográficas luego de haberlas tokenizado.

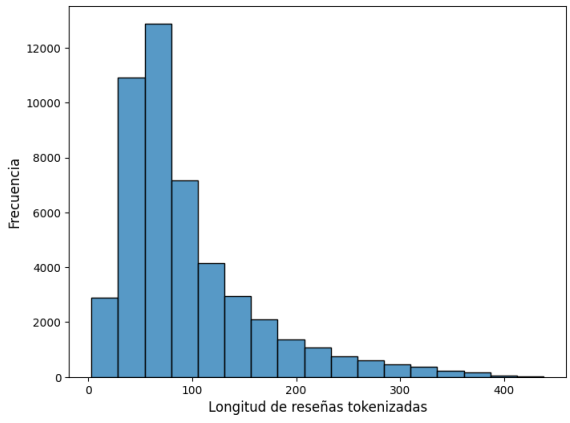


**Fig. 5. Distribución del número de caracteres por reseña cinematográfica.**

Como parte del ciclo del diseño definimos el entorno que se utiliza. Empezamos creando clases especializadas y personalizadas para el texto en español de forma que se pueda preprocesar cada reseña cinematográfica para las arquitecturas CNN y LSTM. Aplicamos varias técnicas dentro

del preprocesamiento, que incluye lematización, retornamos las palabras a su base semántica, eliminamos palabras vacías, acentos o tildes, caracteres especiales y numéricos ya que no aportan contexto en el análisis de sentimientos. Posteriormente, realizamos la tokenización del texto mediant el uso de la librería spaCy, utilizamos el conjunto es_core_news_. Para el manejo de términos desconocidos construimos un vocabulario que incluye palabras especiales que más tarde se convirtieron en índices y tensores para que puedan ser utilizados en los modelos de redes neuronales. Luego, estandarizamos la longitud de cada reseña aplicando la técnica de *padding*, es decir, rellenando los espacios vacíos para obtener una longtud uniforme. Como último paso, creamos *dataloaders*, para establecer una interfaz entre los datos preprocesados y la arquitectura neuronal a fin de optimizar la eficiencia computacional en la fase de entrenamiento (Duong & Nguyen-Thi, 2021).

En la **Tabla *1*** describimos un ejemplo que ilustra el preprocesamiento aplicado a una reseña aleatoria. Mostramos el texto original del conjunto de datos, aplicamos el preprocesamiento personalizado en las clases de Pytorch y finalmente mostramos la versión tokenizada que ingresa a los modelos de redes neurnales. Adicionalmente, se incluye el conteo de tokens por reseña.

**Tabla 1.** Comparación de datos sin procesar versus procesados y tokenizados.

| Datos sin procesar | Datos procesados y tokenizados | Tokens |
|---|---|---|
| Me conmovió particularmente por el coraje y la integridad subestimados de L'Anglaise, en esta película hermosamente actuada, intelectual y visualmente convincente.Muchas gracias, Monsieur Le Directeur Rohmer. | ['conmovio', 'particularmente', 'coraje', 'integridad', 'subestimados', 'l', 'anglaise', 'pelicula', 'hermosamente', 'actuada', 'intelectual', 'visualmente', 'convincente', 'gracias', 'monsieur', 'directeur', 'rohmer'] | 17 |

Para el manejo de valores atípicos realizamos un diagrama de caja y bigote en la **Fig. 6**Fig. 6. Diagrama de caja y bigote para la identificación de valores atípicos., que nos permite examinar la distribución del número de caracteres tokenizados. Mediante el estudio de este gráfico, decidimos eliminar reseñas con más de 350 caracteres, dejando un conjunto final de 47,847 registros (Chatterjee et al., 2021).

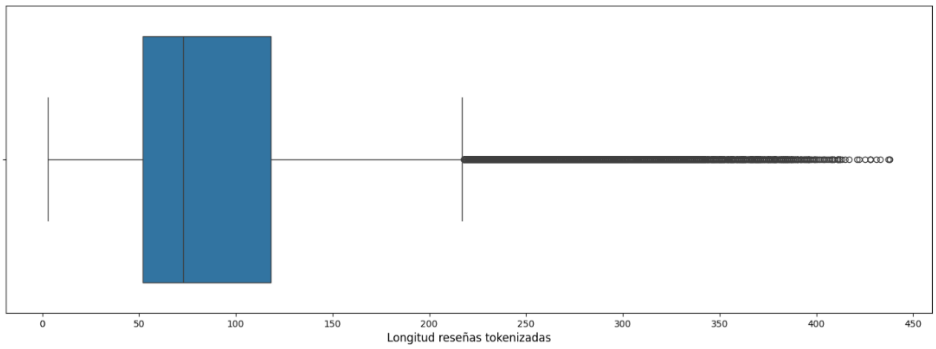

**Fig. 6. Diagrama de caja y bigote para la identificación de valores atípicos.**

Para este estudio seleccionamos tres arquitecturas neuronales: (1) red neuronal convolucional simple (CNN), (2) red LSTM bidireccional y (3) transformadores BERT en su variante para español, BETO. Una vez preprocesado el conjunto de datos, dividimos un 80% para entrenamiento y el 20% restante para pruebas**.** Como primer enfoque, desarrollamos un modelo basado en la arquitectura CNN, que se caracteriza por la aplicación de filtros para realizar la operación de convolución sobre el texto.

$$(f * g)(y) = \int_0^t f(t-u)g(u)du \quad \textbf{1)}$$

Las redes neuronales convolucionales pueden extraer patrones representativos que describen el texto en forma de n-gramas, según lo estudiado por Colón-Ruiz & Segura-Bedmar, 2020. La arquitectura de una CNN cuenta con una capa de convolución, donde diferentes filtros se deslizan a lo largo de la matriz de *word embeddings* de cada reseña cinematográfica, produciendo como salida un mapeo de las características de las reseñas. Tomamos como ejemplo la fracción de una reseña cinematográfica a fin de explicar de forma gráfica el proceso de convolución en la **Fig. 7**.

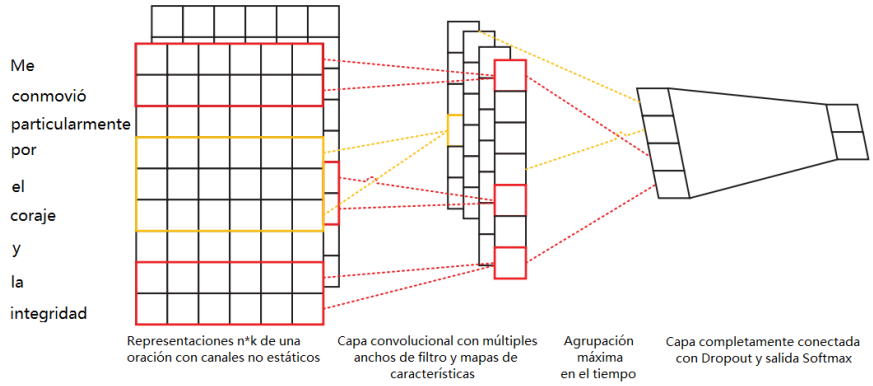


**Fig. 7. Arquitectura de una red neuronal convolucional para la clasificación de texto.**

Como segundo enfoque, manejamos una red neuronal LSTM (Long-Short Term Memory) bidireccional, que, a diferencia de las redes LSTM convencionales, que manejan los datos de manera secuencial unidireccional, las LSTM bidireccionales procesan los datos de entrada tanto en la dirección hacia adelante como hacia atrás de manera concurrente, lo que le permite hacer predicciones más informadas (Pandit, et al., 2024). En la **Fig. 8¡Error! No se encuentra el origen de la referencia.**, describimos la arquitectura Bi – LSTM utilizada en esta investigación. Cabe indicar que, ésta se basa en fórmulas que describen matemáticamente el comportamiento de las compuertas de olvido, entrada y salida, características esenciales que definen este tipo de arquitectura.

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad \textbf{2)}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad \textbf{3)}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad \textbf{4)}$$

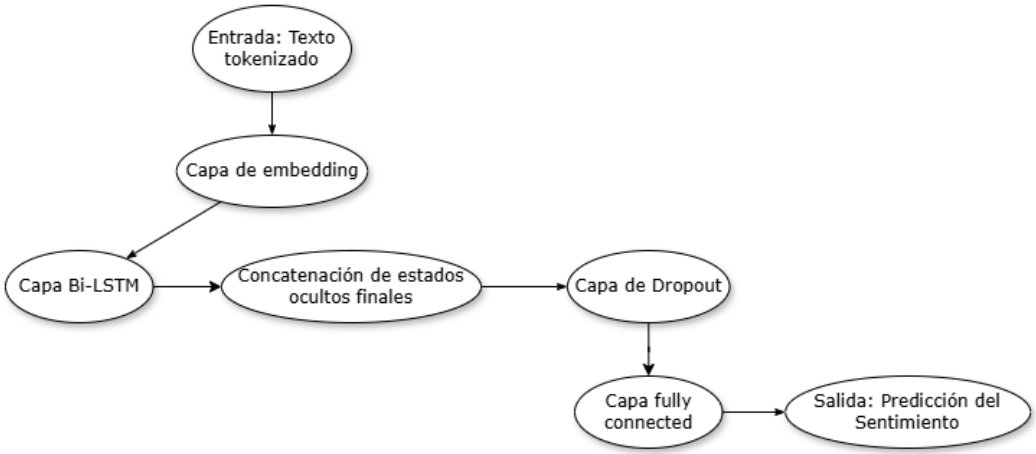


**Fig. 8. Arquitectura Bi – LSTM para el análisis de sentimientos.**

Como tercera opción seleccionamos la arquitectura BERT, en su variante BETO preentrenada para español, que emplea técnicas de self-attention para capturar relaciones contextuales en una oración y procesa el texto de manera simultánea, modelando dependencias de largo alcance (Colón-Ruiz & Segura-Bedmar, 2020). Seleccionamos los hiperparámetros según las características de cada red neuronal y los detallamos en la **¡Error! No se encuentra el origen de la referencia.**. Para las redes CNN y Bi-LSTM configuramos una capa de embedding de 100 dimensiones, suficiente para el análisis de texto, mientras que para el modelo basado en transformadores utilizamos 768 dimensiones, correspondientes a su arquitectura preentrenada. Además, configuramos filtros convolucionales de tamaño 1, 4 y 5. Para Bi-LSTM utilizamos una capa oculta de 128 unidades y dos capas LSTM. Aplicamos técnicas de dropout para evitar sobreajuste, en las redes CNN y LSTM de un 50%, y, para BERT del 10% ya que es una arquitectura más robusta. La tasa de aprendizaje también es diferente entre las arquitecturas neuronales, ya que para CNN y LSTM es de 0.001, pero para BERT es de 0.00002. El número de épocas óptimas para BERT es de 4 según lo estudiado por Devlin et al. (2019), sin embargo, se configuraron 5 épocas para cada uno de los modelos de redes neuronales ya que garantizamos que la carga de trabajo sea la misma y la comparación sea equitativa.

**Tabla 2.** Parámetros configurados en los modelos según la arquitectura neuronal.

| Parámetros | CNN | Bi-LSTM | BERT |
|---|---|---|---|
| Dimensión capa de embedding | 100 | 100 | 768 |
| Número de filtros | 100 | - | - |

| | | | |
|---|---|---|---|
| Tamaño del filtro | 3, 4, 5 | - | - |
| Número de épocas | 5 | 5 | 5 |
| Dimensión capa de salida | 2 | 2 | 2 |
| Dropout | 0.5 | 0.5 | 0.1 |
| Tasa de aprendizaje | 0.001 | 0.001 | 0.00002 |
| Tamaño del lote | 32 | 32 | 16 |
| Optimizador | ADAM | ADAM | AdamW |
| Función de pérdida | Entropía cruzada | Entropía cruzada | Entropía cruzada |
| Tamaño del estado oculto | - | 128 | - |
| Número de capas LSTM | - | 2 | - |
| Tokenizador | Personalizado | Personalizado | BERT |
| Longitud máxima de secuencia | 150 | 150 | 150 |

### 1.3. Ciclo del rigor

Utilizamos varias métricas para evaluar el ciclo de rigor en los algoritmos creados. Por ejemplo, precisión, recall, accuracy, F1-score, curvas ROC y AUC y el tiempo de entrenamiento por modelo. Adicionalmente, generamos matrices de confusión que nos permitieron analizar la distribución de predicciones verdaderas y falsas en cada modelo dado, ya que se trata de un problema de clasificación binaria.

## 3. RESULTADOS Y DISCUSIÓN

A continuación, presentamos los resultados de la evaluación empírica de tres modelos de redes neuronales aplicados a la tarea de clasificación de sentimientos en el conjunto de datos IMDB en español. Las pruebas realizamos en el entorno de Google Colaboratory, utilizando Python versión 3.11.11 y las librerías PyTorch 2.6.0, Transformers 4.50.0, Matplotlib 3.10.0, Scikit-learn 1.6.1, Numpy 2.0.2 y Pandas 2.2.2. Para ello, empleamos un computador con las siguientes características: procesador Intel Core i7-8700T a 2.40GHz, de 16 GB de RAM con tarjeta gráfica NVIDIA-SMI.

El rendimiento de cada arquitectura se detalla en la **Tabla 3** donde se enlista los tres modelos de redes neuronales seleccionados junto con los valores obtenidos en métricas como accuracy, puntaje F1, precisión, recall y el valor AUC que obtuvimos al trazar las curvas ROC. Además, registramos el tiempo de entrenamiento de cada modelo a fin de evaluar su viabilidad en relación con los recursos computacionales disponibles.

**Tabla 3.** Resultados de métricas de evaluación y tiempo de entrenamiento por modelo de red neuronal.

| Modelo | Accuracy | F1-score | Precision | Recall | AUC | Tiempo entrenamiento |
|---|---|---|---|---|---|---|
| LSTM | 86.29% | 86.29% | 86.30% | 86.29% | 0.936 | 1m 22s |

| | | | | | | |
|---|---|---|---|---|---|---|
| CNN | 85.64% | 85.64% | 85.65% | 85.64% | 0.932 | 8m 24s |
| BERT | 87.92% | 87.89% | 88.24% | 87.92% | 0.947 | 2h 14m |

Analizamos los resultados y observamos que la arquitectura basa en transformadores obtuvo una precisión de 88.24%, que corresponde a la más alta. Esto se explica debido a su habilidad para capturar contexto bidireccional y su preentrenamiento en español. En contraste, examinamos que, el modelo LSTM logró una precisión del 86.30%, sobresaliendo particularmente en la identificación de secuencias extensas. La particularidad que encontramos es que este modelo demuestra cierta sensibilidad frente a desafíos como la pérdida del gradiente. Finalmente, obtuvimos que para la arquitectura CNN se logró un 85.65% de precisión, lo que representa la cifra más baja entre las comparaciones. Como explicación, sugerimos que este resultado se debe a que esta arquitectura está más enfocada en identificar patrones locales, por lo que, no reconoce relaciones semánticas más profundas.

Como siguiente análisis realizamos una gráfica que permite identificar la pérdida en el conjunto de prueba, 20% de los datos, según la época de cada modelo de red neuronal. En la **Fig. 9** identificamos que el modelo Bi-LSTM presenta un mejor desempeño ya que tiene un valor menor en cuanto a la pérdida, correspondiente a 0.343, por lo tanto, presenta una convergencia más estable. La línea de color azul representa a la arquitectura CNN, se caracteriza porque también muestra una pérdida baja, de 0.346, y un descenso progresivo durante las cinco épocas de entrenamiento. Por otro lado, el modelo basado en transformadores, graficado de color rojo, inicia con una pérdida inicial alta de 0.872, pero se reduce significativamente para la segunda época y, luego aumenta nuevamente hasta 0.548, lo que podría indicar sobreajuste hacia la última época.

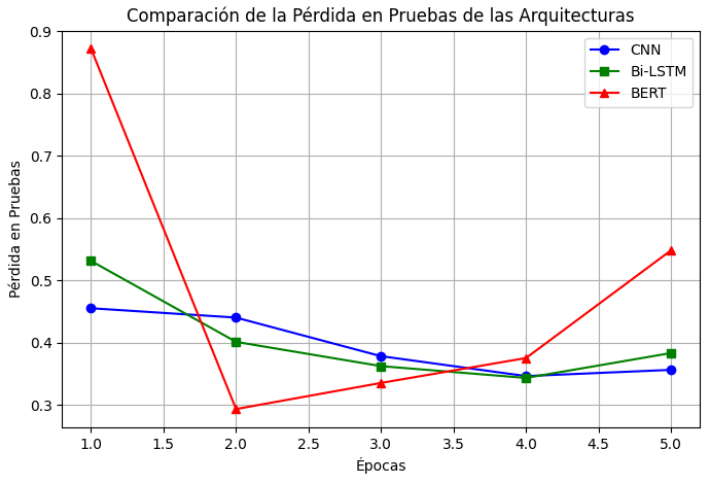


**Fig. 9. Comparativa entre valores de pérdida en el conjunto de prueba.**

De manera análoga, en la **Fig. 10** graficamos la comparativa de la puntuación F1 obtenida en cada época por los modelos de redes neuronales. Analizamos el gráfico y aa arquitectura BERT obtuvo una puntuación F1 alta desde las primeras épocas. En la segunda época llega a un pico de

puntaje F1, que corresponde a 0.894, aunque posteriormente mostró algunas fluctuaciones mínimas. Por otro lado, las líneas de los otros dos modelos describen comportamientos con valores menores al de la arquitectura BERT durante las cinco épocas. Por ejemplo, Bi-LSTM demostró un avance continuo, alcanzando su rendimiento más alto en la cuarta. La arquitectura CNN también mostró un incremento en el puntaje F1, ya que llegó a su máximo valor durante la última etapa del entrenamiento. En síntesis, observamos que la arquitectura BERT se muestra más eficiente desde el inicio de nuestra experimentación, mientras que las redes basadas en las arquitecturas Bi-LSTM y CNN requieren un poco más de tiempo y épocas para alcanzar un mejor nivel de desempeño.

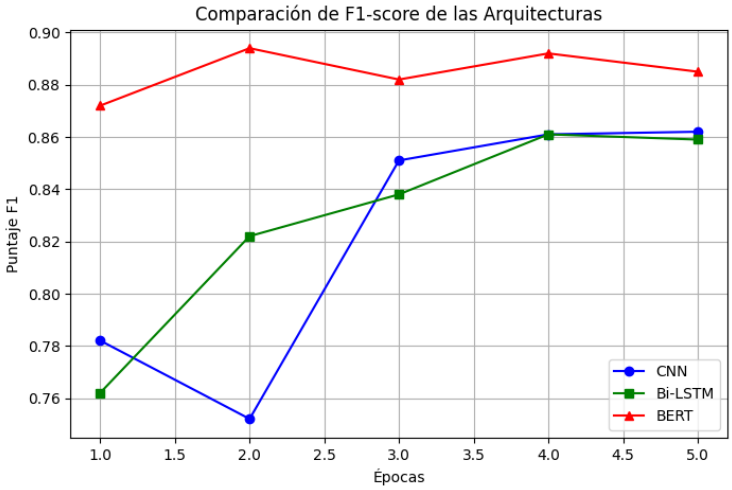


**Fig. 10. Comparativa de la puntuación F1 de las arquitecturas neuronales.**

Finalmente, condensamos los resultados de las matrices en confusión en la ¡Error! No se encuentra el origen de la referencia., donde BERT destaca con el mejor desempeño en la clasificación de reseñas positivas, alcanzando un 92.48%. No obstante, observamos que presenta una debilidad al momento de predecir reseñas negativas. La CNN consigue un balance más equitativo entre ambas clases (86.27% para reseñas positivas y 86.21% para negativas), a pesar de que su rendimiento global es inferior. Finalmente, Bi-LSTM logra un 87.56% en la clase negativa y un 84.79% en la positiva, lo que representa un mayor desafío al categorizar reseñas positivas. Por lo general, BERT muestra preponderancia en la categorización de sentimientos, mientras que CNN y Bi-LSTM son opciones factibles, aunque con un mayor riesgo de error.

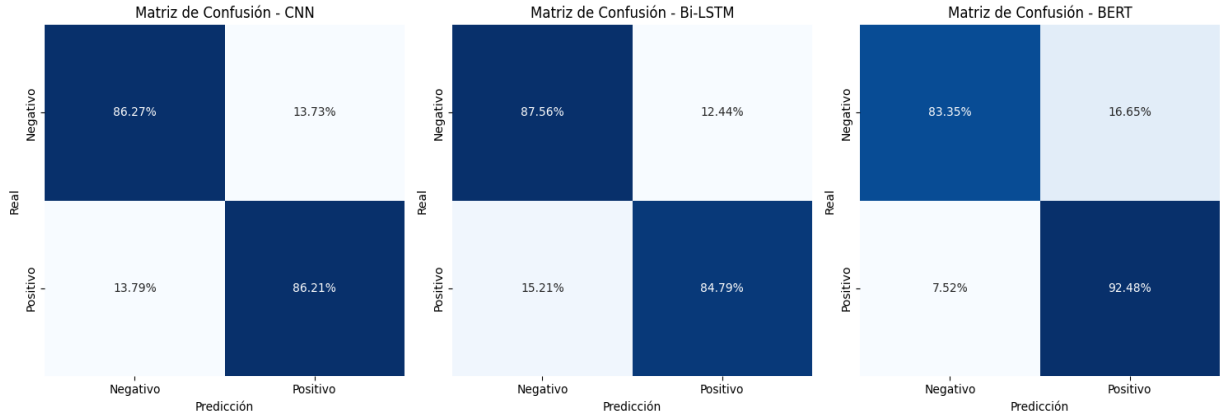


**Fig. 11. Matriz de confusión de los modelos de redes neuronales para el análisis de sentimientos.**

Finalmente, condensamos los resultados de las matrices en confusión en la ¡Error! No se encuentra el origen de la referencia., donde BERT destaca con el mejor desempeño en la clasificación de reseñas positivas, alcanzando un 92.48%. No obstante, observamos que presenta una debilidad al momento de predecir reseñas negativas. La CNN consigue un balance más equitativo entre ambas clases (86.27% para reseñas positivas y 86.21% para negativas), a pesar de que su rendimiento global es inferior. Finalmente, Bi-LSTM logra un 87.56% en la clase negativa y un 84.79% en la positiva, lo que representa un mayor desafío al categorizar reseñas positivas. Por lo general, BERT muestra preponderancia en la categorización de sentimientos, mientras que CNN y Bi-LSTM son opciones factibles, aunque con un mayor riesgo de error.

Luego de la evaluación empírica a las arquitecturas neuronales para la clasificación de sentimientos pudimos evidenciar que BERT superó a LSTM y CNN en precisión, logrando un 88.24%, un F1-score de 87.89% y un AUC-ROC de 0.947. Estos valores nos permitieron confirmar nuestra pregunta de investigación que establece que, el desempeño de las arquitecturas neuronales es diferente para la tarea de clasificación de sentimientos. Pudimos constatar que, el sobresaliente desempeño de BERT, en su variante BETO preentrenada para el español se origina principalmente por su habilidad para representar relaciones bidireccionales y contextuales en los niveles de palabras y oraciones, resultados consistentes con los hallazgos de Rogers et al. (2020), quienes demostraron que los modelos basados en transformadores superan consistentemente a las arquitecturas recurrentes tradicionales en tareas de comprensión del lenguaje natural. De forma similar, Devlin et al. (2019), los creadores originales de BERT, establecieron que el preentrenamiento bidireccional permite al modelo comprender mejor el contexto completo de una oración, lo cual se refleja en nuestros resultados donde se mostró una capacidad superior para discernir matices emocionales en las reseñas cinematográficas.

En la segunda posición situamos a la arquitectura Bi-LSTM en cuanto al rendimiento general, ya que logró una precisión del 86.30% y un AUC-ROC de 0.936. Estos resultados superan

a los obtenidos por CNN en todas las métricas evaluadas, lo que refuerza la eficacia de las redes LSTM para capturar dependencias secuenciales a largo plazo. Hochreiter & Schmidhuber (1997) establecieron teóricamente esta ventaja de las LSTM, y nuestros resultados empíricos la confirman en el contexto específico del análisis de sentimientos. Verificamos que estos resultados son mejores que los obtenidos por CNN, lo que denota la eficacia de las redes LSTM bidireccionales para identificar dependencias secuenciales a largo plazo. Esta es una ventaja esencial considerando la longitud y complejidad habituales de los textos que corresponden a reseñas de películas, ya que generalmente poseen una amplia longitud. No obstante, es esencial que destaquemos la diferencia existente entre los resultados de LSTM y CNN, 85.65%, es menos pronunciada que la que hay entre BERT y LSTM, lo que nos indica que las arquitecturas recurrentes convencionales son efectivas, pero no utilizan completamente la riqueza contextual de las reseñas evaluadas. Kim (2014) fue pionero en demostrar la eficacia de las CNN para clasificación de sentimientos, argumentando que las operaciones de convolución son especialmente útiles para detectar patrones locales en el texto y, Zhang et al. (2015) complementaron estos hallazgos mostrando que las CNN pueden capturar efectivamente n-gramas de diferentes tamaños, lo cual es importante para identificar expresiones idiomáticas y frases con carga emocional.

Los resultados que obtuvimos muestran diferencias con respecto a lo estudiado por Pandit et al.(2024), en donde la arquitectura que más sobresalió es LSTM, con una precisión del 99.91% para datos de entrenamiento. Esta discrepancia puede explicarse por varios factores metodológicos y de datos. Ruder et al. (2016) advirtieron sobre la importancia de considerar las características específicas del conjunto de datos y el dominio de aplicación al comparar arquitecturas neuronales, ya que diferentes tipos de texto pueden favorecer distintas arquitecturas. Particularmente, en nuestra investigación, la mayor precisión global alcanzada corresponde a la arquitectura BERT, que presenta un patrón interesante y tiene excelente capacidad para identificar reseñas positivas, pero cierta tendencia a clasificar erróneamente reseñas negativas como positivas. En contraste con otras arquitecturas, la CNN presenta un desempeño más balanceado entre las clases, aunque su rendimiento global es inferior, lo que sugiere que su uso podría ser beneficioso en situaciones donde sea más importante mantener una precisión balanceada entre las categorías, en vez de perfeccionar la precisión global del modelo. Liu et al. (2019) observaron patrones similares en sus experimentos, sugiriendo que los modelos preentrenados pueden desarrollar sesgos hacia clases más frecuentes durante el preentrenamiento. En síntesis, la arquitectura CNN muestra un rendimiento más equilibrado entre las clases, aunque su desempeño global es menor, lo que indica que su aplicación podría ser útil en contextos donde sea más relevante mantener una precisión equilibrada entre categorías, en lugar de optimizar la precisión global del modelo.

En cambio, al momento de evaluar aplicaciones prácticas es crucial considerar la eficiencia computacional. En este sentido pudimos observar una variación significativa entre las diferentes arquitecturas, particularmente en relación con los períodos de entrenamiento. Pese a que BERT proporciona un desempeño sobresaliente, necesita más recursos y tiempo tanto para su formación como para su validación. En cambio, observamos que modelos como CNN y LSTM resultan más adecuados para sistemas que requieren operar con limitaciones de hardware o recursos restringidos. Es fundamental destacar que este estudio se centró exclusivamente en un conjunto de datos en español, lo que representa una limitación si se busca aplicar el modelo en contextos multilingües.

En resumen, nuestros resultados tienen importantes repercusiones prácticas para la creación de sistemas de análisis de sentimientos en español utilizando PyTorch, confirmando tendencias observadas en la literatura internacional y revelando patrones específicos para el contexto hispanohablante. La convergencia de nuestros hallazgos con estudios previos de Rogers et al. (2020), Sun et al. (2019), y Barbieri et al. (2022) fortalece la validez de los resultados y sugiere que los patrones observados son robustos y generalizables dentro del dominio evaluado. Para las aplicaciones que valoran la máxima exactitud y cuentan con grandes recursos de computación, BERT sobresale como la opción más recomendable. En cambio, pudimos notar que, para situaciones donde existen restricciones computacionales o se necesita inferencia en tiempo real, Bi-LSTM proporciona un balance apropiado entre precisión y eficacia. Igualmente, la alternativa de CNN, a pesar de tener un rendimiento general un poco inferior, demostró ventajas en términos de estabilidad entre clases.

A partir de nuestro estudio, la sugerencia es explorar y crear modelos de arquitecturas híbridas y multilingües que puedan extraer características del texto, además de fusionar mecanismos de atención para el análisis de sentimientos. Esto es un área poco explorada pero prometedora para futuros progresos en el procesamiento del lenguaje natural.

## 4. CONCLUSIONES

En esta investigación, evaluamos de forma empírica las arquitecturas de redes neuronales CNN, LSTM, BERT para clasificación de sentimientos utilizando reseñas de películas en español, a través del entorno de PyTorch. Los resultados de los experimentos nos muestran que BERT presentó el mejor desempeño con una precisión 87.92%, en comparación con Bi-LSTM 86.29% y CNN 85.64%. Por lo tanto, validamos la efectivad de los modelos fundamentados en atención para entender el contexto lingüístico del español, aunque evidentemente demandan mayores recursos computacionales.

En líneas generales, el análisis que realizamos reveló no solo el desempeño, sino también ciertas diferencias importantes entre las arquitecturas. BERT demostró una gran habilidad para

lograr altos niveles de precisión desde el comienzo de las primeras épocas de entrenamiento. En cambio, CNN y Bi-LSTM requirieron significativamente más tiempo para lograr un rendimiento más estable. Finalmente, se observó que los algoritmos utilizan patrones diferentes para la clasificación de opiniones desfavorables, que se pueden aprovechar según el caso de uso.

Luego del análisis de resultados y gráficas concluimos que, la elección del modelo más adecuado dependerá de las necesidades específicas de cada aplicación, como el análisis de redes sociales o el seguimiento de reseñas de productos en línea. La opción correcta, si se cuentan con recursos computacionales y se necesita alta precisión, es la arquitectura de red neuronal BERT. En aplicaciones de tiempo real o con recursos limitados, la arquitectura Bi-LSTM se posiciona como la alternativa más adecuada. Cabe señalar que nuestros hallazgos y su enfoque ayudan al progreso y evolución de sistemas más eficaces para la categorización de sentimientos en español. Sin embargo, la investigación se restringe al idioma español, debido a los datos que se emplearon, y no se analizó diversas variantes de BERT como RoBERTa o DistilBERT ni enfoques híbridos, que son producto de la fusión de distintas redes neuronales.

Como sugerencia, podríamos profundizar más adelante en la fusión de diversas estrategias técnicas, así como en formas de adaptar estos sistemas a contextos particulares, sin dejar de lado la creación de mecanismos que nos ayuden a comprender el motivo de sus decisiones. Todo esto con el objetivo de obtener herramientas más exactas, que se ajusten mejor a diversos contextos y cuyo funcionamiento sea claro al aplicarse a situaciones reales.

## CONFLICTO DE INTERESES (Obligatorio)

Los Autores declaran que no existe conflicto de intereses, o lo que corresponda.

## CONTRIBUCIÓN DE AUTORÍA (Obligatorio)

**(Seleccione con una X según corresponda, los dos últimos criterios es obligatorio participar para ser considerado autor de la obra)**

En concordancia con la taxonomía establecida internacionalmente para la asignación de créditos a autores de artículos científicos (https://credit.niso.org/). Los autores declaran sus contribuciones en la siguiente matriz:

| **Participar activamente en:** | Gabriela Uquillas | Rebeca Moposita |
|---|---|---|
| Conceptualización | X | |
| Análisis formal | X | |
| Adquisición de fondos | X | |

| | | |
|---|---|---|
| Investigación | X | |
| Metodología | X | |
| Administración del proyecto | X | |
| Recursos | X | |
| Redacción –borrador original | X | |
| Redacción –revisión y edición | X | X |
| **La discusión de los resultados** | X | X |
| **Revisión y aprobación de la versión final del trabajo.** | X | X |

## REFERENCIAS


Liu, B. (2020). Sentiment analysis: Mining opinions, sentiments, and emotions (2nd ed.). Cambridge University Press.

Fu, E., Xiang, J., & Xiong, C. (2022). Deep Learning Techniques for Sentiment Analysis. Highlights in Science, Engineering and Technology. 16. 1-7 [Archivo PDF] https://doi.org/10.54097/hset.v16i.2065

Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2023). Emotion recognition in conversation: Research challenges, datasets, and recent advances. IEEE Access, 11, 78347-78372 [Archivo PDF] https://ieeexplore.ieee.org/ielaam/6287639/8600701/8764449-aam.pdf

Zhang, L., Wang, S., & Liu, B. (2022). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(3), e1413 [Archivo PDF] https://doi.org/10.1002/widm.1253

González-Carvajal, S., & Garrido-Merchán, E. C. (2021). Comparing BERT against traditional machine learning text classification. Journal of Computational and Cognitive Engineering, 183, 115345 [Archivo PDF]. https://doi.org/10.47852/bonviewJCCE3202838

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780 [Archivo PDF]. https://doi.org/10.1162/neco.1997.9.8.1735

Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 1746-1751 [Archivo PDF]. https://doi.org/10.3115/v1/D14-1181.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171-4186 [Archivo PDF]. https://doi.org/10.18653/v1/N19-1423

Taylor, JohnMark & Kriegeskorte, Nikolaus. (2023). Extracting and visualizing hidden activations and computational graphs of PyTorch models with TorchLens. Scientific Reports. 13 [Archivo PDF]. https://doi.org/10.1038/s41598-023-40807-0

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning-based text classification: A comprehensive review. ACM Computing Surveys, 54(3), 1-40 [Archivo PDF]. https://doi.org/10.1145/3439726

Moreno-Ortiz, A., García-Gámez, M. (2023) Strategies for the Analysis of Large Social Media Corpora: Sampling and Keyword Extraction Methods. Corpus Pragmatics 7, 241–265 [Archivo PDF]. https://doi.org/10.1007/s41701-023-00143-0

Prottasha, N.J., Mahmud, A., Sobuj, M.S.I. et al. (2024) Parameter-efficient fine-tuning of large language models using semantic knowledge tuning. Sci Rep 14, 30667 (2024) [Archivo PDF]. https://doi.org/10.1038/s41598-024-75599-4

Angel, Sonia & Peña Pérez Negrón, Adriana & Espinoza-Valdez, Aurora. (2021). Systematic literature review of sentiment analysis in the Spanish language. Data Technologies and Applications. ahead-of-print [Archivo PDF]. https://doi.org/10.1108/DTA-09-2020-0200

Alahmari, S.S., Goldgof, D., Mouton, P.R., & Hall, L.O. (2020). Challenges for the Repeatability of Deep Learning Models. IEEE Access, 8, 211860-211868 [Archivo PDF]. https://doi.org/10.1109/ACCESS.2020.3039833

Huang, F., Li, X., Yuan, C., Zhang, S., Zhang, J., & Qiao, S. (2021). Attention-Emotion-Enhanced Convolutional LSTM for Sentiment Analysis. IEEE Transactions on Neural Networks and Learning Systems, 33, 4332-4345 [Archivo PDF]. https://doi.org/10.1109/TNNLS.2021.3056664

Jin, N., Wu, J., Ma, X., Yan, K., & Mo, Y. (2020). Multi-Task Learning Model Based on Multi-Scale CNN and LSTM for Sentiment Classification. IEEE Access, 8, 77060-77072 [Archivo PDF]. https://doi.org/10.1109/ACCESS.2020.2989428

Zhou, X., Li, Y.A., & Liang, W. (2020). CNN-RNN Based Intelligent Recommendation for Online Medical Pre-Diagnosis Support. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 18, 912-921 [Archivo PDF]. https://doi.org/10.1109/TCBB.2020.2994780

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., & Villegas, M. (2022). MarIA: Spanish language models. Procesamiento del Lenguaje Natural, 68, 39-60 [Archivo PDF]. http://doi.org/10.26342/2022-68-3

McClelland, J. L., & Rumelhart, D. E. (1986). Parallel distributed processing: Explorations in the microstructure of cognition. MIT Press [Archivo PDF]. https://doi.org/10.7551/mitpress/5236.001.0001

Baddeley, A. (2023). Working memory: Theories, models, and controversies. Annual Review of Psychology, 74, 1-25 [Archivo PDF]. https://doi.org/10.1146/annurev-psych-120710-100422

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017) [Archivo PDF]. Attention is all you need. Advances in Neural Information Processing

Systems, 30, 5998-6008. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Bengio, Y., Courville, A., & Vincent, P. (2021). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(8), 2519-2535 [Archivo PDF]. https://doi.org/10.1109/TPAMI.2013.50

Shaukat, Zeeshan & Zulfiqar, Abdul Ahad & Xiao, Chuangbai & Azeem, Muhammad & Mahmood, Tariq. (2020). Sentiment analysis on IMDB using lexicon and neural networks. SN Applied Sciences. 2 [Archivo PDF]. https://doi.org/10.1007/s42452-019-1926-x

Henriquez Miranda, Carlos & Guzman, Jaime. (2016). A Review of Sentiment Analysis in Spanish. TECCIENCIA [Archivo PDF]. https://doi.org/10.18180/tecciencia.2017.22.5

Salza, P., Schwizer, C., Gu, J., & Gall, H.C. (2021). On the Effectiveness of Transfer Learning for Code Search. IEEE Transactions on Software Engineering, 49, 1804-1822 [Archivo PDF]. http://doi.org/10.1109/TSE.2022.3192755

Apiola, M., & Sutinen, E. (2020). Design science research for learning software engineering and computational thinking: Four cases. Computer Applications in Engineering Education, 29, 101 - 83 [Archivo PDF]. https://doi.org/10.1002/cae.22291

Schjerven, F. E., Lindseth, F., & Steinsland, I. (2024). Prognostic risk models for incident hypertension: A PRISMA systematic review and meta-analysis. PLOS ONE, 19(3), e0294148 [Archivo PDF]. https://doi.org/10.1371/journal.pone.0294148

Reyes, K., & Aquino, J. (2023) Investigación en las ciencias del diseño: Aplicación en los contextos de computación y tecnología. Entorno de una investigación en ciencias del diseño. Chiclayo: Universidad Católica Santo Toribio de Mogrovejo, 2023 [Archivo PDF]. USAT-LIBRO-INVESTIGACION-CIENCIAS-DEL-DISENO.pdf

Bellar, O., Baina, A., & Ballafkih, M. (2024). Sentiment analysis: Predicting product reviews for e-commerce recommendations using deep learning and transformers. Mathematics, 12(15), 2403 [Archivo PDF]. https://doi.org/10.3390/math12152403

Fernández, D (2021). IMDB Dataset of 50K Movie Reviews (Spanish). https://www.kaggle.com/datasets/luisdiegofv97/imdb-dataset-of-50k-movie-reviews-spanish

Duong, HT., & Nguyen-Thi, TA (2021). A review: preprocessing techniques and data augmentation for sentiment analysis. Comput Soc Netw 8, 1 [Archivo PDF]. https://doi.org/10.1186/s40649-020-00080-x

Chatterjee, I., Zhou, M., Abusorrah, A., Sedraoui, K., & Alabdulwahab, A. (2021). Statistics-based outlier detection and correction method for Amazon customer reviews. Entropy, 23(12), 1645 [Archivo PDF]. https://doi.org/10.3390/e23121645

Colón-Ruiz, Cristóbal & Segura-Bedmar, Isabel. (2020). Comparing deep learning architectures for sentiment analysis on drug reviews. Journal of Biomedical Informatics. 110. 103539 [Archivo PDF]. https://doi.org/10.1016/j.jbi.2020.103539

Pandit, K., Patil, H., Shrimal, D., Suganya, L., & Deshmukh, P. (2024). Comparative analysis of deep learning models for sentiment analysis on IMDB reviews. J. Electrical Systems, 20(2), 424-433 [Archivo PDF]. https://doi.org/10.52783/jes.1345

Jurafsky, Daniel & Martin, James. (2008). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition [Archivo PDF]. https://pages.ucsd.edu/~bakovic/compphon/Jurafsky,%20Martin.-Speech%20and%20Language%20Processing_%20An%20Introduction%20to%20Natural%20Language%20Processing%20(2007).pdf