Evaluación de BERT, CNN y BERT-CNN para la segmentación de tweets en español e inglés

Luis Méndez, Angel Mora
Facultad de Computación
Universidad de Ingeniería y Tecnología
Lima, Perú
{luis.mendez.l,angel.mora}@utec.edu.pe

Abstract

El presente paper evalúa y compara el desempeño de tres modelos de procesamiento del lenguaje natural para la clasificación binaria de tweets en inglés y español. La implementación de un clasificador con una red neuronal convolucional (CNN) sobresale con un Macro F1-Score de 0.91 en inglés y 0.8 en español, resaltando su eficacia en la extracción de características relevantes. BERT es reconocido por su habilidad para comprender contextos semánticos complejos, aunque su rendimiento es un poco más bajo en nuestro experimento, con una puntuación de 0.81 en inglés y 0.7 en español. Una arquitectura híbrida utilizando BERT y CNN para el clasificador muestra resultados competitivos, aunque no superiores a otros modelos. Esto sugiere la necesidad de un enfoque cuidadoso al integrar arquitecturas diversas, considerando posibles beneficios complementarios para mejorar el rendimiento global del modelo.

1 Introducción

Recientemente, se ha observado un gran interés de varios negocios por mejorar la experiencia de usuario en sus plataformas. Una de las herramientas foco del diseño UX es el sistema de recomendaciones y la segmentación de comentarios Furtado et al. (2020). Viendo el auge de herramientas de inteligencia artificial, muchos negocios e investigadores han visto los modelos de machine learning como material de diseño para sus plataformas. Por tanto, portales como X (twitter) y entre otras redes sociales han procurado cuidar la experiencia de sus usuarios realizando análisis sentimental de texto y filtrando discursos en varios idiomas utilizando arquitecturas sofisticadas de procesamiento de lenguaje natural Neethu & Rajasree (2013).

En este marco, la introducción de modelos como BERT y la combinación de BERT con la arquitectura CNN podrían optimizar de manera significativa la realización de tareas de clasificación y análisis de texto en varios idiomas en comparación con modelos convencionales como CNN Safaya et al. (2020). Dependiendo de los datos de entrenamiento y los hiper parámetros seleccionados el desempeño de cada modelo puede ser distinto. En tal sentido, el presente paper busca comparar el desempeño de las arquitecturas mencionadas y evaluar su precisión durante las fases de entrenamiento, validación y pruebas sobre tweets en español e inglés. Además, se discute sobre qué componentes de las arquitecturas y/o métodos de optimización han influenciado en los resultados obtenidos. Finalmente, este experimento nos permite seleccionar el modelo con mejor desempeño, cuyo uso mejoraría la experiencia del usuario de varias plataformas. El modelo seleccionado podría ayudar a muchas plataformas a mejorar su sistema de recomendaciones a partir de la segmentación de reseñas.

Final Paper for the ML3061-2023.2 - Machine Learning Class at UTEC. Lima, Peru.

	Inglés			Español		
	Train	**Validation**	**Test**	**Train**	**Validation**	**Test**
Negativo	75019	15598	15568	81752	10266	10352
Positivo	74966	15401	15431	80672	10037	9951
Total	149985	30999	30999	162424	20303	20303

Table 1: Distribución de los tweets sobre los datasets luego del sampling y split.

2 Trabajo previo

En la literatura académica, se pueden encontrar estudios que evalúan y comparan el rendimiento de estos modelos junto con otros en lo que respecta a tareas de análisis sentimental en texto. Safaya et al. (2020) y Pandey (2023) realizan este tipo de estudio comparativo usando F1 score para la identificación de texto ofensivo en X (twitter), ambos recalcando el outperformance del modelo BERT-CNN.

Otro estudio reciente, el cual involucra un modelo pre entrenado de BERT con un corpus en español y entrenado con medio billón de tweets Pérez et al. (2022). Dado que aún no se ha evaluado el desempeño de los modelos pre-entrenados de BERT y BETO (spanish BERT) en una arquitectura híbrida que incluye CNN, se pretende un experimento similar en los idiomas español e inglés.

Asimismo, Bello et al. (2023) propone el uso de un framework con BERT bastante plausible para el análisis de sentimientos en tweets. Este estudio demuestra que la combinación de BERT con CNN, BERT con RNN y BERT con BiLSTM, todas arquitecturas hibridas, tiene un buen rendimiento en términos de precisión, recuperación y F1 score en comparación con otras técnicas, como el uso de Word2vec y cuando se usa sin ninguna variante.

3 Data

El dataset provisto por Cacharrón (2022) se trata de un conjunto de tweets en inglés, cada tweet etiquetado como 0 (negativo) o 1 (positivo). Por otro lado, los tweets en español son provistos por Ramirez (2023) etiquetados de forma similar. Estos datasets ya habían sido reorganizados en train y test sets de la siguiente forma: En el idioma inglés, el set de entrenamiento y prueba incluye 149,985 y 61,998 tweets respectivamente. En contraste, en el idioma español, el set de entrenamiento y prueba consta de 1,082,821 y 334,641 tweets respectivamente.

Dado que hay una marcada disparidad en los tamaños de los dos conjuntos de datos, tanto de entrenamiento como de pruebas, aplicaremos un muestreo aleatorio del 15% total para ajustar el tamaño del dataset en español de manera que se asemeje al de inglés. Además, procederemos a dividir los datos de prueba para tener además datos de validación. Finalmente, para llevar a cabo el experimento utilizaremos la distribución de datos como sigue en la tabla 1

4 Modelos propuestos

Los siguientes modelos se utilizaron para crear nuestro clasificador de tweets:

4.1 Convolutional Neural Network (CNN)

Se utiliza la arquitectura TextCNN para clasificar textos mediante una serie de capas de convolución unidimensional. La figura 1 muestra a detalle la arquitectura. La red comienza convirtiendo palabras en vectores densos, luego aplica convoluciones con diferentes tamaños de filtro para capturar diversas características en el texto, seguido por ReLU y max pooling para resaltar aspectos relevantes. Este proceso se repite para cada tamaño de filtro especificado.

Después de concatenar los resultados de las operaciones, se aplican a una capa totalmente conectada con función de activación ReLU para prevenir el sobreajuste. Finalmente, la capa de salida genera las predicciones para las clases de salida.

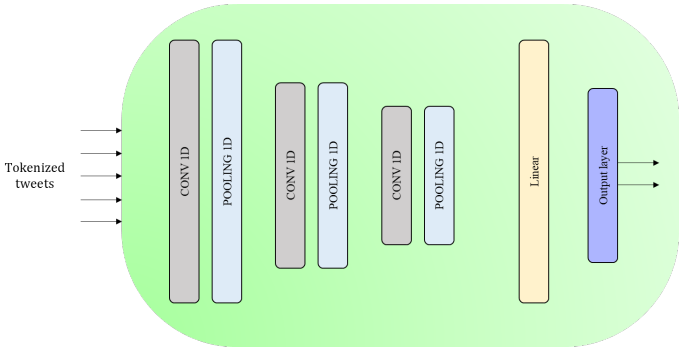

Figure 1: Arquitectura del clasificador con CNN.

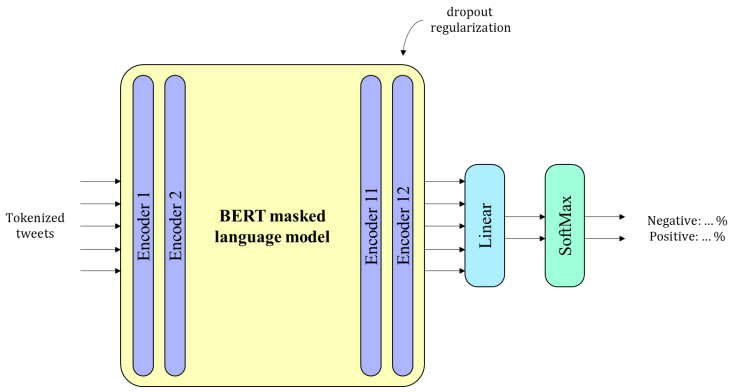

Figure 2: Arquitectura del clasificador con BERT.

4.2 BERT

Utilizaremos el modelo pre entrenado de bert-base-uncased Devlin et al. (2018a) para realizar el respectivo fine-tuning sobre el dataset en inglés y BETO (spanish BERT) que está específicamente entrenado para el idioma español. La figura 2 muestra a detalle la arquitectura. Esta implementación utiliza los ID de palabras y la máscara de atención como entrada y genera una distribución de probabilidad sobre las dos clases de salida. El modelo BERT utiliza una capa de salida para regularización y una capa lineal. Para la clasificación, utiliza una función de activación softmax. El modelo tiene la capacidad de comprender el significado y sentido de las palabras dependiendo del contexto en una oración o frase Devlin et al. (2018b).

4.3 BERT-CNN

La arquitectura en la figura 3 combina el modelo BERT para el procesamiento de lenguaje natural con capas de convolución unidimensional (Conv1d) y una capa lineal para la clasificación de textos. Se utiliza un modelo BERT preentrenado para obtener representaciones contextuales de las palabras en el texto de entrada. Se agregan capas de dropout para reducir el sobreajuste y se aplican convoluciones unidimensionales con diferentes tamaños de filtro para capturar patrones locales en las representaciones de BERT. Después de la convolución, se aplica una capa lineal seguida de una operación softmax para la clasificación binaria.

5 Experimentación y resultados

En relación con los requisitos de experimentación y entrenamiento, para el loss function se empleó Cross Entropy Loss. Los modelos fueron entrenados mediante el método de optimización AdamW, sin incorporar momentum. Para garantizar en la medida de lo posible la convergencia de los loss, se utilizó un learning rate scheduler. El batch size se estableció en 32, y la cantidad de epochs se fijó en

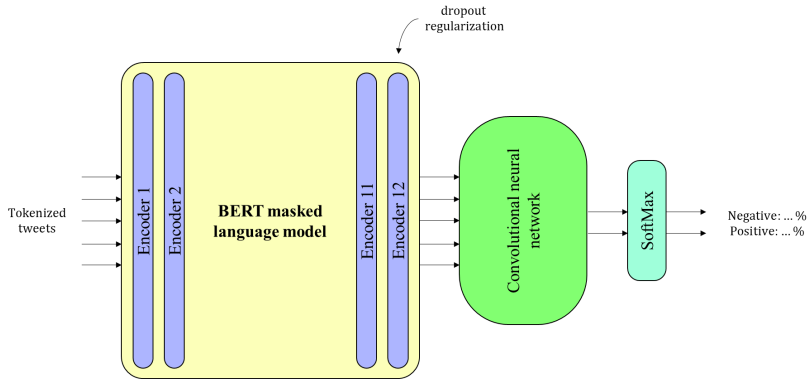

Figure 3: Arquitectura del clasificador con BERT y CNN.

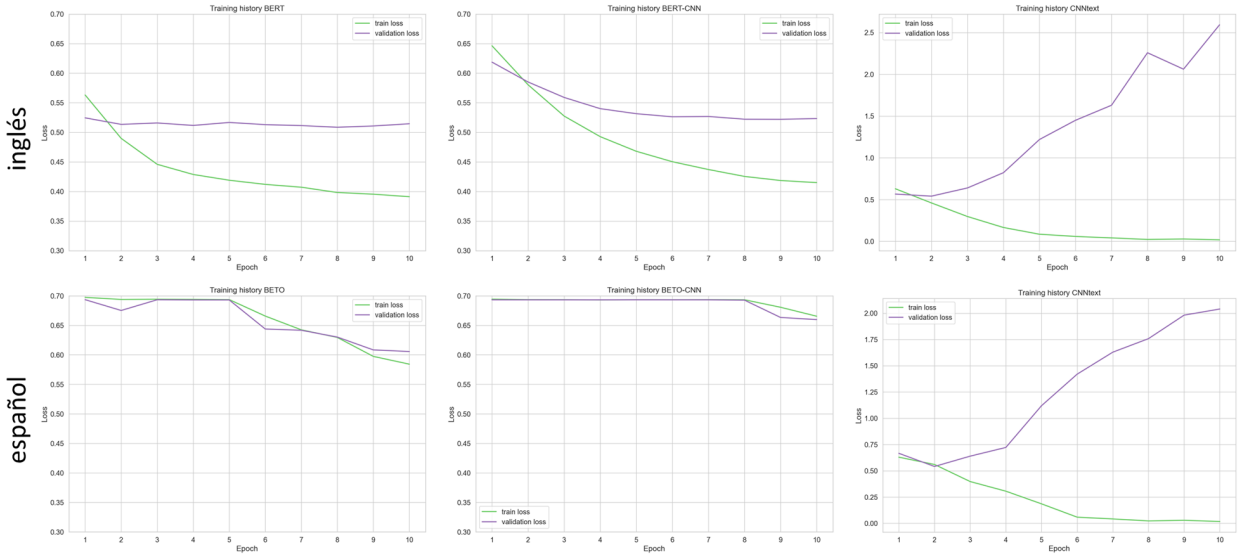

Figure 4: Loss plots por modelo e idioma.

10, ya que durante la experimentación no se observaron mejoras sustanciales en el rendimiento de clasificación de los modelos más allá de este número y para evitar el overfitting. Los resultados se muestran en las figuras 4 y 5

6 Discusión

Según nuestros resultados, en inglés, se observa que BERT y BERT-CNN tienen una curva de loss convergente en el set de entrenamiento, mientras que CNN tiene una curva convergente en el set de entrenamiento pero no en el de validación. Inferimos que una posible razón de este fenómeno es la tendencia del modelo a ajustarse excesivamente a los datos de entrenamiento en lugar de generalizar adecuadamente a los datos de validación. En español, BETO y BETO-CNN tienen curvas

	Inglés	**Español**
CNN	0.91	0.8
BERT	0.81	0.7
BERT-CNN	0.8	0.64

Table 2: Macro F1-Scores promediados de los modelos

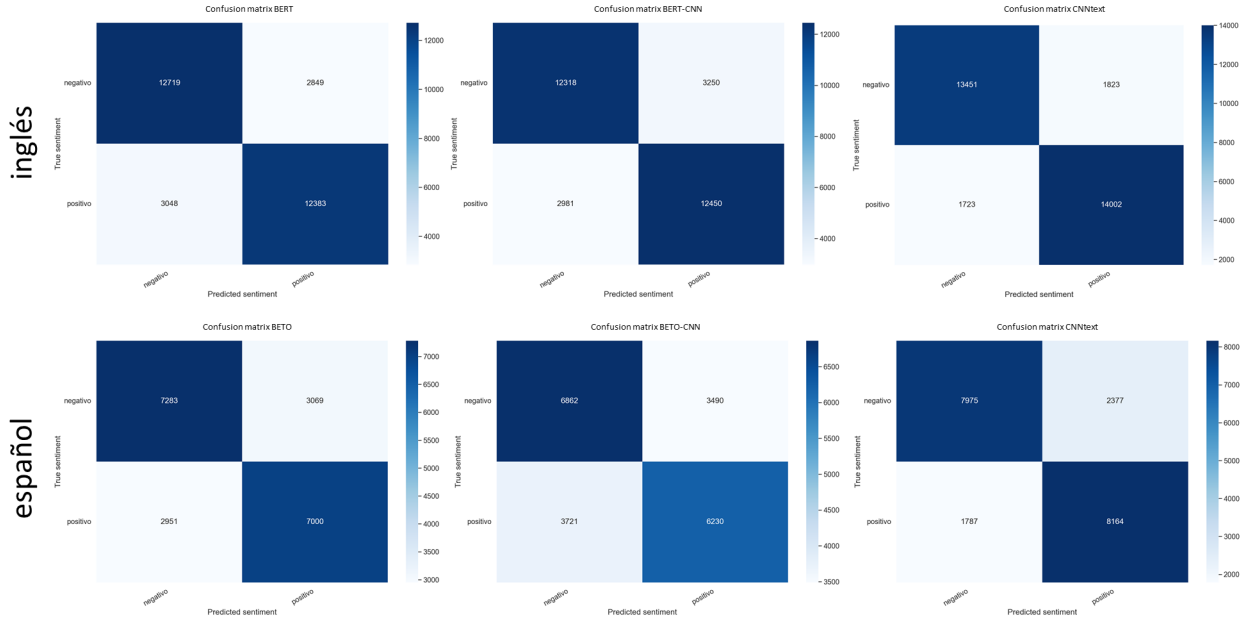

Figure 5: Matrices de confusión por modelo e idioma.

de loss poco convergentes en ambos sets, mientras que CNN tiene una curva convergente en el set de entrenamiento mas no en el de validación. La falta de estabilidad en la optimización del modelo puede haber sido causada por diferencias en el learning rate y la sensibilidad de los hiperparámetros seleccionados, ya que no se exploraron otras combinaciones durante la experimentación. Para mejorar la capacidad de convergencia de la arquitectura BERT-CNN en futuras iteraciones, se requiere una afinación minuciosa de los parámetros y una exploración exhaustiva de técnicas de regularización.

Por otro lado, según las matrices de confusión mostradas en la figura 5, CNN destaca en inglés, mientras que BETO y variantes presentan resultados equilibrados en español. Aunque combinaciones como BERT-CNN muestran mejoras moderadas, todavía hay espacio para las mismas. Se descubrieron problemas para encontrar falsos positivos y negativos en todos los modelos, lo que sugiere áreas de mejora en la optimización de hiperparámetros y técnicas de optimización de idiomas específicas. Estos hallazgos, en conjunto, destacan la importancia de adaptar estrategias según el idioma y optimizar modelos para abordar problemas particulares con la clasificación de tweets.

Finalmente, hemos resumido los classification reports de cada modelo en la tabla 2. Se observa que CNN destaca con un Macro F1-Score de 0.91 en inglés y 0.8 en español, evidenciando su eficacia en la extracción de características relevantes. Aunque BERT presenta un rendimiento ligeramente inferior, con puntuaciones de 0.81 en inglés y 0.7 en español, su capacidad para comprender el contexto semántico complejo sigue siendo notable. La arquitectura basada en la combinación de BERT y CNN demuestra resultados competitivos, mas no superiores a los otros modelos.

7 Conclusiones

En conclusión, nuestros hallazgos demuestran que nuestro clasificador exclusivo de CNN funciona mejor en inglés y español que otros modelos, demostrando su habilidad para identificar características relevantes en tweets. Aunque BERT y BERT-CNN muestran un desempeño competitivo en ambos idiomas, se recomienda implementar estrategias diferentes para maximizar el rendimiento. Para abordar los problemas de clasificación de tweets, la optimización de hiperparámetros y técnicas de idioma específicas son esenciales. Además, las matrices de confusión destacan el buen desempeño de CNN en inglés, mientras que BETO todavía tiene problemas para clasificar tweets en español. Los F1 scores confirman la eficacia de CNN en la tarea especificada, mientras que BERT y BERT-

CNN demuestran capacidades notables, subrayando su capacidad de generalización y robustez, característica que le falta a CNN durante el proceso de validación.

References

Abayomi Bello, Sin-Chun Ng, and Man-Fai Leung. A bert framework to sentiment analysis of tweets. *Sensors*, 23(1):506, 2023.

Miguel Carlos Blanco Cacharrón. twitter-sentiment-analysis, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018a. URL `http://arxiv.org/abs/1810.04805`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018b.

Elizabeth S Furtado, Marilia S Mendes, Denilson C Oliveira, and Lanna Lima. An analysis of ux based on users' emotional intention and values, both expressed through twitter posts. In *Software Ecosystems, Sustainability and Human Values in the Social Web: 8th Workshop of Human-Computer Interaction Aspects to the Social Web, WAIHCWS 2017, Joinville, Brazil, October 23, 2017 and 9th Workshop, WAIHCWS 2018, Belém, Brazil, October 22, 2018, Revised Selected Papers 8*, pp. 79–98. Springer, 2020.

MS Neethu and R Rajasree. Sentiment analysis in twitter using machine learning techniques. In *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*, pp. 1–5. IEEE, 2013.

SHASANK SEKHAR Pandey. A comparative study of bert-cnn and gcn for hate speech detection. B.S. thesis, University of Twente, 2023.

Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M. Luque. RoBERTuito: a pre-trained language model for social media text in Spanish. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 7235–7243, Marseille, France, June 2022. European Language Resources Association. URL `https://aclanthology.org/2022.lrec-1.785`.

Bruno Gil Ramirez. Twitter$_s entiment_a nalysis_t rain_c orpus_i n_s panish$, 2023.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. *arXiv preprint arXiv:2007.13184*, 2020.