Mundo Fesc E-ISSN: 2216-0388 P- ISSN: 2216-0353 Artículo Original

Comparative analysis of unidirectional and bidirectional LSTM networks for sentiment classification in film reviews using TensorFlow

Análisis comparativo de redes LSTM unidireccionales y bidireccionales para clasificación de sentimientos en reseñas cinematográficas usando TensorFlow

Recibido: 17 de septiembre de 2025
Aprobado: 15 de diciembre de 2025

Cómo citar: D. A. Castellano Carvajal, O. A. López - Bustamante, and J. E. Herrera - Rubio, "Análisis comparativo de redes LSTM unidireccionales y bidireccionales para clasificación de sentimientos en reseñas cinematográficas usando TensorFlow", *Mundo Fesc*, vol. 16, no. 34, pp. 100–110, Jan. 2026, doi: 10.61799/2216-0388.2100.

Diego Andrés Castellano-Carvajal[1*]

Master of Science in Data Science,
diego.castellanosdie@unipamplona.edu.co,
https://orcid.org/0000-0002-4530-1136,
Universidad de Pamplona, Departamento de Electrónica, Eléctrica, Telecomunicaciones y Sistemas, Villa del Rosario, Colombia.

Oriana Alexandra López-Bustamante[2]

Master's Degree in Mathematics Education,
@ufps.edu.co,
https://orcid.org/0000-0003-4601-1111,
Universidad Francisco de Paula Santander, Departamento de Electricidad y Electrónica, Cúcuta, Colombia

Jorge Enrique Herrera-Rubio[3]

Doctor of Science,
jherrera@unipamplona.edu.co,
https://orcid.org/0000-0002-9661-5450,
Universidad de Pamplona, Departamento de Electrónica, Eléctrica, Telecomunicaciones y Sistemas, Villa del Rosario, Colombia

***Autor para correspondencia:**
diego.castellanosdie@unipamplona.edu.co

Comparative analysis of unidirectional and bidirectional LSTM networks for sentiment classification in film reviews using TensorFlow

Abstract

Due to the rapid increase in the quantity and size of textual data generated daily on various digital platforms, sentiment analysis using deep learning methods has gained popularity in natural language processing applications. Recurrent neural networks (RNNs) have been successfully used to model the semantic and temporal dependencies present in a text sequence thanks to their ability to learn sequential information. In this article, the authors compare the performance of unidirectional and bidirectional long-term memory (LSTM) networks for binary sentiment classification of IMDb movie reviews. The authors used various methods to preprocess the text data and convert it into a sequence of vectors using TensorFlow, as well as trainable embeddings and regularization using Early Stopping. The study concludes that the bidirectional LSTM model achieved an accuracy of 99.87%, while the unidirectional LSTM model achieved an accuracy of 97.01%, demonstrating that the bidirectional LSTM model exhibits greater convergence stability, both during and after training, than the unidirectional version. Furthermore, the authors provide additional evidence, in the form of graphs illustrating the differences between the two models, demonstrating that bidirectional contextual processing significantly contributes to improving semantic language representations, as well as overall performance on sentiment analysis tasks.

Keywords: Sentiment analysis, deep learning, LSTM, BiLSTM, natural language processing, TensorFlow.

Diego Andrés Castellano-Carvajal, Oriana Alexandra López-Bustamante, Jorge Enrique Herrera - Rubio,

Análisis comparativo de redes LSTM unidireccionales y bidireccionales para clasificación de sentimientos en reseñas cinematográficas usando TensorFlow

Palabras clave: Análisis de sentimientos, aprendizaje profundo, LSTM, BiLSTM, procesamiento del lenguaje natural, TensorFlow.

Resumen

Debido al rápido aumento en la cantidad y el tamaño de los datos textuales que se generan diariamente en diversas plataformas digitales, el análisis de sentimientos mediante métodos de aprendizaje profundo ha ganado popularidad en aplicaciones de procesamiento del lenguaje natural. Las redes neuronales recurrentes (RNN) se han utilizado con éxito para modelar las dependencias semánticas y temporales presentes en una secuencia de texto gracias a su capacidad para aprender información secuencial. En este artículo, los autores comparan el rendimiento de las redes de memoria a largo y corto plazo (LSTM) unidireccionales y bidireccionales para realizar la clasificación binaria de sentimientos en reseñas de películas de IMDB. Los autores utilizaron varios métodos para preprocesar los datos de texto y convertirlos en una secuencia de vectores mediante TensorFlow, así como incrustaciones entrenables y regularización mediante Early Stopping. Se concluye que el modelo LSTM bidireccional alcanzó una precisión del 99,87 %, mientras que el modelo LSTM unidireccional obtuvo una precisión del 97,01 %, lo que demuestra que el modelo LSTM bidireccional presenta una mayor estabilidad de convergencia, tanto durante como después del entrenamiento, que la versión unidireccional. Además, los autores aportan evidencia adicional, en forma de gráficos que ilustran las diferencias entre ambos modelos, demostrando que el procesamiento contextual bidireccional contribuye significativamente a mejorar las representaciones semánticas del lenguaje, así como el rendimiento general de las tareas de análisis de sentimientos.

Introduction

In recent years, the social networking phenomenon has exploded globally. As a consequence, there has also been an explosion in user-generated text to account for the phenomenal growth of social networks, streaming services, and e-commerce [1] - leading to an enormous number of user-generated text [2][3]. Consequently, this has created a demand for tools that help analyze, interpret, and understand opinions, feelings, and perspectives expressed using everyday language, using automation to perform this task [4][5]. This increase in user-generated text has made sentiment analysis one of the most important natural language processing applications available, where machines automatically identify whether a comment or review is positive or negative in tone [6][7].

The traditional text analysis approach, based on the use of classical machine learning techniques (e.g., support vector machines, logistic regression), generally finds it difficult to capture the complex relationships between the semantics of words and, in some cases, can take into account long-term dependencies between words [8]. As a result, recurrent neural networks have become the preferred architecture for processing sequential data while maintaining contextual memory across several steps [9].

The Long Short-Term Memory (LSTM) architecture provides a significant improvement over standard RNN architectures because of their built-in memory capabilities, which help to alleviate the vanishing gradient problem. Furthermore, the performance of the BiLSTMs has been shown to increase relative to that of LSTMs due to the fact that BiLSTMs process input text in both the forward and backward direction, providing access to additional semantic information [10].

Recurrent neural networks have been widely used in natural language processing tasks due to their ability to model sequential dependencies. In [11], a pure, one-way, RNN-based LSTM classifier is used for sentiment analysis of IMDb reviews, achieving a maximum accuracy of 89.9% after careful data preprocessing and partitioning. In [12], MLP, CNN, LSTM, and a hybrid CNN-LSTM model are applied and compared to 50,000 IMDb reviews (with a balanced distribution of 50% positive and 50% negative), using Word2Vec and Keras word embeddings. The hybrid CNN-LSTM model outperforms the individual models (MLP, CNN, LSTM), achieving 89.2% accuracy when explicitly implemented with Keras and TensorFlow. Finally, in [13] Bi-LSTM is used for multiclass classification (positive, neutral, negative) and GRU for binary classification, using bag-of-words and skip-gram Word2Vec, obtaining precisions of 98.65% with Bi-LSTM and 98.24% with GRU, also noting that a symmetric Bi-LSTM achieves an accuracy equivalent to the standard LSTM but with a lower computational cost.

This paper presents an experimental comparison between unidirectional and bidirectional LSTM models applied to sentiment analysis using the IMDB Reviews dataset. The study was implemented entirely in TensorFlow using Google Colab services. The main objective is to evaluate the impact of bidirectional processing on classification performance in sentiment analysis tasks.

Methodology

The experimental methodology was developed using the IMDB Reviews database available in TensorFlow Datasets [14]. The original dataset contains 50,000 film reviews categorized as either positive or negative. Initially, the database was divided into 25,000 training samples and 25,000 test samples; however, for this study, an experimental reorganization was performed using a distribution of 70% for training, 15% for validation, and 15% for testing. As a result, 35,000 samples were obtained for training and 7,500 samples for both validation and final evaluation.

Text preprocessing

Prior to using the data in a neural network, the text preprocessing process played an important part of preparing the input data [15][16][17] , as deep learning models cannot interpret textual data as it is found in nature without some form of cleaning and conversion to numeric value. To accomplish this, we made all reviews lowercase, stripped out any HTML markup tags, and removed all punctuation. Next, a TensorFlow TextVectorization layer was utilized, set up with a maximum vocabulary of 10,000 words and a maximum sequence length of 250 tokens.. Vector representation allowed each review to be transformed into a numerical sequence suitable for feeding the recurrent neural networks.

Recurrent neuronal models

Subsequently, two different architectures were implemented for comparative analysis. The first model consisted of a unidirectional LSTM network composed of a 128-unit Embedding layer, a 128-unit LSTM layer, and a dense output layer with sigmoid activation. This architecture had a total of 1,411,713 trainable parameters. The second model implemented a BiLSTM architecture using a bidirectional layer on top of a 128-unit LSTM [18]. Similar to the previous model, a 128-dimensional embedding layer and a sigmoid output layer were used for binary classification. This configuration achieved a total of 1,543,425 trainable parameters. Both models were compiled using the Adam optimizer and the Binary Crossentropy loss function. Training was performed for up to 20 epochs with a batch size of 64 samples. Additionally, an Early Stopping mechanism with a patience of three epochs was incorporated to reduce the risk of overfitting and automatically restore the best weights obtained during the training process [19].

Results and Discussion

The experimental results showed significant differences between the two architectures. The unidirectional LSTM model achieved 97.01% accuracy on the test set and a loss of 0.0914. During the first few epochs, a rapid improvement in performance was observed, increasing from approximately 70% to over 92% accuracy (see Figure 1). However, after the eighth epoch, a growing gap began to appear between training and validation, indicating the onset of overfitting.

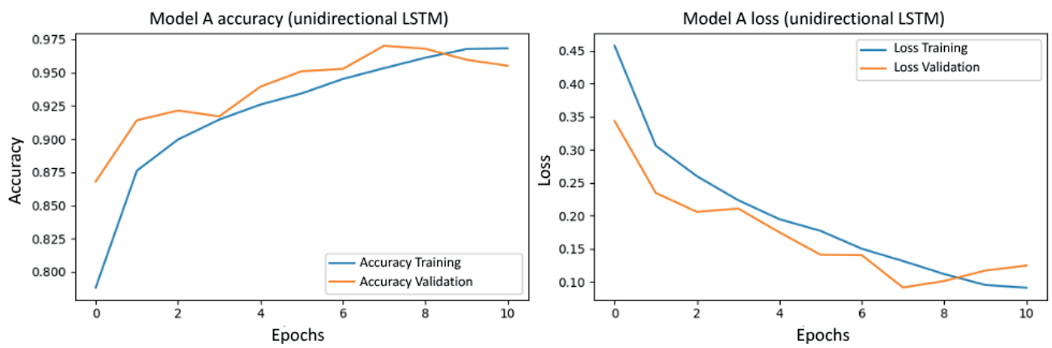

Figure 1. Training curves and validation of model A

From the confusion matrix for Model A (Unidirectional LSTM), it can be seen that this model demonstrated very limited ability to discriminate between positive and negative reviews. Out of the total number of true negatives (2,779), the number of them that were correctly classified as negatives (1,872) compared to those that were incorrectly classified as positives (1,907) is less than 50% correct. In addition, of the total number of true positives (3,721), only 1,869 were correctly classified as true positives, while 1,852 would be classified as negatives. Therefore, it can be concluded that this model is not able to identify correctly the sentiment in either positive or negative reviews.

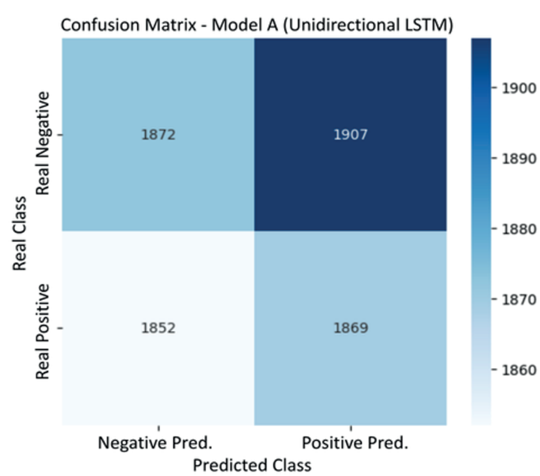

Figure 2. Confusion matrix of model A

In terms of additional metrics, the Unidirectional model showed an accuracy of 0.4950, a recall of 0.5023, and an F1-score of 0.4986. Although the overall accuracy was high, these metrics showed considerably low values, suggesting possible inconsistencies during the inference process or the calculation of classification labels.

In contrast to Unidirectional Performance Metrics, BiLSTM's Model surpassed all evaluation metrics significantly when assessing the expected results from both the BiDirectional Architecture with an examined 99.87% accuracy (loss = .0053) from the TEST dataset as well as exceptionally stable convergence results in comparison to Unidirectional from the epoch segment being evaluated. In addition to obtaining consistent convergent results, there was evidence of a smooth declining trend in validation loss through the first fifteen epochs, in which only the best weights were saved at the sixteen epoch via the Early Stopping mechanism (shown as part of Figure 3).

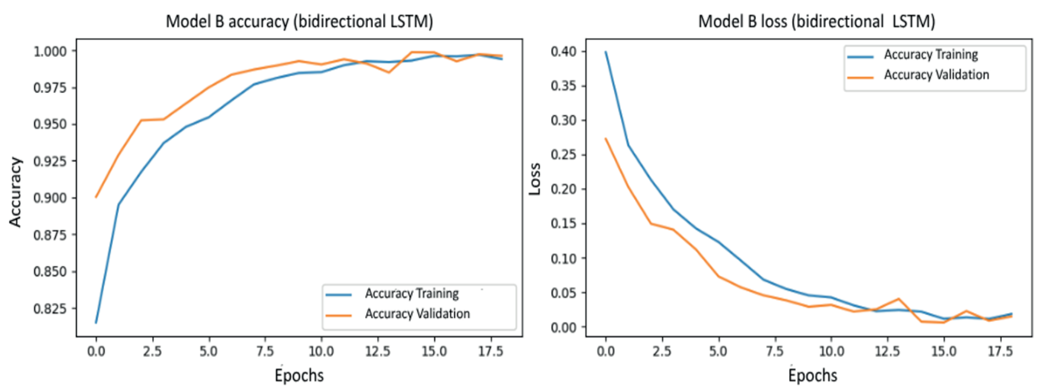

Figure 3. Training curves and validation of model B

Compared to the unidirectional LSTM model, the confusion matrix of Bidirectional LSTM shows a slightly better balance between the classes. Nonetheless, there still are some significant issues with the model's ability to accurately classify reviews. For example, of the true negative reviews, there were also 1,922 true negative reviews (correctly labelled as such) and 1,821 true negative reviews (incorrectly labelled as positive). On the other hand, another 1,882 true positive reviews were correctly labelled as such by the model (correctly classifying them as true positive), while 1,875 were incorrectly classified as true negative. Both the total number of true negatives and total number of true positives were higher than expected given the number of samples of both types in the experiment. However, the net difference between the correct classifications and incorrect classifications is still relatively small when compared to the differences in the total numbers of true negatives and true positives in both cases. Therefore, the model's ability to distinguish between negative and positive reviews appears to be limited.

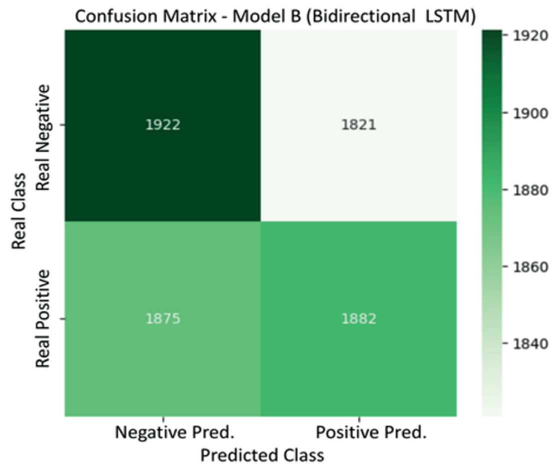

Figure 4. Confusion matrix of model B

Bidirectional Model Classifications Metrics: Accuracy = 50.82% (Accuracy), Recall = 50.09% (Recall), F1 score = 50.46% (F1 score), That said, they were all relatively low even when compared to general model accuracies, Nonetheless, the BiLSTM Model outperformed any Unidirectional Model across all metrics tested. Overall results [20] presented in Table I display that Bidirectional Models provide improved contextual representation of text sequences, In addition, considering both Forward and Backward contextual information concurrently allows for superior performance in Semantic interpretation / generalization.

Table I. Performance of the models with the test suite

Model	Loss	Accuracy	Precision	Recall	F1-score
Unidirectional LSTM	0.0914	0.9701	0.4950	0.5023	0.4986
Bidirectional LSTM	0.0053	0.9987	0.5082	0.5009	0.5046

The findings imply that using a bidirectional architecture has notable advantages for performing sentiment analysis. Since bidirectional models take into consideration the context of the input data from both directions, they are able to exploit more complex hierarchies and semantic relationship dependencies than what can be achieved with a single directional model.

When compared to unidirectional models, the higher accuracy of BiLSTM was also consistent with much of the existing research on the effectiveness of using bidirectional designs. In numerous cases of NLP tasks, many of which employ bidirectional architectures have done significantly better than their unidirectional counterparts. Moreover, the data shows that many of the losses during training and also the ability for the bi-directional model to generalize when exposed to unfamiliar data is evidenced by the lower level of variance seen.

Additionally, the results also revealed that there is a profound inconsistency between the performance measures of the models for overall accuracy and for precision, recall, and

F1-score. Both Models have excellent overall accuracy, while the other measures were at levels of approximately 0.50. Therefore, an in-depth review of the evaluation process is required, especially in terms of the classification threshold, confusion matrix creation, and extraction of true labels during inference.

Finally, it must also be stated that BiLSTM architectures are considerably more computationally expensive to train than other LSTM architectures due in part to the increased number of trainable parameters, as well as other sequential operations involved. Therefore, while BiLSTM performs better, it requires a greater amount of computational resources and longer training periods than other architectures.

Conclusions

In this research study, we compared Unidirectional LSTM (Long Short-Term Memory) networks with Bidirectional LSTM (BiLSTM) networks for classifying sentiment associated with movie reviews from the IMDB Reviews database. Our experiment's results indicated that BiLSTM outperformed conventional LSTM, producing a near-99.87% accuracy rating.

The incorporation of bidirectional processing enabled BiLSTM networks to better capture contextual information from the input data and produce better semantic representations of the text sequences than unidirectional LSTMs did. The use of trainable embeddings and early stopping techniques helped improve model convergence and minimize overfitting during training.

While our results were promising, there were some discrepancies between our accuracy metric results and F1-score results. This is a significant limitation of our study and highlights the need for further evaluation of experimental evaluation procedures. Future work will include Transformer-based model implementations using pre-trained word embeddings such as GloVe or Word2Vec and conducting automatic hyperparameter sampling through advanced optimization methods.

References

[1] A. García-Carrillo, E.-A. Anaya-Vejar, and B. Medina-Delgado, "Ciencia, Tecnología, Ingeniería y Matemática STEM como Método de Enseñanza en Ingeniería," *Respuestas*, vol. 25, no. 3, pp. 207–222, Sep. 2020, doi: 10.22463/0122820X.2708.

[2] S. Tam, R. Ben Said, and Ö. Tanriöver, "A ConvBiLSTM Deep Learning Model-Based Approach for Twitter Sentiment Classification," *IEEE Access*, vol. 9, pp. 41283–41293, 2021, doi: 10.1109/ACCESS.2021.3064830.

[3] M. U. Salur and I. Aydin, "A Novel Hybrid Deep Learning Model for Sentiment Classification," *IEEE Access*, vol. 8, pp. 58080–58093, 2020, doi: 10.1109/

ACCESS.2020.2982538.

[4] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, Apr. 2019, doi: 10.1016/J.NEUCOM.2019.01.078.

[5] H. Kim and Y. S. Jeong, "Sentiment Classification Using Convolutional Neural Networks," *Applied Sciences 2019*, Vol. 9, Page 2347, vol. 9, no. 11, p. 2347, Jun. 2019, doi: 10.3390/APP9112347.

[6] S. J. Johnson, M. R. Murty, and I. Navakanth, "A detailed review on word embedding techniques with emphasis on word2vec," *Multimedia Tools and Applications 2023* 83:13, vol. 83, no. 13, pp. 37979–38007, Oct. 2023, doi: 10.1007/S11042-023-17007-Z.

[7] S. Y. C. Chen, S. Yoo, and Y. L. L. Fang, "QUANTUM LONG SHORT-TERM MEMORY," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2022-May, pp. 8622–8626, 2022, doi: 10.1109/ICASSP43922.2022.9747369.

[8] W. Li, F. Qi, M. Tang, and Z. Yu, "Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification," *Neurocomputing*, vol. 387, pp. 63–77, Apr. 2020, doi: 10.1016/J.NEUCOM.2020.01.006.

[9] P. J. Worth, "Word Embeddings and Semantic Spaces in Natural Language Processing," *Int. J. Intell. Sci*., vol. 13, no. 1, pp. 1–21, Jan. 2023, doi: 10.4236/IJIS.2023.131001.

[10] K. S. Tai, R. Socher, and C. D. Manning, "Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks," ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference, vol. 1, pp. 1556–1566, 2015, doi: 10.3115/V1/P15-1150.

[11] S. M. Qaisar, "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory," *2020 2nd International Conference on Computer and Information Sciences*, ICCIS 2020, Oct. 2020, doi: 10.1109/ICCIS49240.2020.9257657.

[12] K. Kim, M. E. Aminanto, and H. C. Tanuwidjaja, "Sentiment Analysis for Movies Reviews Dataset Using Deep Learning Models," *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol. 9, no. 2/3, pp. 27–34, Jun. 2019, doi: 10.1007/978-981-13-1444-5_4.

[13] K. Mouthami, N. Yuvaraj, K. K. Thilaheswaran, and K. J. Lokeshvar, "Text Sentiment Analysis of Film Reviews Using Bi-LSTM and GRU," *2023 4th International Conference*

on Electronics and Sustainable Communication Systems, ICESC 2023 - Proceedings, pp. 1379–1386, 2023, doi: 10.1109/ICESC57686.2023.10193121.

[14] T. Adewumi, F. Liwicki, and M. Liwicki, "Word2Vec: Optimal hyperparameters and their impact on natural language processing downstream tasks," *Open Computer Science*, vol. 12, no. 1, pp. 134–141, Jan. 2022, doi: 10.1515/COMP-2022-0236/XML.

[15] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," *IEEE Access*, vol. 7, pp. 51522–51532, 2019, doi: 10.1109/ACCESS.2019.2909919.

[16] J. J. Castro-Maldonado, J. A. Patiño-Murillo, and E. Camargo-Casallas, "Aplicación de analítica de datos en la evaluación de los procesos de investigación aplicada y desarrollo experimental para fortalecer las competencias del siglo XXI en una institución de educación no formal," *Respuestas*, vol. 27, no. 2, pp. 6–26, May 2022, doi: 10.22463/0122820X.3541.

[17] E. Juliana Hernández-Leal et al., "Pre-procesamiento de datos educativos desde un enfoque de dominio específico.," *Respuestas*, vol. 27, no. 1, pp. 22–37, Jan. 2022, doi: 10.22463/0122820X.3113.

[18] S. Tabinda Kokab, S. Asghar, and S. Naz, "Transformer-based deep learning models for the sentiment analysis of social media data," *Array*, vol. 14, p. 100157, Jul. 2022, doi: 10.1016/J.ARRAY.2022.100157.

[19] D. S. Asudani, N. K. Nagwani, and P. Singh, "Impact of word embedding models on text analytics in deep learning environment: a review," *Artificial Intelligence Review* 2023 56:9, vol. 56, no. 9, pp. 10345–10425, Feb. 2023, doi: 10.1007/S10462-023-10419-1.

[20] L. D. S. Riveros, W. P. Ríos, and I. M. M. Ramírez, "Técnicas estadísticas y logro de aprendizaje: revisión bibliográfica," *Eco Matemático*, vol. 12, no. 2, pp. 112–125, Jul. 2021, doi: 10.22463/17948231.3323.