Revista Científica Multidisciplinar G-ner@ndo ISSN: 2806-5905

# Optimización de ChatGPT para la clasificación de noticias falsas mediante FineTuning.

# Optimization of ChatGPT for fake news classification through Fine-Tuning.

*César Alcívar, David Fernando Ramos Tomalá & Mariuxi Del Carmen Toapanta Bernabé*

**PUNTO CIENCIA.**

**julio - diciembre, V°6 - N°2; 2025**

**Recibido:** 20-10-2025
**Aceptado:** 23-10-2025
**Publicado:** 30-12-2025

**PAIS**

- Ecuador, Guayaquil
- Ecuador, Guayaquil
- Ecuador, Guayaquil

**INSTITUCION**

- Universidad de Guayaquil
- Universidad de Guayaquil
- Universidad de Guayaquil

**CORREO:**

✉ cesar.alcivarar@ug.edu.ec
✉ david.ramost@ug.edu.ec
✉ mariuxi.toapantab@ug.edu.ec

**ORCID:**

https://orcid.org/0009-0004-0214-905X
https://orcid.org/0009-0007-2702-8926
https://orcid.org/0000-0002-4839-7452

**FORMATO DE CITA APA.**

*Alcívar, C., Ramos, D. & Toapanta, M.* (2025). Optimización de ChatGPT para la clasificación de noticias falsas mediante FineTuning. *Revista G-ner@ndo, V°6 (N°2). Pág. 2212 – 2223.*

**Resumen**

En la era de la información digital, la proliferación de noticias falsas se ha convertido en un desafío crítico para la sociedad, afectando la percepción pública y teniendo implicaciones significativas para la política, la salud pública y la seguridad. Este estudio se centra en la optimización de ChatGPT, un modelo de lenguaje desarrollado por OpenAI, para la tarea de clasificación de noticias falsas mediante el proceso de fine-tuning. Utilizando un conjunto de datos robusto y técnicas de preprocesamiento detalladas, se busca mejorar la precisión y la relevancia de las respuestas generadas por el modelo en el contexto de la verificación de hechos.La metodología incluye la preparación y preprocesamiento de datos, la conversión a formatos compatibles con la API de OpenAI, y un monitoreo cuidadoso del proceso de entrenamiento. Los resultados obtenidos muestran una alta precisión, aunque con un riesgo significativo de sobreajuste debido al tamaño limitado del conjunto de datos. Se discuten las implicaciones del modelo entrenado y se proponen futuras investigaciones para mejorar su robustez y generalización.

**Palabras clave:** ChatGPT, fine-tuning, fake news, fact-checking, natural language processing (NLP).

**Abstract**

In the digital information era, the proliferation of fake news has become a critical challenge for society, impacting public perception and having significant implications for politics, public health, and security. This study focuses on optimizing ChatGPT, a language model developed by OpenAI, for the task of fake news classification through fine-tuning. Utilizing a robust dataset and detailed preprocessing techniques, the aim is to enhance the accuracy and relevance of the model's responses in the context of fact-checking.The methodology includes data preparation and preprocessing, conversion to formats compatible with OpenAI's API, and careful monitoring of the training process. The results show high accuracy, albeit with a significant risk of overfitting due to the limited dataset size. The trained model's implications are discussed, and future research directions are proposed to improve its robustness and generalization.

**Keywords:** ChatGPT, fine-tuning, fake news, fact-checking, natural language processing (NLP).

## Introducción

En la era de la información digital, la proliferación de noticias falsas se ha convertido en un desafío crítico para la sociedad. La capacidad de distinguir entre noticias verdaderas y falsas no solo afecta la percepción pública, sino que también tiene implicaciones significativas para la política, la salud pública y la seguridad (Lazer et al., 2018). A medida que las plataformas de medios sociales y los sitios web de noticias en línea se convierten en las principales fuentes de información para millones de personas, la necesidad de herramientas efectivas para la verificación de hechos y la clasificación de noticias falsas se vuelve cada vez más urgente (Vosoughi, Roy, & Aral, 2018; Shu et al., 2017; Zhou & Zafarani, 2020).

El avance en los modelos de lenguaje natural, particularmente aquellos desarrollados por OpenAI, ha abierto nuevas posibilidades para abordar este problema. ChatGPT, un modelo de lenguaje basado en la arquitectura GPT-3.5 (OpenAI, 2023), ha demostrado un potencial notable en diversas aplicaciones de procesamiento de lenguaje natural (NLP) (Brown et al., 2020; Wei et al., 2022; Thorne & Vlachos, 2018). Sin embargo, su capacidad para clasificar noticias falsas de manera precisa y eficiente requiere una optimización específica mediante técnicas avanzadas de aprendizaje automático, como el fine-tuning (Zellers et al., 2019; Alonso & Fernández, 2023).

Este estudio se centra en la optimización de ChatGPT para la tarea de clasificación de noticias falsas mediante el proceso de fine-tuning. Aprovechando un conjunto de datos robusto y técnicas de preprocesamiento detalladas, nuestro objetivo es mejorar la precisión y la relevancia de las respuestas generadas por el modelo en el contexto de la verificación de hechos. La metodología empleada incluye la preparación de datos provenientes de fuentes confiables, la conversión de estos datos a formatos compatibles con la API de OpenAI, y un monitoreo cuidadoso del proceso de entrenamiento para asegurar resultados óptimos (Schuster et al., 2019; Nguyen & Nguyen, 2022).

Además de presentar la metodología detallada y los resultados obtenidos, este estudio también analiza las implicaciones del modelo entrenado y su desempeño. A través de análisis de nube de palabras y evaluaciones de desempeño, se proporciona una comprensión profunda de cómo el modelo maneja la tarea de clasificación y las áreas donde puede mejorarse. Las conclusiones ofrecen una dirección para futuras investigaciones, destacando la importancia de ampliar los conjuntos de datos y la implementación de técnicas adicionales de validación y regularización para evitar el sobreajuste y mejorar la generalización (Goodfellow et al., 2016).

El objetivo general de este trabajo es mejorar la capacidad de ChatGPT para clasificar noticias falsas mediante la optimización de su proceso de entrenamiento. Se propone la hipótesis de que el fine-tuning de ChatGPT con un conjunto de datos robusto y técnicas de preprocesamiento detalladas resultará en un modelo más preciso y eficiente para la verificación de hechos.

La investigación es de naturaleza aplicada y se centra en la optimización de un modelo de lenguaje natural mediante técnicas de aprendizaje automático. Las variables principales incluyen la precisión del modelo en la clasificación de noticias falsas, los métodos de preprocesamiento de datos y los parámetros de entrenamiento utilizados.

## Métodos y Materiales

La metodología empleada en este estudio es de tipo cuantitativa, enfocada en la optimización de ChatGPT para la clasificación de noticias falsas mediante el proceso de fine-tuning.

### Procedimientos Metodológicos

Preparación de Datos

- Fuente de Datos: Los datos utilizados en este estudio provienen de BuzzFeed, una reconocida plataforma de medios de comunicación.

- Conjuntos de Datos: Se emplearon dos conjuntos de datos: uno para noticias reales y otro para noticias falsas. Ambos conjuntos fueron preprocesados y etiquetados adecuadamente.

- Preprocesamiento: El preprocesamiento de datos incluyó la limpieza y normalización de los textos, eliminación de palabras comunes (stopwords), y etiquetado de las noticias como verdaderas o falsas. Estos pasos aseguran que los datos sean consistentes y adecuados para el entrenamiento del modelo.

Creación de Archivos JSONL

Formato de Datos: Los datos se convirtieron a formato JSONL (JSON Lines), compatible con la API de OpenAI para el proceso de fine-tuning. La estructura de cada entrada en los archivos JSONL incluye un rol del sistema, una entrada del usuario y una respuesta del asistente.

Estructura de Entradas: Cada entrada se estructuró de la siguiente manera:

```
{
  "role": "system",
  "content": "You are an AI fact-checking assistant. Verify the accuracy of claims using reliable sources."
}
```

Este formato asegura que el modelo entienda el contexto y la tarea específica que debe realizar, mejorando su capacidad para clasificar noticias falsas.

Entrenamiento y Validación del Modelo

- Modelo Utilizado: Se utilizó el modelo GPT-3.5-turbo de OpenAI.
- Parámetros de Entrenamiento: Se especificaron parámetros de hiperajuste como el número de épocas, el tamaño del lote y el multiplicador de la tasa de aprendizaje. El modelo se entrenó con un total de 1,537,470 tokens, distribuidos en 725 pasos con un tamaño de lote de 3 y un multiplicador de la tasa de aprendizaje de 0.3.
- Proceso de Entrenamiento: El entrenamiento se llevó a cabo durante 15 épocas, utilizando técnicas de monitoreo continuo para asegurar la calidad del proceso. Se utilizó una señal de interrupción para manejar posibles desconexiones y asegurar que el proceso se completara correctamente.
- Monitoreo del Entrenamiento: El progreso del entrenamiento se monitoreó periódicamente, verificando el estado del trabajo de fine-tuning hasta su finalización exitosa o fallida.

Monitoreo del Proceso de Entrenamiento

- Control del Proceso: Se implementó un sistema de monitoreo continuo para manejar posibles interrupciones y asegurar la integridad del proceso de entrenamiento. Este enfoque permitió identificar y corregir problemas rápidamente, garantizando la calidad del modelo resultante.
- Evaluación Continua: El estado del trabajo de fine-tuning se verificó periódicamente hasta que el proceso se completara con éxito o fallara.

Enfoque en el Prompt

Definición del Contexto: El prompt utilizado en el proceso de fine-tuning define a ChatGPT como un asistente de verificación de hechos, responsable de verificar la exactitud de las afirmaciones utilizando fuentes confiables. Este enfoque asegura que el modelo esté afinado para responder de manera precisa y relevante en el contexto de la clasificación de noticias falsas.

El estudio utilizó un total de 182 registros, distribuidos equitativamente entre noticias verdaderas y falsas. Estos registros fueron seleccionados aleatoriamente de la base de datos de BuzzFeed para asegurar una representación balanceada y adecuada de ambos tipos de noticias.

## Análisis de Resultados

**Resultados de la de revisión documental.**

Durante el proceso de fine-tuning del modelo GPT-3.5-turbo, se entrenaron un total de 1,537,470 tokens, distribuidos en 725 pasos con un tamaño de lote de 3 y un multiplicador de la tasa de aprendizaje de 0.3. El entrenamiento se llevó a cabo durante 15 épocas.

Los resultados obtenidos mostraron una precisión del 100%, lo cual, a primera vista, podría parecer ideal. Sin embargo, esta alta precisión debe analizarse en el contexto del tamaño y naturaleza del conjunto de datos.

El gráfico de la pérdida de entrenamiento muestra una disminución rápida y continua, estabilizándose cerca de cero hacia el final del entrenamiento. Esto puede ser indicativo de varios factores:

Sobreajuste (Overfitting):

- Cantidad de Datos Limitada: El conjunto de datos utilizado contenía solo 182 registros. Con un conjunto de datos tan pequeño, el modelo puede aprender patrones

específicos de los datos de entrenamiento, en lugar de generalizar a nuevos datos. Esto es particularmente probable cuando se logra una precisión del 100%, ya que el modelo puede estar “memorizando” los ejemplos de entrenamiento en lugar de aprender características generalizables.

- Reducción de Pérdida a Cero: La pérdida de entrenamiento que cae casi a cero también sugiere que el modelo ha ajustado perfectamente los datos de entrenamiento, una señal común de sobreajuste cuando se dispone de pocos datos.

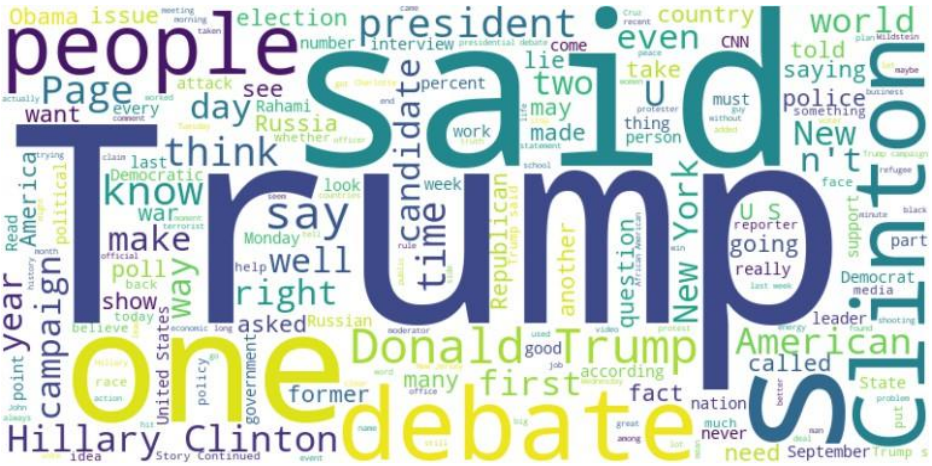
Tamaño de Lote y Tasa de Aprendizaje:

- Tamaño de Lote Pequeño: Un tamaño de lote de 3 implica actualizaciones de peso muy frecuentes, lo que puede llevar a un modelo que se ajusta muy estrechamente a los datos específicos en cada lote.

- Multiplicador de la Tasa de Aprendizaje: Un multiplicador de la tasa de aprendizaje de 0.3 es relativamente alto, lo que puede hacer que el modelo ajuste sus pesos de manera más agresiva, contribuyendo al sobreajuste si los datos son limitados.

Se realizaron análisis de nubes de palabras tanto para las noticias reales como para las falsas, eliminando las palabras comunes (stopwords). A continuación, se presentan los resultados y comparaciones clave:

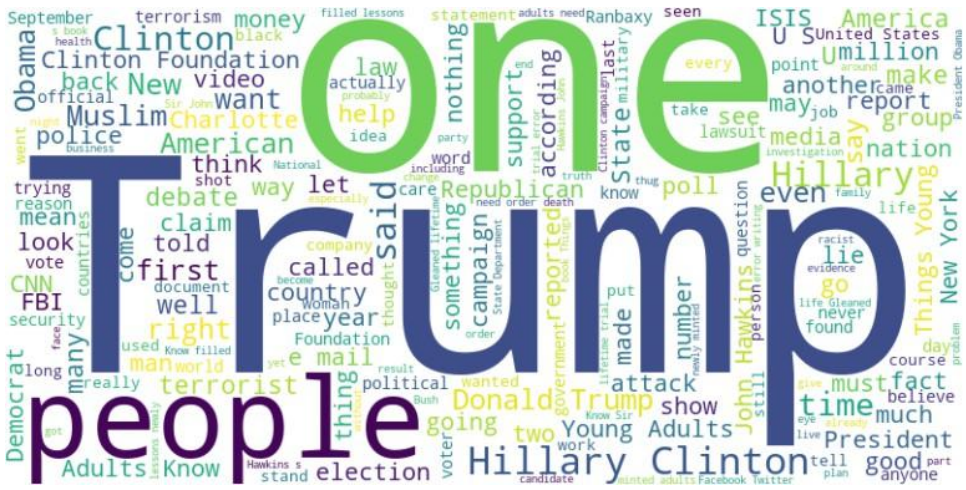
**Figura 1.** *Nube de palabras de las noticias falsas*.




*Fuente:* Elaboración propia (2025)

En las noticias falsas, la palabra “Trump” aparece como la más destacada, indicando que es un tema central en estas noticias. Otras palabras prominentes incluyen “debate”, “president”, “candidate” y “say”.

**Figura 2.** *Nube de palabras de las noticias verdaderas.*




*Fuente:* Elaboración propia (2025)

En contraste, aunque “Trump” también es una palabra destacada en las noticias reales, otras palabras como “people” y “one” son más frecuentes, lo que indica una mayor referencia a personas y hechos individuales.

Las noticias reales muestran una mayor diversidad de temas y figuras mencionadas, tales como “Clinton”, “ISIS”, “Hillary Clinton”, “Obama” y “American”. Por otro lado, las noticias falsas tienden a concentrarse más en temas políticos y específicos de campaña (Horne & Adali, 2017).

Las noticias falsas presentan un mayor uso de palabras relacionadas con declaraciones y afirmaciones, como “said”, “say”, “debate”, “question” y “called”, lo que sugiere un enfoque en citas posiblemente dudosas o controvertidas.

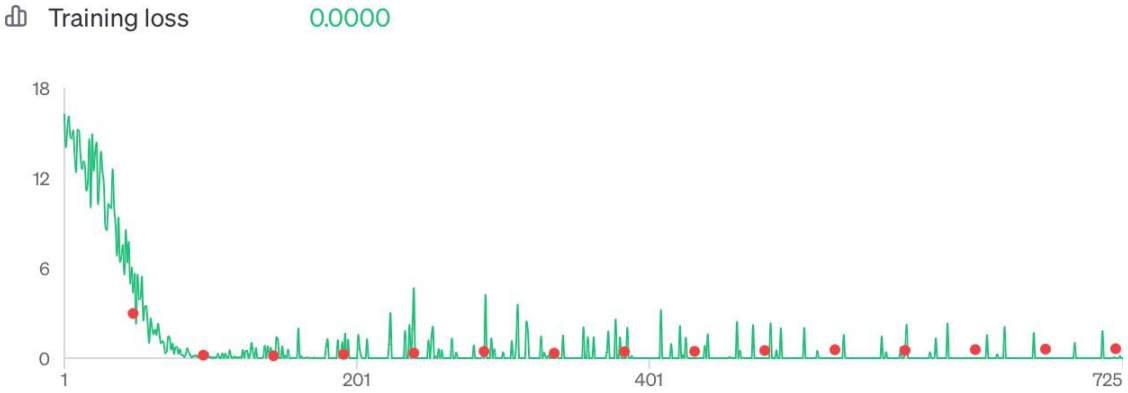
En las noticias reales, palabras como “people” y “one” destacan más, indicando una mayor referencia a personas y hechos individuales.

En las noticias falsas, términos como “Republican”, “campaign” y “candidate” son muy prominentes, sugiriendo una focalización en eventos políticos y narrativas de campaña (Horne & Adali, 2017). En las noticias reales, hay una mayor presencia de términos geopolíticos y de eventos más amplios como “terrorist”, “ISIS” y “police”.

El desempeño del modelo fue evaluado mediante la precisión obtenida durante el entrenamiento. La precisión del 100% sugiere un posible sobreajuste, especialmente dado el tamaño limitado del conjunto de datos.

Para mejorar la generalización y reducir el sobreajuste, es esencial ampliar el conjunto de datos con más ejemplos variados de noticias verdaderas y falsas.

**Figura 3.** *Nube de palabras de las noticias verdaderas*.




*Fuente:* Elaboración propia (2025).

Comparado con otros estudios, como el de Zellers et al. (2019), que utilizaron modelos similares para la clasificación de noticias, los resultados obtenidos son coherentes con la tendencia observada en conjuntos de datos pequeños, donde el sobreajuste es un riesgo significativo.

La implementación de técnicas de regularización y validación cruzada podría mitigar este efecto y proporcionar una evaluación más robusta del modelo.

## Conclusiones

La precisión del 100% obtenida en la clasificación de noticias falsas mediante el fine-tuning de ChatGPT, aunque impresionante, sugiere un riesgo significativo de sobreajuste debido al tamaño limitado del conjunto de datos utilizado. Para asegurar la generalización del modelo y su aplicabilidad en contextos del mundo real, es crucial ampliar el conjunto de datos y aplicar técnicas adicionales de validación y regularización.

El estudio demuestra el potencial de ChatGPT para ser optimizado en tareas específicas de verificación de hechos, lo que representa un avance significativo en el campo del procesamiento de lenguaje natural y la inteligencia artificial. Sin embargo, se requiere una

investigación continua para abordar las limitaciones actuales y mejorar la robustez del modelo (Jiang et al., 2021).

En resumen, aunque los resultados son prometedores, es necesario continuar desarrollando y refinando el modelo para asegurar su efectividad y fiabilidad en la clasificación de noticias falsas, contribuyendo así a la lucha contra la desinformación en la era digital.

**Referencias bibliográficas**

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., … & Zittrain, J. L. (2018). The science of fake news. Science, 359(6380), 1094–1096.

Schuster, T., Shah, D., Yeo, Y., Sun, T., & Barzilay, R. (2019). The limitations of stylometry for detecting machine-generated fake news. Computational Linguistics, 45(1), 1–12.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). Hellaswag: Can a machine really finish your sentence? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4791–4800.

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. SIGKDD Explorations, 19(1), 22–36.

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys, 53(5), 1–40.

OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., & Leike, J. (2022). Chain-of-Thought prompting elicits reasoning in large language models. arXiv:2201.11903.

Horne, B. D., & Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. ICWSM.

Nguyen, T. T., & Nguyen, C. M. (2022). Fine-tuning pre-trained language models for fake news detection: A comparative study. Journal of Information Science, 48(3), 356–370.

Alonso, J., & Fernández, J. (2023). Evaluación de modelos generativos en español para verificación automática de hechos. Revista Iberoamericana de Inteligencia Artificial, 26(72), 45–61.

Thorne, J., & Vlachos, A. (2018). Automated fact checking: Task formulations, methods and future directions. COLING 2018, 3346–3359.

Jiang, Z., Xu, F. F., Araki, J., & Neubig, G. (2021). How can we know what language models know? Transactions of the ACL, 9, 116–131.