

# Configuración y Puesta en Operación de Infraestructura para Analítica de Datos en la Nube

CARLIN GUZMAN, Axel Gael<sup>1</sup>, HERNÁNDEZ CARBAJAL, Alejandro<sup>2</sup>, ISAZA BOHORQUEZ, Cesar Augusto<sup>3</sup>.

<sup>1</sup>*Instituto Tecnológico Superior de Monclova, Carretera No. 57Los 90's C.P 25733 Monclova, Coahuila. México. [122050326@monclova.tecnm.mx](mailto:122050326@monclova.tecnm.mx)*

<sup>2</sup>*Universidad Politécnica de Querétaro, Redes y Telecomunicaciones Carretera Estatal 420 S/N, El Rosario, 76240, Santiago de Querétaro, Qro. [122042758@upq.edu.mx](mailto:122042758@upq.edu.mx)*

<sup>3</sup>*Universidad Politécnica de Querétaro, Redes y Telecomunicaciones Carretera Estatal 420 S/N, El Rosario, 76240, Santiago de Querétaro, Qro. [cesar.isaza@upq.edu.mx](mailto:cesar.isaza@upq.edu.mx)*

*[International Identification of Science - Technology and Innovation](#)*

ID 1<sup>er</sup> Autor: Axel Gael, CARLIN GUZMAN (ORC ID 0009-0008-8031-5550)

ID 1<sup>er</sup> Coautor: Alejandro, HERNÁNDEZ CARBAJAL (ORC ID 0009-0001-1143-5646)

ID 2<sup>do</sup> Coautor: Cesar Augusto, ISAZA BOHÓRQUEZ (ORC ID 0000-0002-0995-6231)

**Resumen —** Este trabajo presenta un sistema independiente para analizar entrevistas usando el modelo Mistral (derivado de TULU), redes neuronales convolucionales (CNN) y embeddings para procesar datos localmente, evaluando métricas como relación pregunta-respuesta, calidad del texto y comprensión. Combina infraestructura física y en la nube para escalabilidad y rendimiento, ofreciendo una herramienta autónoma y adaptable para periodismo e investigación, superando limitaciones de soluciones comerciales.

**Palabras clave —** Procesamiento de Lenguaje Natural, Análisis de Datos, Inteligencia Artificial, Redes Neuronales Convolucionales, Mistral, Sistemas Distribuidos, TULÜ, XGBoost, Boosting, Embeddings, CNN, Grandes Modelos de Lenguaje, LLM, PLN, Python, SSH, Secure Shell, Conexión remota, IA, Algoritmos, Entrevistas.

**Abstract —** This work presents an independent system for analyzing interviews using the Mistral model (a TULU variant), convolutional neural networks (CNNs), and embeddings to process data locally, evaluating metrics such as question-answer ratio, text quality, and comprehension. It combines physical and cloud infrastructure for scalability and performance, offering an autonomous and adaptable tool for journalism and research, surpassing the limitations of commercial solutions.

**Keywords —** Natural Language Processing, Data Analysis, Artificial Intelligence, Convolutional Neural Networks, Mistral, Distributed Systems, TULÜ, XGBoost, Boosting, Embeddings, CNN, Large Language Models (LLM), NLP, Python, SSH, Secure Shell, Remote Connection, AI, Algorithms, Interviews.

## I. INTRODUCCIÓN

El proyecto desarrolla un sistema independiente basado en Mistral para analizar entrevistas en Excel mediante redes neuronales convolucionales y embeddings, evaluando métricas clave como relación pregunta-respuesta, calidad y comprensión. Combina servidores físicos y en la nube en una arquitectura distribuida, gestionada de forma remota.

## II. MARCO TEÓRICO

### A. Antecedentes

Este proyecto integra modelos de lenguaje basados en transformers, como TULÜ y Mistral, con CNNs y embeddings para analizar entrevistas. Estas tecnologías permiten procesar texto de manera eficiente, capturando patrones semánticos y representando palabras en espacios vectoriales. El objetivo es desarrollar una herramienta autónoma y precisa para el procesamiento de datos textuales, respondiendo a la necesidad de soluciones avanzadas en el análisis de lenguaje natural.

### B. Bases Teóricas

El modelo utiliza Mistral para la generación de texto y una CNN para procesar embeddings (GloVe) de entrevistas, extrayendo características semánticas mediante convoluciones. Posteriormente, XGBoost se emplea para clasificar o mejorar el análisis de estas características.

El aprendizaje combina:

1. *Supervisado*: Entrenamiento para mejorar el proceso de evaluación con XGBoost.
2. *No supervisado*: Uso de *embeddings* preentrenados y Mistral para análisis inicial.

El sistema cuantifica:

- **Relación pregunta-respuesta**: Similitud coseno entre *embeddings*.
- **Calidad**: Claridad y riqueza léxica (entropía del texto).
- **Comprensión**: Coherencia con el contexto de la entrevista.

## III. MATERIALES Y MÉTODOS

### A. Materiales utilizados

A continuación, se presenta la lista de los materiales y equipos utilizados en el desarrollo de este proyecto. Cabe destacar que gran parte del equipamiento fue cedido, por lo que no se incluyen los costos de adquisición.

1. Hardware:

- Servidor Hp DL360 G9 Xeon 2630 V3 32GB RAM 1TB
- Servidor Contabo: Cloud VPS 10

2. Software:

- Python v3.12.3
- Visual Studio Code v1.101
- Git v2.50.1
- Ngrok v3.26.0
- Remote SSH (extensión) v0.120.0
- Mistral 7B Instruct v0.2

## B. Metodología de trabajo

El trabajo se desarrolló siguiendo la metodología ágil SCRUM, con una duración total de cinco semanas. Se establecieron sprints semanales, reuniones de seguimiento cada semana y la asignación de tareas por rol mediante un tablero Kanban para facilitar la organización y el control del progreso.

## IV. DESARROLLO

### 4.1 Inventario

Se analizó el equipo donado y sus especificaciones para identificar las herramientas viables, encontrando que los servidores disponen de 32 GB de RAM pero carecen de GPU, lo que limita significativamente la capacidad para trabajar con grandes modelos de inteligencia artificial.



**Fig 2.** Inspección del equipo

### 4.2 Instalación de Ubuntu Server

Tras recopilar el inventario, se eligió la última versión LTS de Ubuntu Server por su soporte prolongado, se creó un usuario, se configuraron reglas de firewall y se actualizaron los paquetes para garantizar el funcionamiento óptimo del sistema.



**Fig 3.** Instalación de Ubuntu server

```
[sudo] password for dremel:  
Status: active  
Logging: on (low)  
Default: deny (incoming), allow (outgoing), disabled (routed)  
New profiles: skip  
  
To          Action    From  
--          -----  
22/tcp      ALLOW IN  Anywhere  
13687/tcp   ALLOW IN  Anywhere  
8000/tcp    ALLOW IN  Anywhere  
22/tcp (v6)  ALLOW IN  Anywhere (v6)  
13687/tcp (v6) ALLOW IN  Anywhere (v6)  
8000/tcp (v6) ALLOW IN  Anywhere (v6)
```

**Fig 4.** Configuración de reglas del firewall

### 4.3 Instalación de mistral

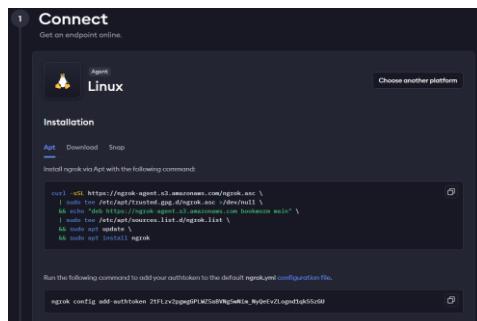
Debido a las limitaciones de hardware, TULÜ3 no fue una opción y se optó por Mistral, un modelo basado en TULÜ2 en versión cuantizada que opera a nivel de bits en lugar de números flotantes, lo que plantea un reto para lograr un sistema escalable y mantenable. Su instalación se realizó desde el repositorio oficial de Hugging Face utilizando GIT como sistema de control de versiones.

```
dremel@ProLiant-DL360-Gen9:~$ git clone https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GGUF
Cloning into 'Mistral-7B-Instruct-v0.2-GGUF'...
remote: Enumerating objects: 57, done.
remote: Total 57 (delta 0), reused 0 (delta 0), pack-reused 57 (from 1)
Unpacking objects: 100% (57/57), 14.33 KiB | 862.00 KiB/s, done.
```

**Fig 5.** Instalación de Mistral

#### 4.4 Instalación de NGROK

Para habilitar la colaboración remota directa con el servidor sin necesidad de un dominio ni estar en la misma red, se instaló NGROK para crear túneles SSH, siguiendo las instrucciones oficiales para su implementación en Ubuntu Server.



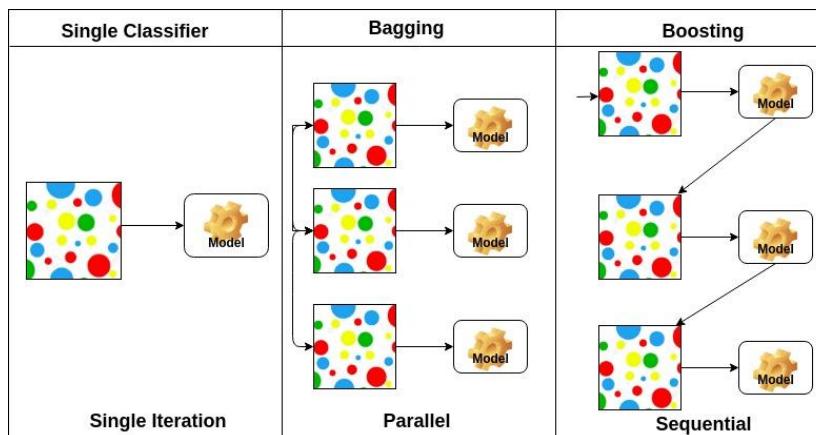
**Fig 6.** Página oficial de NGROK



**Fig 7.** Instalación en el servidor

#### 4.5 Desarrollo del algoritmo XGBoost

Dado que la versión cuantizada de Mistral limitaba la meta de desarrollar un modelo escalable y de crecimiento exponencial, se optó por implementar un algoritmo de aprendizaje capaz de analizar y evaluar datos de entrevistas previas.



**Fig 8.** Representación del método de aprendizaje

##### Esquema de **Boosting**:

- Entrenamiento **secuencial** de modelos.
- Cada nuevo modelo intenta corregir los errores del anterior.
- La combinación final pondera los modelos para mejorar el rendimiento.

## IV. Resultados

Después de semanas de investigación y desarrollo, logramos crear el sistema esperado, trabajando en todo momento sobre los servidores proporcionados. Este sistema es capaz de aprender de cada iteración y adaptarse a las limitaciones que enfrentamos durante el desarrollo del proyecto.

Si bien Mistral continúa utilizándose a modo de ejemplo y con datos simples, el verdadero peso recae en el algoritmo XGBoost y en su capacidad para procesar embeddings y redes neuronales convolucionales, logrando así una mayor comprensión semántica de los datos a evaluar.

## V. Análisis de resultados

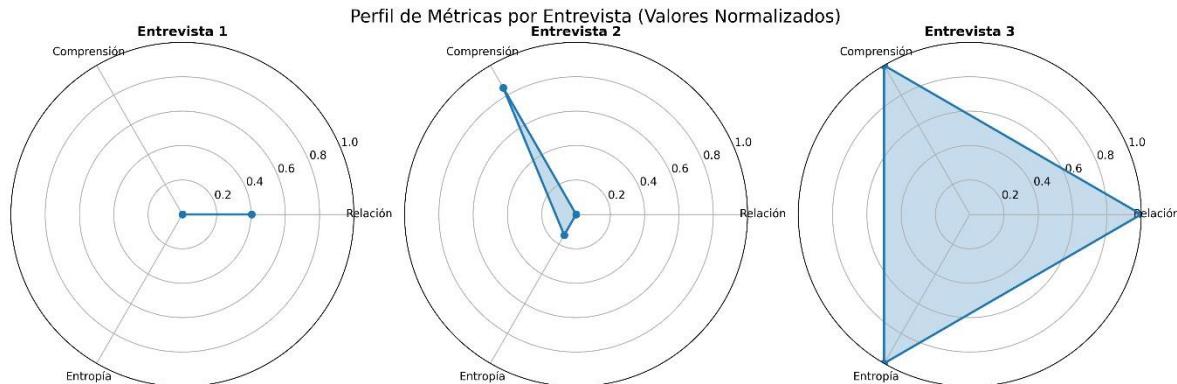
### 5.1 Conclusiones

Este proyecto representó un desafío importante, ya que implicó avanzar en áreas que inicialmente no dominábamos. Sin embargo, a lo largo de su desarrollo adquirimos nuevos conocimientos tanto en el análisis de datos mediante modelos como en el manejo y comprensión de los servidores físicos disponibles en la Universidad Politécnica de Querétaro (UPQ).

### 5.2 Reconocimientos

Parte de estos aprendizajes fueron posibles gracias al apoyo del Dr. César Augusto Isaza Bohórquez, quien nos brindó orientación en la implementación y retroalimentación del uso de modelos en servidores, así como en el manejo adecuado de estos equipos.

## VI. Referencias



- **Ejes:** Tres ejes (uno por métrica: Relación, Compresión, Entropía) que parten del centro y van hasta 1 (valor máximo normalizado).
- **Polígono:** Cada entrevista forma un polígono que conecta los valores de las métricas. Un polígono grande y equilibrado indica una entrevista fuerte en todo.
- **Múltiples subgráficas:** Si hay varias entrevistas, cada una tiene su propio radar chart en una cuadrícula (máximo 3 por fila para no saturar).
- **Normalización:** Las métricas se escalan (0 a 1) para que sean comparables, porque Longitud tiene valores mucho más grandes que las demás.

## REFERENCIAS

- [1] C. Gao, H. Lin, S. Huang, X. Huang, X. Han, J. Feng, C. Deng y J. Chen, "Understanding LLMs' Cross-Lingual Context Retrieval: How Good It Is And Where It Comes From," 2025. Disponible en: <https://arxiv.org/pdf/2504.10906v1.pdf>.
- [2] N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, F. Brahman, L. J. V. Miranda, A. Liu, N. Dziri, X. Lyu, Y. Gu, S. Malik, V. Graf, J. D. Hwang, J. Yang, R. Le Bras, O. Tafjord, C. Wilhelm, L. Soldaini, N. A. Smith, Y. Wang, P. Dasigi y H. Hajishirzi, "Tülu 3: Pushing Frontiers in Open Language Model Post-Training," 2025. Disponible en: <https://arxiv.org/pdf/2411.15124v5.pdf>.
- [3] Equipo de GeeksforGeeks, "XGBoost en aprendizaje automático: Fundamentos y aplicaciones," GeeksforGeeks, 2025. Disponible en: <https://www.geeksforgeeks.org/machine-learning/xgboost/>.
- [4] J. R. F. Junior, "XGBoost: Una introducción a su uso en aprendizaje automático," LinkedIn Pulse, 2025. Disponible en: <https://pt.linkedin.com/pulse/xgboost-jose-r-f-junior>.
- [5] J. B. Mendoza, "Tutorial de XGBoost en Python: Implementación práctica," Medium, 2025. Disponible en: <https://medium.com/@jboscomendoza/tutorial-xgboost-en-python-53e48fc58f73>.
- [6] TheBloke, "Mistral 7B v0.1-AWQ: Modelos cuantizados para inferencia eficiente en aprendizaje automático," Hugging Face, 2025. Disponible en: <https://huggingface.co/TheBloke/Mistral-7B-v0.1-AWQ>.
- [7] Software de ngrok, "ngrok: Fundamentos de túneles seguros para aplicaciones web," ngrok, 2025. Disponible en: <https://ngrok.com/docs>.
- [8] Equipo de Contabo, "Cloud VPS 10: Especificaciones y rendimiento de servidores virtuales de bajo costo," Contabo, 2025. Disponible en: <https://contabo.com/en/vps/cloud-vps-10>.
- [9] Equipo de Servidor, "Servidor HP ProLiant DL360 G9: Configuración con procesador Xeon E5-2630 V3, 32 GB RAM y 2 TB de almacenamiento," HPE 2025. Disponible en: <https://www.hpe.com/us/en/product-catalog/servers/proliant-servers/pip.hpe-proliant-dl360-gen9-server.755258.html>.