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

ISAZA BOHORQUEZ, Cesar Augusto, CARLIN GUZMAN, Axel Gael, HERNÁNDEZ CARBAJAL, ALEJANDRO.

Universidad Politécnica de Querétaro, Redes y Telecomunicaciones Carretera Estatal 420 S/N, El Rosario, 76240, 76240 Santiago de Querétaro, Qro. cesar.isaza@upq.edu.mx

International Identification of Science - Technology and Innovation

Cesar Augusto, ISAZA BOHÓRQUEZ (ORC ID 0000-0002-0995-6231)
Axel Gael, CARLIN GUZMAN
Alejandro, HERNÁNDEZ CARBAJAL

Resumen — En este trabajo se presenta la configuración e implementación de una variante del modelo de TULU (Procesador de lenguaje natural) que permitirá realizar consultas hacer de información almacenada de forma local para así consultar datos de interés.

Palabras clave — Procesador de Lenguaje natural.

Abstract — This work presents the configuration and implementation of a variant of the TULU model (Natural Language Processor), which enables querying locally stored information in order to retrieve relevant data.

Keywords — Natural Language Processor

I. INTRODUCCIÓN

El auge de la inteligencia artificial (IA) ha revolucionado el análisis de datos en campos como el periodismo y la investigación científica, gracias a los modelos de lenguaje de gran escala (LLM). Sin embargo, soluciones comerciales como ChatGPT o Grok están limitadas por suscripciones de pago y falta de personalización para tareas específicas, como el análisis de entrevistas. Este proyecto propone un sistema independiente basado en Mistral (derivado de TULU), utilizando redes neuronales convolucionales (CNN) y embeddings para procesar entrevistas almacenadas en un archivo Excel. El objetivo es cuantificar métricas como la relación pregunta-respuesta, calidad y comprensión, generando promedios y tendencias útiles para el periodismo y la ciencia. Este enfoque autónomo busca superar las restricciones de modelos comerciales, ofreciendo una herramienta escalable y adaptada a necesidades específicas gracias a una propia arquitectura basada en la nube y sistemas distribuidos.

II. MARCO TEÓRICO

1. Antecedentes

Los modelos de lenguaje como TULU y Mistral, basados en transformers, han avanzado el procesamiento del lenguaje natural (PLN), destacando por su capacidad para entender y generar texto. Mistral, derivado de TULU, es ideal para aplicaciones personalizadas debido a su eficiencia y flexibilidad. Las redes neuronales convolucionales (CNN) son efectivas para capturar patrones semánticos en texto, mientras que los embeddings representan palabras en espacios vectoriales. Este proyecto combina estas tecnologías para analizar entrevistas, motivado por la necesidad de herramientas independientes que procesen datos textuales de manera precisa y autónoma.

2. **Bases Teóricas**

El modelo utiliza Mistral para generación de texto y una CNN para procesar embeddings de entrevistas. La CNN toma embeddings (GloVe o generados desde cero) y extrae características semánticas mediante convoluciones.

El aprendizaje combina:

- **Supervisado**: Entrenamiento para mejorar similitudes semánticas.
- **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.

3. **Marco Conceptual**

La metodología es mixta:

- **Cuantitativa**: Procesamiento de entrevistas (Excel) con CNN y embeddings.
- **Cualitativa**: Interpretación contextual para validar resultados.

[Espacio para la infraestructura de red]

III. MATERIALES Y MÉTODOS

La lista de materiales que se ocuparon para la elaboración de este proyecto fue:

Material	Imagen
Laptop Gamer Acer Nitro 5 Core I5 8g Ram 512ssd Gtx1650 W 11	
Servidor Hp Dl360 G9 Xeon 2630 V3 Ram 32gb 2 Dd 1tb Rack	

Servidor Contabo: Cloud VPS 10	

Gran parte de los materiales ya se tenían en posesión por eso no se adjuntaron precios.

IV desarrollo (4.1 instalacion del SO, 4.2 configuracion de ssh, 4.3 instalacion de mistral)

IV. Resultados

V. Analisis de resultados
VI. Conclusiones
VII. Reconocimientos
VIII. Referencias

A. Tipos y tamaño de letra

Tod el documento es redaactado usando letra Arial. Las diferentes partes del documento tendrán tamaño de letra de acuerdo con lo indicado en la Tabla 1, mostrada más adelante.

B. Formato general del documento

- El documento deberá tener una extensión de 4 páginas a 6 páginas, tamaño carta. El archivo deberá ser en .doc (MS Word) (no se aceptarán .pdf), con un máximo de 6 Mb.
- El documento debe redactarse a una columna, con márgenes izquierdo y derecho de 3 cm, y superior e inferior de 2 cm y 2.5 cm, respectivamente. Sólamente los márgenes izquierdo y derecho del Resumen y Abstract serán de 3.75 cm.
- El interlineado será de 1.15, con un espacio al término y al inicio de las secciones o subsecciones. Sólo el texto del Resumen y Abstract tendrá interlineado sencillo.
- Todos los párrafos deben tener sangría de 0.5 cm en el primer renglón, y con justificación hacia la izquierda. Debe haber una separación de 6 pt entre párrafos.
- Los títulos de las secciones, subsecciones y sub-subsecciones no deben tener sangría.
- **No deben numerarse las páginas.**

C. Título, autores e instituciones de adscripción

El título del documento, autores (estudiante e investigador), institución de adscripción y correos electrónicos deben estar centrados, en la parte superior de la primera página. Los nombres de los autores no deben mostrar ningún título profesional como PhD, MSc, Dr., etc.

D. Secciones

El reporte debe contener mínimamente los nombres de las siguientes secciones, escritas en mayúsculas y en negritas, en el siguiente orden, y con numeración en números romanos.

I. INTRODUCCIÓN
II. MARCO TEÓRICO (OPCIONAL)
III. MATERIALES Y MÉTODO
IV. RESULTADOS
V. ANÁLISIS DE RESULTADOS (O ANÁLISIS DE RESULTADOS)
VI. CONCLUSIONES
VII. RECONOCIMIENTOS (O AGRADECIMIENTOS)
VIII. REFERENCIAS.

Pueden intercalarse más secciones, dependiendo del trabajo en particular y del area de conocimiento, de acuerdo con el criterio de los autores. Cada sección puede tener una o varias subsecciones.

E. Subsecciones y sub-subsecciones

Cada sección deberá dividirse como máximo en 2 niveles:

Subsección: Numerada usando letras mayúsculas en orden consecutivo, seguidas por un punto y alineada a la izquierda, sin sangría. El tamaño de letra es de 10 puntos y en itálicas.

Sub-subsección: Numerada usando números seguidos por un paréntesis y alineados a la izquierda. El tamaño de letra es de 10 puntos y en itálicas, con sangría de 0.5 cm.

1) Ejemplo de sub-subsección

Nunca debe colocarse un punto final en los título secciones, subsecciones o sub-subsecciones.

F. Tablas y figuras

Las tablas y figuras deben llevar numeración arábiga en negritas, ejemplos **Tabla 1, Fig. 1**, de acuerdo con su orden de aparición y, al igual que las ecuaciones, se hará referencia a ellas en el párrafo más cercano a las mismas.

Las tablas deben estar centradas horizontalmente en la página. En la parte superior de cada tabla, debe indicarse el número y nombre de la tabla, justificados al centro, como se indica en la tabla 1. El tipo y tamaño de letra de los datos de las tablas debe ser Arial 10 pt.

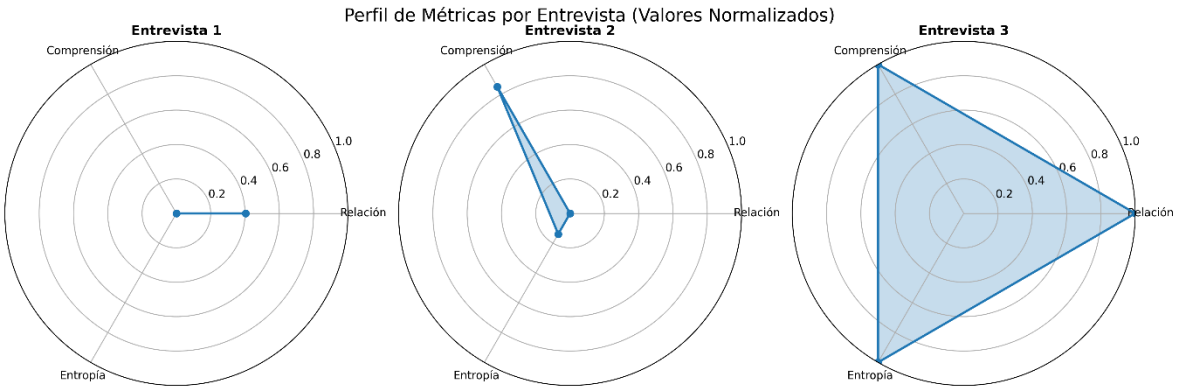
Las figuras deben estar centradas horizontalmente en la página. El tamaño de la figura debe ser el adecuado, a criterio de los autores, de modo que no aparezca demasiado pequeña o demasiado grande. Debe indicarse el número y nombre de la figura en la parte

inferior de la figura, justificados al centro, como se indica en la figura 1. La figura 2 muestra un conjunto de subfiguras.

Las tablas y figuras deben estar colocadas en la parte superior o inferior de la página. Nunca deben estar colocadas entre párrafos.

Tabla 1. Tipos de letra, justificación y tamaño

Letra	Tamaño	Letra	Tipo	Justificación
Titulo	20 pt	Arial	Negrita	Centrada
Nombre de autores	10 pt	Arial	Negrita	Centrada
Datos de institución y correos	9 pt	Arial	Normal	Centrada
Resumen y abstract	10 pt	Arial	Negritas / normal	Justificada
Secciones	11 pt	Arial	Negrita	Justificada
Subsecciones y sub-subsecciones	11 pt	Arial	Itálica	Justificada
Texto de párrafos	11 pt	Arial	Normal	Justificada
Número de ecuación	10 pt	Arial	Normal	Justificada
Número y nombre de figuras	10 pt	Arial	Negritas / Normal	Centrada
Número y nombre de tablas	10 pt	Arial	Negritas / Normal	Centrada
Referencias	10 pt	Arial	Normal / itálicas	Justificada



- **Ejes**: Tres ejes (uno por métrica: Relación, Comprensión, 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.

Fig. 1. Correlación de imagen original [4].

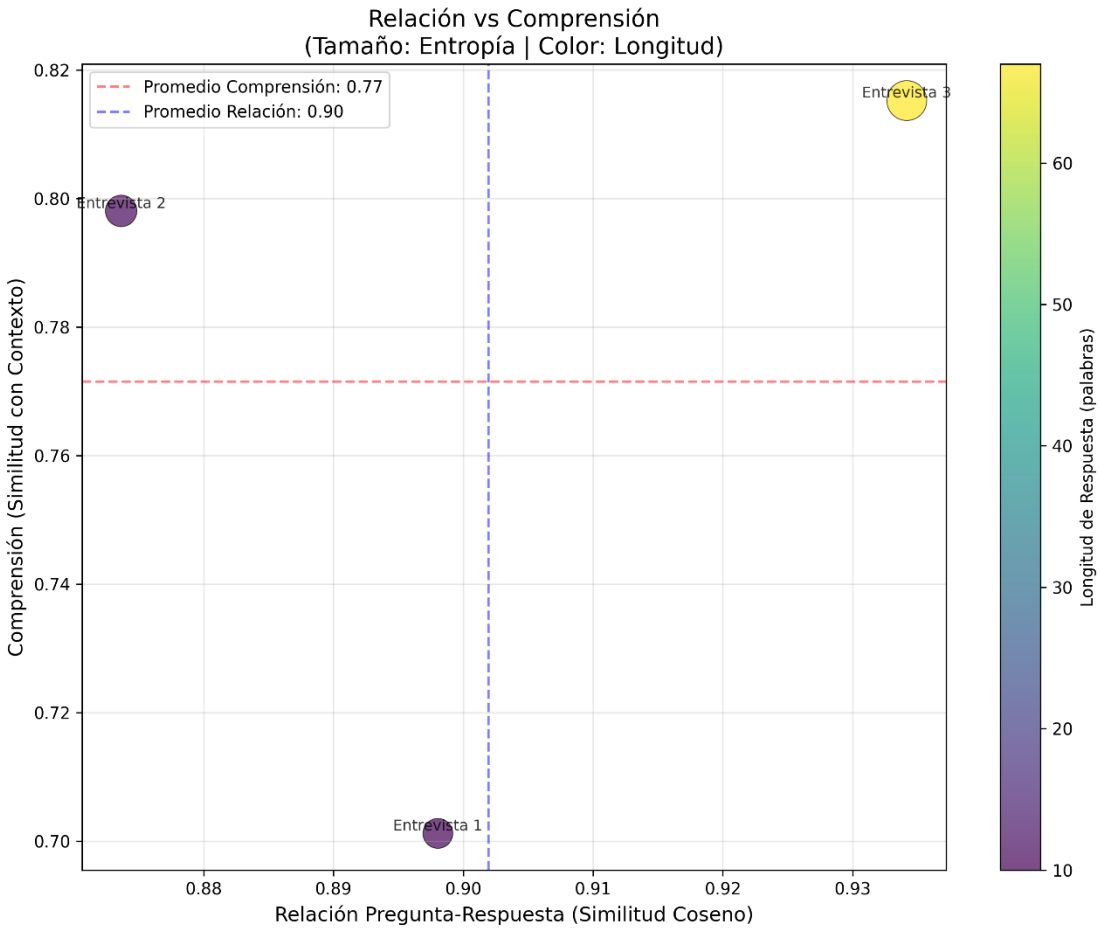


Fig. 2. Correlación de imagen a) original y b) cifrada [4].

Eje X: Relación (similitud coseno entre pregunta y respuesta, de 0 a 1).
Eje Y: Comprensión (similitud coseno entre respuesta y contexto, de 0 a 1).
Tamaño de los puntos: Más grande si la Entropía es alta (respuesta con palabras más variadas), más pequeño si es baja.

G. Ecuaciones

Las ecuaciones deben estar centradas en la página, y numeradas en orden consecutivo en paréntesis normal colocado en el margen derecho (Arial 10 pt). Para escribir la ecuación, se recomienda usar el editor de ecuaciones de MSWord o MathType. Es importante que las variables de la ecuación se definan antes o inmediatamente después de que aparece la ecuación, como se muestra en las Ec. (1) y (2).

$$SimilitudCoseno(A,B) = A \cdot B \parallel A \parallel\parallel B \parallel$$

donde:
A = Area, en m^2
r = radio, en m
π = 3.1416

$$I = V/R \tag{2}$$

donde I es la corriente, en amperes, V es voltaje, en volts, y R es resistencia, en ohms.

IX. RESULTADOS

Los resultados deben ser presentados en una secuencia lógica en el texto, tablas y figuras, evitando la repetición de los mismos datos en diferentes formas. Al describir los resultados la redacción debe ser utilizando en tiempo pasado.

X. DISCUSIÓN (O ANÁLISIS DE RESULTADOS)

La discusión debe considerar los resultados en relación con las hipótesis formuladas en la introducción y el lugar del estudio en el contexto de otros trabajos. Las secciones de Resultados y Discusión (o análisis de resultados) pueden ser combinadas.

XI. CONCLUSIONES Y RECOMENDACIONES

Las conclusiones deben ser claras y concisas entendiendo que no debe repetirse lo indicado en el resumen. Deben expresar el balance final del trabajo desrrollado, comentando sobre los resultados y la relevancia que tiene para el área del conocimiento.

En esta sección se suelen mencionar también los trabajos futuros que se pueden realizar en el tema.

XII. RECONOCIMIENTOS (O AGRADECIMIENTOS)

En esta sección se describen de manera beve los reconocimientos de personas, instituciones, fondos, etc. Los autoresr hacen un reconocimiento / agradecimiento a las personas o instituciones que apoyaron en el desarrollo del trabajo.

REFERENCIAS

[1] *Formato IEEE para presenter artículos*. Disponible en: http://www.unisecmexico.com/archivosPDF/Formato_IEEE.pdf [consultado en junio 2019].

[2] Revista Politécnica, *Formato de artículos*. Disponible en: https://www.google.com.mx/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&ved=2ahUKEwj2pKHT-eriAhUEb60KHXXKD60QFjACegQIBhAC&url=https%3A%2F%2Fwww.politecnicojic.edu.co%2Fimages%2Fdownloads%2Fdocs%2Fformato_articulo_revista_normas_publicacion.doc&usg=AOvVaw0Ev5Nc1-0FCEnx5pil2hLF [consultado en junio 2019].

[3] *Citas y elaboración de bibliografía: el plagio y el uso ético de la información: Estilo IEEE.* Disponible en: https://biblioguias.uam.es/citar/estilo_ieee [consultado en mayo 2019].

[4] S. Ávila Martínez y J. S. Murguía, "Cifrado de imágenes basados en sistemas dinámicos". *Memorias del 20º Verano de la Ciencias de la región Centro, Vol. Ingeniería y Tecnología*, Querétaro, Qro. Junio-agosto 2018.

NOTA SOBRE REFERENCIAS BIBLIOGRÁFICAS

La sección de referencias siempre aparece al final del documento, y el título de la sección no debe tener número de sección. Las referencias bibliográficas son un listado del material bibliográfico consultado para desarrollar el documento y deberán describirse según el orden de aparición en que se van citando al interior del documento (párrafos).

Existen varios estilos internacionales para la elaboración de citas y referencias bibliográficas; no hay un estilo único. Normalmente, cada área del conocimiento utiliza uno en particular. Los reportes del 21º Verano de la Ciencia de la Región Centro podrán utilizar el estilo preferido por los autoes, de acuerdo con el área del conocimiento, pero cualquiera que sea el estilo utilizado deberá seguir las normas de ese mismo estilo.

Para ver la manera en que se utlizan los diferentes estilos se recomienda consultar a la liga http://blogs.ujaen.es/biblio/wp-content/uploads/2015/07/citas-bibliograficas.pdf.

En este documento se utilizó el estilo IEEE de referencias bibliográficas, donde la identificación de las referencias es con números arábigos consecutivos, iniciando con 1, encerrado en paréntesis cuadrados (por ejemplo [1]).

Todas las referencias listadas deben ser citadas al interior del documento (párrafos), iniciando con la referencia [1] y en orden consecutivo hasta citar la última referencia. **Nunca deben incluirse referencias que no estén citadas al interior del documento.**

Si se cita en los párrafos a varias referencias juntas, estas deben estar separadas por comas; por ejemplo [1,3,5]. Si varias referencias consecutivas van a ser citadas, entonces se debe mencionar sólo la primera y la última, separadas por un guión; por ejemplo [1-5].

La descripción de cada referencia tiene una estructura específica. Dependiendo del tipo de documento, la descripción debe ser como lo señala el estilo de referencias utilizado. En el estilo IEEE, los autores deben llevar los nombres abreviados y luego los apellidos. Enseguida se muestran la manera de describir diferentes tipos de referencias, en el estilo IEEE.

Libro:

C. Para, J. Pelzl, *Understanding Cryptography: A Textbook for Students and Practitioners*, Ed. Springer-Verlag. Berlin, 2010.

Artículo de revista indizada:

S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, “A novel ultrathin elevated channel low-temperature poly-Si TFT,” IEEE Electron *Device Lett.*, Vol. 20, pp. 569–571, Nov. 1999.

Artículo de revista indizada bajado de página web:

J. A. Aboytes-González, J. S. Murguía, M. Mejía-Carlos, et al., “Design of a strong S-box based on a matrix approach”, *Nonlinear Dynamics*, Vol. 94, Issue 3, pp 2003–2012. Disponible en https://doi.org/10.1007/s11071-018-4471-z [consultado en 2018].

Artículo de memorias de congreso:

S. Ávila Martínez y J. S. Murguía, “Cifrado de imágenes basados en sistemas dinámicos”. *Memorias del 20º Verano de la Ciencias de la región Centro, Vol. Ingeniería y Tecnología,* Querétaro, Qro. junio-agosto 2018.

Conferencia:

J. C. Garzón, “Más allá de las decisiones económicas”. *II Jornada de Análisis Económico*, La Habana, Cuba, marzo de 2000.

Reporte técnico:

U.S. EPA. *Status of Pesticides in Registration: an Special Review*. EPA 738-R-94-008. Environmental Protection Agency, Washington, DC, 1994.

Manual

Motorola, *FLEXChip Signal Processor (MC68175/D)*, 1996.

Estándar

Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11, 1997.

Tesis:

J. Jacobs, *Regulation of Life History Strategies Within Individuals in Predictable and Unpredictable Environments* [PhD Thesis]. University of Washington, Seattle, WA, 1996.

Patente:

R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, “High-speed digital-to-RF converter,” *U.S. Patent 5668842*, Sept. 16, 1997.