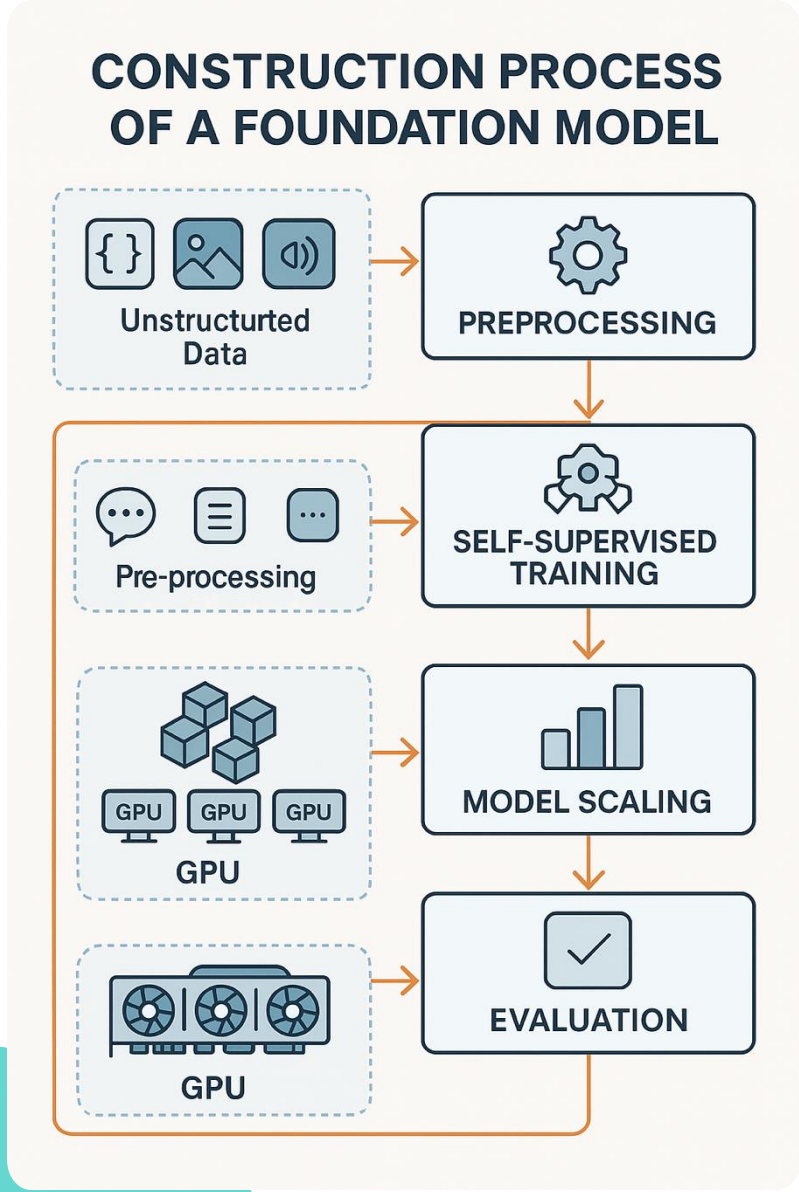
El Uso de los Modelos Fundacionales en los Procesos de Producción de Información Estadística y Geográfica

INEGI

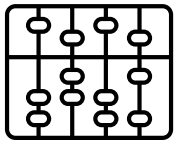
inegi.org.mx



¿Qué son los modelos fundacionales?

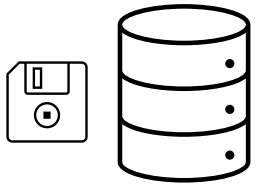
Un **modelo fundacional** es entrenado usando **grandes volúmenes de datos**, generalmente no estructurados y provenientes de diversas fuentes, con el objetivo de capturar patrones amplios y variados del lenguaje, imágenes u otros tipos de datos. Estos modelos **no están diseñados para una tarea específica** desde el inicio, sino que se *fundamentan* como base para múltiples aplicaciones.

Características principales de los modelos fundacionales

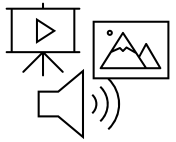


Preentrenamiento generalizado: Se entrenan de manera auto-supervisada en tareas generales (como predecir la siguiente palabra o contrastar texto e imagen) sin necesidad de etiquetado manual.

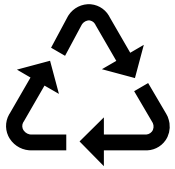
Adaptabilidad: Se pueden ajustar mediante fine-tuning o prompting para tareas específicas como clasificación, traducción, generación de texto, segmentación de imágenes, etc.



Escalabilidad: Poseen cientos de millones a billones de parámetros, lo que les permite capturar representaciones amplias y transferibles, y mejorar su desempeño en la media que crecen.



Multimodalidad (en algunos casos): Pueden procesar y generar texto, imágenes, audio o combinaciones de ellos.



Capacidad de generalización y reutilización: Se diseñan como una base única para múltiples tareas, dominios y contextos con mínimos ajustes..

EJEMPLOS DE MODELOS FUNDACIONALES

Dominio	Modelos principales	Capacidades destacadas
Procesamiento de Lenguaje Natural (NLP)	- GPT (OpenAI) - LLaMA (Meta) - Falcon (TII) - Mistral - BERT / BETO	Generación y comprensión de texto, resumen, traducción, codificación automática
Visión por Computadora (CV)	- SAM (Segment Anything Model, Meta) - DINOv2 - MAE (Masked Autoencoders)	Segmentación, detección de objetos, extracción de características sin supervisión
Multimodales	- CLIP (OpenAI) - PALIGEMMA (Google) - GPT-4o (OpenAI) - Gemini (Google)	Comprensión conjunta de texto e imagen, interacción con múltiples tipos de datos

(Grandes) Modelos de Lenguaje

1. Representación densa de palabras (embeddings):

- Cada palabra, frase o documento es representado como un vector en un espacio de alta dimensión. Estas representaciones no se basan en coincidencias exactas de palabras, sino en su significado contextual aprendido durante el entrenamiento.

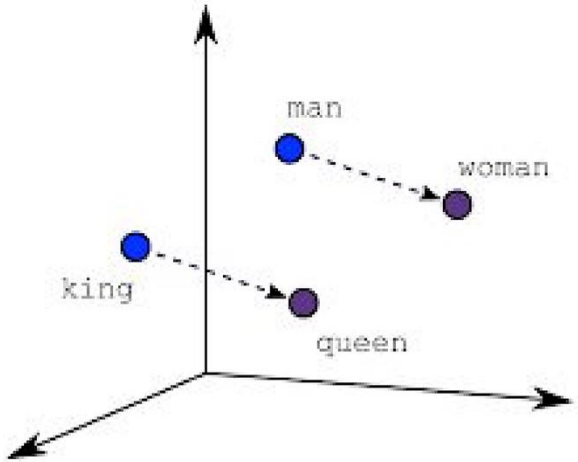
2. Espacio semántico multidimensional:

- Los vectores generados por el modelo habitan un espacio matemático donde la distancia entre vectores refleja la afinidad semántica entre los términos. Es decir:
  - Palabras con significados similares (e.g., doctor y médico) estarán cercanas entre sí.
  - Palabras con relaciones análogas forman estructuras consistentes (e.g., Rey - Hombre + Mujer ≈ Reina).

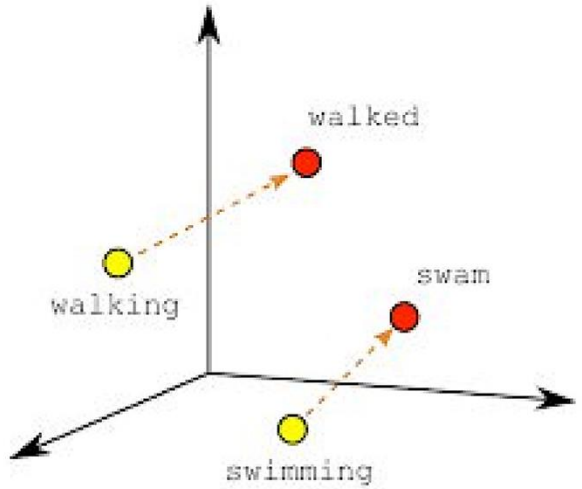
3. Contextualización dinámica:

- A diferencia de modelos tradicionales, los modelos fundacionales de lenguaje generan representaciones que varían según el context.

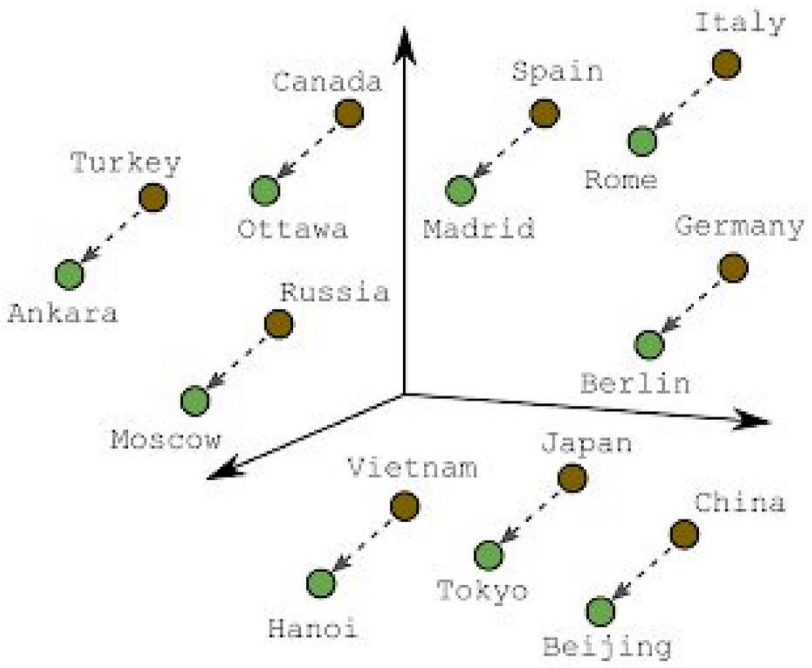
Word Embeddings



Male-Female



Verb Tense



Country-Capital

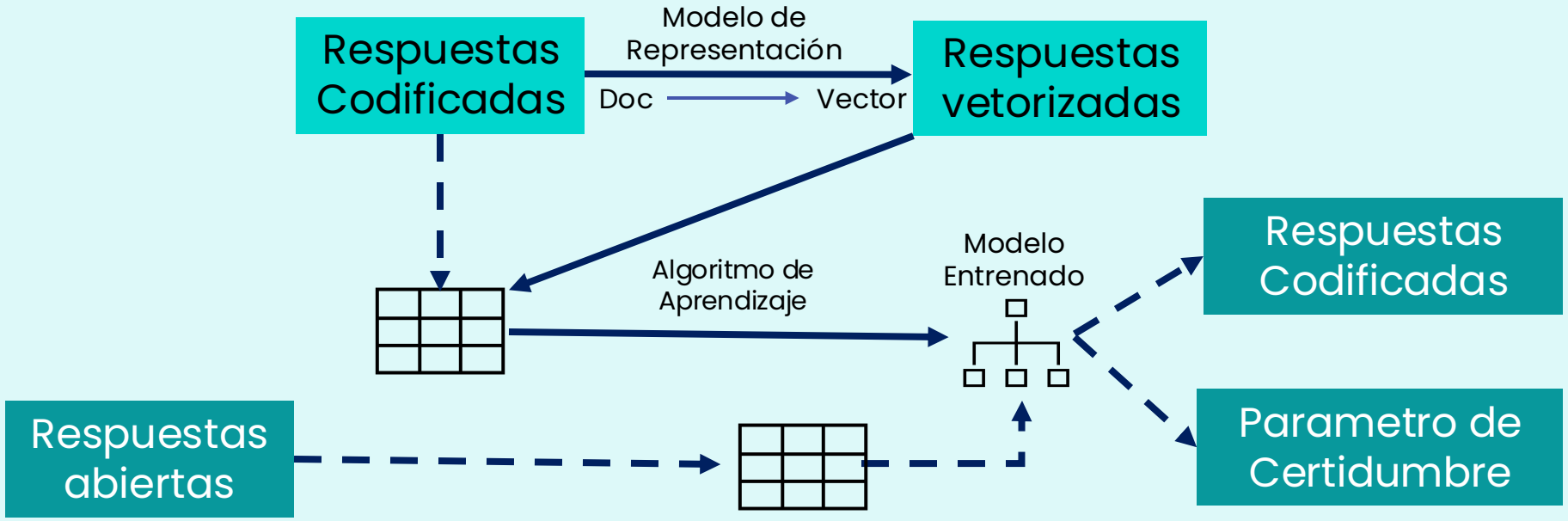
Aplicaciones (experimentales) en el INEGI

- Procesamiento: condificaión automática
- Difusión: generación de documentos informativos.
- Difusión: desarrollo de buscadores semánticos.
- Captación: extracción de información en colecciones de documentos.
- Procesamiento: Identificación de objetos en imágenes satelitales.

Codificación automática

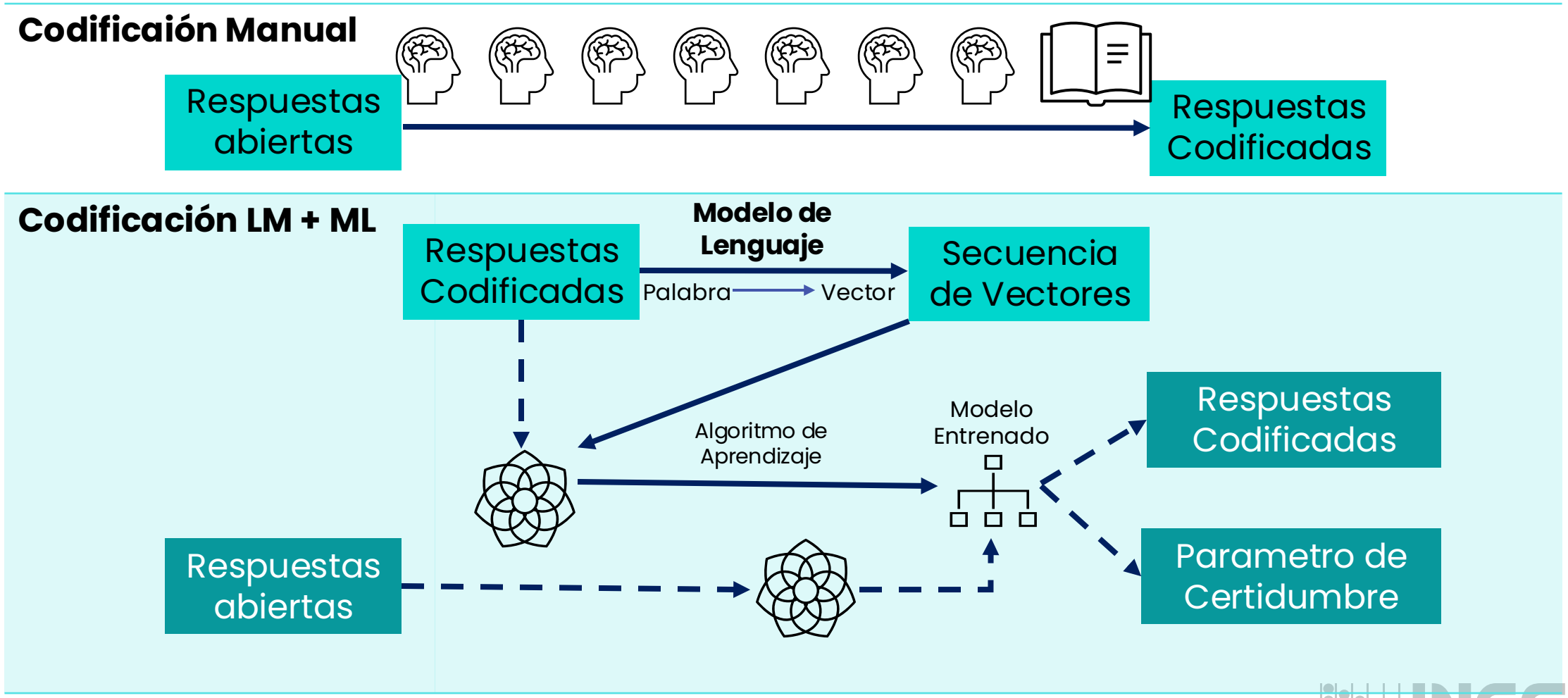
Codificaión Manual

Codificación con ML



inegi.org.mx

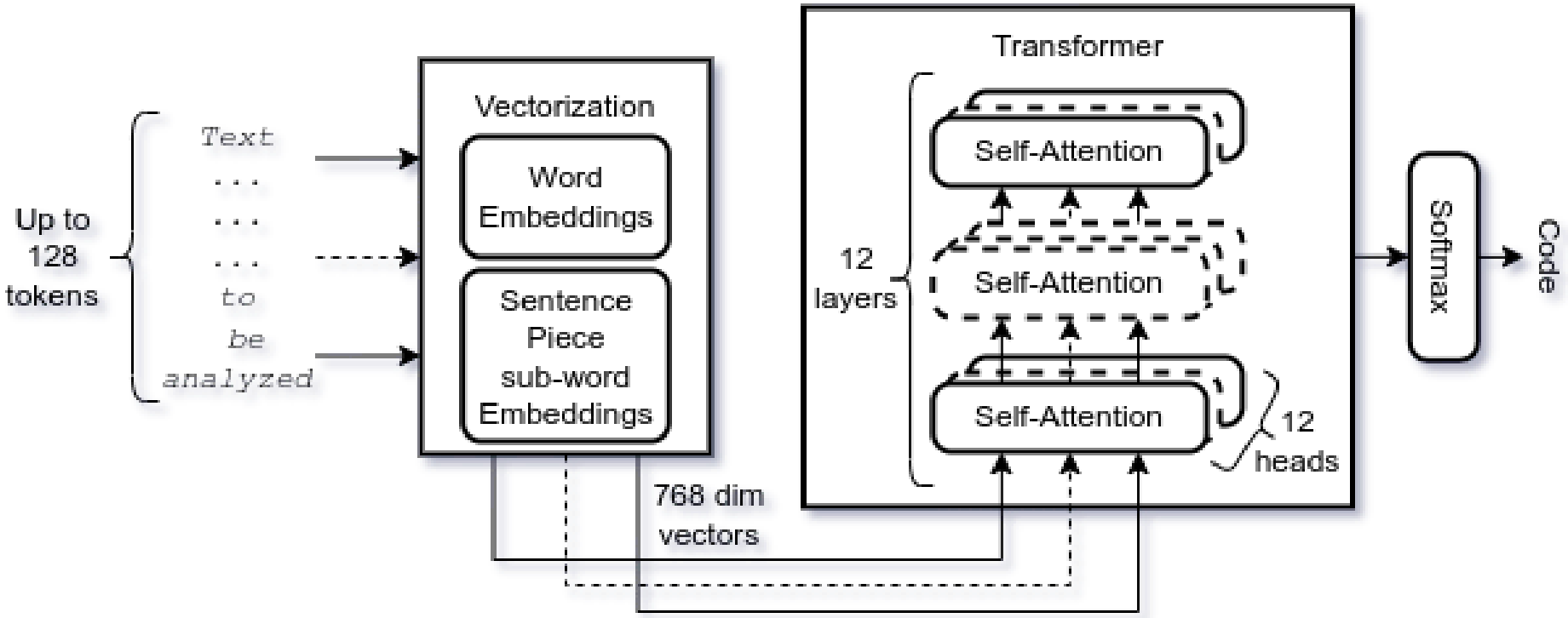
Codificación automática



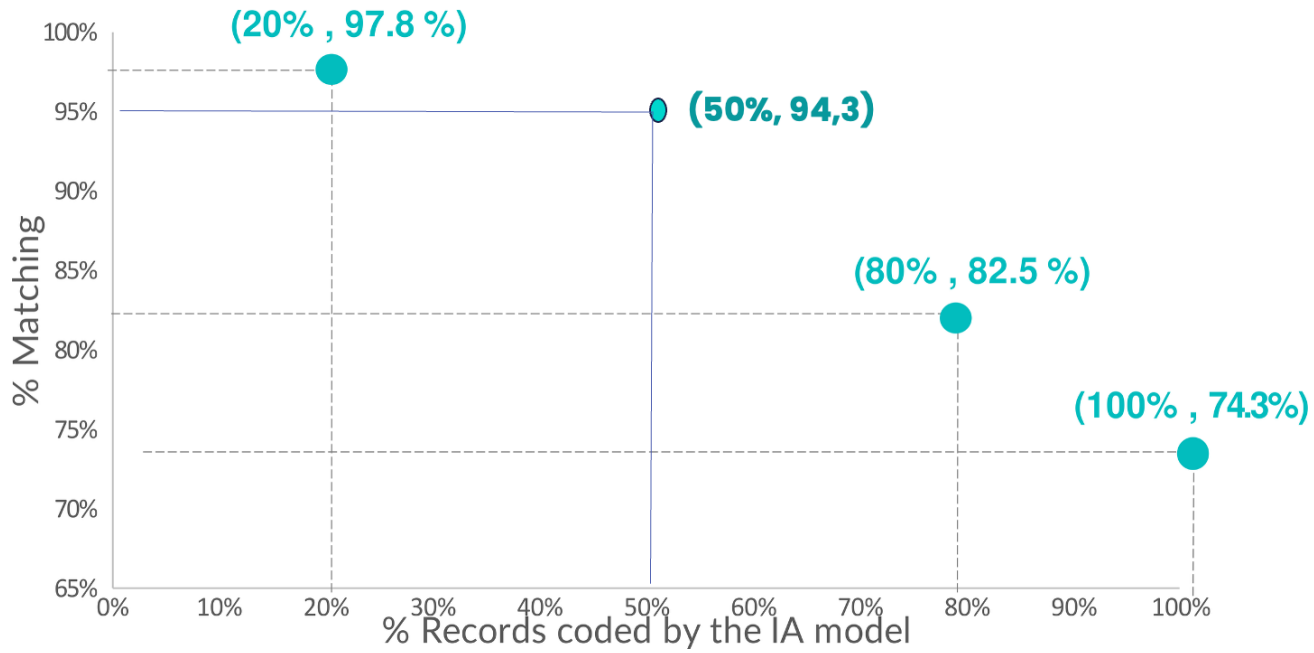
inegi.org.mx

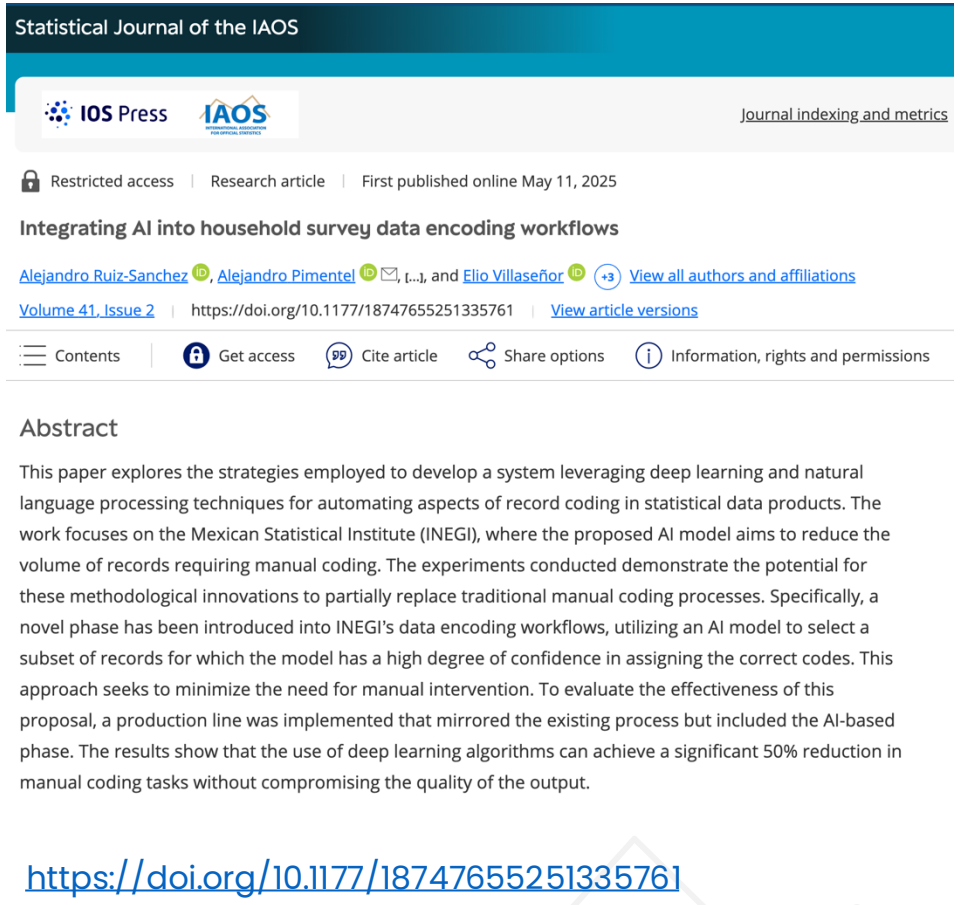
INEGI

Codificación automática



Codificación automática





Statistical Journal of the IAOS

IOS Press IAOS

Journal indexing and metrics

Restricted access | Research article | First published online May 11, 2025

Integrating AI into household survey data encoding workflows

Alejandro Ruiz-Sanchez, Alejandro Pimentel, [...], and Elio Villaseñor +3 View all authors and affiliations

Volume 41, Issue 2 | https://doi.org/10.1177/18747655251335761 | View article versions

Contents | Get access | Cite article | Share options | Information, rights and permissions

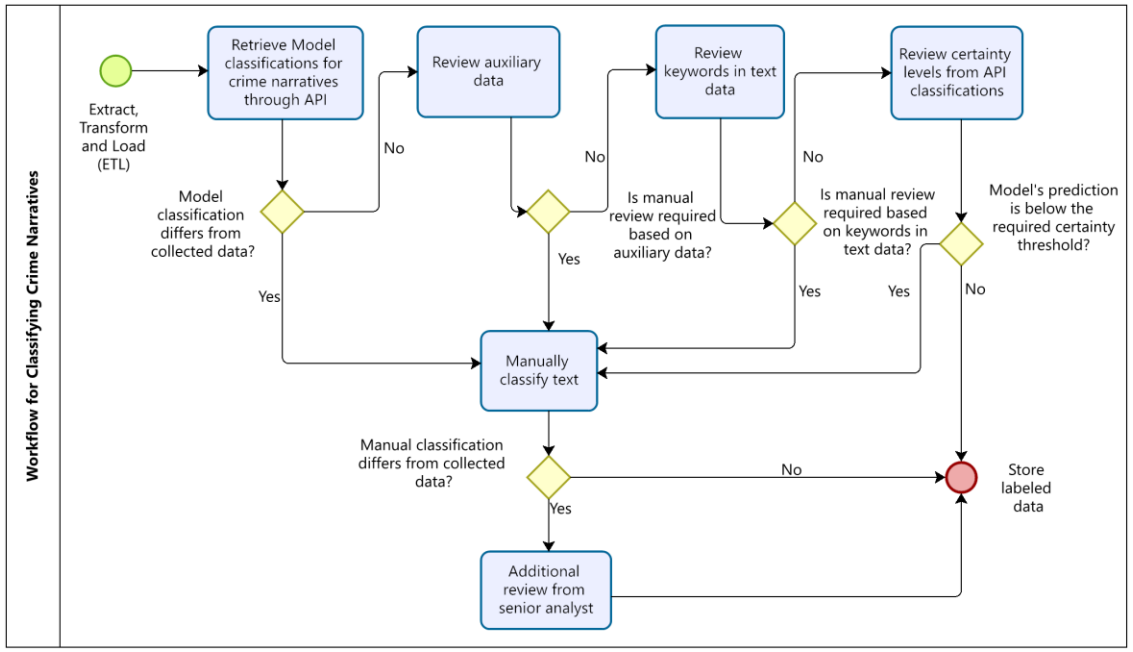
Abstract

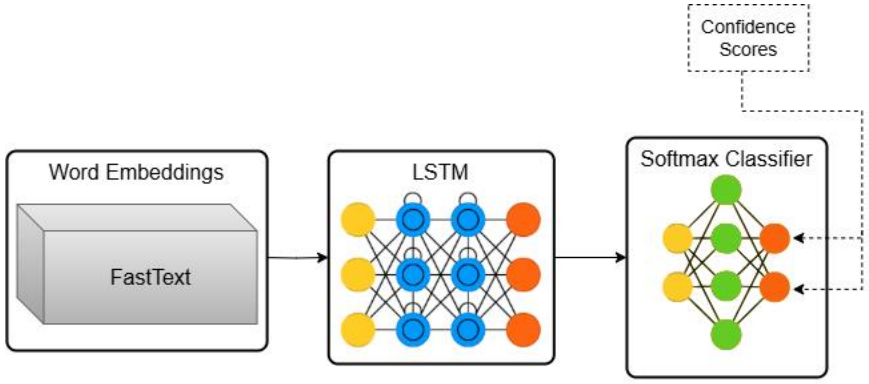
This paper explores the strategies employed to develop a system leveraging deep learning and natural language processing techniques for automating aspects of record coding in statistical data products. The work focuses on the Mexican Statistical Institute (INEGI), where the proposed AI model aims to reduce the volume of records requiring manual coding. The experiments conducted demonstrate the potential for these methodological innovations to partially replace traditional manual coding processes. Specifically, a novel phase has been introduced into INEGI's data encoding workflows, utilizing an AI model to select a subset of records for which the model has a high degree of confidence in assigning the correct codes. This approach seeks to minimize the need for manual intervention. To evaluate the effectiveness of this proposal, a production line was implemented that mirrored the existing process but included the AI-based phase. The results show that the use of deep learning algorithms can achieve a significant 50% reduction in manual coding tasks without compromising the quality of the output.

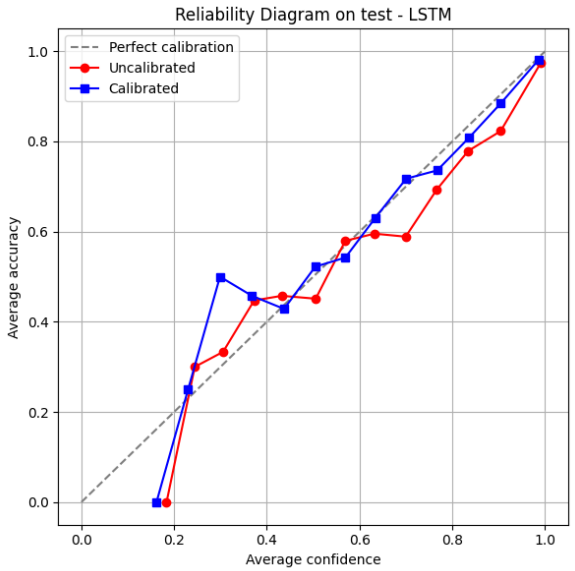
https://doi.org/10.1177/18747655251335761

inegi.org.mx

Automated Classification of Crime Narratives Using Machine Learning and Language Models in Official Statistics



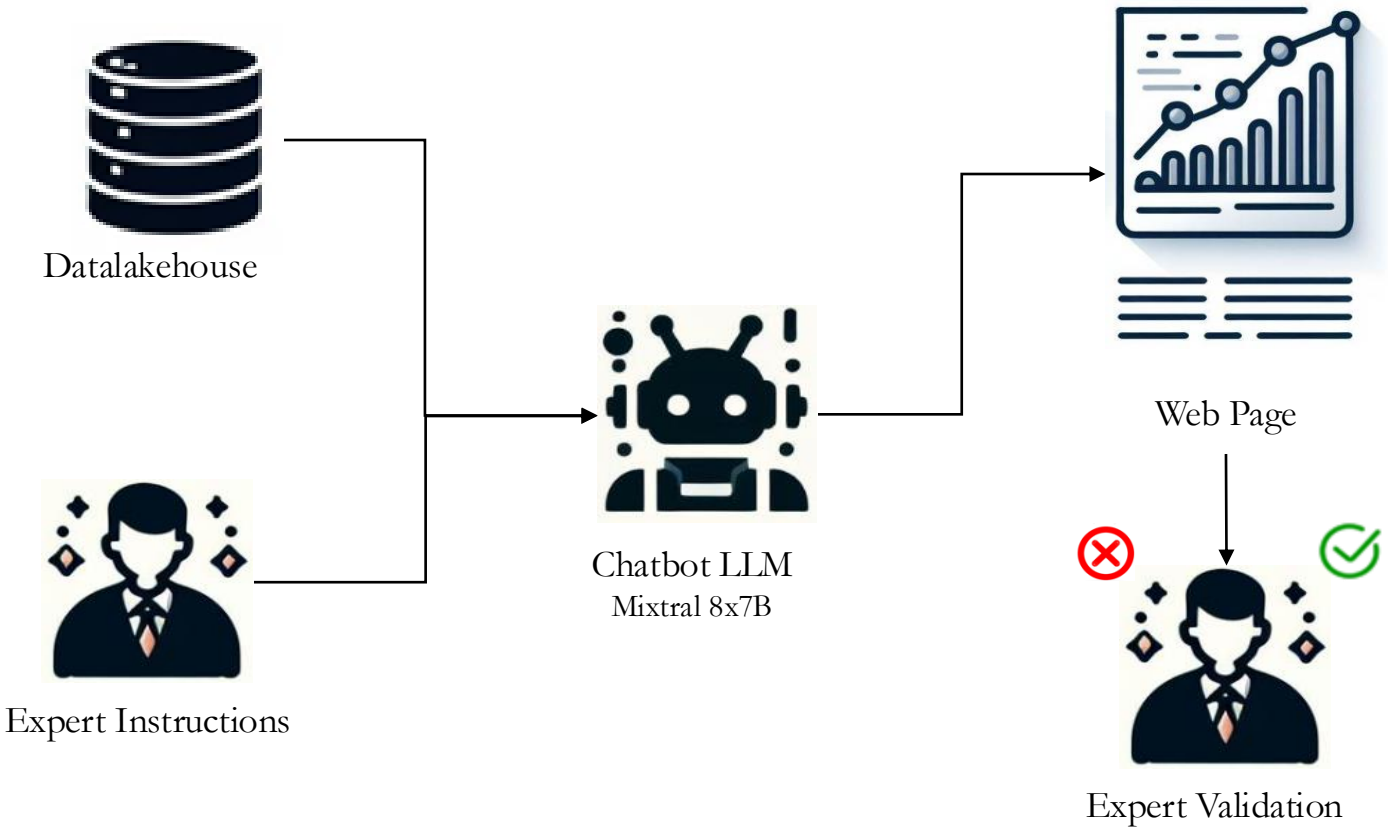




Metric / Model	XGB-BOW	XGB-EMB	LSTM	BETO
Accuracy	0.854	0.872	0.914	0.922
Precision (macro)	0.830	0.821	0.904	0.915
Recall (macro)	0.761	0.773	0.858	0.871
F1-Score (macro)	0.781	0.790	0.860	0.880
Precision (weighted)	0.856	0.871	0.914	0.923
Recall (weighted)	0.854	0.872	0.914	0.922
F1-Score (weighted)	0.852	0.870	0.913	0.922
Eval Loss	0.529	0.412	0.288	0.304
Support	10126	10126	10126	10126

https://gitlab.com/chile-mexico/crime-classifier

Generación de Documentos Informativos



Conociendo la Industria

Conociendo la Industria es una colección de estudios sectoriales y regionales del Instituto Nacional de Estadística y Geografía (INEGI) que brinda un panorama económico de diversos sectores industriales. Estos estudios incluyen información sobre actividades productivas, empleo, salarios, cadenas productivas e inflación. La información se presenta en un lenguaje sencillo y está respaldada por gráficos y tablas. Este trabajo se realiza en colaboración con las respectivas cámaras industriales.

La industria de la vivienda

La industria automotriz

La industria aeroespacial

etc.

Generación de Documentos Informativos

INEGI Piloto Prototipo

English Otros idiomas Contacto +A

Temas | Programas de información | Sistemas de Consulta | Infraestructura | Acerca del INEGI | Buscar... | Enviar

Inicio / Sistema de consultas / Tableros estadísticos / Conociendo la Industria

Industria textil y del vestido

Introducción

La industria textil y del vestido en México es un sector económico importante que ha experimentado un crecimiento significativo en los últimos años. Con una larga tradición en la producción de textiles, México se ha posicionado como el quinto productor mundial de algodón y el décimo en hilados y tejidos. Esta industria emplea directamente a más de 600 mil personas y genera ingresos por más de 25 mil millones de dólares anuales. La industria se caracteriza por su diversidad, abarcando desde la producción de materias primas hasta la confección de prendas de vestir y artículos textiles para el hogar. A pesar de los desafíos globales, como la competencia de países asiáticos y las fluctuaciones en el precio de las materias primas, la industria textil y del vestido en México continúa siendo un pilar fundamental de la economía nacional.

Producto Interno Bruto

94,908	-5.4
Millones de pesos a precios reales	Variación con respecto al mismo trimestre del año anterior

4° trimestre 2023

Puestos de trabajo

519,313	-0.2
puestos de trabajo	Variación porcentual anual

Año 2022

Inflación

3.3%

Prendas de vestir
Variación con respecto al mismo mes del año anterior

Enero 2024

Comercio exterior

Exportaciones

841	0.1%
millones de dólares	Variación con respecto al mismo mes del año anterio

Diciembre 2023

Comercio exterior

Importaciones

1,025	3.2%
millones de dólares	Variación con respecto al mismo mes del año anterio

Diciembre 2023

Índice

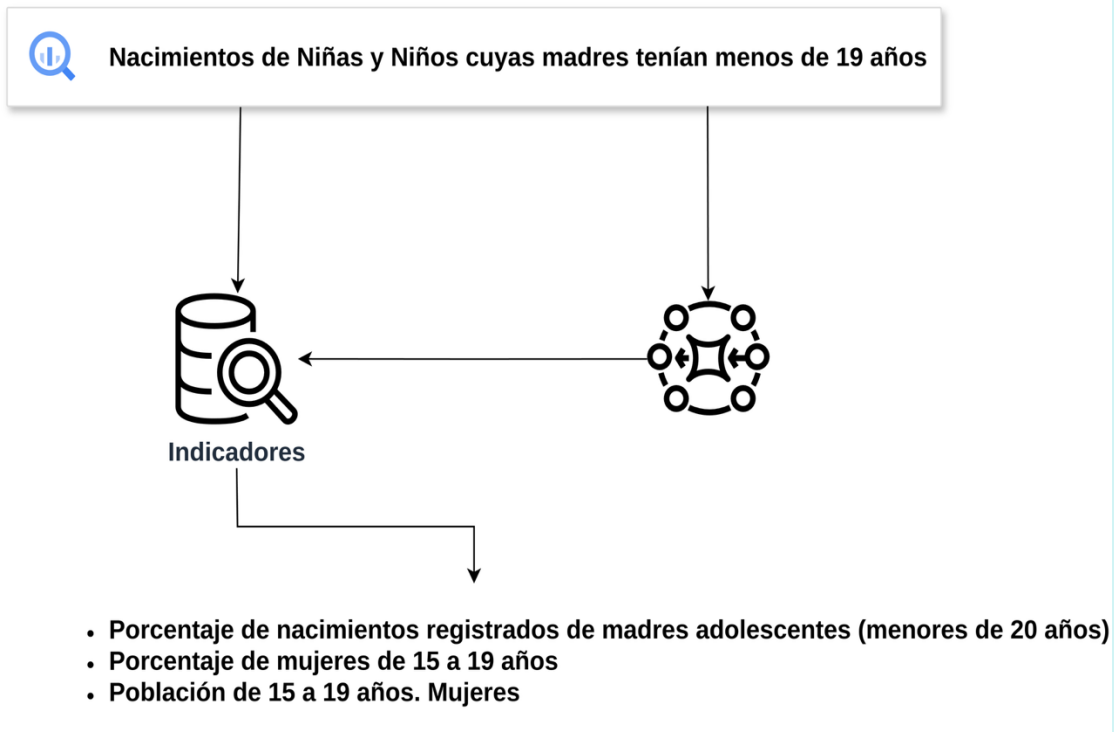
Delimitación
Producción

INEGI

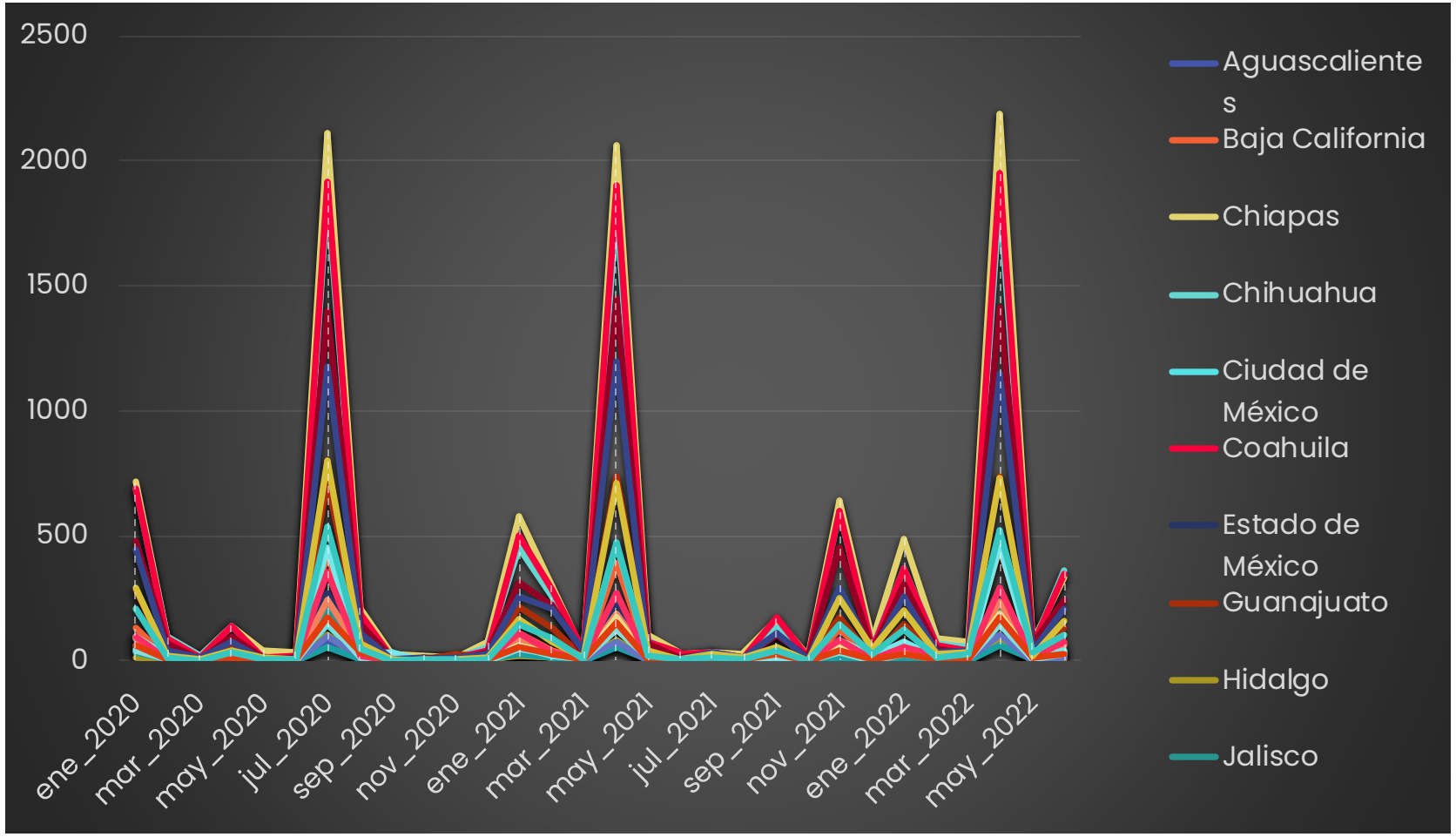
MEJORAS EN EL BUSCADOR DEL SITIO DEL INEGI

Busqueda basada en Palabras Clave

Buscador Semántico



EXTRACCIÓN DE INFORMACIÓN EN COLECCIONES DE DOCUMENTOS



ANÁLISIS CON MODELOS DE LENGUAJE DE GRAN TAMAÑO

Prompt:

*Analize the following text:'+**texto_noticia**+'*

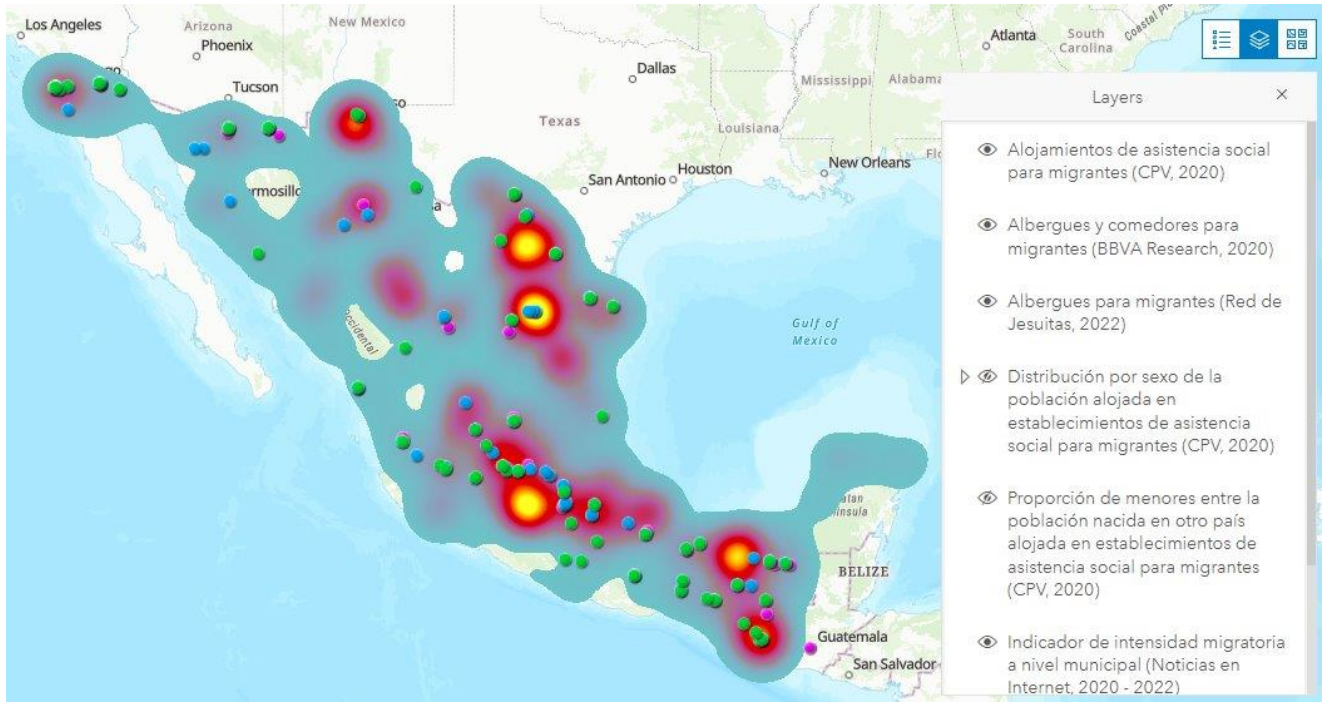
¿Podrías identificar si hay presencia de mujeres o niñas migrantes en el texto? Limite su respuesta a un solo número, por ejemplo, si está presente use 1, si no está presente use 2.

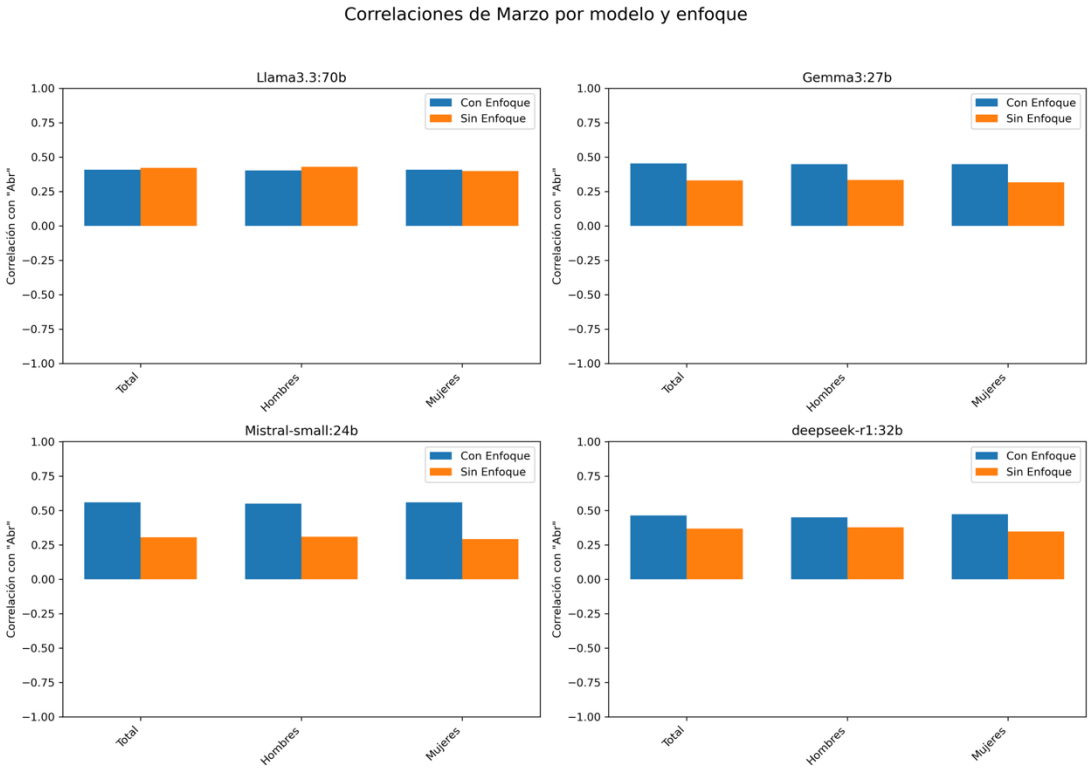
Numero de noticias que mencionan mujeres o niñas migrantes

CVE_GEO	Municipality	Llama3.3:70b		Mixtra:8x7b		Gemma3:27b		Mistral-small:24b		deepseek-r1:32b	
		1	2	1	2	1	2	1	2	1	2
26002	agua-prieta	17	33	21	29	21	29	10	40	13	37
1001	aguascalientes	41	58	39	61	47	53	14	85	22	78
9010	alvaro-obregon	39	56	47	49	47	49	22	72	29	67
30014	amatlan	32	62	23	71	26	68	12	82	12	82
15010	apaxco	2	57	4	55	6	53	1	57	1	58
29005	apizaco	33	66	24	75	33	66	13	86	18	81
7009	arriaga	42	51	49	46	51	44	19	74	25	70

ANÁLISIS CON MODELOS DE LENGUAJE DE GRAN TAMAÑO

Albergues e indicador de intensidad migratoria basado en noticias



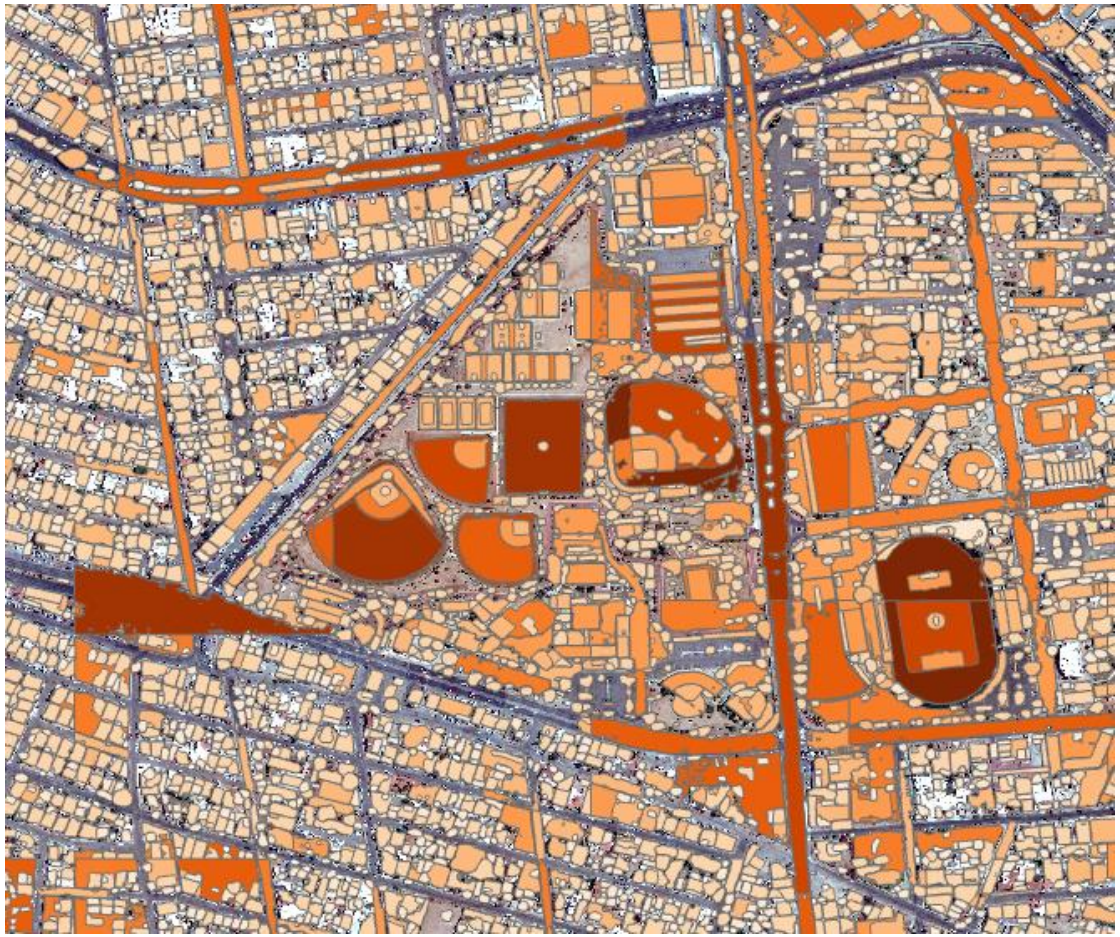


INEGI

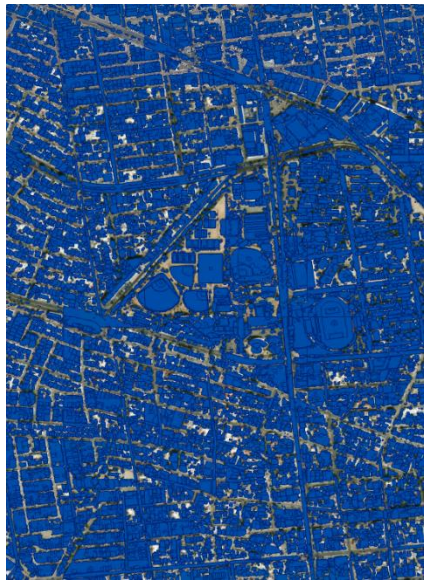
IDENTIFICACIÓN DE OBJETOS EN IMÁGENES SATELITALES

Modelo preentrenado Segment Anything Model (SAM)

SAM de Meta AI es un modelo que segmenta cualquier cosa sin necesidad de entrenamiento adicional, utilizando un escenario de Machine learning, SAM puede identificar y diferenciar objetos, incluso aquellos no vistos anteriormente, siendo útil para extraer máscaras de múltiples objetos en cualquier imagen.



IDENTIFICACIÓN DE OBJETOS EN IMÁGENES SATELITALES

Imagen Satelital HD	Segmentación (SAM)	Clustering	LLM Multimodal (PALIGEMMA)	Reconocimiento
				

inegi.org.mx

PODER COMPUTACIONAL

Cluster and Grid Sandbox-ITo (Pre-production Training Sandboxes – 4 nodes) Processing 160 cores in cpu's, Ram memory 1.5 TB, Storage 16 TB

Cluster and Grid Sandbox (LLM & SML – 4 nodes) Processing 96 cores in cpu´s, Ram memory 2 TB, Storage 48 TB, GPU´s 1,152 Tensor cores (4 NVIDIA RTX 6000 ADA - 48 Gb).

Cluster and Grid HPC (High Performance Computing), Processing 448 cores in CPUs and 4 GPUs [Tensor Core + TeraFlops], 3 TB RAM Memory, 30 TB Storage, 2,560 Tensor Core GPUs.

Grid Storage Raid (Data Lake)

NAS (Network Attached Storage)
50 TB

SAN (Storage Area Network)
20 TB

INEGI

LINEAMIENTOS ESTRATÉGICOS

- Impulsar la innovación en el uso de fuentes alternativas y métodos innovadores, integrándolos en la producción estadística y geoespacial para responder oportunamente a las crecientes demandas de información.
- Acelerar la interoperabilidad de datos y sistemas para facilitar un uso intensivo y extensivo de la información estadística y geográfica, asegurando su disponibilidad y reutilización por parte de humanos y sistemas de inteligencia artificial.
- Modernizar de manera prioritaria la infraestructura tecnológica y de datos, dotándola de capacidades para procesar e integrar eficientemente fuentes diversas, y para aplicar métodos basados en inteligencia artificial de forma escalable y sostenible.
- Fortalecer la colaboración regional e internacional para ampliar y validar el uso de fuentes de datos alternativas, promoviendo el intercambio de experiencias, metodologías y buenas prácticas, bajo principios compartidos de ética y calidad.

¡Gracias!

CONOCIENDO MÉXICO

800 111 46 34
www.inegi.org.mx
atencion.usuarios@inegi.org.mx

INEGIINFORMA