

Decoding Economic Insights: The Analytical Power of News Content

Descodificando Perspectivas Económicas: El Poder Analítico de los Contenidos de Noticias

Pierre A. Delice

Ingeniero de Inteligencia Artificial en
Northware

CIBERTIC 2025

Guadalajara, México

22 Mayo 2025

Contenido

1. Antecedentes
2. Planteamiento
3. Objetivos
4. Propuesta metodológica
5. Resultados
6. Conclusión

Antecedentes

- El propósito de buscar inferir los indicadores económicos mediante textos escritos en lenguaje natural tiene que ver con el interés de usar los avances del Procesamiento de Lenguaje Natural (PLN) para la generación de fuentes de información alternativas para el análisis económico.
- Las noticias abundan en información crucial sobre eventos inesperados, cambios en políticas, desarrollos tecnológicos y variaciones en el sentimiento público, factores que los datos numéricos por sí solos pueden no captar.
- Integrar noticias en las tareas de estimación económica enriquece los análisis ya que refleja mejor las complejidades del comportamiento humano y los cambios sociales.

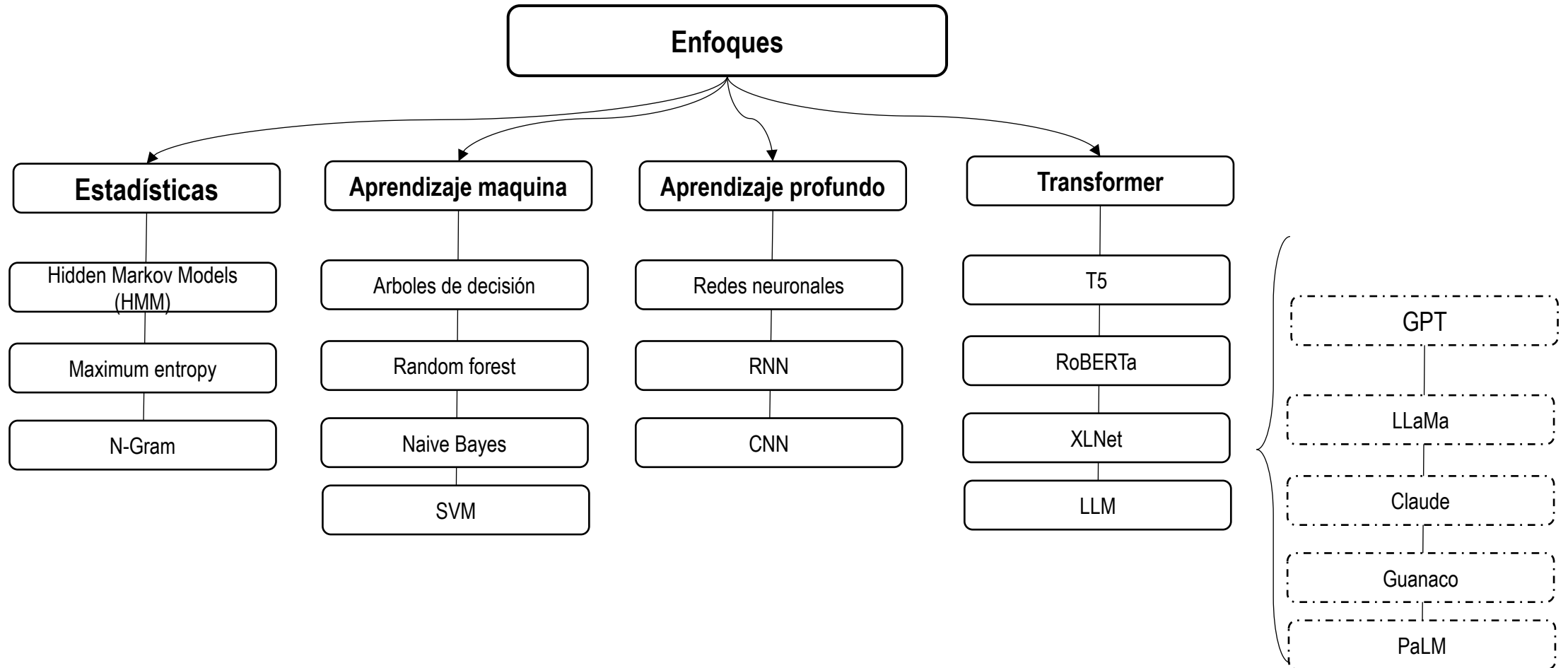
Antecedentes

Diferentes áreas y tareas que involucra el uso de PLN

Área	Tarea
Lingüístico	Resumir y extraer información clave de documentos financieros. Extraer narrativas financieras
Análisis de sentimiento	Cuantificar el sentimiento del mercado con base en expresiones de los medios de comunicación
Serie de tiempo	Tendencia del mercado, proyección, detección de anomalías y clasificación de datos financieros
Razonamiento	Ayuda en la planeación financiera, recomendación de inversión y en la toma de decisión.
Modelos basados en agentes	Usa la capacidad de razonamiento de los modelos para su interacción con el entorno

Antecedentes

Diferentes enfoques metodológicos en la implementación de PLN



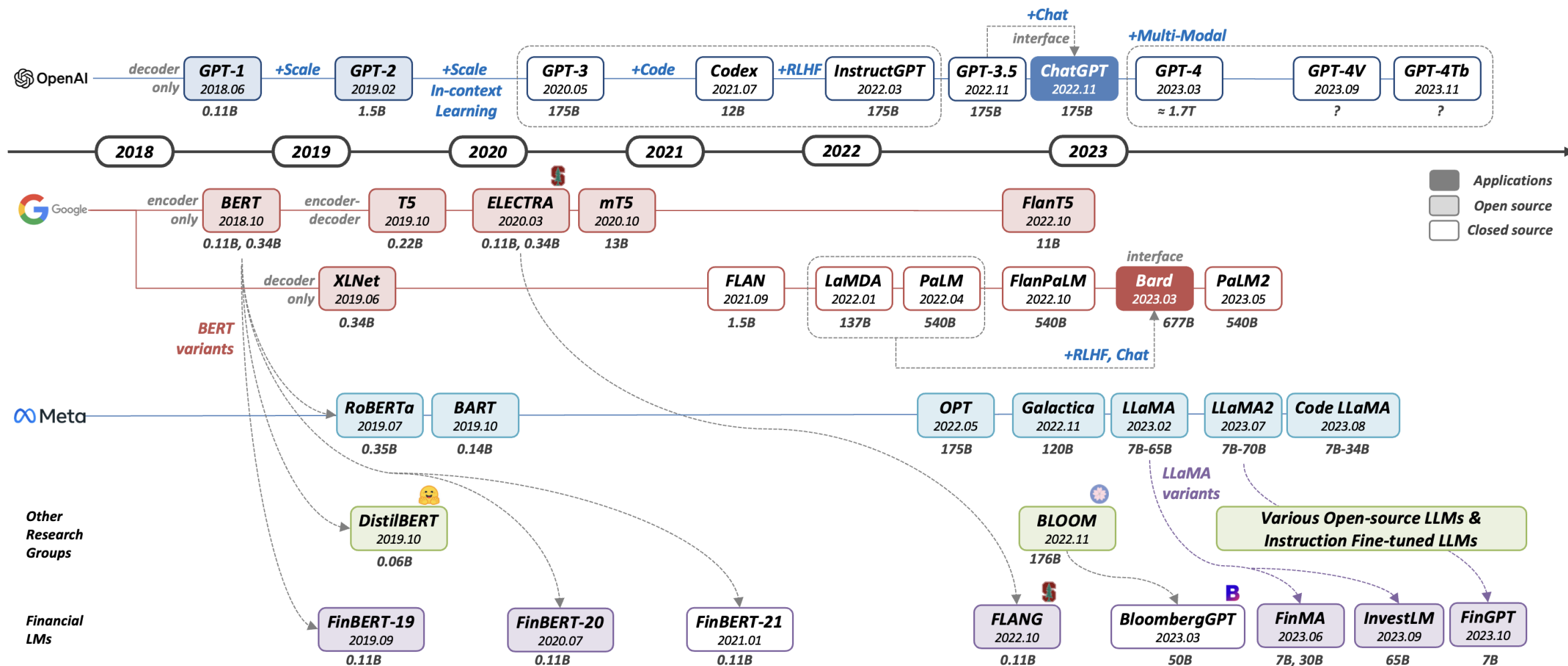
Antecedentes

Crecimiento del uso de PLN para el análisis económico

Modelos	Tarea	Modalidad	Model	Idioma	Abierto
PIXIU (2023)	Múltiples tareas de PLN a finanzas incluyendo, predicción de stock	Texto, tabla, series de tiempo	FinMA	Ingles, Mandarín	Sí
FLUE (2022)	Múltiples tareas PLN	Texto	FLANG	Inglés	Sí
AlphaFin (2024)	PR financieras, predicción de stock	Texto	Stock-chain	Mandarín, Ingles	Sí
DOCMATH-EVAL (2023)	Razonamiento numérico	Texto y tablas		Inglés	Sí
EconLogicQA (2024)	PR financieras	Texto		Inglés	Sí
BBT-Fin (2023)	Múltiples tareas de PLN en finanzas	Texto	BBT-FinT5	Mandarín	Sí
MultiLing (2019)	Resumen de narrativas financieras	Texto		Inglés	Sí
Hirano (2024)	Múltiples tareas de PLN en finanzas	Texto		Japonés	Si
R-Judge (2024)	Identificación de riesgo y análisis de seguridad	Texto		Inglés	Sí
FLARE-ES (2024)	Múltiples tareas de PLN en finanzas	Texto, tablas y series de tiempo	FinMA-ES	Español, Inglés	Sí

Antecedentes

Evolución de los modelos de lenguaje y su aplicación en finanzas



Fuente: Lee et al., 2024

Planteamiento

Observación

- Un crecimiento vertiginoso de los modelos de lenguajes en finanzas.
- Los modelos de lenguajes abarcan las tareas más relevantes.
- Una concentración de estos modelos en idioma inglés.
- La mayoría usa un enfoque supervisado y/o semi-supervisado.

Limitación

- Aplicación limitada en otras áreas de la economía.
- Escasos estudios sobre uso de textos para inferir tendencias.
- Escasos estudios en el idioma español.
- Aplicación limitada de modelos de lenguaje no supervisados

Preguntas

- ¿Es posible utilizar técnicas de Procesamiento de Lenguaje Natural (PLN) para inferir y caracterizar variables económicas en México durante el periodo 1999-2024 a partir de textos de opinión?
- ¿Qué tan representativas y precisas son las inferencias económicas derivadas de textos de opinión comparadas con las estadísticas oficiales de tradicionales del INEGI o el Banco de México?

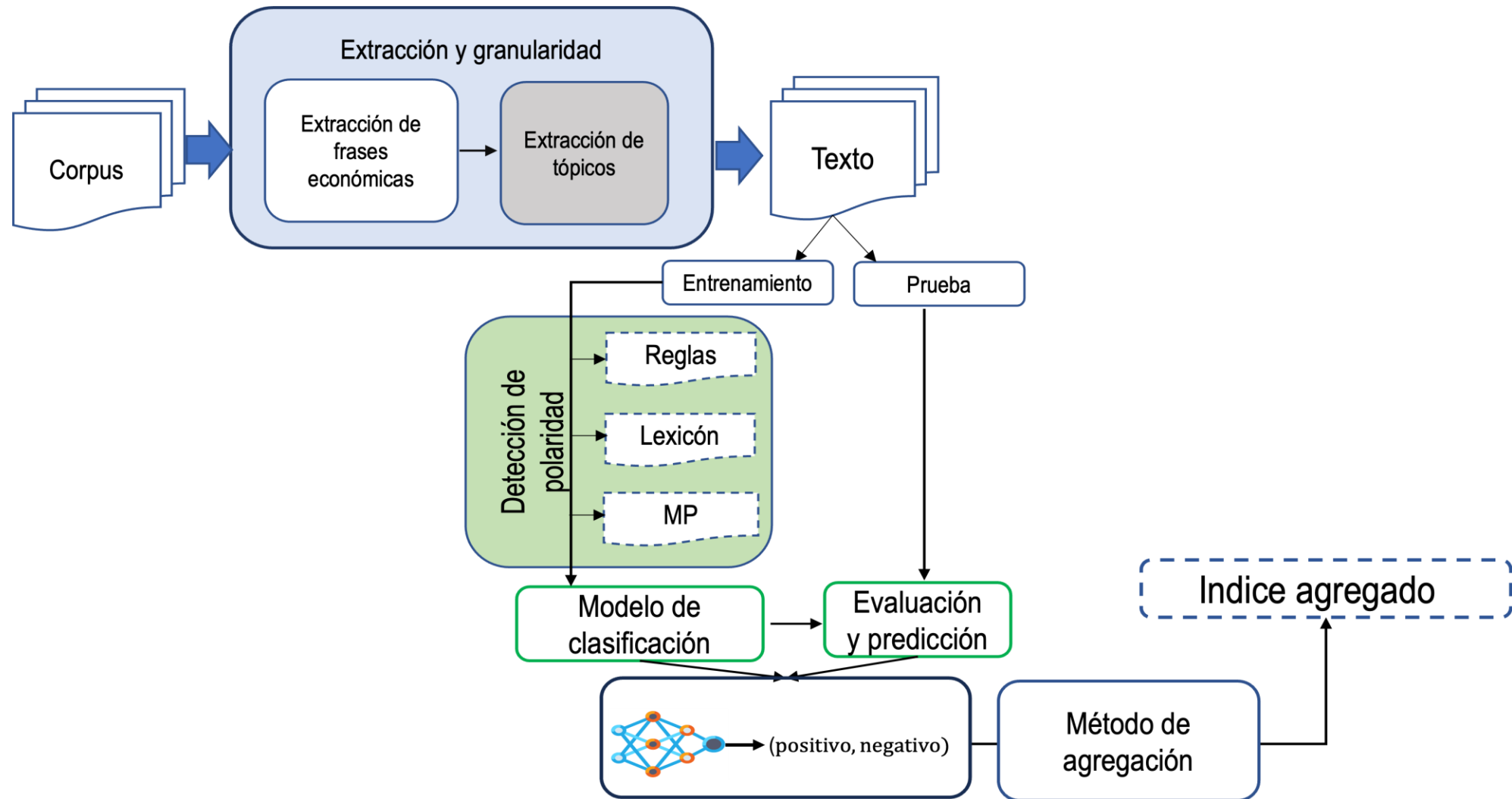
Preguntas

- ¿Qué técnicas de representación de textos permiten identificar de manera efectiva los sub-tópicos y eventos económicos relevantes dentro del corpus?
- ¿Qué métodos son adecuados para clasificar y agregar la polaridad en el tiempo?
- ¿En qué medida es generalizable el marco desarrollado para la inferencia de variables económicas a otros dominios o contextos?

Objetivo

Este trabajo consiste en analizar y aplicar técnicas de Procesamiento de Lenguaje Natural (PLN) para inferir y caracterizar variables económicas en México durante el periodo 1999-2024, utilizando textos de opinión como principal fuente de información. Este enfoque busca identificar tendencias, patrones y relaciones en las actividades económicas, proporcionando una herramienta innovadora para el análisis económico a partir de datos no estructurados.

Propuesta metodológica

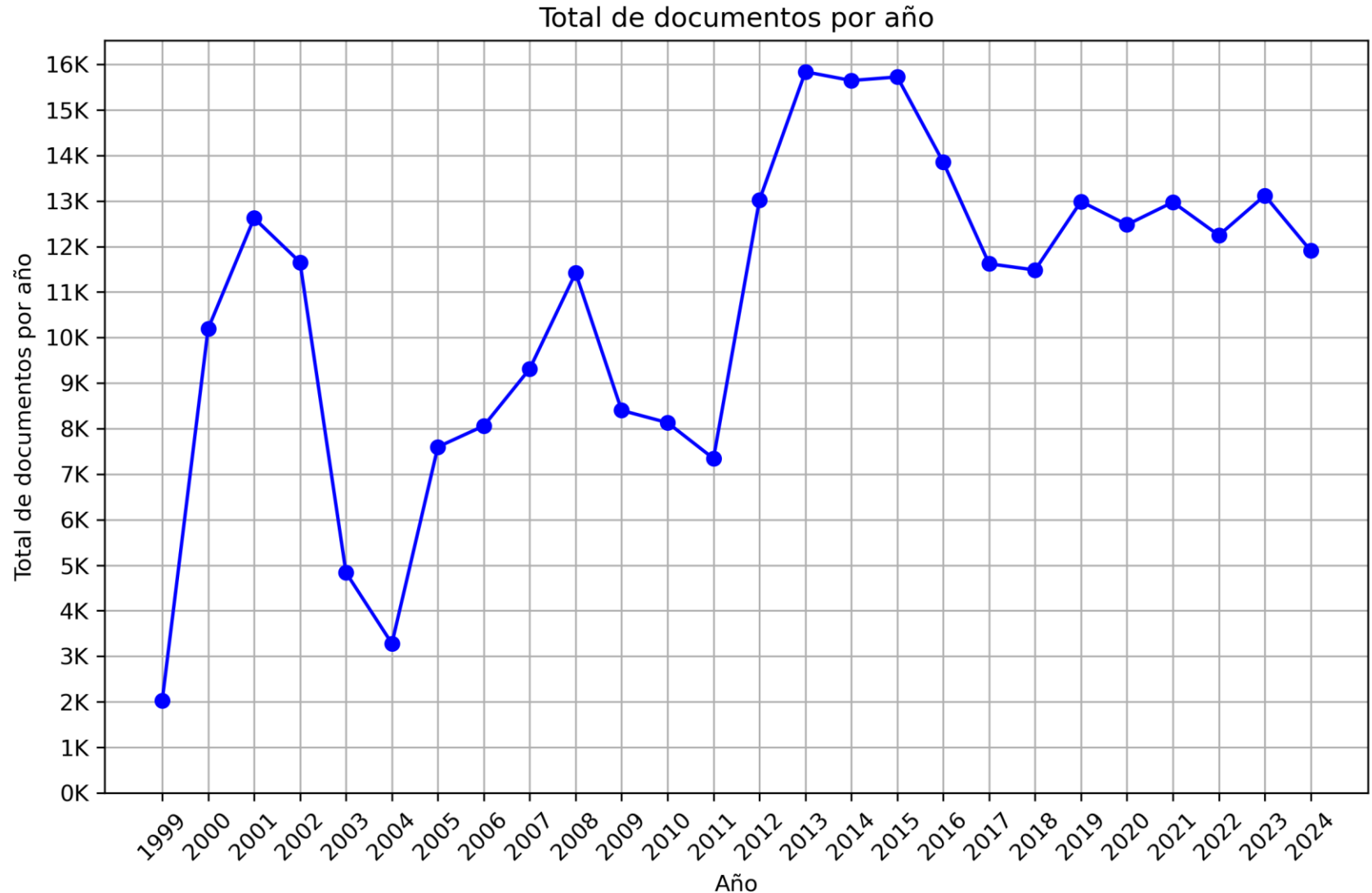


Fuente de información - corpus

Un total de 277,789 noticias económicas de 1999 a 2024.

Los noticieros son:

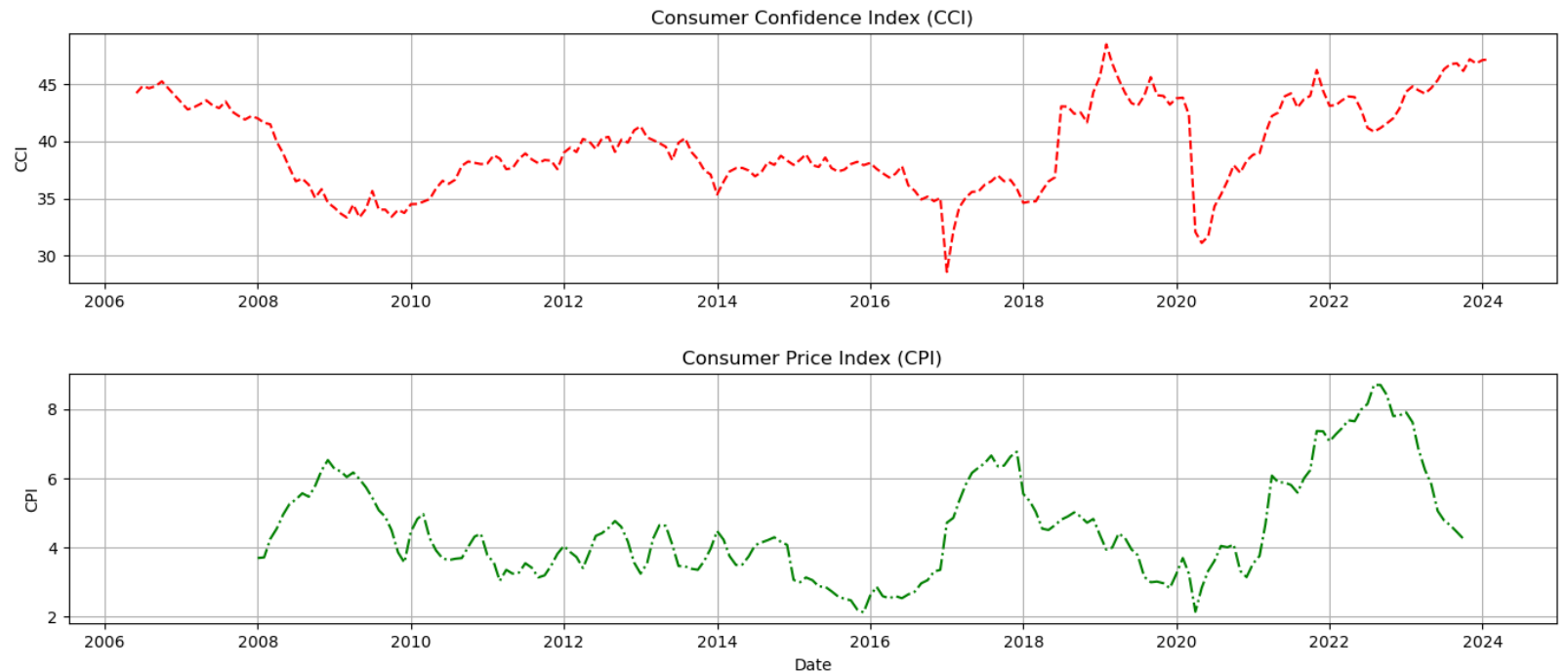
- El Financiero
- El Economista
- Excélsior
- Reforma
- LaJornada



Fuente de información - Inegi

De acuerdo con INEGI:

- **Índice de confianza del consumidor (ICC)** mide el grado de confianza del consumidor respecto a la economía.
- **Índice de Precios al Consumidor (IPC)** mide el precio ponderado de la canasta de bienes y servicios en el país.



Extracción de oraciones

Algoritmo 1: Pipeline de limpieza con múltiples LLMs

Input: Texto original `text`

Output: Texto limpio seleccionado

Paso 1: Cargar prompt de limpieza

```
| prompt  $\leftarrow$  load_prompt_template();
```

Paso 2: Refinar en paralelo con cada LLM

```
| responses  $\leftarrow$  {};  
| foreach m  $\in$  { "llama3.3", "mistral", "deepseek-r1" } do  
| | responses[m]  $\leftarrow$  call_llm(m, prompt, text);
```

Paso 3: Evaluar cada respuesta

```
| scored  $\leftarrow$  {};  
| foreach (model, candidate) en responses do  
| | sim  $\leftarrow$  measure_cosine_similarity(text, candidate);  
| | cleaness  $\leftarrow$  measure_cleanliness(candidate);  
| | scored[model]  $\leftarrow$   $0,6 \times sim + 0,4 \times cleaness$ ;
```

Paso 4: Seleccionar la mejor respuesta

```
| best_model  $\leftarrow$   $\arg \max_m scored[m]$ ;  
| return responses[best_model];
```

Prompt de limpieza

Eres un experto en limpieza y normalización de textos periodísticos con enfoque en contenido económico.

1. Eliminar todo contenido no económico (publicidad, menús, pies de página, avisos legales, etc.).
2. Extraer sólo frases y párrafos con información económica.
3. Corregir errores ortográficos y gramaticales.
4. Normalizar puntuación redundante y espacios.
5. Mantener coherencia y números decimales intactos.
6. Si no puedes corregir lógicamente una parte, devuelve UNKNOWN.

Devuelve únicamente un JSON con el campo "text".

Ejemplo de la extracción y limpieza

Fecha	contenido	texto limpio
2015-11-13	No comprometer más de 30 % del ingreso en las compras, aconseja la Condusef. La compra de bienes sólo subió 0.4 % mensual en 8 meses de 2015, dice Inegi. El juez federal Guillermo Campos aceptó el nombramiento del abogado Javier Navarro Velasco como síndico responsable de la liquidación de activos de Ficrea. El Instituto Federal de Telecomunicaciones (Ifetel) ordenó acceso gratuito al 911 sin importar saldo y dispositivos accesibles para personas con discapacidad. Migración de servicios especiales: 060 policía local, 061 Policía Judicial, 065 Cruz Roja, 066 Sistema Nacional de Emergencias, 068 Bomberos, 080 Seguridad y Emergencia. La depreciación del peso y el retiro de capitales por política de la Fed mermaron 19,149 mdd de reservas, al nivel de hace dos años. Entre enero y septiembre se importaron 127 161 autos usados (13.5 % de ventas nuevas) según AMDA. Viajeros nacionales gastaron 7 354.1 mdd a septiembre. La junta de gobierno del Banxico reconoció un crecimiento económico moderado, similar al promedio de los últimos 15 años.	No comprometer más de 30 % del ingreso en las compras, aconseja la Condusef. La compra de bienes subió sólo 0.4 % mensual en los primeros ocho meses de 2015. El juez Guillermo Campos nombró a Javier Navarro Velasco síndico de la liquidación de Ficrea. El Ifetel dispone acceso gratuito y dispositivos accesibles al 911; migración de códigos: 060, 061, 065, 066, 068, 080. La depreciación del peso y el retiro de capitales por la Fed redujeron las reservas en 19 149 mdd. Se importaron 127 161 autos usados (13.5 % de ventas nuevas) y los viajeros gastaron 7 354.1 mdd a septiembre. El Banxico destacó un crecimiento moderado, en línea con el promedio de 15 años.

Extracción de tópicos (LDA + LLM)

5.2.1. Generación de clusteres económicos

Dado un corpus de oraciones $\mathcal{S} = \{s_i\}_{i=1}^N$, se aplica un encoder semántico $f : \mathcal{S} \rightarrow \mathbb{R}^d$ (por ejemplo, Sentence-BERT Reimers y Gurevych, 2019) para obtener vectores $z_i = f(s_i)$. A continuación, se reduce la dimensionalidad con UMAP McInnes et al., 2018 preservando la estructura de vecindad:

$$\{y_i\}_{i=1}^N = \text{UMAP}(\{z_i\}_{i=1}^N), \quad y_i \in \mathbb{R}^2. \quad (5.1)$$

Extracción de tópicos (LDA + LLM)

5.2.2. Clustering de Oraciones

Se agrupan las representaciones reducidas mediante K-Means MacQueen, 1967, resolviendo:

$$\{\mu_k\}_{k=1}^K = \arg \min_{\{\mu_k\}} \sum_{i=1}^N \min_{1 \leq k \leq K} \|y_i - \mu_k\|^2, \quad (5.2)$$

donde μ_k son los centroides de los K clusters. Cada oración s_i se etiqueta con su índice de cluster $c(i) = \arg \min_k \|y_i - \mu_k\|$.

Extracción de tópicos (LDA + LLM)

5.2.3. Selección de Oraciones Representativas

Para cada cluster k , se eligen las k_r oraciones más cercanas al centroide μ_k mediante distancia euclídea:

$$\text{rep}_k = \arg \min_{i: c(i)=k} \|y_i - \mu_k\|. \quad (5.3)$$

Estas oraciones resumen el contenido semántico de cada cluster y sirven de entrada para la etapa de clasificación.

Extracción de tópicos (LDA + LLM)

5.2.4. Clasificación de Clusters con LLM

Se construye un único prompt por cluster incluyendo las oraciones representativas. Un LLM (por ejemplo, Llama 3.3 a través de Ollama) clasifica cada cluster como económico o no económico. Esto reduce el número de llamadas al modelo a K en lugar de N .

Extracción de tópicos (LDA + LLM)

Input: `topic_summary`

Definiciones esenciales:

- *Inflation*: precios ↑, erosión de poder adquisitivo.
- *CCI*: optimismo y expectativas de los consumidores.

Categorías clave:

Inflation | Consumer Confidence Index | Public Spending | ... |
Other

Instrucciones:

1. Si menciona precios o poder de compra → **Inflation**.
2. Si menciona sentimiento u expectativas → **Consumer Confidence Index**.
3. En otro caso, elegir la categoría más adecuada.

Extracción de tópicos (LDA + LLM)

5.2.5. Filtrado de Clusters Económicos

Definiendo un umbral τ (p.ej., 0.7), se retienen únicamente los clusters cuya proporción de oraciones económicas supere τ . Si $\mathbb{K}_{econ}(s_i)$ indica la predicción de LLM para la oración s_i , entonces para cluster k :

$$\rho_k = \frac{1}{|C_k|} \sum_{i: c(i)=k} \mathbb{K}_{econ}(s_i), \quad (5.4)$$

y se conserva k si $\rho_k \geq \tau$.

Extracción de tópicos (LDA + LLM)

5.2.6. Modelado de Tópicos Final con LDA

Sobre el subconjunto filtrado se entrena un modelo LDA Blei et al., [2003b](#). Cada documento se asume generado por:

$$\theta_d \sim \text{Dir}(\alpha), \quad (5.5)$$

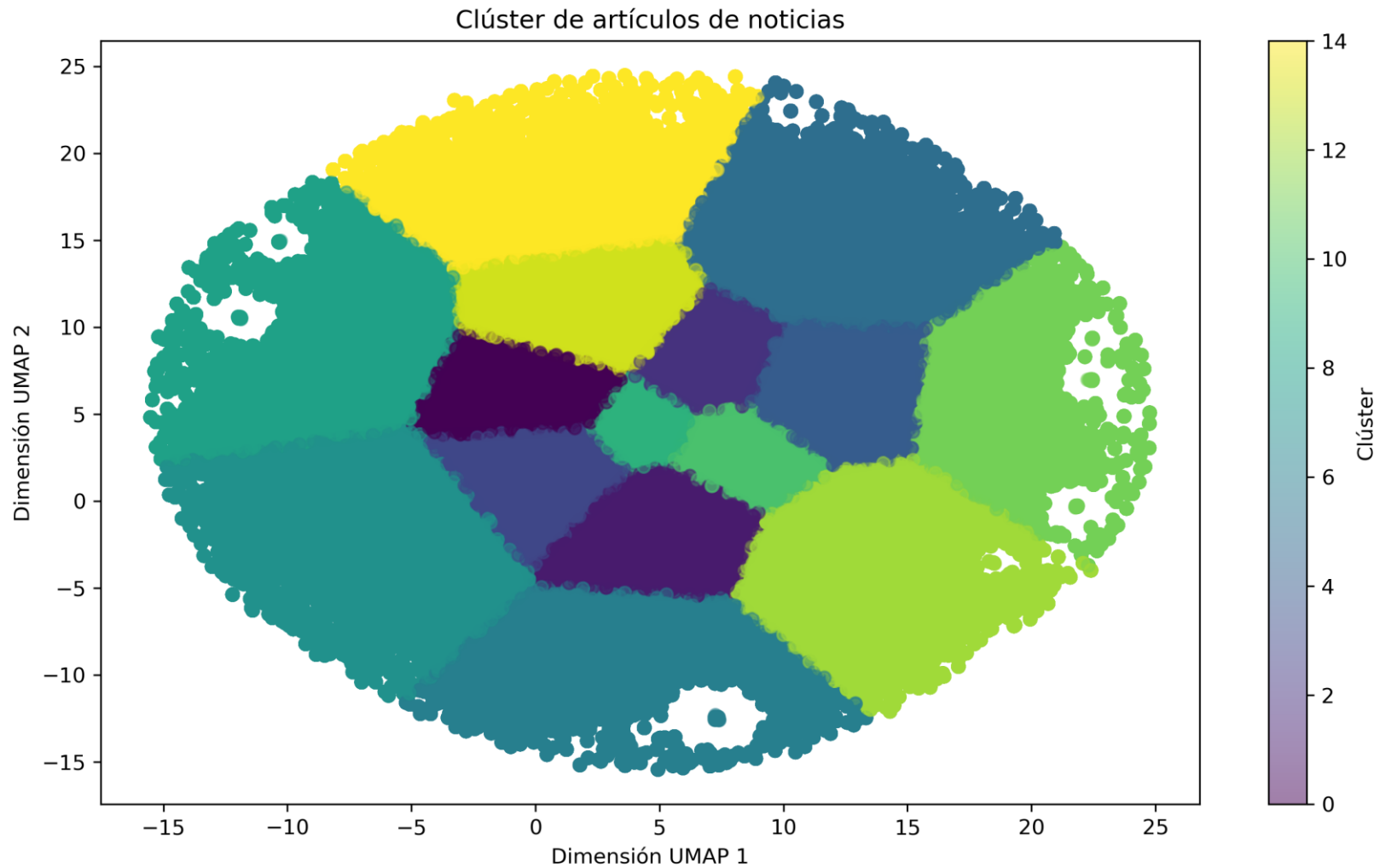
$$\phi_k \sim \text{Dir}(\beta), \quad (5.6)$$

$$w_{dn} \sim \text{Multinomial}(\phi_{z_{dn}}), \quad (5.7)$$

$$z_{dn} \sim \text{Multinomial}(\theta_d). \quad (5.8)$$

Se optimizan α, β y el número de tópicos T mediante métricas de coherencia Greene et al., [2015](#).

Representación de las oraciones



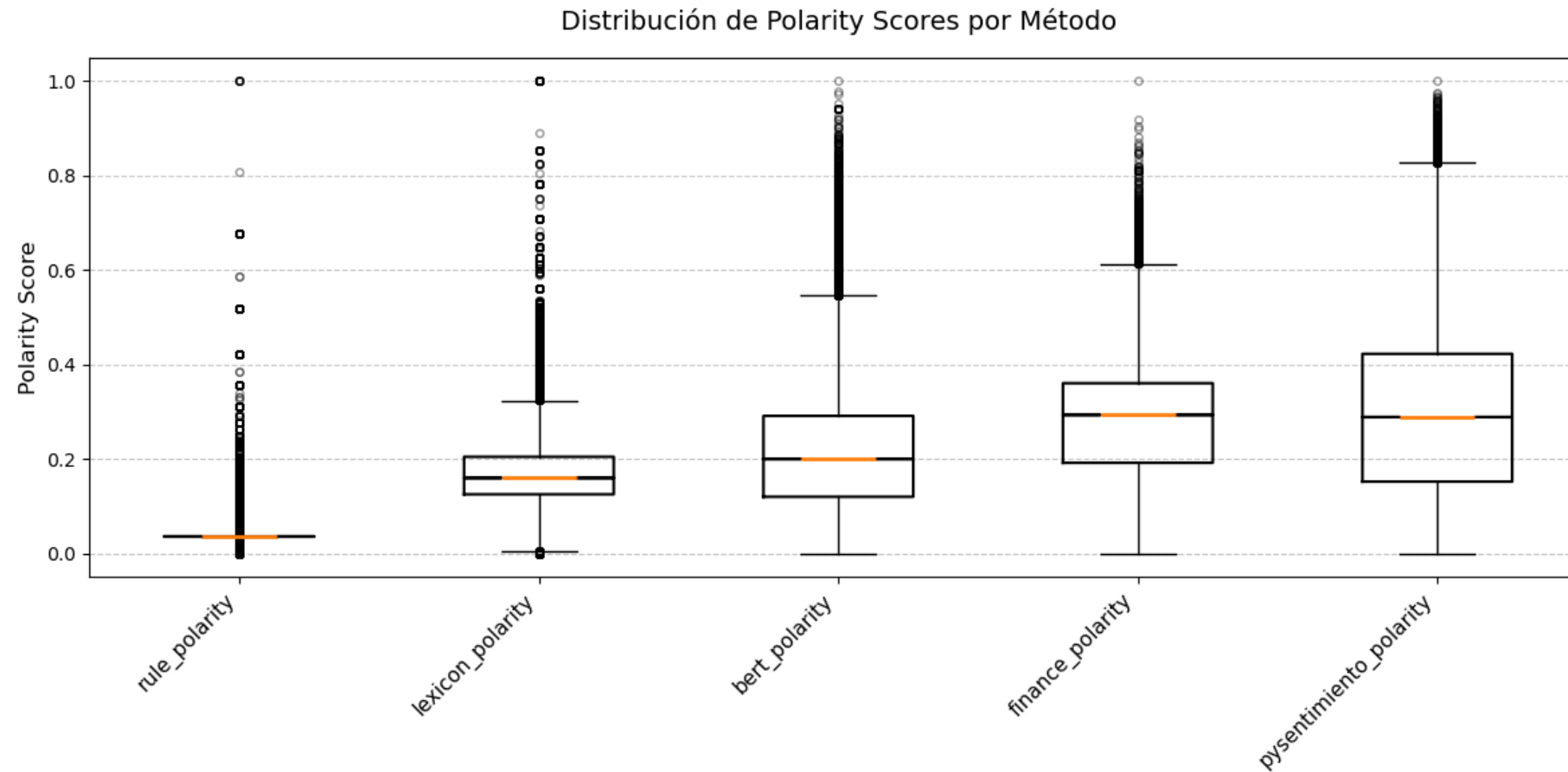
Extracción de oraciones

.....

	count	mean	std	min	50 %	max
n_chars	2293493.0	263.458112	214.093606	11.0	206.0	8457.0
n_words	2293545.0	42.449696	32.428318	0.0	34.0	384.0

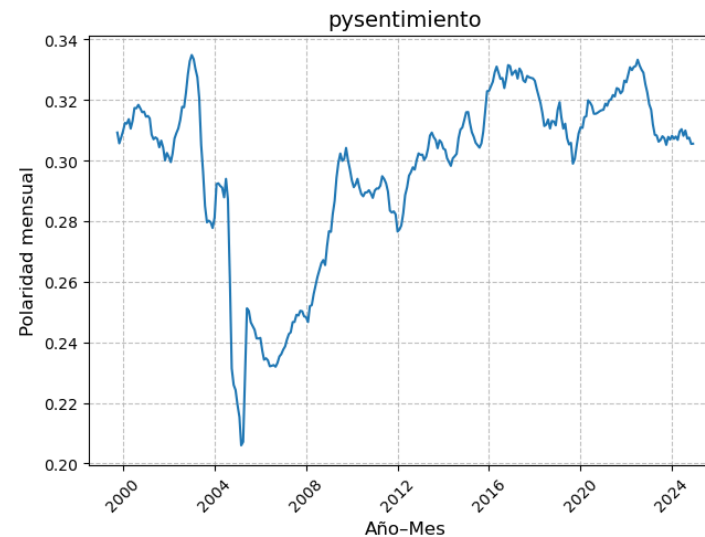
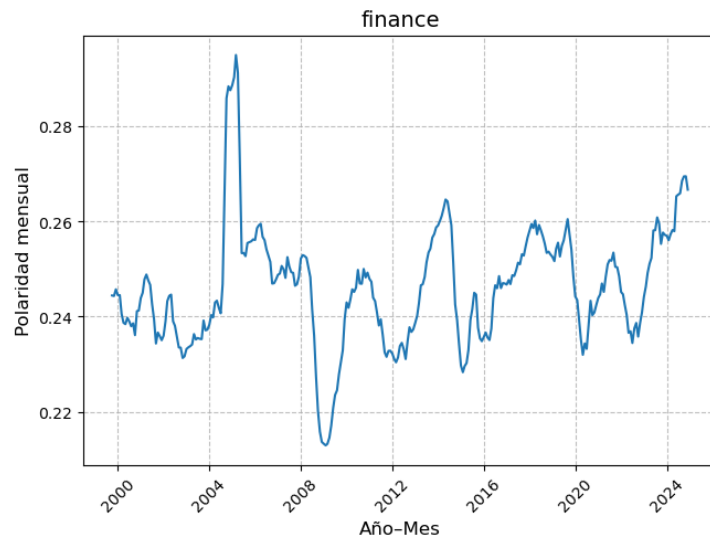
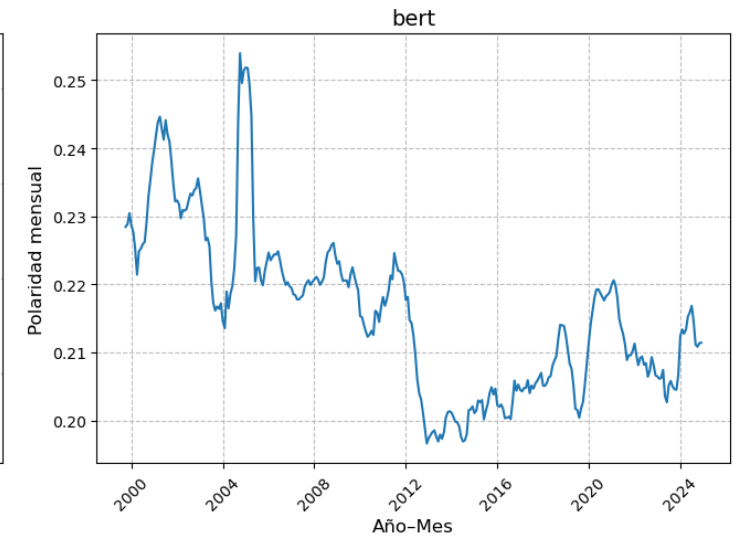
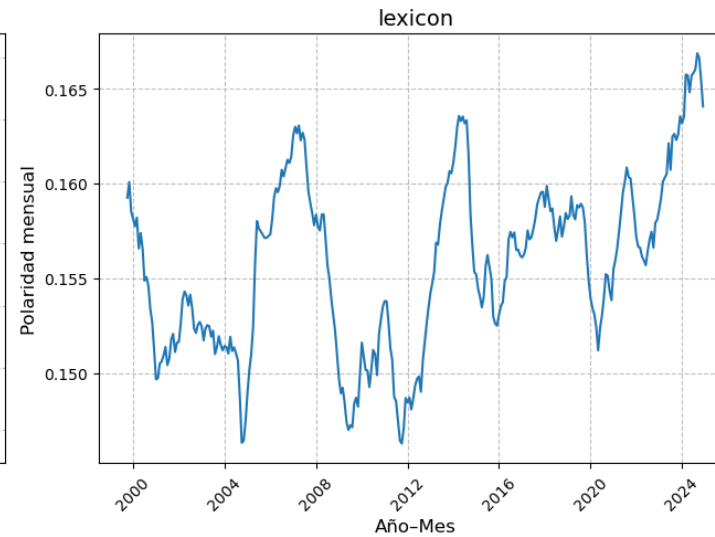
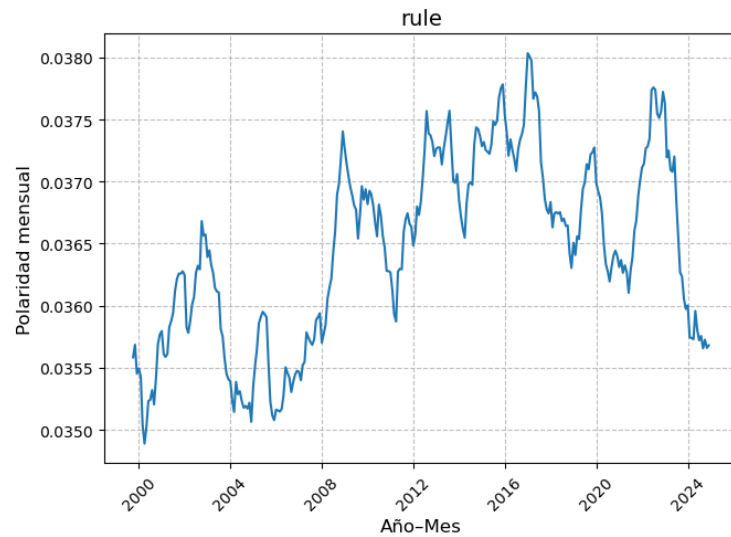
Distribución de la polaridad

Comparación entre la tendencia inferida y las variables económicas observadas

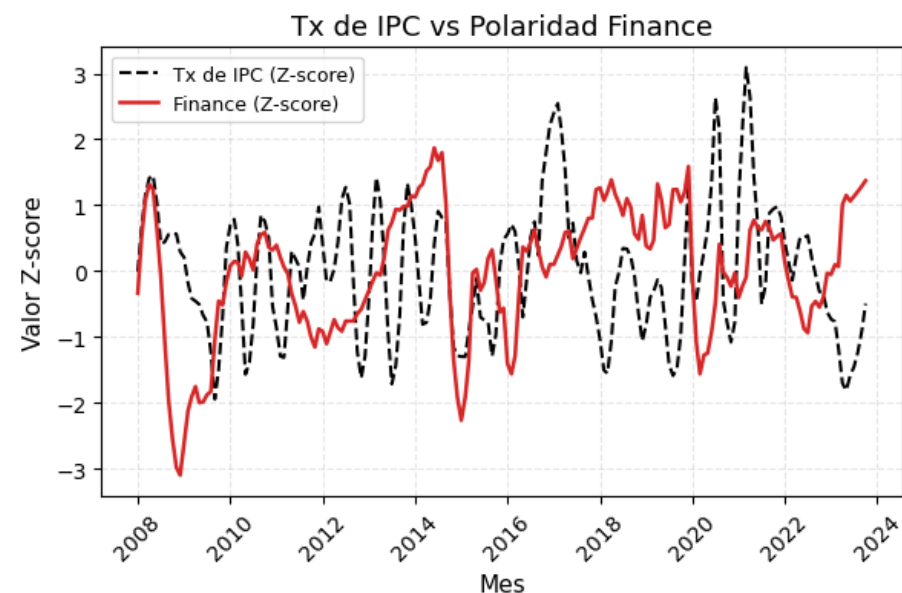
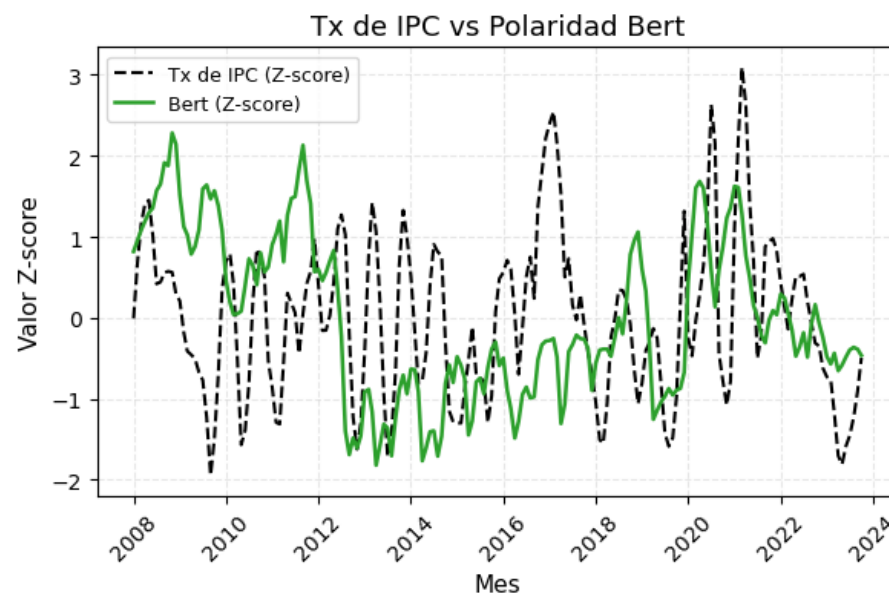
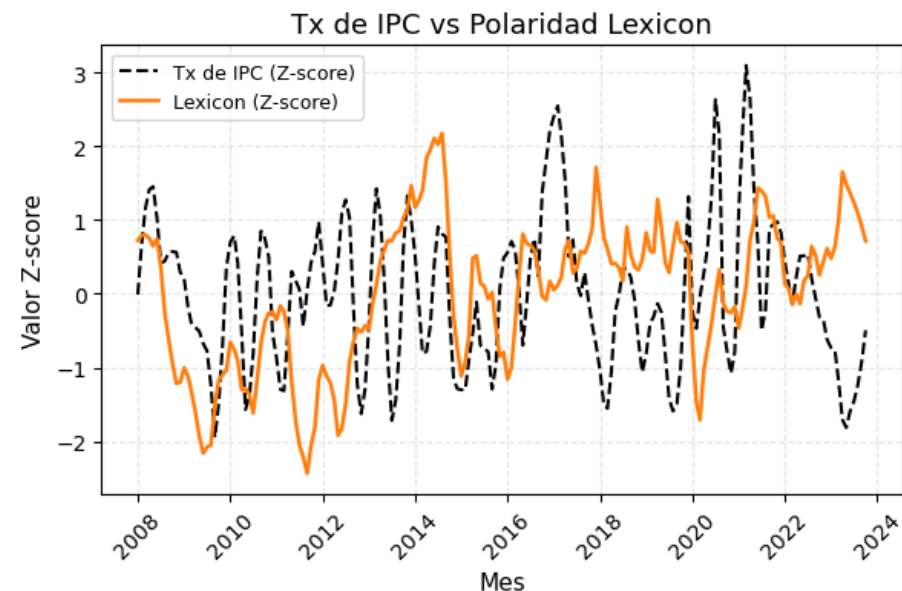
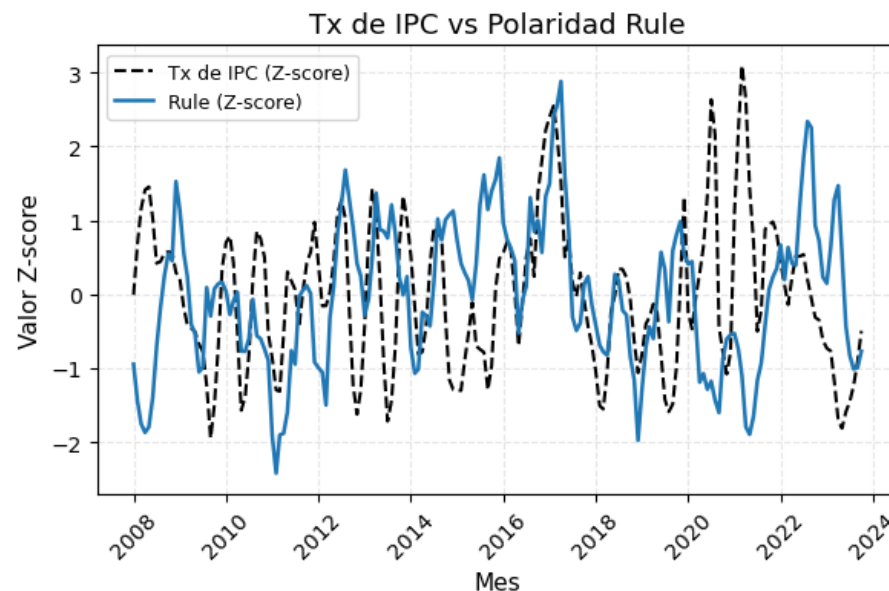


Distribución de la polaridad

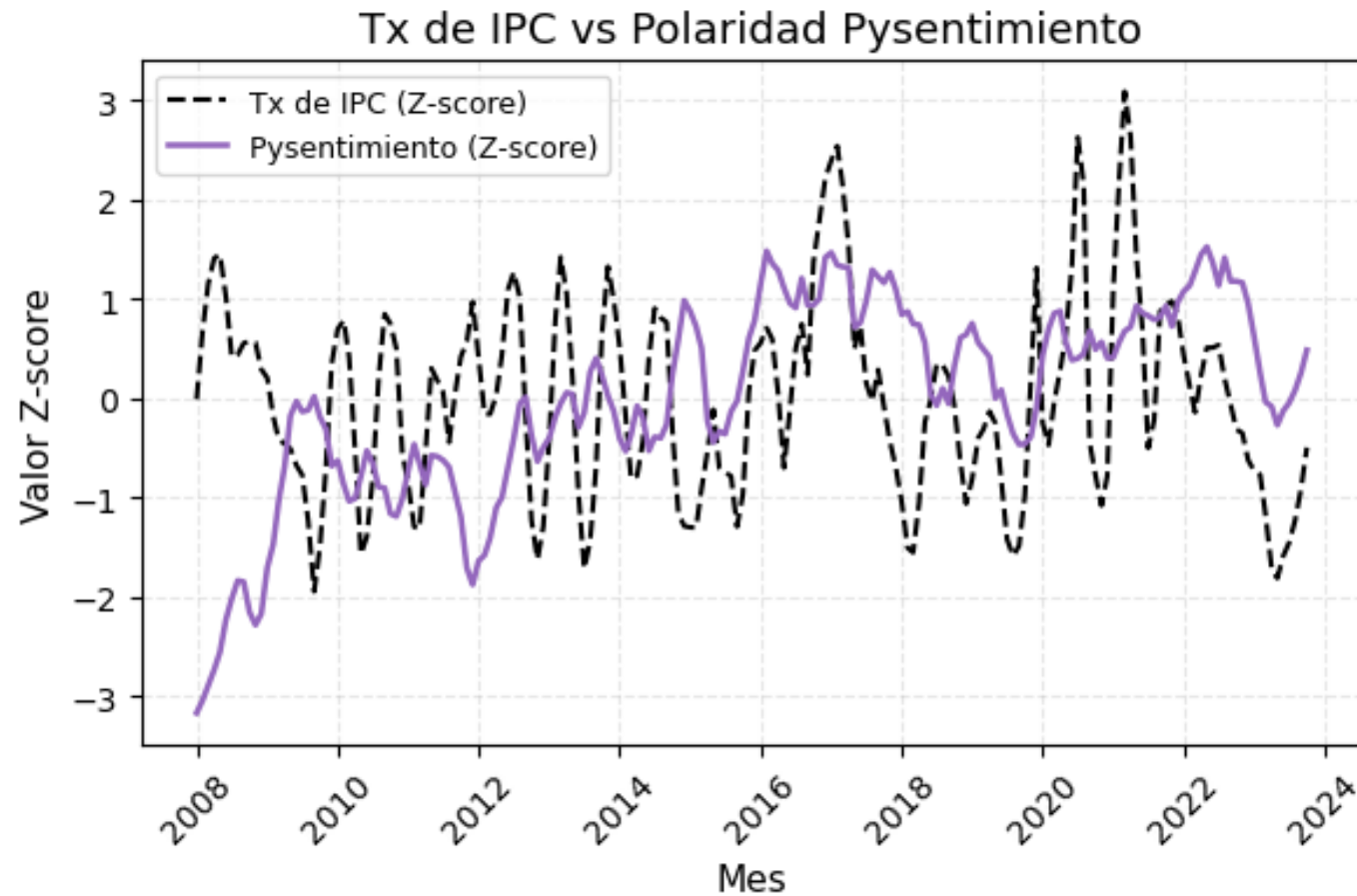
Polaridad de los textos económicos por mes, año y método



Comparación de la polaridad vs IPC



Distribución de la polaridad



Evaluación – función de divergencia (KL)

La función de pérdida de los grupos es definida como una divergencia (KL) entre P y Q , donde es aproximada mediante una función de distribución de t -student.

P es la función objetivo derivada de Q .

$$L_c = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Evaluación

Correlación parcial y prueba de causalidad de Granger

Método	KL(TxIPC Polaridad)	KL(Polaridad TxIPC)	KL simétrico
lexicon_polarity	0.559	0.152	0.356
rule_polarity	0.556	0.277	0.417
bert_polarity	0.715	0.137	0.426
pysentimiento_polarity	0.858	0.155	0.506
finance_polarity	1.012	0.201	0.607

Evaluación – Granger test

- SSR_t y SSR_{ur} son las sumas de cuadrados de los residuos del modelo restringido y del no restringido,
- p es el número de rezagos de X añadidos,
- T es el número de observaciones,
- k es el número total de parámetros estimados en el modelo no restringido.
- Si F excede el valor crítico de la distribución $F_{p,T-k}$ rechazamos H_0 y concluimos que X Granger-causea a Y .

Modelo restringido:

$$Y_t = \sum_{i=1} \alpha_i Y_{t-i} + u_t$$

Modelo no restringido:

$$Y_t = \sum_{j=1} \alpha_j Y_{t-j} + \sum \beta_j X_{t-j} + \varepsilon_t$$

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

(X no Granger-causea Y)

$$F = \frac{(SSR_r - SSR_{ur})}{SSR_{ur} (T - k)} \sim F_{p, T-k}$$

$$F = \frac{(SSR_r - SSR_{ur})}{SSR_{ur}} - \frac{S\hat{\sigma}^2}{T - k}$$

Evaluación

“Ningún enfoque de polaridad muestra evidencia estadística de Granger-causalidad sobre la inflación en lags de 1 a 4 meses; rule_polarity lidera en correlación parcial, pero sin llegar a ser causal.”

Método	Correlación Parcial	p-Valor Granger (Lag 1)	p-Valor Granger (Lag 2)	p-Valor Granger (Lag 3)	p-Valor Granger (Lag 4)
rule_polarity	0.086	0.3551	0.6018	0.6144	0.7217
lexicon_polarity	0.013	0.6435	0.7364	0.6261	0.4667
bert_polarity	0.056	0.0658	0.1380	0.2377	0.3474
finance_polarity	0.020	0.6419	0.9031	0.5557	0.6375
pysentimiento_polarity	0.031	0.2637	0.1244	0.2176	0.2911

Conclusión

- Desarrolló un corpus que recopile opiniones y análisis económicos de México entre 2008-2024, implementando técnicas automatizadas y semi-automatizadas para la limpieza, normalización de la fuente de información y extracción de tópicos, con el fin de buscar garantizando su representatividad y calidad.
- Desarrolló técnicas de representación de textos para identificar sub-tópicos/eventos relacionados con el objeto de estudio.
- Clasificó la relación de estos tópicos con el objeto de estudio con el fin de su comparabilidad en el tiempo.

Conclusión

- Implementó métodos para identificar tendencias y patrones entre los indicadores derivados de los tópicos y las variables económicas observadas.
- Comparó las inferencias obtenidas mediante los modelos de extracción de texto con fuentes oficiales en el tiempo, como los indicadores del INEGI o del Banco de México.
- Mostró la importancia del uso de PLN para el análisis económico.