Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

Uso de técnicas de lenguaje como herramienta para la caracterización e inferencia de dinámicas económicas en México entre 1999-2024

Tesis presentada en cumplimiento parcial de los requisitos para obtener el grado académico de

Doctorado en Ingeniería del Lenguaje y del Conocimiento

Presenta:

Pierre Antoine Delice

Bajo la supervisión del:

Dr. David Eduardo Pinto Avendaño,

Benemérita Universidad Autónoma de Puebla

Puebla, 10 de noviembre de 2025

Resumen

El tema principal de esta tesis doctoral consiste en hacer uso de las técnicas y modelos de lenguaje como herramientas para la caracterización e inferencia de dinámicas económicas en México durante el período que va de 1999 a 2024. Las principales contribuciones se basan en explorar los recursos lingüísticos y computacionales disponibles, así como, los modelos de lenguaje para caracterizar el comportamiento de variables económicas y el desarrollo de metodologías para inferir patrones y tendencias económicas a partir de fuentes alternativas como son las noticias económicas.

El trabajo sera dividido en capítulos, el primero consistirá en describir los objetivos, motivos, contexto, las hipotésis de trabajo, así como, los resultados esperados. Esto es importante para guiar al lector no solo respecto a la estructura de la tesis, sino también en los objetivos y alcances de la misma. Luego, se presenta el estado del arte que consiste en un recuento de los distintos trabajos que vincula los métodos computacionales con el análisis económico. Esta sección ayudará al lector a entender mejor las contribuciones de este trabajo y como se piensa participar en el debate que impulsan los métodos computacionales en témáticas de bajo recursos como es el ámbito de la economía.

Adicionalmente, una gran parte del desarrollo de este trabajo se enfo-

cará en las técnicas de mineria de textos necesarias para constituir los datos para el análisis, determinar la granularidad de análisis, extraer los tópicos relacionados con el fenómeno a estudiar, por ejemplo: inflación, consumo, entre otros. Para eso, se utilizarán una gama de técnicas de procesamiento de lenguaje natural que trasciende la literatura, empezando por técnicas basadas en reglas, pasando por métodos no supervisado, probabilísticos y los basados en modelos de lenguaje.

En los subsiguientes capítulos se exploraran el uso de los modelos de lenguaje en tareas específicas de caracterización e inferencia económica, tales como la predicción de indicadores económicos, el análisis de políticas públicas y la detección de tendencias de mercado, así como las técnicas de agrupamiento y clasificación para la selección de los tópicos. La parte neuralgica de esta tesis consiste en clasificar de manera automática los textos usando un modelo enmascarado y las técnicas de agrupamiento. Esto permite ahorrar el desarrollo de metodos de minería de textos que son computacionalmente complejos y costosos.

Finalmente, el último capítulo de esta tesis se enfoca en analizar los principales resultados de los distintos experimentos realizados, y se discuten las implicaciones teóricas y prácticas asociadas al uso de modelos de lenguaje en el análisis económico y las principales perspectivas del trabajo.

Agradecimientos

Me gustaría empezar este trabajo por agradecer de manera especial a todos los que me apoyaron durante mis estudios de doctorado.

Las palabras no son suficientes para expresar mi gratitud a mis padres y mi familia quienes nunca cesaron de apoyarme y de creer en mi.

Me gustaría agradecer de manera particular a mi director Dr. David Pinto por confiarme en mi y darme la oportunidad de trabajar bajo su supervisión, por su apoyo, paciencia y por estar siempre. Espero poder seguir colaborando con él en un futuro. ¡Gracias!

También, quiero extender los agradecimientos a mis lectores quienes fueron muy pacientes conmigo y me han apoyado a ver las cosas de manera diferente.

- Prof. Darnes Vilariño Ayala
- Prof. Helena Adorno Gómez

- Prof. Manuel Montes y Gómez

No puedo cerrar este ciclo sin olvidar a dos personas que hicieron posible mi estancia en la universidad:

- Prof. Maria Josefa Somodevilla
- Prof. Blanca Bermúdez Juárez (¡En paz descanse!)

Quiero agradecer de manera especial a estas personas por sus apoyos, discusiones, aportaciones y sus orientaciones que tuvimos durante mis estudios:

- Prof. Víctor Mireles Chávez
- Prof. Sergio Hernández López

En está lista, no se puede olvidar a:

- Dr. Adrián Pastor López-Monroy
- Dr. Fernando Sánchez-Vega

Tanto desde su esfuerzo personal como mediante la cobertura institucional que me brindaron, empezando por la institución que otorgó la beca, Consejo Nacional de Humanismo, Ciencias y Tecnología (CONAHCYT), la Benemérita Universidad Autónoma de Puebla (BUAP), pasando por la Facultad de Ciencias de la Computación (FCC), así como, el Laboratorio

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Contexto

Varias instituciones gubernamentales y privadas de la rama financiera y económica han estado invirtiendo en el uso de tecnologías digitales y analítica de datos para mejorar tanto la producción de estadísticas oficiales como la generación de información financiera. A estas se suman instituciones como el Instituto Nacional de Estadística y Geografía (INEGI), el Banco de México, la Secretaría de Hacienda y Crédito Público (SHCP), así como la banca privada [1] [2,3,4]

La adopción de la Inteligencia Artificial (IA) en el sector financiero, particularmente en las áreas de investigación y desarrollo, ha impulsado mejoras significativas en los procesos internos y en la eficiencia operativa (for International Settlements, 2020). La aplicación de estas tecnologías

[1] En el caso de la banca privada, BBVA ha desarrollado la "BBVA AI Factory" con presencia en México, para crear productos basados en inteligencia artificial y analítica de datos (véase https://www.bbva.com/en/innovation/bbva-internationalizes-its-artificial-intelligence-factory-to-grow-its-ai-product-range/)

[2] HSBC México ha incorporado el servicio DiMo® en su aplicación, como parte de su estrategia de transformación digital y plataformas de pago móvil (véase https://www.about.hsbc.com.mx/-/media/mexico/es/news-and-media/240911-comunicado.pdf?sc_lang=es-MX).

[3] En el sector público financiero, la Secretaría de Hacienda y Crédito Público (SHCP) pone a disposición portales de datos abiertos con indicadores fiscales y financieros, en formatos reutilizables para análisis automatizado (véase https://datos.gob.mx/dataset/estadisticas_oportunas_finanzas_publicas_principales_indicadores-fiscales).

[4] La Instituto Nacional de Estadística y Geografía (INEGI) ha fortalecido sus procesos estadísticos mediante lineamientos en materia de tecnologías de la información (TIC) para la producción de datos oficiales (véase https://sc.inegi.org.mx/repositorioNormateca/Lci19Dic18.pdf).

ha fortalecido los mecanismos de apoyo a la toma de decisiones mediante modelos predictivos y sistemas analíticos avanzados (Company, 2021). A su vez, el uso de IA ha contribuido a reducir costos operativos mediante la automatización de procesos, la optimización de flujos de trabajo y la mejora en la gestión del back-office (Fund, 2018). Asimismo, sus aplicaciones se han extendido de manera notable en áreas como la detección de fraude, gracias al uso de algoritmos de aprendizaje automático capaces de identificar patrones anómalos en tiempo real (for Economic Co-operation & Development, 2021). Finalmente, la IA se ha convertido en una herramienta esencial para el análisis y la gestión del riesgo, permitiendo modelos más robustos para la evaluación de solvencia, exposición crediticia y resiliencia financiera (of England, 2020).

En cuanto a las agencias de gobierno, las aplicaciones abarcan en general tareas de análisis, como la proyección de indicadores, seguimiento de tendencias en redes sociales, y en gran medida la generación de reportes. Los métodos usados se extienden desde el uso de aprendizaje máquina para los ejercicios de predicción del crecimiento económico (Jokubaitis et al., 2020; Richardson et al., 2021), aprendizaje profundo (Zheng et al., 2023), modelos generativos, así como, de Procesamiento de Lenguaje Natural (PLN) y por último del uso de los Grandes Modelos de Lenguaje (LLMs por su sigla en Inglés).

Más allá de que las estadísticas oficiales es el medio por el cual el Estado se nutre de la realidad para tomar decisiones, la sociedad también se informa mediante este ejercicio, considerado de transparencia y de rendición de cuentas. En este sentido, toda adopción de nueva tecnología para la producción, análisis, medición y proyección de estadísticas oficiales debe ser en miras de la mejora de los resultados para el consumo de los actores antes mencionados.

Bajo esta perspectiva, esta tesis busca como objetivo el de explorar el uso de los métodos computacionales basados en PLN para la inferencia de indicadores económicos en México, con especial énfasis en la inflación. El interés de un tema así como un proyecto doctoral, radica en la posibilidad de explorar las herramientas computacionales de análisis de textos en un contexto de producción de estadísticas oficiales, de análisis económico y sobre todo, de la posibilidad de contribuir al debate relacionado con los métodos de extracción y clasificación de información basada en textos.

Además considerar solo los métodos computacionales desde un enfoque de aprendizaje no supervisado o aprendizaje débil es una apuesta muy elevada para una area teoricamente compleja como los fenómenos económicos. Pero, también es una oportunidad para mostrar el potencial de la IA como herramienta de extracción de información, reconocimiento de patrones y análisis de datos.

1.2. Planteamiento

Medir la inflación es una de las principales tareas para el Banco Central, pero también para la sociedad en general, porque los efectos de la inflación suelen ser devastadores para casi todos los sectores de la economia. Aun cuando existen controversias relacionadas con los efectos reales de la inflación, debido a que produce ganadores y perdedores, en general no es bien vista. Tanto los académicos como los analistas financieros alertan de los peligros de una situación de alta inflación.

Definida como un aumento generalizado de los precios, y medida por el Indice Nacional del Precio al Consumidor (IPC), es en efecto uno los principales indicadores macroeconómicos que refleja en promedio el costo de vida de una población y que tiende a ofrecer un panorama de su vida

cotidiana.

El Instituto Nacional de Estadística y Geografía (INEGI), encargado de medirla, procede a levantar una encuesta de ingreso-gasto de los familiares a nivel nacional para saber qué es lo que se suele consumir. Una vez que se tienen identificados los bienes y servicios, recopilan sus precios en tiendas de todo el país. Esta información se compara de manera quincenal, mensual y anualmente para medir la variación en los precios.

Cada mes el INEGI monitorea alrededor de 235 mil precios en 46 ciudades del país. La información se procesa tomando en cuenta qué tanto se gasta en ello, para así saber cuál de los rubros tiene mayor importancia en el consumo de las familias.

Si bien los indicadores económicos recoletados por INEGI tienden a ser disponibles en un lapso de tiempo razonable y de alta confiabilidad, no siempre son suficientes para captar los cambios en los precios que se producen en el mercado.

Tampoco refleja la percepción de los agentes económicos sobre la economía, ni los efectos de las políticas económicas en la inflación. Razón por la cual nos preguntamos si la información contenida en los textos, y en particular en los medios de comunicación, podría ser una fuente de información útil para inferir la inflación y por lo tanto suplir las carencias que tienen las fuentes tradicionales.

Por otro lado, aunque se observa un creciente interés en el uso de técnicas de aprendizaje automático y profundo, así como, de los modelos de lenguaje para el análisis de fenómenos económicos, donde se presume una mejora en los resultados obtenidos (Mulvey et al., 2022); solo algunas áreas específicas de la economía han sido testigos del uso de dichas metodologías, principalmente el sector financiero. Lo que eleva el interés en entender el comportamiento de los recursos computacionales en la economia real como

son los indicadores de inflación, consumo, índice de confianza del consumidor y otros.

Por otro lado, la mayoría de los trabajos encontrados en la literatura utilizan conocimiento de expertos para la generación de los corpora de entrenamiento. Por ejemplo, el Financial PhraseBank y FinQA que son utilizados para análisis de sentimiento en el contexto financiero fueron generados por expertos (Z. Chen et al., 2021; Malo et al., 2013). Otros trabajos que permiten realizar tareas de reconocimiento de entidades nombradas hacen uso extensivo de anotadores profesionales para la generación de los datasets (Sang & Meulder, 2003; Wang et al., 2020).

Por eso, nos hemos preguntado ¿qué tanto los métodos computacionales pueden contribuir en la extracción y clasificación de los textos económicos sin el uso de anotadores profesionales? En particular, nos interesa saber si es posible utilizar métodos no supervisados o de aprendizaje débil para extraer información relevante de los textos económicos y si estos métodos pueden ser utilizados para inferir indicadores económicos como la inflación.

En este sentido, esta tesis doctoral se centra en el análisis de textos económicos mediante el uso de modelos de lenguaje y técnicas de procesamiento de lenguaje natural, con el objetivo de inferir indicadores económicos en México. El enfoque principal es explorar las posibilidades que ofrecen los métodos computacionales para extraer información relevante de los textos sin depender de conocimiento experto o de anotadores profesionales.

1.3. Preguntas de investigación

Con base en las distintas limitaciones observadas anteriormente, se plantean las siguientes preguntas de investigación:

- ¿Es posible caracterizar la tendencia de fenómenos económicos solo

a partir de noticias económicos en México en un periodo de tiempo bien delimitado?

- ¿Cómo generar los recursos cómputo-lingüísticos necesarios para hacer de los textos una fuente representativa para el estudio de los fenómenos económicos?

- ¿Qué tan representativas y precisas son las inferencias económicas derivadas de los textos de opinión comparadas con las estadísticas oficiales y tradicionales levantadas por INEGI?

1.4. Objetivos de la investigación

En esta sección se presentan los objetivos generales y específicos de esta investigación. Estos objetivos proporcionan una descripción clara y concisa de las acciones propuestas para resolver un problema de investigación.

- **Objetivo general:** analizar y aplicar técnicas de PLN para inferir y caracterizar la tendencia de variables económicas en México durante el periodo 1999-2024, utilizando textos de opinión como principal fuente de información.

- **Objetivos específicos:**

  - Construir un corpus que recopile opiniones y análisis económicos en México entre 1999-2024, implementando técnicas automatizadas y semi-automatizadas para la normalización de la fuente de información y extracción de tópicos, garantizando su representatividad y calidad.

  - Aplicar técnicas de representación de textos para identificar subtópicos/eventos relacionados con el objeto de estudio.

- Clasificar estos tópicos en función de su relación con el objeto de estudio para permitir su comparabilidad en el tiempo.
- Implementar métodos para identificar tendencias y patrones entre los indicadores derivados de los tópicos y las variables económicas observadas.
- Comparar 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.

1.5. Contribuciones

Las contribuciones de esta investigación se resumen de la siguiente manera:

- Primero, la introducción de técnicas avanzadas de PLN y modelos de lenguaje en la construcción de un corpus específico para el análisis económico. Esto abarca desde la normalización de la fuente de información, pasando por la extracción de tópicos relevantes, hasta la clasificación de estos tópicos en función de su relación con fenómenos económicos específicos.

- Segundo, la introducción de un enfoque no supervisado que combina técnicas de representación de textos y modelos de lenguaje para identificar sub-tópicos/eventos relacionados con fenómenos económicos. Este enfoque permite una comprensión más profunda de las relaciones semánticas y léxicas en los textos económicos, sin depender de anotadores profesionales.

- Tercero, un marco metodológico que permite utilizar la propuesta mencionada anteriormente para inferir patrones y tendencias en otro ámbito de estudio, como la economía, a partir de textos de opinión.

Esto incluye la implementación de métodos para identificar tendencias y patrones entre los indicadores derivados de los tópicos y las variables económicas observadas, así como la comparación de las inferencias obtenidas con fuentes oficiales en el tiempo.

1.6. Estructura de la tesis

La tesis está organizada en capítulos, cada uno de los cuales aborda un aspecto específico de la investigación. Eso permite una lectura fluida y ágil, facilitando a los lectores las secciones de su interés.

En el Capítulo 2 se presenta el marco téorico, así como, los conceptos clave relacionados con las técnicas, métodos y teorias subyacentes que fundamentan el desarollo de la tesis.

El Capítulo 3 examina en detalle la literatura relacionada con el uso de textos para el análisis económico, los distintos enfoques adoptados, los objetivos alcanzados, las fuentes de información empleadas y los métodos desarrollados. Este análisis conforma el estado del arte de la investigación y permite delinear los límites, así como las oportunidades de mejora de los métodos existentes.

El Capítulo 4 hace referencia al desarrollo metodológico que constituye la propuesta de investigación, con el fin de mejorar los enfoques encontrados en la literatura.

En el Capítulo 5 se presenta los resultados de las distintos experimentos realizados y su evaluación.

El Capítulo 6 presenta las conclusiones de la investigación, así como las limitaciones y las líneas de investigación futuras.

Capítulo 2

Marco téorico

Me gustaría comenzar este capítulo, al igual que Bhargav Srinivasa-Desikan (2018) en su libro *Natural Language Processing and Computational Linguistics*, planteando la siguiente pregunta: *¿qué se entiende por análisis de texto?*

El análisis de texto, o *text analysis*, se refiere a un conjunto de métodos computacionales diseñados para extraer información estructurada y significativa a partir de datos textuales no estructurados. Esto comprende diversas etapas como la *tokenización*, la *representación semántica*, la *detección de patrones* y la *interpretación* de los resultados en función de un objetivo analítico específico (Srinivasa-Desikan, 2018). Para el autor, lo importante es el proceso mediantel el cual se lleva a cabo el análisis de textos, destacando su capacidad para transformar datos textuales en información útil.

Desde la misma perspectiva, Feldman y Sanger (2007) en su libro *The Text Mining Handbook* define el análisis de texto como un proceso de conocimiento intensivo entre los intereses del investigador y una serie de documentos. Esto con el fin de proceder a la extracción de información útil mediante la identificación y exploración de patrones (Feldman & Sanger, 2006). En ese contexto, el autor busca mostrar cómo el análisis de textos, combinado con la creatividad del investigador, es capaz de generar nuevas fuentes de información y contribuir a superar la crisis actual.

Desde otra área como la literatura, uno de los textos de referencia escri-

to por Jockers (2013) define de manera somera el análisis de texto como un conjunto de métodos que combina el uso computacional y cuantitativo en grandes corpora de textos literarios (Jockers, 2013). Destacó, la importancia del análisis exploratorio y cuantitativo, como conocer la distribución de palabras, la medición y uso de modelos, así como la transformación de los textos en datos.

Por su parte, en el ámbito de las ciencias sociales y política, Grimmer et al. (2022) aborda el análisis de texto como un conjunto de métodos para estudiar grandes volumenes tratando las palabras como datos cuantitativos, permitiendo un análisis sistemático con el fin de estudiar el comportamiento político, las actitudes y las instituciones (Grimmer et al., 2022). Desde está perspectica, el autor muestra la capacidad de este enfoque para producir evidencia empírica a partir del lenguaje.

De estas definiciones genéricas, que solo ubican el análisis de texto en un plano general, podemos rescatar no solo su importancia en esta nueva era de la inteligencia artificial —marcada por la demanda de más fuentes de información—, sino también su carácter técnico, sujeto a distintos enfoques para una implementación adecuada. Esto nos lleva a considerar los trabajos de Jurafsky y Martin (2023), quienes se suman a este debate al presentar las tareas que sustentan el análisis de texto, como la clasificación de textos, la extracción de información (IE), el reconocimiento de entidades (NER), el análisis de sentimiento y el análisis de tópicos entre otros (Jurafsky & Martin, 2023).

Estas definiciones, aunque provienen de diferentes disciplinas, coinciden en que el análisis de textos es una herramienta metodológica poderosa para convertir información lingüística en datos útiles, ya sea con fines descriptivos, explicativos o predictivos. Esto a su vez ofrece un cambio de paradigma en la forma en qué se producen estadísticas oficiales, se reportan los eventos

socioeconómicos y sobre todo en la toma de decisiones.

Para eso, la investigación se basa en 3 pilares fundamentales para lograr los objetivos.

1. La teoria económica subyacente que orienta el análisis de texto, así como, los recursos cómputo-lingüísticos.

2. El marco teorico que constituye los recursos cómputo-lingüísticos, lo que implica definir la combinación óptima de los métodos computacionales y lingüísticos.

3. Las técnicas de procesamiento de lenguaje necesarios.

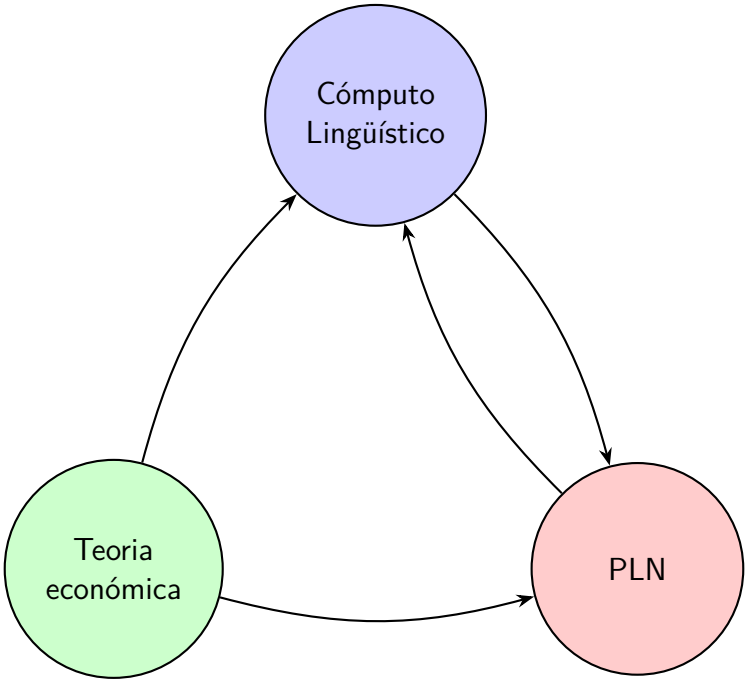

Figura 2.1: Marco teórico conceptual de la tesis (elaboración propia)

Lo anterior se puede visualizar a través de la gráfica anterior (véase Figura 2.1). Consideramos que el marco teórico es fundamental ya que

guiará toda la investigación mostrando los alcances, delimitación téorica y metodológica.

Primero, nos permite entender ¿qué entidades, frases, párrafos o documentos? son importantes para captar las relaciones léxicas y/o semánticas del objeto de estudio.

Por otro lado, los recursos cómputo-lingüísticos se refieren a los algoritmos y/o métodos computacionales necesarios para procesar, analizar y modelar el lenguaje natural. Esto abarca una variedad de formas lingüísticas como la semántica, morfología, entre otros. En cuanto a los métodos computacionales se deben considerar los tipos de algoritmos y/o enfoques de aprendizaje necesario.

2.1. Conceptos básicos

Ahora que contamos con el marco teórico que guiará la investigación, estaremos presentando los conceptos que constituyen cada pilar, así como sus principales enfoques. Empezaremos con la teoría económica, definiendo los conceptos clave, seguido por los métodos computacionales necesarios para el análisis de los textos económicos, finalmente, los modelos.

2.1.1. Teoría económica

La economía es la ciencia que estudia cómo las sociedades utilizan recursos escasos para producir bienes y servicios y distribuirlos entre sus miembros. Se divide en dos ramas principales: la microeconomía, que analiza el comportamiento de individuos y empresas, y la macroeconomía, que estudia la economía en su conjunto, incluyendo temas como el crecimiento económico, la inflación y el desempleo (Mankiw, 2015).

Entre los indicadores económicos más relevantes para la sociedad se

encuentran el Producto Interno Bruto (PIB que mide la producción total de bienes y servicios), el desempleo que refleja la cantidad de personas en búsqueda de empleo, las exportaciones e importaciones que indican el grado de integración o dependencia de una economía con respecto al exterior, la inflación que refleja el aumento generalizado de los precios, entre otros. Estos indicadores son esenciales para la toma de decisiones para la planeación económica.

En esta investigación, nos centraremos en la inflación, un fenómeno económico que afecta a todos los sectores de la economía y que refleja el costo de vida de las personas (Tang & Lei, 2023). Se define como el aumento generalizado y sostenido de los precios de bienes y servicios en un país durante un período de tiempo. Se mide comúnmente a través del INPC en México.

De esta definición tan general de la inflación, se puede observar diferentes posturas teóricas que buscan explicar sus causas y efectos. Por ejemplo, la teoría monetarista sostiene que la inflación es causada por un exceso de oferta monetaria en relación con la demanda de bienes y servicios, es decir que los bancos centrales tienden a imprimir más dinero del que la economía puede absorber[1].

Por otro lado, los keynesianos argumentan que la inflación puede ser causada por un aumento en la demanda agregada, lo que significa que los agentes económicos (consumidores y empresas) tienden a gastar más, infligiendo una presión al alza de los precios. Desde esta perspectiva, también puede ser causada por un aumento en los costos de producción, como los salarios o los precios de las materias primas.

Por último, se identifica la inflación estructural, desarrollada por economistas como Julio H. G. Olivera y Aldo Ferrer en América Latina, que

[1]Exploring Economics: La inflación en la teoría económica, consultado el 28 de septiembre 2025

sostiene que la inflación es causada por problemas estructurales en la economía, como la falta de competencia en ciertos sectores, la rigidez del mercado laboral y la dependencia de las importaciones (Gutiérrez Andrade, 2006). Como podemos observar, independientemente de las causas, la inflación suele tener efectos negativos en la economía y en la sociedad.

Las causas pueden ser diversas, desde un aumento en la demanda de bienes y servicios, es decir que los agentes económicos tienden a comprar más de lo que se produce, como también en los costos de producción, los salarios, los precios de las materias primas, entre otros. Esta diversidad de causas implica que los efectos pueden ser variados, desde la pérdida del poder adquisitivo de los agentes económicos, la incertidumbre para los inversionistas, hasta agudizar las desigualdades sociales.

Revisamos algunas relaciones entre la inflación y las variables económicas antes de presentar su impacto en la sociedad. Por ejemplo, la inflación tiende a afectar el consumo de los hogares, ya que un aumento en los precios reduce el poder adquisitivo de las personas, lo que puede llevar a una disminución en el gasto de consumo. Esto a su vez puede afectar la producción y el empleo en la economía.

También, puede afectar las tasas de interés, desanimando la inversión y el crecimiento económico, ya que los bancos centrales tienden a aumentar las tasas para controlar la inflación. Por otro lado, el comercio internacional también puede verse afectado, ya que una alta inflación puede provocar que los productos nacionales sean menos competitivos en el mercado global.

Especialmente en los últimos años, debido a la pandemia del COVID-19, la inflación se ha generalizado en muchas economías del mundo, incluyendo México. Según El País(2022), la inflación es tan global como la COVID-19 debido a las políticas subsidiarias de los gobiernos para proteger a los más

desfavorecidos por la pandemia[2].

Numerosas fuentes periodísticas documentaron varios disturbios a causa de la inflación en el mundo, por ejemplo, la agencia Reuters reportó en 2024 las protestas por el costo de vida que causaron 3 muertos en el estado de Kaduna en Nigeria a causa de un aumento en el costo de vida[3]. El País(2024), por su lado, documentó al menos 39 personas muertas y más de 300 heridas en Kenia por movilización contra la subida de los impuestos[4]. En México, hemos sido testigos de actos de saqueos y vandálicos en 2017 por el aumento del precio de la gasolina, conocido como el *gasolinazo*[5]. Así, podemos relatar un sinnúmero de casos que muestran el impacto que puede causar la inflación en la sociedad.

Bajo esta perspectiva, la inflación ha tomado un lugar central en los medios de comunicación, los debates políticos y en los textos de opinión buscando explicar no sólo sus causas, sino también sus efectos. Es por eso que las instituciones financieras como el Banco de México tienden a monitorear de cerca este fenómeno económico.

La presencia de la inflación en los medios constituye una fuente de información relevante para entender su evolución y sus efectos en la economía. ¿Tendrá la cobertura mediática suficiente para inferir la inflación? ¿Qué tanto la secuencia de los tópicos económicos nos permiten inferir la tendencia de la inflación? Estas son algunas preguntas que no estamos en capacidad de responder, pero que constituyen el eje central de esta investigación; ya que la construcción del corpus contempla directamente las noticias económicas relacionadas con la inflación y temas afines.

[2]El País: La escalada de la inflación asfixia a América, consultado el 27 de septiembre del 2025

[3]Reuters: At least 3 killed in Nigeria at protests over high cost of living, consultado el 27 de septiembre del 2025

[4]El País: Las manifestaciones por una subida de impuestos en Kenia derivan en una protesta contra el presidente

[5]El País: Saqueos y actos vandálicos por el alza del precio de la gasolina

A la luz de lo anterior, me gustaría definir algunos conceptos clave que serán utilizados a lo largo de la tesis:

- **Bienes y servicios:** Son los productos que se consumen en una economía. Los bienes son tangibles, como alimentos, ropa, casa, mientras que los servicios son intangibles, como educación, salud y transporte.

- **Canasa básica:** Es la composición de bienes y servicios que una familia promedio consume en un período determinado. Se utiliza para medir la inflación y el costo de vida.

- **Consumo:** Se refiere a la cantidad de bienes y servicios que los hogares adquieren para satisfacer sus necesidades. Es importante para medir el nivel de actividad económica y el bienestar de la población.

- **Demanda agregada:** Es la cantidad total de bienes y servicios que los agentes econónomicos (consumidores, empresas, gobierno y sector externo) están dispuestos a comprar a un nivel de precios determinado.

- **Oferta agregada:** Es la cantidad total de bienes y servicios que las empresas están dispuestas a ofrecer a un nivel de precios dado los factores de producción disponibles.

- **Política monetaria:** Comprenden todas las acciones que desde el Banco Central se llevan a cabo para controlar la oferta monetaria y las tasas de interés con el fin de lograr objetivos macroeconómicos como la estabilidad de precios, el crecimiento económico y el empleo.

- **Empleo:** Se refiere a la cantidad de personas que tienen un trabajo remunerado en la economía. Es un indicador importante del bienestar económico y social.

- **Tasa de interés:** Es el costo del dinero, dicho de otra manera, es el precio que se paga por el uso del dinero y se considera un instrumento fundamental en la política monetaria.

En la siguiente sección, revisamos los principales conceptos relacionados con los recursos cómputo-lingüísticos necesarios para el análisis de textos económicos.

2.1.2. Recursos cómputo-lingüísticos

Los primeros trabajos relacionados con análsis de texto remontan desde los años 50 como un sub-área de la inteligencia artificial y la lingüística computacional con el principal objetivo de estudiar los problemas derivados de la generación automática de contenidos y del entendimiento del lenguaje natural[6].

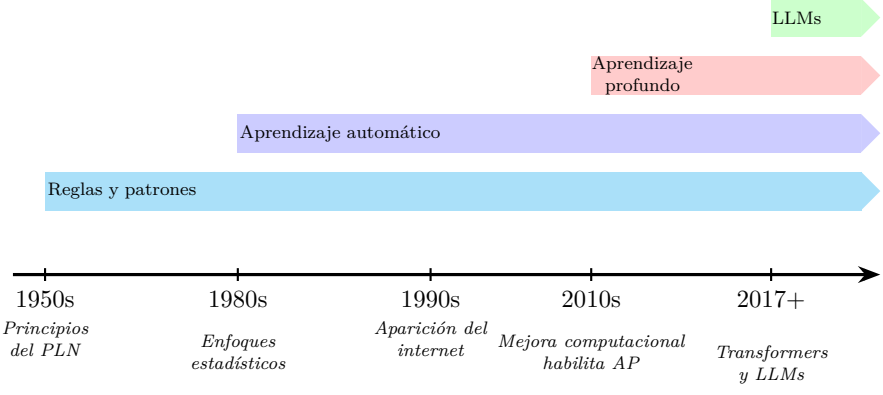

Figura 2.2: Línea de tiempo del PLN con diferentes de enfoque (Fuente: Kochmar, 2022)

Durante este periodo, se observa que el análisis de texto fue sujeto a diversos paradigmas, cada uno con sus fortalezas y limitaciones. A continuación, revisamos los enfoques más relevantes en la historia del PLN que

[6]Medium: History and present of Natural Language Processing, consultado el 17 de abril 2025

van desde los métodos basados en reglas y patrones hasta los modelos de lenguaje.

2.1.3. Reglas y patrones

Los enfoques basados en reglas y reconocimiento de patrones sentaron las bases del análisis de texto. Los pioneros en PLN de los años 50 y 60 se centraron en la construcción manual de reglas gramaticales y diccionarios para el análisis sintáctico y el reconocimiento de patrones léxicos. Revisamos los principales hitos que marcaron el desarrollo de este enfoque:

- **1950s–1960s**: Gramáticas formales inspiradas en la lingüística generativa de Chomsky; surgimiento de primeros analizadores sintácticos basados en gramáticas libres de contexto (Jurafsky & Martin, 2023).
- **1966**: ELIZA, un programa computacional que establece la interacción humano-computadora desarrollado por Weizenbaum, hace uso de patrones y sustitución de cadenas para simular diálogo terapéutico, demostrando el potencial de este enfoque (Weizenbaum, 1966).
- **1970s–1980s**: La incorporación de expresiones regulares y autómatas finitos permitió avances significativos en tareas de tokenización y segmentación de texto. Además, se desarrollaron sistemas basados en Prolog para la producción de reglas, lo que facilitó la creación de modelos más complejos (Kochmar, 2022).

Estos enfoques alcanzan un alto nivel de precisión al ser utilizados en dominios específicos, ya que su estructura está cuidadosamente diseñada para capturar las características particulares de los contextos en los que se aplican. Además, generan resultados predecibles, lo que asegura que, al ingresar los mismos datos, siempre se obtendrá el mismo resultado, lo cual facilita la transparencia y la trazabilidad en todo el proceso.

No obstante, presentan ciertas restricciones en términos de flexibilidad, ya que es complicado adaptarlos a nuevos dominios o contextos que cambian. Además, el mantenimiento de estos sistemas puede resultar costoso y complicado, especialmente cuando las necesidades evolucionan o surgen nuevas circunstancias que requieren la modificación o reescritura de numerosas reglas.

2.1.3.1. Ejemplo de aplicación

A continuación se presentan ejemplos representativos de aplicación en sistemas de Traducción Automática Basados en Reglas (RBMT, por su sigla en Inglés) documentados por (Hutchins, 1986; Somers, 2003):

- **Extracción de Información (EI):** Los primeros enfoques para la extracción de información, como los que se desarrollaron en el marco de la Message Understanding Conference (MUC) en la década de 1990, utilizaban reglas lingüísticas explícitas para identificar entidades nombradas, eventos y relaciones en textos noticiosos y documentos oficiales (Grishman, 1997).

- **Análisis morfosintáctico y semántico**: Los etiquetadores gramaticales basados en reglas, como el propuesto por Brill (1995), asignaban categorías gramaticales a las palabras utilizando un conjunto de reglas sintácticas y léxicas, lo cual resultaba especialmente útil cuando los datos de entrenamiento eran limitados. En este proceso, se realiza un análisis detallado de la estructura de la oración para extraer información léxica, categorías gramaticales y relaciones sintácticas. En este punto, se suelen emplear diccionarios y analizadores morfológicos avanzados para resolver ambigüedades en los homógrafos y establecer las dependencias sintácticas correspondientes (Kipper

et al., 2005).

- **Traducción automática:** Los primeros sistemas de traducción automática utilizaban reglas estructurales y léxicas para realizar traducciones directas entre idiomas, apoyándose en diccionarios y transformaciones sintácticas previamente definidas (Hutchins, 1986). Además, mediante un conjunto de reglas formales, se transformaba la representación abstracta del texto original en una estructura equivalente en el idioma de destino. Estas reglas operaban a nivel de sintagmas, cláusulas o relaciones semánticas, manteniendo los roles temáticos y el orden canónico de los constituyentes (Arnold et al., 1994).

2.1.4. Enfoque estadístico

Con la explosión de datos textuales y mayores capacidades de cómputo, surgieron técnicas basadas en modelos probabilísticos y de conteo, como los modelos de n-gramas, bolsa de palabras (BoW) y TF–IDF(Kochmar, 2022).

Estos enfoques cuantificaban la frecuencia de términos y su relevancia, lo cual mejoró tareas de recuperación de información y clasificación básica. No obstante, seguían sin capturar adecuadamente la semántica ni las dependencias a largo plazo en el texto, y su efectividad decaía en presencia de palabras polisémicas o sintagmas complejos.

A principio de los años 80, se empieza a apreciar el uso de los métodos basado en estadística como una alternativa al enfoque de reglas. Esto ocurrió en un contexto de desarrolllo de mejores capacidades computacionales y mayor disponibilidad de grandes corpus.

A diferencia de los métodos basados en reglas, los métodos estadísticos eviten plantear hipotesis sobre el corpus, el análisis del lenguaje está sujeto

a la modelización de distribuciones de probabilidad y la observación de patrones de frecuencia donde se aprende lo bueno y lo malo del corpus.

Los algoritmos más frecuentes suelen ser los modelos de n-gramas, el análisis de co-ocurrencia de palabras y métodos como Naïve Bayes y regresión logística (Manning et al., 2008). Estos modelos sentaron las bases para muchos desarrollos posteriores al permitir que las máquinas "aprendieran" directamente de los datos lingüísticos sin una codificación manual.

Algunos ejemplos de algoritmos y modelos estadísticos de este enfoque pueden ser:

- Modelo de bolsa de palabras (BoW)
- Modelos n-grama
- TF-IDF (Term Frequency–Inverse Document Frequency)
- Modelos ocultos de Markov (HMM)
- Modelos de tópicos como LDA (Latent Dirichlet Allocation)

2.1.4.1. Modelo de Bolsa de Palabras

El *Bag of Words* (BoW) es un modelo clásico y de implementación no compleja para la representación vectorial de documentos textuales en tareas de minería de texto y recuperación de información. Bajo este enfoque, cada documento se reduce a una "bolsa" de términos, esto es, a un vector de conteos o ponderaciones que ignora por completo el orden y la estructura sintáctica de las palabras (Manning et al., 2008).

A pesar de su aparente simplicidad, BoW constituye la base de métodos más avanzados como el TF–IDF y variantes ponderadas, y funciona sorprendentemente bien con clasificadores lineales y modelos de aprendizaje supervisado.

Sea un corpus

$$\mathcal{D} = \{\, d_1, \ldots, d_N \,\} \quad \text{y un vocabulario} \quad V = \{\, w_1, \ldots, w_{|V|} \,\}.$$

Entonces cada documento d_i se codifica como el vector

$$\mathbf{x}_i = \big(x_{i1}, x_{i2}, \ldots, x_{i|V|}\big)^{\mathsf{T}} \in \mathbb{N}^{|V|}, \quad x_{ij} = \mathrm{freq}\big(w_j, d_i\big),$$

donde $\mathrm{freq}(w_j, d_i)$ es el número de apariciones del término w_j en d_i. Para mitigar el sesgo hacia documentos largos, a menudo se normaliza:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sum_{k=1}^{|V|} x_{ik}} \quad \text{o bien} \quad \tilde{x}_{ij} = x_{ij} \times \log \frac{N}{\mathrm{df}(w_j)},$$

introduciendo así, el componente de *frecuencia inversa de documento* (IDF).

En resumen, el método de BoW destaca por su simplicidad y eficiencia al transformar texto en vectores dispersos, su amplia compatibilidad con modelos lineales y basados en conteo, así como su papel fundamental como base para técnicas avanzadas como TF–IDF, hashing y embeddings de co-ocurrencia.

Sin embargo, presenta limitaciones relevantes, pues ignora el orden y contexto del texto, lo que impide modelar fenómenos como la negación o expresiones idiomáticas; además, su alta dimensionalidad requiere técnicas de reducción, y no captura la polisemia ni relaciones semánticas profundas, ya que trata cada palabra como independiente.

2.1.4.2. Frecuencia Inversa de Documentos (TF–IDF)

El enfoque basado en *Term Frequency–Inverse Document Frequency* (TF–IDF) extiende el modelo BoW incorporando conocimiento global del corpus

para penalizar términos muy frecuentes y favorecer aquellos más distintivos, mejorando la representatividad y discriminación de los vectores de características (Salton & Buckley, 1988).

Sea un corpus

$$\mathcal{D} = \{d_1, \ldots, d_N\}, \quad V = \{w_1, \ldots, w_{|V|}\}$$

y un documento d_i.

Definimos:

$$tf_{ij} = \frac{f_{ij}}{\sum_{k=1}^{|V|} f_{ik}}, \quad idf_j = \log\left(\frac{N}{df_j}\right),$$

donde, f_{ij} es la frecuencia absoluta de w_j en d_i y df_j el número de documentos en que aparece.

El peso TF–IDF se obtiene como:

$$\text{tfidf}_{ij} = tf_{ij} \times idf_j.$$

Para mayor robustez, se emplean variantes:

- *TF sublineal*: $tf'_{ij} = 1 + \log(f_{ij})$.
- *Suavizado de IDF*: $idf'_j = \log\big((N+1)/(df_j+1)\big) + 1$.
- *Normalización* (L2) del vector $\mathbf{t}_i$ para aplicar similitud de coseno.

El método TF–IDF suele ser usado ampliamente en minería de texto dada su capacidad para mejorar la discriminación de términos relevantes al atenuar el impacto de palabras vacías y resaltar términos informativos.

Presenta flexibilidad mediante diversos esquemas de normalización adaptables a distintos dominios, y su compatibilidad lo hace indispensable como insumo para motores de búsqueda, agrupamiento y clasificación de texto (Manning et al., 2008; Salton & Buckley, 1988).

Sin embargo, TF–IDF tiene limitaciones importantes. No captura semántica profunda, pues solo mide frecuencias y no distingue sentidos ni relaciones entre palabras. Además, genera vectores de alta dimensionalidad y esparsidad, lo que puede requerir técnicas de reducción o hashing.

Es sensible al ruido y a términos poco frecuentes o con errores ortográficos, y no modela el contexto ni el orden de las palabras, ignorando dependencias sintácticas y semánticas dentro del texto (Manning et al., 2008).

A pesar de estas limitaciones, TF–IDF sigue siendo fundamental en aplicaciones como la recuperación de información, donde se utiliza para ponderar y clasificar documentos, en algoritmos de ranking basados en similitud coseno, y como entrada principal para clasificadores supervisados y análisis de sentimiento en minería de texto.

2.1.4.3. Modelos Ocultos de Markov

Los Modelos Ocultos de Markov (HMM) son modelos probabilísticos generativos que describen secuencias observadas a partir de un proceso subyacente que permanece oculto, el cual está determinado por una cadena de Markov (Rabiner, 1989).

Está compuesto por lo siguiente:

- N: cantidad de estados ocultos $\mathcal{S} = \{s_1, \ldots, s_N\}$.
- M: número de símbolos observables (p. ej., palabras).
- $\pi_i = P(q_1 = s_i)$: parametros de distribución inicial de estados.
- $A = [a_{ij}]$: matriz de transiciones, $a_{ij} = P(q_{t+1} = s_j \mid q_t = s_i)$.
- $B = [b_j(k)]$: probabilidades de emisión, $b_j(k) = P(o_t = v_k \mid q_t = s_j)$.

La probabilidad conjunta de una secuencia observable $\boldsymbol{o} = (o_1, \ldots, o_T)$ y una trayectoria de estados $\boldsymbol{q} = (q_1, \ldots, q_T)$ viene dada por:

$$P(\boldsymbol{o}, \boldsymbol{q}) = \pi_{q_1}\, b_{q_1}(o_1) \prod_{t=2}^{T} a_{q_{t-1},q_t}\, b_{q_t}(o_t). \tag{2.1}$$

Los HMM son fundamentales en el modelado de secuencias temporales, permitiendo capturar dependencias probabilísticas entre estados ocultos y observaciones. Su estructura probabilística facilita la evaluación de la probabilidad de una secuencia observada mediante el algoritmo hacia adelante, la decodificación de la secuencia de estados más probable con el algoritmo de Viterbi, y el entrenamiento eficiente de parámetros mediante el algoritmo de Baum–Welch basado en Expectation-Maximization (Rabiner, 1989).

A pesar de su gran utilidad, los HMM tienen ciertas limitaciones, como la suposición de independencia condicional restringida y su incapacidad para capturar dependencias a largo plazo o características contextuales complejas, lo cual puede afectar su rendimiento en tareas con estructuras lingüísticas más ricas y variadas (Jurafsky & Martin, 2009).

Sin embargo, su uso en el procesamiento de lenguaje natural (PLN) sigue siendo extenso y eficiente. Los HMM se aplican con éxito en tareas como el etiquetado de partes del discurso (POS tagging), el reconocimiento de entidades nombradas (NER) y la segmentación de oraciones y palabras, donde su equilibrio entre eficiencia y precisión lo mantiene como un enfoque clásico y la base para modelos más sofisticados (Jurafsky & Martin, 2009; Rabiner, 1989).

2.1.4.4. Latent Dirichlet Allocation (LDA)

El LDA toma un enfoque probabilístico generativo introducido por Blei et al. (2003), diseñado para identificar tópicos subyacentes en vastas coleccio-

nes de documentos. Su principal objetivo es modelar cada documento como una combinación de diferentes tópicos y cada tema como una distribución de palabras, facilitando la extracción de estructuras semánticas implícitas en un conjunto de documentos.

Se asume que:

- Cada documento está compuesto por varios tópicos.
- Cada tópicos está representado por una distribución probabilística sobre un conjunto fijo de palabras.
- Las palabras observadas en un documento son generadas seleccionando un tópico latente y luego una palabra de este mismo.

Este enfoque permite representar documentos en un espacio de baja dimensión, donde cada una corresponde a un tópico inferido automáticamente, lo cual resulta útil en tareas de minería de texto, reducción de dimensionalidad y análisis exploratorio.

Sea:

- D el número de documentos.
- K el número de tópicos.
- V el tamaño del vocabulario.
- $\boldsymbol{w}_d = (w_{d1}, \ldots, w_{dN_d})$ el conjunto de palabras del documento d.

LDA define el siguiente proceso generativo para cada documento $d \in \{1, \ldots, D\}$:

1. Elegir una distribución de tópicos $\boldsymbol{\theta}_d \sim \text{Dirichlet}(\boldsymbol{\alpha})$
2. Para cada palabra w_{dn} en el documento:

a) Elegir un tema $z_{dn} \sim \text{Multinomial}(\boldsymbol{\theta}_d)$

b) Elegir una palabra $w_{dn} \sim \text{Multinomial}(\boldsymbol{\beta}_{z_{dn}})$

Donde:

- $\boldsymbol{\theta}_d$ es la distribución de temas para el documento d.
- z_{dn} es el tema asignado a la palabra w_{dn}.
- $\boldsymbol{\beta}_k$ es la distribución de palabras asociada al tema k.
- $\boldsymbol{\alpha}$ es el parámetro de la distribución Dirichlet sobre temas.
- $\boldsymbol{\eta}$ es el parámetro de la Dirichlet sobre palabras (a veces denotado $\boldsymbol{\beta}$ en la literatura).

La probabilidad conjunta del modelo LDA para un documento d y sus variables latentes es:

$$p(\boldsymbol{w}_d, \boldsymbol{z}_d, \boldsymbol{\theta}_d \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) \prod_{n=1}^{N_d} p(z_{dn} \mid \boldsymbol{\theta}_d)\, p(w_{dn} \mid z_{dn}, \boldsymbol{\beta}) \tag{2.2}$$

Y para el corpus completo $\mathcal{D}$ de D documentos:

$$p(\mathcal{D} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{d=1}^{D} \int p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \boldsymbol{\theta}_d)\, p(w_{dn} \mid z_{dn}, \boldsymbol{\beta}) \right) d\boldsymbol{\theta}_d \tag{2.3}$$

El objetivo es calcular la distribución posterior de los temas y sus asignaciones:

$$p(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \tag{2.4}$$

Dado que esta distribución es intratable de forma exacta, se utilizan métodos aproximados como:

- Inferencia variacional
- Muestreo colapsado de Gibbs

2.1.4.5. Aplicaciones de LDA en minería de texto

LDA ha sido ampliamente utilizado en tareas como:

- Agrupamiento de tópicos en documentos
- Visualización de grandes corpus de texto mediante reducción de dimensionalidad
- Extracción de tópicos en noticias, artículos científicos, discursos y redes sociales
- Representación semántica de textos para tareas de clasificación o recomendación

En el contexto económico, LDA permite extraer temas latentes relacionados con inflación, crecimiento, política monetaria u otros fenómenos, lo que facilita la detección de patrones en noticias o reportes institucionales.

2.1.5. Aprendizaje Automático

2.1.5.1. Enfoque clásico (1995–2012)

La disponibilidad de grandes corpus y bibliotecas de ML impulsó la adopción de algoritmos supervisados como Naïve Bayes, Máquinas de Vectores de Soporte (SVM) y árboles de decisión. Joachims demostró la eficacia de SVM para categorización de texto con miles de características relevantes (Joachims, 1998).

Estos modelos superaron ampliamente a los métodos puramente estadísticos en clasificación y filtrado de documentos, pero dependían de ingeniería de características manual (n-gramas, selecciones léxicas) y no aprendían representaciones profundas del lenguaje.

El aprendizaje automático (ML) en PLN busca inferir patrones complejos en textos a partir de características extraídas automáticamente. A diferencia de los modelos puramente estadísticos, ML permite clasificar, agrupar o predecir etiquetas a partir de entrenamiento supervisado o no supervisado.

El auge de SVMs, árboles de decisión y redes neuronales simples marcó una nueva etapa en tareas como análisis de sentimientos, clasificación de noticias o reconocimiento de entidades nombradas (Cambria et al., 2017).

Ejemplos de algoritmos clásicos de ML:

- Support Vector Machines (SVM)
- Naïve Bayes
- Árboles de decisión y Random Forest
- k-Nearest Neighbors (k-NN)
- K-means y agrupamiento jerárquico
- Redes neuronales feed-forward simples

2.1.5.2. Naïve Bayes

Naïve Bayes es un clasificador probabilístico basado en el teorema de Bayes con la suposición de independencia condicional entre características (Rish, 2001). Dado un vector de características $\mathbf{x} = (x_1, \ldots, x_n)$ y un conjunto de clases $\{C_k\}$.

La probabilidad posterior se calcula de la siguiente manera:

$$P(C_k \mid \mathbf{x}) = \frac{P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)}{P(\mathbf{x})}. \tag{2.5}$$

Puesto que $P(\mathbf{x})$ es constante para todas las clases, la decisión se reduce a:

$$\hat{C} = \arg\max_{C_k} \; P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k).$$

También se puede observar en la siguiente gráfica 2.3 una situación ideal de clasificación basada en dicho método.

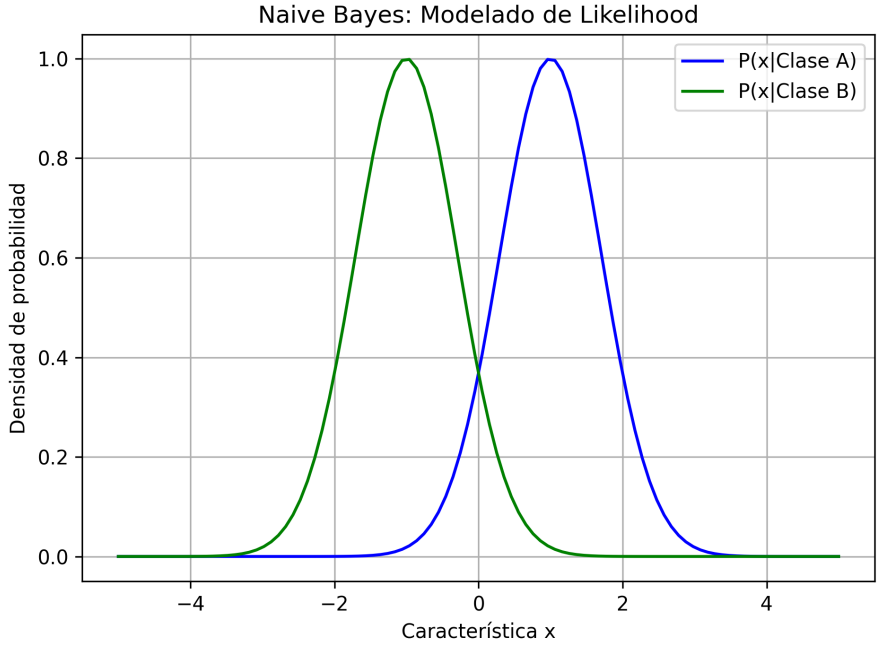

Figura 2.3: Ejemplo de clasificación con Naïve Bayes (elaboración propia)

Existen diferentes enfoques del método de NB, como se puede apreciar la combinación de estas distribuciones probabilísticas como:

- **Gaussian NB:** diseñado para atributos continuos, asume que $P(x_i \mid C_k)$ sigue una distribución normal.

- **Multinomial NB:** ideal para conteos de características (p. ej., frecuencia de palabras en documentos).
- **Bernoulli NB:** modela características binarias (presencia o ausencia de un término).

En general, el uso que se suele dar al método abarca lo siguiente:

- **Clasificación de texto:** análisis de sentimientos, categorización temática.
- **Filtrado de correo:** detección de correo no deseado.
- **Diagnóstico médico:** predicción de enfermedades a partir de síntomas discretos.
- **Detección de fraude:** análisis de patrones de transacciones.

El clasificador Naïve Bayes es ampliamente valorado por su simplicidad y eficiencia, ya que requiere pocos datos para entrenamiento y no es tan computacionalmente costoso, además de ser escalable a grandes volúmenes de datos. Su desempeño es robusto con datos dispersos como los obtenidos mediante Bolsa de Palabras o TF–IDF, lo que lo hace una opción popular en tareas de clasificación de texto (McCallum & Nigam, 1998; Sebastiani, 2002).

No obstante, su principal limitación radica en la suposición de independencia condicional entre características, que rara vez se cumple en escenarios reales, afectando su precisión. Además, es sensible a probabilidades cero, por lo que requiere técnicas de suavizado como Laplace. En comparación con métodos discriminativos, suele presentar menor exactitud cuando las características están correlacionadas (Rish, 2001).

A pesar de estas limitaciones, Naïve Bayes continúa siendo una herramienta eficaz para tareas de clasificación rápida y escalable, especialmente en sistemas donde la interpretabilidad y la velocidad son prioritarias, como en filtrado de spam, categorización de documentos y análisis preliminar de sentimientos.

2.1.5.3. Máquinas de Vectores de Soporte

Las *Support Vector Machines* (SVM) son clasificadores supervisados basados en la teoría de minimización del riesgo estructural, que buscan el hiperplano óptimo que separa dos clases con el mayor margen posible (Cortes & Vapnik, 1995).

Dado un conjunto de entrenamiento:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}, \quad \mathbf{x}_i \in \mathbb{R}^d,\ y_i \in \{-1, 1\},$$

la SVM resuelve el problema minimización de los márgenes duros de la siguiente forma:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \tfrac{1}{2}\|\mathbf{w}\|^2 \\ \text{sujeto a} \quad & y_i\big(\mathbf{w}^\top \mathbf{x}_i + b\big) \geq 1, \quad i = 1, \ldots, n. \end{aligned} \tag{2.6}$$

Para datos no separables, se introducen variables de holgura $\xi_i \geq 0$ y un parámetro de penalización $C > 0$:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \\ \text{sujeto a} \quad & y_i\big(\mathbf{w}^\top \mathbf{x}_i + b\big) \geq 1 - \xi_i,\ \xi_i \geq 0. \end{aligned} \tag{2.7}$$

Además, mediante un *kernel* $K(\mathbf{x}_i, \mathbf{x}_j)$ (p. ej., RBF, polinomial), la SVM proyecta los datos a espacios de mayor dimensión sin calcular explí-

citamente la transformación.

En la siguiente gráfica 2.4 se puede apreciar una solución basada en SVM.

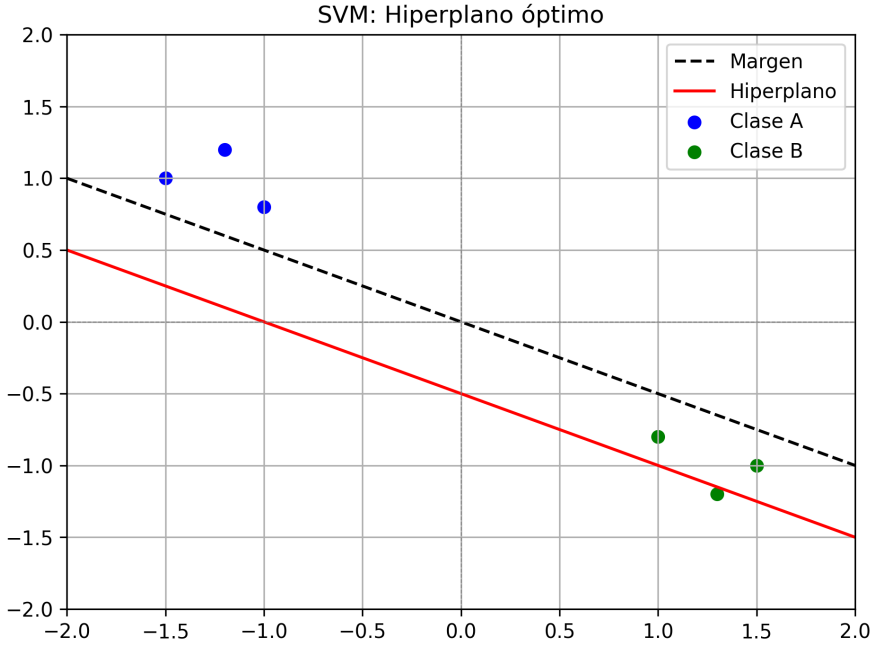

Figura 2.4: Ejemplo de clasificación basado en SVM (elaboración propia)

Las máquinas de vectores de soporte (SVM) son algoritmos de aprendizaje supervisado ampliamente utilizados por su eficacia en espacios de alta dimensión y su flexibilidad para modelar relaciones complejas mediante funciones kernel.

Su sólida base teórica en la teoría de la generalización garantiza buenos resultados en problemas de clasificación, siendo aplicadas exitosamente en áreas como clasificación de texto (spam, análisis de sentimientos, categorización temática), reconocimiento de voz e imágenes, bioinformática y detección de anomalías en transacciones y redes (Cortes & Vapnik, 1995; Schölkopf & Smola, 2002).

Entre sus limitaciones, el entrenamiento puede ser costoso en conjuntos

de datos muy grandes debido a su complejidad computacional, además la selección y ajuste de parámetros críticos como el margen C, el tipo de kernel y el parámetro γ requiere cuidadosa validación para evitar sobreajuste o bajoajuste. También, de manera nativa, las SVM no proporcionan probabilidades de clasificación, lo que puede ser una desventaja en aplicaciones que requieren una estimación probabilística directa (Burges, 1998; Hastie et al., 2009).

A pesar de estas limitaciones, las SVM siguen siendo una herramienta potente y versátil en aprendizaje automático, especialmente en dominios donde la precisión y la interpretabilidad de la frontera de decisión son prioritarias.

2.1.5.4. Árboles de Decisión

Los árboles de decisión segmentan el espacio de entrada mediante reglas de decisión jerárquicas, particionando recursivamente los datos para maximizar una medida de pureza como la entropía o el índice de Gini. Cada nodo interno corresponde a una prueba sobre una característica, y cada hoja asigna una etiqueta de clase o un valor de regresión (Breiman et al., 1984).

Una medida de impureza muy utilizada es el índice de Gini:

$$Gini = 1 \; - \; \sum_{i=1}^{C} p_i^2, \tag{2.8}$$

donde p_i es la proporción de ejemplos de la clase i en el nodo. El algoritmo selecciona en cada paso la división que maximiza la reducción de impureza.

En la siguiente gráfica 2.5 podemos apreciar la representación de una solución basada en árbol de decisión.

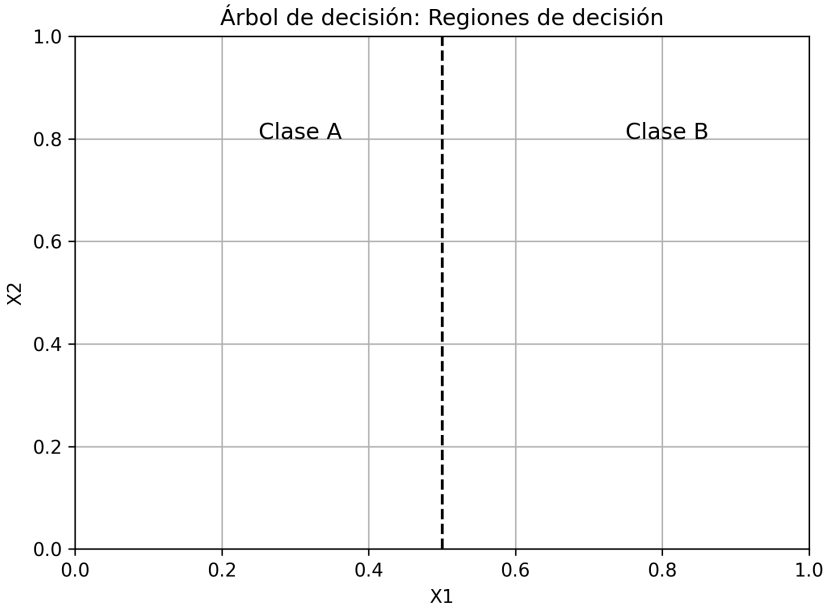

Figura 2.5: Ejemplo de clasificación basado en árbol de decisión (elaboración propia)

2.1.5.5. Random Forest

Random Forest consiste en un conjunto de árboles de decisión entrenados sobre muestras bootstrap del conjunto de datos y seleccionando aleatoriamente subconjuntos de características en cada división. Luego, combina sus predicciones por votación (clasificación) o promedio (regresión) (Breiman, 2001), reduciendo varianza y mejorando la generalización.

A continuación se muestra de manera gráfica 2.6 una representación basada en Random Forest.

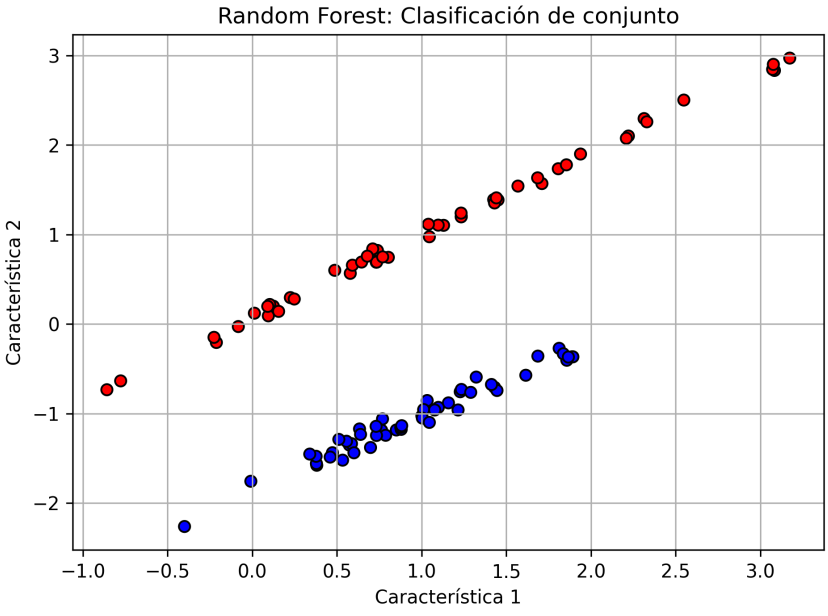

Figura 2.6: Ejemplo gráfico de clasificación con Random Forest (elaboración propia)

Los algoritmos de Random Forest combinan múltiples árboles de decisión para mejorar la precisión y robustez en tareas de clasificación y regresión. Son ampliamente aplicados en la categorización de texto, como la detección de spam y clasificación de noticias, así como en predicciones continuas en tareas de regresión.

En bioinformática, se emplean para la selección de características genómicas, y en detección de fraudes y anomalías en finanzas y redes, donde su capacidad para manejar datos complejos y variados es especialmente valiosa (Breiman, 2001; Liaw & Wiener, 2002).

Entre sus ventajas, destacan la interpretabilidad de árboles individuales, el manejo eficiente de datos mixtos numéricos y categóricos, y su alta precisión combinada con una robustez significativa frente al sobreajuste. Sin embargo, los árboles individuales pueden sobreajustar sin técnicas de poda adecuadas.

En conjunto, el Random Forest sacrifica parte de la interpretabilidad global debido a la agregación de muchos árboles y presenta un mayor costo computacional. Además, su desempeño puede verse afectado por conjuntos de datos desequilibrados si no se ajustan los pesos o técnicas de balanceo (Balle et al., 2016; Cutler et al., 2007).

A pesar de estas limitaciones, Random Forest sigue siendo un método versátil y potente, especialmente útil cuando se busca un equilibrio entre precisión, robustez y manejo de diferentes tipos de datos.

2.1.5.6. K vecinos próximos

El algoritmo k vecinos próximos (k-NN) clasifica una instancia según la mayoría de clases de sus k vecinos más cercanos en el espacio de características, usando métricas de distancia como la Euclideana o el coseno (Cover & Hart, 1967). No requiere fase de entrenamiento explícito (modelo "perezoso"), pero su complejidad recae en la búsqueda de vecinos al clasificar.

Para un punto $\mathbf{x}$, su predicción es

$$\hat{y} = \text{mode}\big\{y_j : \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x})\big\}, \tag{2.9}$$

donde $\mathcal{N}_k(\mathbf{x})$ son los k puntos del conjunto de entrenamiento más cercanos a $\mathbf{x}$.

A continuación se presenta graficamente 2.7 el método basado en KNN.

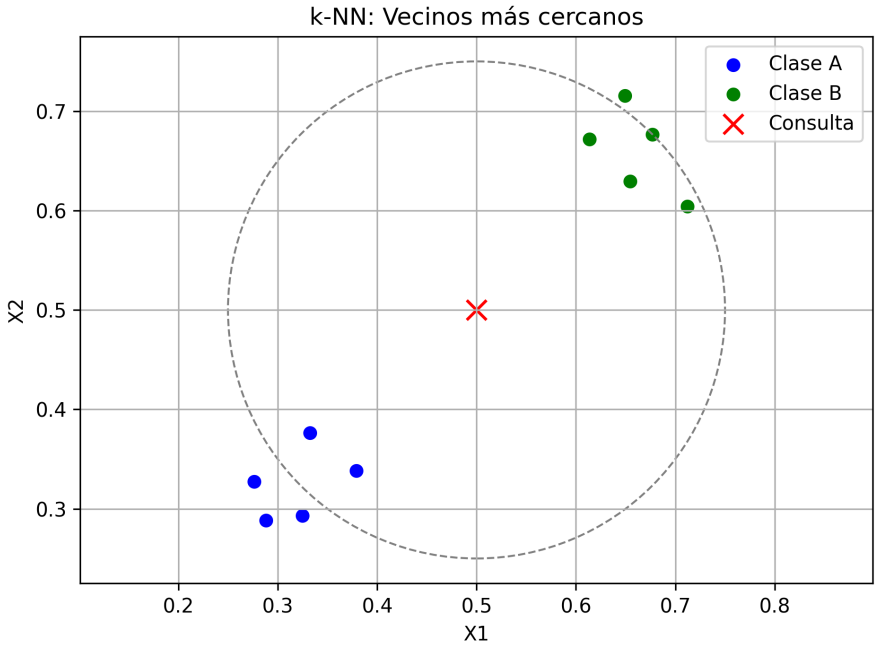

Figura 2.7: Ejemplo de clasificación de k-NN con $k = 5$ (elaboración propia)

El k-NN es un algoritmo sencillo y no paramétrico utilizado en tareas de reconocimiento de patrones, como la clasificación de imágenes y señales, sistemas de recomendación basados en preferencias de vecinos, detección de anomalías en datos de sensores, y clasificación rápida de texto mediante vectores TF–IDF (Altman, 1992; Cover & Hart, 1967).

Entre sus ventajas destacan su simplicidad conceptual y facilidad de implementación, así como su capacidad para adaptarse a distribuciones arbitrarias sin suponer un modelo paramétrico.

Sin embargo, presenta desventajas importantes, como el alto costo computacional en la fase de predicción, la sensibilidad a la escala de las características y a la presencia de variables irrelevantes, lo que exige una adecuada normalización y selección de características.

Además, su rendimiento disminuye considerablemente en espacios de alta dimensionalidad debido a la “maldición de la dimensionalidad” (Beyer

et al., 1999; Weinberger & Saul, 2006).

A pesar de estas limitaciones, k-NN sigue siendo una herramienta útil y efectiva en aplicaciones donde la interpretabilidad y la simplicidad son prioritarias, y cuando se cuenta con un volumen de datos manejable para la búsqueda eficiente de vecinos.

2.1.5.7. K medias

K medias (K-means) es un algoritmo de agrupamiento que busca particionar los datos en k clústeres minimizando la suma de distancias cuadradas de cada punto a su centroide (MacQueen, 1967):

$$\arg\min_{\{C_i\}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2, \tag{2.10}$$

donde $\boldsymbol{\mu}_i$ es el centroide del clúster C_i.

A continuación se puede apreciar de manera gráfica el método en cuestión.

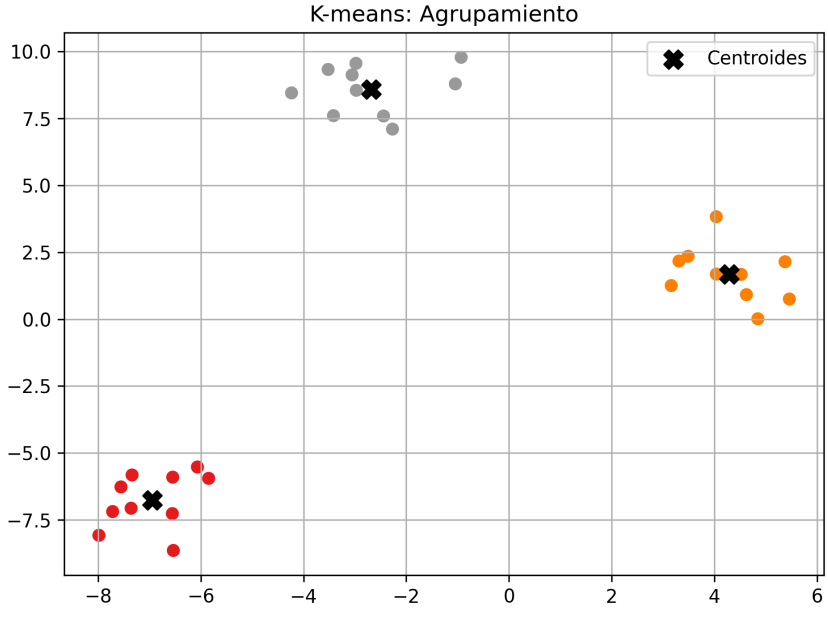

Figura 2.8: Ejemplo de clasificación usando K-means (elaboración propia)

El algoritmo k-means es uno de los métodos de agrupamiento más populares debido a su escalabilidad a grandes conjuntos de datos y su rápida convergencia. Se utiliza comúnmente para segmentación de clientes, agrupamiento de documentos y temas, análisis de expresión génica y exploración de patrones en datos de sensores y series temporales (Lloyd, 1982; MacQueen, 1967).

Entre sus ventajas destaca la eficiencia computacional y la facilidad de implementación. Sin embargo, presenta limitaciones como la necesidad de predefinir el número de clústeres k, la sensibilidad a la inicialización de centroides, y la incapacidad para detectar clústeres no esféricos o con formas complejas (Arthur & Vassilvitskii, 2007).

A pesar de estas limitaciones, k-means sigue siendo ampliamente utilizado debido a su simplicidad y buen desempeño en datos con estructuras claras y bien separadas.

2.1.5.8. Cluster jerárquico

El agrupamiento jerárquico construye una jerarquía de clústeres mediante métodos aglomerativos o divisivos, representada con un dendrograma. En el enfoque aglomerativo se inicia con cada punto como clúster y se fusionan iterativamente los pares más cercanos según un criterio de enlace (simple, completo, promedio) (S. C. Johnson, 1967).

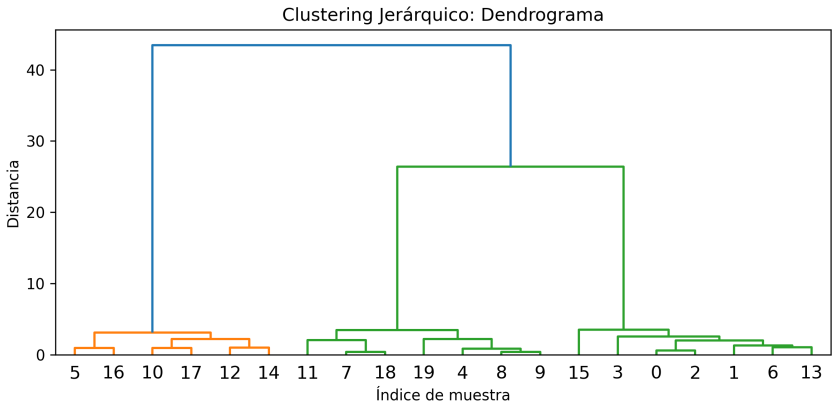

Figura 2.9: Ejemplo de dendrograma jerárquico (elaboración propia)

El agrupamiento jerárquico es un método que no requiere la especificación previa del número de clústeres y proporciona una representación en forma de dendrograma, facilitando la interpretación de las relaciones jerárquicas entre datos. Es aplicado en segmentación de clientes, agrupamiento de documentos, análisis de expresión génica y exploración de patrones en series temporales (S. C. Johnson, 1967; Murtagh, 2012).

Sus ventajas incluyen la flexibilidad para descubrir estructuras anidadas y la ausencia de necesidad de definir k. Sin embargo, su complejidad computacional elevada, que puede llegar a $\mathcal{O}(n^3)$, limita su uso en grandes conjuntos de datos. Además, es sensible al ruido y a la elección de la métrica de distancia, lo que puede afectar la calidad de los clústeres (Ward Jr, 1963).

A pesar de estas limitaciones, el agrupamiento jerárquico es una herramienta valiosa para análisis exploratorios donde la comprensión de la estructura jerárquica es crucial.

2.1.6. Aprendizaje Profundo

Aprendizaje profundo (2013–2017)

La introducción de *word embeddings* como Word2Vec (Mikolov et al., 2013) y GloVe (Pennington et al., 2014) permitió representar palabras en espacios vectoriales donde la proximidad refleja similitud semántica. Esto facilitó el desarrollo de arquitecturas de redes neuronales recurrentes (LSTM) (Hochreiter & Schmidhuber, 1997) y convolucionales (CNN) (Kim, 2014) capaces de procesar secuencias y extraer rasgos contextuales. A pesar de sus ventajas, estos modelos seguían siendo costosos de entrenar para secuencias muy largas y requerían ajustar numerosas hiperparámetros.

El aprendizaje profundo (DL) revolucionó el campo del PLN con la capacidad de aprender representaciones jerárquicas del lenguaje directamente desde los datos sin requerir ingeniería manual de características. Modelos como las redes recurrentes (RNN), LSTM y más recientemente Transformers, lograron avances notables en tareas complejas como traducción automática, resumen y respuesta a preguntas (Otter et al., 2020; Young et al., 2018).

Ejemplos de arquitecturas de DL:

- Convolutional Neural Networks (CNN)
- Redes neuronales recurrentes (RNN)
- Long Short-Term Memory (LSTM)
- Gated Recurrent Units (GRU)
- Transformers

2.1.6.1. Convolutional Neural Networks

Las Redes Neuronales Convolucionales (CNN), inicialmente diseñadas para visión por computadora, han revolucionado el PLN al permitir la detección eficiente de patrones locales en secuencias textuales. Mientras que los métodos tradicionales basados en n-gramas o bolsa de palabras carecían de la capacidad para modelar relaciones contextuales complejas, las CNN permiten extraer características jerárquicas y relevantes automáticamente (LeCun et al., 2015; Mikolov et al., 2013).

Una CNN aplicada a texto recibe como entrada una matriz $X \in \mathbb{R}^{n \times d}$, donde n es la longitud de la secuencia y d la dimensión del embedding de cada palabra. Un filtro convolucional $w \in \mathbb{R}^{h \times d}$, con ventana de tamaño h, se desliza sobre ventanas contiguas de h palabras. La operación convolucional para la posición i se define como:

$$c_i = f\left(\langle X_{i:i+h-1}, w\rangle + b\right) = f\left(\sum_{j=0}^{h-1}\sum_{k=1}^{d} X_{i+j,k} \cdot w_{j,k} + b\right)$$

donde $\langle \cdot, \cdot \rangle$ denota el producto punto entre la submatriz de embeddings y el filtro, $b \in \mathbb{R}$ es un sesgo escalar, y f es una función de activación no lineal, típicamente ReLU.

El vector de activaciones $\mathbf{c} = [c_1, c_2, \ldots, c_{n-h+1}]$ conforma un mapa de características que captura patrones locales en la secuencia. Vea la siguiente gráfica 2.10 para una representación visual.

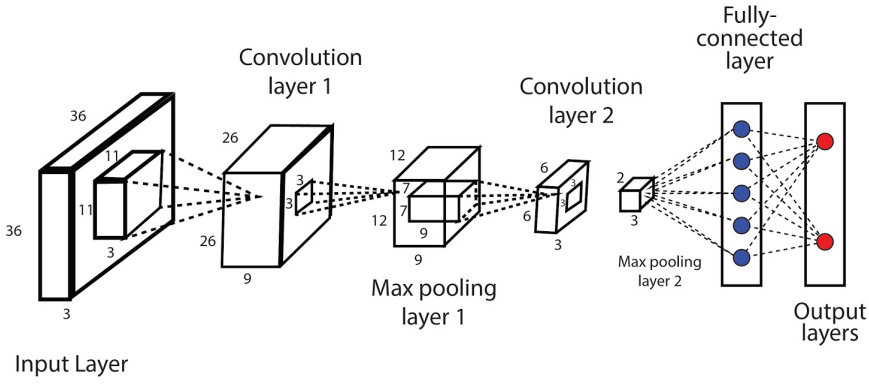

Figura 2.10: Ejemplo de arquitectura de CNN (Mlyahilu et al., 2019)

Posteriormente, se aplican operaciones de pooling, como max-pooling, para reducir dimensionalidad y extraer las características más salientes. Estas representaciones son procesadas por capas totalmente conectadas para realizar tareas específicas como clasificación o etiquetado (Kalchbrenner et al., 2014; Kim, 2014).

Este enfoque ha demostrado alta eficacia en tareas de PLN, incluyendo análisis de sentimientos (Dos Santos & Gatti, 2014), reconocimiento de entidades nombradas (Collobert et al., 2011) y clasificación de texto, al superar modelos basados en características manuales y facilitar la extracción automática de patrones relevantes.

En resumen, la incorporación de CNN en PLN representa un avance metodológico crucial que combina la potencia del aprendizaje profundo con la detección local de patrones contextuales en texto, facilitando modelos robustos y escalables para el análisis lingüístico.

2.1.6.2. Redes Neuronales Recurrentes (RNN)

Las Redes Neuronales Recurrentes (RNN) son una clase especializada de redes neuronales diseñadas para procesar datos secuenciales, donde la dependencia temporal y el orden de los elementos son fundamentales. Esto las hace especialmente adecuadas para tareas en Procesamiento de Lenguaje Natural (PLN), series temporales y señales de audio (Elman, 1990).

Dado un vector de entrada secuencial $\{x_1, x_2, \ldots, x_T\}$, la RNN actualiza su estado oculto h_t en cada instante t utilizando la siguiente función recursiva:

$$h_t = \tanh(W_h h_{t-1} + W_x x_t + b)$$

donde W_h, W_x son matrices de pesos entrenables y b un vector de sesgo. La salida estimada $\hat{y}_t$ se calcula comúnmente mediante una capa softmax:

$$\hat{y}_t = \text{softmax}(W_y h_t)$$

donde W_y es otra matriz de pesos aprendidos. Esta arquitectura permite que la información se propague a lo largo de la secuencia, capturando dependencias temporales y contextuales. Vea la representación gráfica en 2.11.

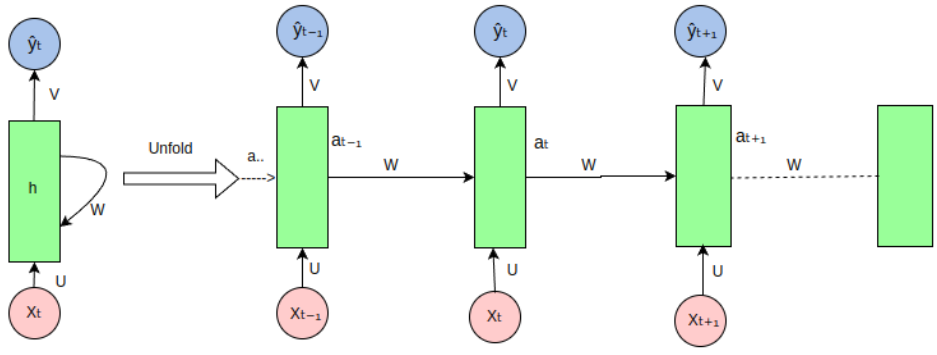

Figura 2.11: Red neuronal recurrente (RNN).
Fuente: Rahul Tawar, "Recurrent Neural Networks Explained," *Medium*, 2020. Disponible en: https://rahul-tawar.medium.com/recurrent-neural-networks-explained-b78db8c92810 (consulta: 19 de noviembre de 2025).

Sin embargo, las RNN tradicionales presentan dificultades para modelar dependencias a largo plazo debido a problemas de desvanecimiento o explosión del gradiente durante el entrenamiento (Bengio et al., 1994; Hochreiter & Schmidhuber, 1997). Para superar estas limitaciones, se han desarrollado variantes como las LSTM y GRU, que incluyen mecanismos de memoria y puertas para regular el flujo de información.

Las RNN han sido ampliamente aplicadas en tareas de PLN tales como modelado de lenguaje, traducción automática, reconocimiento de voz y análisis de sentimientos, demostrando un avance significativo frente a modelos de ventana fija y basados en características manuales (Mikolov et al., 2010; Sutskever et al., 2014).

2.1.6.3. Long Short-Term Memory

Las redes neuronales recurrentes (RNN) clásicas son capaces de modelar secuencias y dependencias temporales, pero presentan dificultades para aprender relaciones a largo plazo debido al problema del desvanecimiento o explosión del gradiente durante el entrenamiento (Bengio et al., 1994). Este problema limita la capacidad de las RNN tradicionales para capturar

dependencias que ocurren a gran distancia en la secuencia.

Para superar esta limitación, Hochreiter y Schmidhuber propusieron en 1997 la Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997), una arquitectura que introduce un sistema de compuertas especializadas para regular el flujo de información y mantener una memoria a largo plazo efectiva.

Transición metodológica desde RNN a LSTM Mientras que una RNN estándar actualiza su estado oculto h_t directamente en función del estado previo y la entrada actual, las LSTM incorporan un estado de celda c_t que funciona como una "memoria interna". Además, emplean tres compuertas con funciones sigmoides para decidir qué información conservar, actualizar o olvidar:

- Compuerta de olvido (f_t): decide qué información previa eliminar de la memoria.
- Compuerta de entrada (i_t): regula qué nueva información se almacena en la memoria.
- Compuerta de salida (o_t): determina qué parte del estado de la celda se expone como salida.

Este mecanismo permite que la LSTM retenga información relevante durante muchos pasos temporales y evite la degradación del gradiente durante el entrenamiento.

Sea $x_t \in \mathbb{R}^d$ la entrada en el tiempo t, $h_{t-1} \in \mathbb{R}^h$ el estado oculto previo y $c_{t-1} \in \mathbb{R}^h$ el estado de la celda previa, con h el tamaño de la memoria.

Las compuertas y actualizaciones se definen como sigue:

$$
\begin{aligned}
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) && \text{(compuerta de olvido)} \\
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) && \text{(compuerta de entrada)} \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) && \text{(compuerta de salida)} \\
\tilde{c}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c) && \text{(candidata a nuevo estado de celda)} \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t && \text{(actualización del estado de celda)} \\
h_t &= o_t \odot \tanh(c_t) && \text{(salida o estado oculto)}
\end{aligned}
$$

Donde:

- σ es la función sigmoide que restringe valores a $[0, 1]$, funcionando como filtro o compuerta.
- tanh es la función tangente hiperbólica que escala valores a $[-1, 1]$.
- $\odot$ denota producto elemento a elemento (Hadamard).
- W_*, U_* son matrices de pesos aprendibles, y b_* son vectores de sesgo.

La célula LSTM está diseñada para controlar el flujo de información mediante sus compuertas: la compuerta de olvido f_t decide qué parte del estado de celda previo c_{t-1} se conserva o elimina, permitiendo que la memoria descarte información irrelevante o desactualizada.

Por su parte, la compuerta de entrada i_t regula qué nueva información, representada por el vector candidato $\tilde{c}_t$, debe incorporarse a la memoria, actualizando el estado de celda c_t como una combinación ponderada entre la memoria anterior y la información actual. Finalmente, la compuerta de salida o_t determina qué parte de la memoria interna se expone como estado oculto h_t, que sirve como salida para la capa o para pasos futuros.

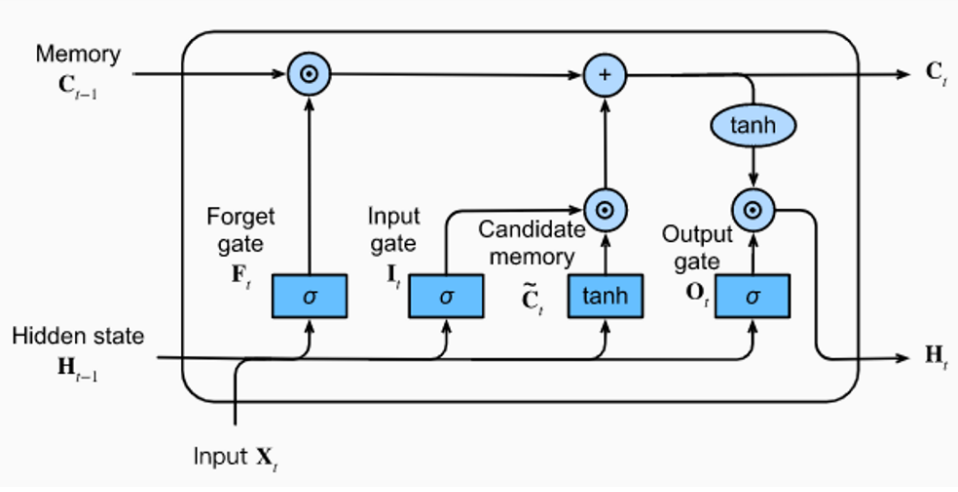

Figura 2.12: Célula LSTM que mantiene una memoria a largo plazo mediante compuertas especializadas. Ottavio Calzone, "An Intuitive Explanation of LSTM," *Medium*, 2020. Disponible en: https://medium.com/@ottaviocalzone/an-intuitive-explanation-of-lstm-a035eb6ab42c (consulta: 19 de noviembre de 2025).

Esta arquitectura tiene varias ventajas clave: permite mitigar eficazmente el problema del desvanecimiento y explosión del gradiente, facilitando el aprendizaje de dependencias a largo plazo en secuencias. Gracias a su mecanismo de memoria interna y compuertas, las LSTM pueden modelar secuencias más largas y complejas con mayor precisión que las RNN tradicionales.

Además, su flexibilidad las hace aplicables a diversas tareas en procesamiento de lenguaje natural, como modelado de lenguaje, análisis de sentimiento y traducción automática. Aunque su estructura es más compleja y computacionalmente costosa, la mejora en desempeño y capacidad justifica ampliamente su uso en la mayoría de aplicaciones secuenciales.

2.1.6.4. Gated Recurrent Units (GRU)

Los **Gated Recurrent Units** (GRU) son una variante simplificada de las LSTM que reducen la complejidad computacional al tener menos compuertas, pero manteniendo un rendimiento competitivo en el modelado de

secuencias (Cho, 2014). Esta arquitectura fue propuesta para acelerar el entrenamiento y reducir la cantidad de parámetros sin sacrificar la capacidad de capturar dependencias a largo plazo.

Sea $x_t \in \mathbb{R}^d$ la entrada en el tiempo t y $h_{t-1} \in \mathbb{R}^h$ el estado oculto previo, las actualizaciones en una unidad GRU se definen mediante:

$$
\begin{aligned}
z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) && \text{(compuerta de actualización)} \\
r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) && \text{(compuerta de reinicio)} \\
\tilde{h}_t &= \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) && \text{(candidato a nuevo estado)} \\
h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t && \text{(estado oculto actualizado)}
\end{aligned}
$$

donde:

- σ es la función sigmoide que restringe valores a $[0, 1]$.
- tanh es la función tangente hiperbólica que escala valores a $[-1, 1]$.
- $\odot$ denota el producto elemento a elemento (Hadamard).
- W_*, U_* son matrices de pesos aprendibles y b_* vectores de sesgo.

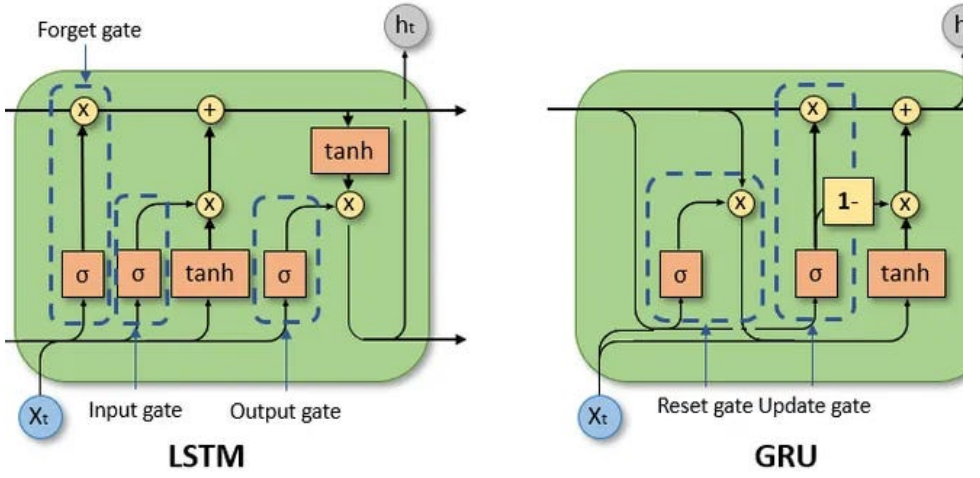

Figura 2.13: Unidad GRU simplificada que combina las compuertas en un único flujo.

"Página de producto," *bestrussm.click*, s.f. Disponible en: https://bestrussm.click/product_tag/67892128_.html (consulta: 19 de noviembre de 2025).

En una unidad GRU, la compuerta de actualización z_t controla en qué proporción se mezcla el estado oculto previo h_{t-1} con el nuevo candidato $\tilde{h}_t$, actuando como un balance entre retener información antigua y agregar información nueva. La compuerta de reinicio r_t decide cuánto de la información pasada se utiliza para calcular el nuevo candidato, permitiendo olvidar parte del estado previo cuando sea necesario.

Esta estructura simplificada elimina la memoria explícita de celda que tienen las LSTM, fusionando los mecanismos de olvido y entrada en la compuerta de actualización, lo que reduce el número de parámetros y el coste computacional. A pesar de esta simplificación, las GRU han demostrado rendimiento comparable al de las LSTM en muchas tareas de modelado secuencial, siendo preferidas cuando se requiere un entrenamiento más rápido o menor uso de memoria.

Gracias a su arquitectura eficiente, las GRU son especialmente útiles en aplicaciones donde el balance entre rendimiento y eficiencia computacional

es crítico, como en dispositivos con recursos limitados o en modelos con grandes volúmenes de datos.

2.1.7. Transformers

2.1.7.1. Modelos preentrenados (2017–2019)

Antes del auge de los Transformers, los primeros avances en preentrenamiento contextual —como *ELMo* y *ULMFiT*— mostraron que reutilizar conocimiento lingüístico aprendido de forma general podía mejorar sustancialmente tareas específicas de PLN: ELMo produce representaciones contextuales a partir de modelos bidireccionales basados en LSTM, mientras que ULMFiT introduce un esquema de ajuste fino gradual y robusto para transferir ese conocimiento a distintos dominios (Howard & Ruder, 2018; Peters et al., 2018).

El punto de inflexión llegó con *Attention Is All You Need*, que presentó el *Transformer* (Vaswani, Shazeer, Parmar et al., 2017a). Su contribución central es la *autoatención* —un mecanismo que modela dependencias de largo alcance sin recurrencia— y una arquitectura totalmente paralelizable que sustituye las RNN tradicionales. Sobre esta base, arquitecturas como *BERT* (Devlin et al., 2019a) y *GPT-2* (Radford et al., 2019) consolidaron el paradigma de *preentrenamiento masivo* sobre texto no etiquetado seguido de *ajuste fino*, alcanzando resultados de frontera en comprensión y generación de lenguaje.

La operación de autoatención se define como:

$$\text{Attention}(Q, K, V) \;=\; \text{softmax}\!\left(\frac{QK^{\top}}{\sqrt{d_k}}\right) V,$$

donde Q son las *queries*, K las *keys*, V los *values* y d_k es la dimensión de las *keys*. En la práctica, la variante de *multi-cabeza* (*multi-head*) proyecta

Q, K y V a varios subespacios y concatena las atenciones resultantes, lo que permite capturar relaciones complementarias en paralelo (Vaswani, Shazeer, Parmar et al., 2017a).

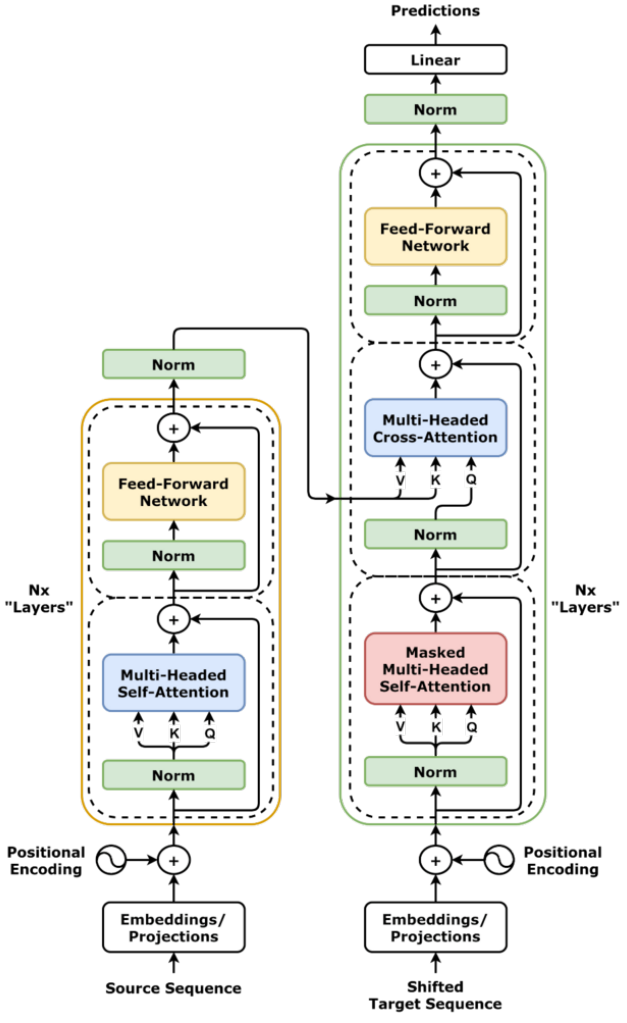

Figura 2.14: Arquitectura Transformer(Sádaba-Campo & Gómez-Moreno, 2025)

2.1.7.2. Modelos de lenguaje (2019–presente)

En la literatura reciente, el término *large language model* (LLM) suele reservarse para modelos *generativos* de propósito general, típicamente *decoder-only* basados en Transformer y entrenados a gran escala (centenas de miles

de millones de parámetros), capaces de *in-context learning* y razonamiento emergente (Brown et al., 2020; Kaplan et al., 2020; Wei et al., 2022a). En la práctica, el umbral de adopción del término se consolida con *GPT-3* (Brown et al., 2020); trabajos posteriores como *PaLM* profundizan el paradigma de escala y muestran mejoras sistemáticas en múltiples tareas (Chowdhery et al., 2022). Estos sistemas han redefinido el análisis de texto por su coherencia, versatilidad y capacidad de generalización, aunque su tamaño trae consigo retos de eficiencia, interpretabilidad y sesgos del corpus de entrenamiento.

Es importante distinguir que modelos *encoder-only* como *BERT* y *RoBERTa* no suelen clasificarse como LLMs: son *modelos de lenguaje preentrenados* (PLMs) optimizados para comprensión textual (p. ej., clasificación, extracción) y no para generación abierta (Devlin et al., 2019a; Y. Liu et al., 2019). Por su parte, *T5* es una arquitectura *encoder–decoder* (texto-a-texto) de gran escala que sirve de puente entre ambas familias (Raffel et al., 2020). En cambio, la línea *GPT-2/3/4* representa la rama de LLMs generativos *decoder-only*, donde el aumento de datos, parámetros y cómputo impulsa habilidades de few-shot y razonamiento emergente (Brown et al., 2020; Radford et al., 2019; Wei et al., 2022a).

En economía aplicada —donde abundan textos no estructurados— los LLMs habilitan clasificación y extracción sin etiquetado extensivo, generación de resúmenes y respuestas, así como *inferencia exploratoria* (p. ej., señales de inflación) mediante *prompting* e integración con recuperación (*RAG*) (Bommasani, Hudson et al., 2021; Brown et al., 2020). No obstante, para tareas focalizadas y con restricciones de cómputo, los PLMs *encoder-only* siguen siendo competitivos, especialmente con ajuste fino ligero (adapters, LoRA) o destilación.

A continuación enlistamos algunos ejemplos de aplicación de ambos

enfoques:

- **PLMs (encoder-only, no LLM):** BERT (Devlin et al., 2019a), RoBERTa (Y. Liu et al., 2019), ALBERT (Lan et al., 2020), DistilBERT (Sanh et al., 2019).
- **Texto-a-texto (encoder–decoder):** T5 (Raffel et al., 2020).
- **LLMs generativos (decoder-only):** GPT-2 / GPT-3 / GPT-4 (Brown et al., 2020; Radford et al., 2019), PaLM (Chowdhery et al., 2022).

En síntesis, a partir de *GPT-3* la comunidad comienza a hablar de LLMs en sentido estricto: modelos generativos, de propósito general y gran escala. En la práctica, la elección entre PLMs y LLMs debería guiarse por el *ajuste a la tarea, los recursos y los riesgos*: los LLMs amplían el alcance analítico, mientras que los PLMs siguen ofreciendo soluciones eficientes y controlables; ambos enfoques son complementarios en un flujo moderno de PLN (Bommasani, Hudson et al., 2021).

Capítulo 3

Revisión de literatura

Esta tesis plantea la problemática de analizar textos económicos no estructurados con el fin de extraer información relevante que caracteriza la tendencia de la inflación. En este capítulo, revisaremos los avances recientes en el uso del análisis de texto en economía en general y en particular relacionado con el objeto de estudio, destacando las contribuciones metodológicas y aplicaciones prácticas.

Este planteamiento esta sujeto a varios desafíos técnicos y conceptuales que podemos abordar mediante la literatura existente. En este sentido, se busca dar respuesta a las siguientes preguntas:

- ¿Cuáles son los enfoques más relevantes de PLN en economía?
- ¿Cómo ha sido la evolución de las técnicas de PLN y su aplicación en el área?
- ¿Cuáles son los principales corpora usados en la literatura relacionado con economía?
- ¿Cuáles son las técnicas más usadas para extraer, procesar y analizar textos económicos?
- Por último, ¿qué métricas se utilizan para evaluar la calidad de los resultados obtenidos?

3.1. Análisis de texto en economía

El análisis de texto constituye un campo amplio y heterogéneo que integra múltiples técnicas, enfoques y aplicaciones, cuya elección depende de la naturaleza del fenómeno que el investigador busca comprender. En el caso de la economía, su incorporación es relativamente reciente y se encuentra en constante evolución, lo que ha dificultado la construcción de un marco teórico y metodológico común que oriente su aplicación. Como señalan Ash y Hansen (2023): «existe una diversidad metodológica considerable y aún no se cuenta con un marco unificado —ni siquiera con un vocabulario compartido— que permita guiar las decisiones de modelado».

Lo primeros estudios de PLN inician en el área de las finanzas, principalmente con Tetlock et al. (2008) quien buscaba entender ¿cómo una simple medición a partir de las noticias financieras podía predecir movimientos en el mercado de valores? para eso se utilizó un diccionario de palabras anotado con valores positivas y negativas para medir el sentimiento derivado de las noticias. Este trabajo mostró que el análisis de texto puede capturar información relevante para la toma de decisiones económicas y financieras, sentando las bases para investigaciones posteriores en diversas áreas de la economía.

Por otro lado, Antweiler y Frank (2004) utilizaron los reportes financieros de Yahoo para mostrar que las noticias financieras tienen un gran impacto en el mercado, de hecho concluyó que ayudan a predecir la volatilidad. Metodológicamente, usaron técnicas de conteo de palabras y análisis de sentimiento basados en diccionarios, similares a los de Tetlock et al. (2008), pero aplicados a un corpus más amplio y diverso. Se estimó que usaron más de 1.5 millones de mensajes de los foros de Yahoo-Finanzas.

Bajo la misma perspectiva, Tetlock (2007) logró medir el impacto de las

noticias en la volatilidad del mercado de valores, que le permitió concluir que altos niveles de pesimismo anticipan presión a la baja a corto plazo. Al igual que los trabajos anteriores, usó técnicas de conteo de palabras y Análisis en Componente Principal (PCA, por su sigla en Inglés) para reducir la dimensionalidad del texto y extraer temas relevantes.

A diferencia de los estudios previos, Koppel y Schler (2004), en su trabajo titulado "Good News or Bad News? Let the Market Decide", incorporaron técnicas de clasificación supervisada para categorizar noticias financieras como positivas o negativas, alcanzando una precisión aproximada del 70%. Para alcanzar dicho resultado, emplearon un conjunto de datos etiquetado manualmente y entrenaron sus modelos mediante algoritmos de aprendizaje automático. Este enfoque representó un avance significativo al evidenciar que los modelos basados en aprendizaje automático pueden superar las limitaciones inherentes a los métodos tradicionales basados en diccionarios.

A partir del trabajo de Loughran y Mcdonald (2011), la construcción de diccionarios comenzó a orientarse hacia una mayor especialización, particularmente en contextos financieros. En su propuesta, los autores enfatizan la necesidad de adaptar las categorías léxicas al lenguaje propio del mercado, lo que permitió mejorar de manera significativa la precisión del análisis de sentimiento frente a los diccionarios de uso general. Como señalan: «un diccionario genérico puede arrojar resultados erróneos al ignorar el léxico específico del objeto de estudio, por lo que nuestra propuesta busca reflejar las particularidades semánticas del lenguaje financiero».

Un trabajo destacado es el de Baker et al. (2015), quienes desarrollaron un índice de incertidumbre de política económica a partir de un diccionario de términos especializados y análisis textual. Sus resultados muestran que una mayor incertidumbre se asocia con una reducción significativa en la

inversión y el empleo. Para construir el índice y evaluar su efecto sobre variables macroeconómicas, emplearon técnicas de conteo de palabras y análisis de series de tiempo.

En conjunto, estos antecedentes evidencian la incorporación progresiva de técnicas de PLN en el análisis económico, mediante enfoques basados en conteo, bolsas de palabras, uso de diccionarios, análisis de sentimiento y clasificación supervisada. Dichas metodologías permiten caracterizar con mayor precisión fenómenos económicos específicos y, a partir de ello, realizar inferencias sobre variables relevantes o anticipar su comportamiento futuro.

Pese a sus aportes, estos enfoques aún enfrentan limitaciones importantes, entre ellas la dificultad para capturar el contexto y las sutilezas del lenguaje, así como su dependencia de diccionarios específicos y la necesidad de desarrollar repertorios léxicos más especializados. Además, la mayoría de los estudios se han concentrado en el ámbito financiero, dejando relativamente inexploradas otras áreas de la economía, como la microeconomía, la economía del comportamiento, el desarrollo económico y la macroeconomía.

3.2. Uso de contexto en los análisis económicos

A medida que avanzan las técnicas de PLN, también se diversifican y sofistican los análisis de texto aplicados a la economía. En este sentido, destaca el uso de modelos basados en representaciones vectoriales de las palabras (word embeddings), estos permiten capturar el contexto y las relaciones semánticas entre las palabras. La hipótesis fundamental de este enfoque sostiene que las palabras con significados similares tienden a aparecer en

contextos similares, lo que posibilita una modelización más profunda del significado lingüístico y su relación con fenómenos económicos.

Podemos ilustrar este enfoque con el uso de Word2Vec (Mikolov et al., 2013), que define un modelo vectorial cuya representación se encuentra en un espacio de alta dimensión, donde la distancia entre los vectores refleja la similitud semántica entre las palabras. Por ejemplo, en un modelo entrenado con textos económicos, las palabras *inflación*, *precios* y *costos* estarían más cercanas entre sí que de palabras como *empleo* o *crecimiento*. Este enfoque permite predecir palabras a partir de su contexto, mejorando así las tareas de clasificación de textos y análisis de sentimiento.

A modo de ejemplo, uno de los estudios que demuestra la incorporación del contexto mediante representación vectorial de las palabras en economía es el de Sehrawat (2019). En este trabajo, los autores propusieron un enfoque de extracción de eventos económicos basado en aprendizaje profundo para la predicción del comportamiento del mercado de valores.

Las noticias financieras fueron representadas en un espacio vectorial denso utilizando modelos Word2Vec, lo que permitió identificar de manera automática eventos similares y superar la precisión alcanzada por los métodos tradicionales basados en diccionarios o reglas. De este modo, el estudio evidenció una mejora sustancial en la capacidad predictiva del mercado al capturar relaciones semánticas complejas entre eventos económicos y fluctuaciones bursátiles.

Otro ejemplo relevante es el trabajo de Kraus y Feuerriegel (2017) quienes, haciendo hincapié en el uso de arquitecturas de redes profundas como RNN/LSTM y transferencia de aprendizaje, tuvieron como objetivo principal la predicción de precios de acciones a partir de noticias financieras. Para eso, combinaron embeddings preentrenados sobre grandes corpus (139.1 millones de palabras) para asegurar la transferencia de conocimiento, con

modelos LSTM para capturar dependencias temporales en las noticias. Los resultados muestran una mejora significativa en la precisión de las predicciones de precios respecto a modelos tradicionales.

Otro trabajo necesario de comentar es el del Banco Central Europeo (BCE), que emplea Word2Vec para analizar las conferencias de prensa realizadas por las autoridades del banco. En este estudio, se procesaron los textos de las conferencias para generar representaciones vectoriales de las palabras, permitiendo identificar patrones semánticos y cambios en el discurso oficial sobre política monetaria y expectativas económicas.

El uso de Word2Vec facilitó la detección de términos clave y la evolución temática en las comunicaciones del BCE, contribuyendo a una mejor comprensión del impacto de la comunicación institucional en los mercados financieros y en la percepción pública sobre la inflación y otras variables macroeconómicas (Silva et al., 2025).

La evolución de los modelos de representación semántica continuó con la introducción de arquitecturas más avanzadas, especialmente los modelos transformer, que marcaron un punto de inflexión en el procesamiento del lenguaje natural. A diferencia de los embeddings estáticos como Word2Vec o GloVe, los modelos transformer —entre los que destacan BERT y GPT— generan representaciones contextuales dinámicas que varían según el entorno lingüístico de cada palabra.

Esta capacidad de contextualización profunda ha permitido capturar matices semánticos y sintácticos más complejos, lo que amplía significativamente las posibilidades de aplicación en el análisis económico, particularmente en la detección de eventos, la inferencia de sentimiento y la predicción de tendencias del mercado a partir de grandes volúmenes de texto.

3.3. Uso de Transformer en economía

Los métodos basados en transformer marcaron un punto de inflexión en el desarrollo del PLN, al superar las limitaciones de los modelos previos. Desde su introducción en 2017 por Vaswani, Shazeer, Parmar et al. (2017a), estos modelos han mostrado una capacidad sin precedentes para modelar dependencias a largo plazo y capturar relaciones contextuales complejas dentro de los textos.

Entre los primeros trabajos que aplicaron arquitecturas transformer en economía destaca el de Araci (2019), quien desarrolló FinBERT (Financial BERT), una adaptación de BERT orientada al análisis del lenguaje financiero. Su propuesta parte del reconocimiento de que los modelos de propósito general presentan limitaciones frente al léxico especializado del ámbito económico, y plantea que los modelos preentrenados pueden superar este problema al requerir menos datos etiquetados y adaptarse mediante fine-tuning a corpus específicos del dominio. Los resultados muestran que FinBERT supera de forma consistente los métodos tradicionales de aprendizaje automático en tareas de análisis de sentimiento financiero, incluso con conjuntos de entrenamiento reducidos.

Posteriormente, Yang et al. (2020) et al. ampliaron este enfoque al preentrenar el modelo FinBERT de Araci (2019) utilizando un extenso corpus de comunicaciones financieras —incluyendo reportes corporativos, documentos regulatorios y noticias económicas— con el objetivo de generar una representación contextual más precisa del lenguaje del sector. Sus experimentos en tres tareas de clasificación de sentimiento confirmaron la ventaja del modelo especializado frente a las versiones genéricas de BERT, consolidando así la utilidad de los modelos contextuales preentrenados para el procesamiento del lenguaje financiero y estableciendo una base para

futuras aplicaciones en el análisis económico automatizado.

En la misma línea, Mishev et al. (2020) desarrollaron una plataforma de evaluación sistemática para comparar distintos enfoques de análisis de sentimiento en finanzas, desde el uso de diccionarios especializados y representaciones estáticas de palabras hasta los modelos más recientes basados en transformers. A través de más de cien experimentos con conjuntos de datos anotados por expertos, demostraron que los embeddings contextuales ofrecen una eficiencia superior para extraer señales accionables del texto financiero, incluso en ausencia de grandes volúmenes de datos etiquetados. Además, observaron que las versiones distiladas de los modelos transformer alcanzan un rendimiento comparable al de sus versiones completas, lo que las hace especialmente adecuadas para entornos de producción y aplicaciones en tiempo real.

Complementariamente, Zhao et al. (2020) propusieron un enfoque basado en BERT para el análisis de sentimiento y la detección de entidades clave en textos financieros en línea, con especial atención a la información negativa presente en medios digitales y redes sociales. Su método combina la comprensión lectora automática (Machine Reading Comprehension, MRC) con técnicas de ensemble learning para identificar entidades relevantes y extraer señales de riesgo con mayor precisión que los modelos tradicionales como SVM, LR o NBM. Los resultados experimentales evidencian un rendimiento superior en tareas de clasificación de sentimiento y detección de entidades financieras, demostrando el potencial de los modelos transformer para aplicaciones dinámicas de minería de texto y análisis de opinión pública en tiempo real.

Cuadro 3.1: Evolución de los modelos basados en *transformers* aplicados a textos especializados en economía y finanzas

Año	Autores	Modelo / Contribución	Enfoque metodológico	Principales resultados / hallazgos
2019	Araci (2019)	**FinBERT (Financial BERT)**	Ajuste fino de *BERT* sobre corpus financiero (noticias y reportes).	Supera a los métodos tradicionales en análisis de sentimiento financiero con menos datos etiquetados.
2020	Sousa et al. (2019)	**BERT para análisis de sentimiento bursátil**	Ajuste fino de *BERT* con 582 noticias financieras clasificadas manualmente (positivas, negativas, neutras) para apoyar la toma de decisiones en tiempo real.	Alcanzó un 72.5% de F-score en clasificación de sentimiento; mostró correlación entre las predicciones del modelo y los movimientos posteriores del índice Dow Jones.
2020	Yang et al. (2020)	**Preentrenamiento financiero de BERT**	Preentrenamiento desde cero con un gran corpus de comunicaciones financieras.	Confirma la superioridad del modelo especializado frente a *BERT* genérico; consolida el uso contextual en finanzas.
2021	Mishev et al. (2020)	**Evaluación comparativa de PLN financiero**	Análisis de diccionarios, embeddings estáticos y *transformers* (*distilados*).	Los embeddings contextuales son más eficientes; los modelos distilados mantienen precisión con menor costo computacional.
2021	Zhao et al. (2020)	**BERT para sentimiento y detección de entidades**	Uso de *BERT* para minería financiera en línea con comprensión lectora automática (MRC) y *ensemble learning*.	Mejor rendimiento que SVM y LR; destaca la detección de información negativa y entidades clave en redes sociales.
2021	P. Liu, Wang et al. (2021)	**RoBERTa-Econ**	Adaptación de *RoBERTa* a textos macroeconómicos (discursos de bancos centrales).	Mejora la inferencia de tono y predicción de indicadores frente a métodos tradicionales.
2022	Y. Chen et al. (2022)	**BERT4ECON**	Modelo *BERT* entrenado con corpus económico general (reportes, artículos y documentos oficiales).	Extiende el uso de *transformers* a dominios macroeconómicos; mejora la detección de narrativas y proyecciones de tendencia.
2022	R. S. Shah et al. (2022)	**FLANG / FLUE**	Modelo de lenguaje financiero con enmascaramiento por palabras clave y objetivos adicionales (span boundary, infilling); evaluación con *FLUE*.	Supera modelos previos en cinco tareas de PLN financiero; establece nuevos estándares de evaluación abierta para el dominio.

La evolución de los modelos basados en *transformers* en economía se consolidó inicialmente con trabajos orientados al análisis de noticias financieras en tiempo real. Sousa et al. (2019) propusieron un modelo de análisis de sentimiento basado en *BERT* ajustado con un corpus de noticias bur-

sátiles clasificadas manualmente, demostrando su utilidad para predecir movimientos inmediatos del índice Dow Jones. Este enfoque evidenció que los modelos contextuales podían ofrecer ventajas significativas en entornos de decisión rápida.

A partir de esta línea, los desarrollos posteriores —como *FinBERT* (Araci, 2019), *RoBERTa-Econ* (P. Liu, Wang et al., 2021), *BERT4ECON* (Y. Chen et al., 2022) y *FLANG* (R. S. Shah et al., 2022)— ampliaron las capacidades del PLN económico al incorporar corpus especializados, nuevas estrategias de enmascaramiento y plataformas de evaluación abiertas, consolidando una tendencia hacia la modelización semántica profunda de la información económica.

3.4. Uso de LLM en economía

La evolución reciente de PLN ha dado paso a una nueva generación de modelos conocidos como LLM, los cuales representan una ampliación sustancial de las arquitecturas transformer especializadas descritas previamente. Estos modelos, entrenados con billones de parámetros y grandes volúmenes de texto general, han demostrado una notable capacidad para realizar tareas complejas de comprensión, inferencia y generación de lenguaje sin necesidad de un ajuste fino específico.

En el ámbito económico, su incorporación marca una transformación profunda, ya que permiten analizar grandes corpus de noticias, discursos y reportes financieros con un grado de autonomía y contextualización inédito. En este contexto, Brown et al. (2020) evidencian cómo *GPT-3*, con su entrenamiento masivo y su arquitectura autorregresiva, puede ejecutar tareas de análisis de sentimiento, clasificación temática y razonamiento económico con una mínima intervención humana, superando las limitaciones

de modelos previos como *BERT* o *RoBERTa*.

Posteriormente, Wu et al. (2023) introdujeron *BloombergGPT*, un modelo de 50 mil millones de parámetros entrenado sobre un extenso conjunto de 363 mil millones de tokens provenientes de fuentes financieras de Bloomberg, complementado con 345 mil millones de tokens de datos generales. Este modelo representa el primer LLM especializado en finanzas, capaz de mejorar significativamente el desempeño en tareas financieras sin sacrificar rendimiento en evaluaciones generales. Su desarrollo marca un punto de inflexión al combinar el entrenamiento mixto —entre datos especializados y de propósito general— con metodologías de evaluación estandarizadas, estableciendo así una nueva referencia para el uso de LLM en economía y tecnología financiera.

Más recientemente, Xie et al. (2024) presentaron *FinBen*, el primer marco de evaluación integral y de código abierto para modelos de lenguaje en el dominio financiero. Este banco de pruebas reúne 36 conjuntos de datos que abarcan 24 tareas en siete dimensiones clave: extracción de información, análisis textual, respuesta a preguntas, generación de texto, gestión de riesgos, pronóstico y toma de decisiones. *FinBen* introduce innovaciones como la evaluación de agentes, el uso de Retrieval-Augmented Generation (RAG) y los primeros conjuntos abiertos para resumen, question answering y negociación bursátil.

Su análisis de 15 LLM representativos —incluidos *GPT-4*, *ChatGPT* y *Gemini*— muestra que estos modelos sobresalen en extracción y análisis textual, pero presentan limitaciones en tareas de razonamiento avanzado, generación de texto y predicción. La iniciativa de *FinBen* consolida un nuevo estándar para la evaluación comparativa de LLM financieros, impulsando la innovación en el uso de inteligencia artificial para la investigación y la toma de decisiones económicas.

En paralelo, y reconociendo la escasez de recursos lingüísticos en otros idiomas, X. Zhang et al. (2024) desarrollaron *Toisón de Oro*, el primer marco bilingüe para modelos financieros en español e inglés. Este trabajo introduce *FinMA-ES*, un LLM diseñado para aplicaciones financieras bilingües, entrenado con un conjunto de instrucciones que combina más de 144 000 muestras de ambos idiomas procedentes de 15 conjuntos de datos y siete tareas.

Además, proponen *FLARE-ES*, el primer banco de evaluación bilingüe con 21 conjuntos de datos y nueve tareas, revelando una brecha de rendimiento multilingüe significativa en los LLM existentes. Los modelos *FinMA-ES* superan a LLM de última generación como *GPT-4* en tareas financieras en español, destacando el impacto positivo del ajuste por instrucciones multilingües y la transferencia cruzada entre lenguas. Este avance marca un paso decisivo hacia la democratización del análisis económico basado en lenguaje en contextos no anglófonos.

3.4.1. Análisis de sentimiento

El análisis de sentimiento aplicado a textos económicos y financieros enfrenta desafíos particulares debido a la naturaleza contextual del lenguaje en este dominio. Las palabras y expresiones utilizadas suelen adquirir connotaciones distintas según el entorno macroeconómico o la orientación del discurso, lo que dificulta la detección automática de polaridad mediante enfoques tradicionales. Esta problemática se intensifica en idiomas con recursos limitados, como el español, donde la disponibilidad de corpus especializados y herramientas de anotación es escasa.

En este contexto, García-Díaz et al. (2023) realizaron un estudio exhaustivo sobre análisis de sentimiento en textos financieros en español, explorando la combinación de distintos conjuntos de características para mejorar

la precisión de los modelos. Para ello, construyeron un corpus de 15,915 tuits anotados manualmente con las categorías positiva, negativa y neutral, y evaluaron el impacto de las representaciones basadas en embeddings contextuales y no contextuales junto con rasgos lingüísticos. Los mejores resultados —con un F1 ponderado de 73.16%— se obtuvieron al integrar de forma conjunta los diferentes conjuntos de características, demostrando que la combinación de conocimiento lingüístico y representaciones distribuidas potencia significativamente el rendimiento del análisis de sentimiento en textos financieros en español.

En una línea complementaria, Pan et al. (2023) abordan el análisis de sentimiento financiero en español desde una perspectiva dirigida a entidades económicas específicas. Reconociendo que la información financiera en redes sociales se ha convertido en un insumo clave para anticipar movimientos bursátiles, los autores plantean que la creciente diversidad de actores y opiniones sobre un mismo evento económico exige modelos capaces de identificar el objetivo principal del texto y estimar su polaridad hacia dicho objetivo, hacia otras empresas y hacia la sociedad en general.

Para ello, compilaron un nuevo corpus de tuits y titulares financieros en español, constituyendo un recurso de alto valor para la comunidad investigadora hispanohablante. Asimismo, realizaron una comparación del desempeño de distintos modelos de lenguaje específicos del español, destacando que MarIA y BETO obtuvieron los mejores resultados, con un desempeño global de 76.04%, 74.16% y 68.07% en *macro F1-score* para la clasificación de sentimiento hacia el objetivo económico principal, la sociedad y otras compañías, respectivamente, y una exactitud de 69.74% en la detección del objetivo. Además, la evaluación de modelos de clasificación multietiqueta alcanzó un rendimiento de 71.13%, demostrando la viabilidad del enfoque de análisis de sentimiento dirigido para capturar matices

contextuales y diferencias de polaridad entre distintos agentes económicos.

A pesar de los avances recientes, la literatura sobre análisis de sentimiento económico continúa enfrentando limitaciones estructurales. En primer lugar, la mayoría de los estudios depende de datos etiquetados manualmente, lo que restringe la escalabilidad de los modelos y acota su capacidad para adaptarse a contextos cambiantes del discurso económico.

Además, los corpus disponibles —incluso los más recientes en español— se concentran en el ámbito financiero o bursátil, dejando de lado dimensiones macroeconómicas relevantes como la inflación, la política monetaria o las expectativas de precios. Esta falta de recursos temáticos impide capturar los matices del lenguaje económico institucional y mediático que reflejan dinámicas inflacionarias o de incertidumbre.

En segundo lugar, el predominio de enfoques supervisados limita la exploración de estructuras latentes y relaciones emergentes entre lenguaje y variables económicas. Los modelos actuales se centran en clasificar textos según una polaridad predeterminada, sin traducir de manera explícita las señales discursivas en tendencias económicas observables. Este vacío metodológico plantea la necesidad de enfoques no supervisados o híbridos que permitan inferir dinámicas económicas a partir del comportamiento semántico del texto.

En este sentido, la metodología propuesta en este trabajo busca superar dichas limitaciones mediante la construcción de un corpus económico orientado a la inflación y la aplicación de técnicas de representación y segmentación textual capaces de transformar la información lingüística en indicadores interpretables y comparables con las series económicas oficiales.

Capítulo 4

Metodología

En este capítulo se presenta la propuesta metodológica, detallando el diseño general, la construcción del corpus, el preprocesamiento, la extracción de tópicos, las fases de procesamiento y el análisis de resultados. En particular, se describen los métodos utilizados para consolidar la propuesta y se explica cómo cada uno de ellos contribuye al análisis y la interpretación de los textos económicos.

El objetivo es establecer un método robusto para extraer y analizar la polaridad en los textos económicos, de modo que esta refleje el comportamiento de los fenómenos económicos reportados oficialmente por las instituciones de referencia.

Para alcanzar este propósito, se propone una estrategia metodológica estructurada en etapas sucesivas, que pueden describirse de la siguiente manera:

- conformación del corpus
- homologación y determinación del nivel de granularidad
- extracción de tópicos
- procesamiento de texto
- determinación del conjunto de entrenamiento
- detección de polaridad

- métodos de agregación de la polaridad
- evaluación del método.

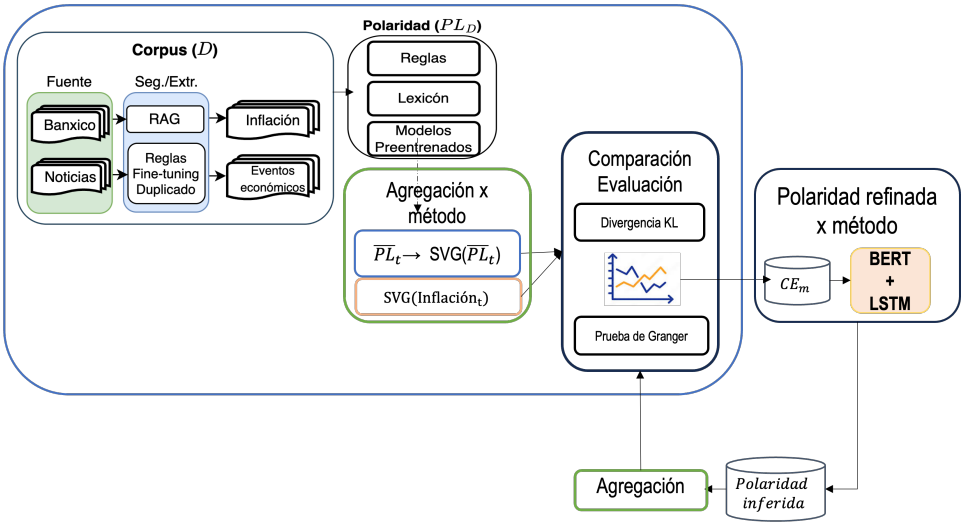

Figura 4.1: Propuesta métodologica de procesamiento y etiquetado

En la Figura 4.1 presentamos parte del proceso descrito anteriormente organizada en cuatro principales fases.

La primera fase del proceso consiste en la conformación del corpus, donde se recopilan los textos de interés que forman parte del corpus a analizar. Este conjunto de textos económicos es sometido a una serie de transformación y ajustes para su posterior análisis.

La fase de preprocesamiento comprende la normalización del corpus, la segmentación en oraciones y la extracción con base en tópicos económicos relevantes. Para eso, se emplean técnicas como LDA, así como de agrupamiento semántico que permiten reunir documentos basados en similitudes temáticas (Silge & Robinson, 2020).

Una vez extraídas las oraciones con contenido estrictamente económico, se obtiene un subconjunto del corpus más depurado y coherente, que cons-

tituye la base para describir y modelar la dinámica económica observada en los textos.

La siguiente etapa de la metodología corresponde a la detección de polaridad, componente esencial para evaluar el tono y la orientación del discurso económico presente en el corpus depurado. Para este fin, se implementaron tres enfoques complementarios, aplicados sobre el mismo conjunto de datos, con el propósito de comparar sus resultados y analizar su capacidad para capturar los matices del sentimiento económico:

- **Reglas**: Este primer enfoque se fundamenta en la definición de reglas morfosintácticas que combinan sustantivos y verbos de relevancia económica. A cada combinación se le asigna un valor de polaridad dentro del rango $[-1,1]$, lo que permite estimar si la oración expresa una valoración positiva, negativa o neutra respecto a la dinámica económica descrita.

- **Léxicon**: En el segundo enfoque se emplea un repertorio léxico en español previamente clasificado según su carga afectiva. La polaridad de cada oración se determina mediante la presencia de términos asociados a sentimientos positivos o negativos, lo que facilita una asignación directa del tono predominante en el texto.

- **Modelos preentrenados**: Finalmente, se utilizan modelos de lenguaje avanzados —como *BERT*— ajustados al dominio económico mediante técnicas de refinamiento o *fine-tuning*. Estos modelos aprovechan representaciones semánticas profundas para captar la relación contextual entre las palabras y asignar automáticamente la polaridad correspondiente.

La fase final del proceso metodológico consiste en agregar la información de polaridad en el tiempo, con el fin de observar la evolución del sentimiento

económico y detectar posibles tendencias o cambios en la narrativa a lo largo del periodo analizado.

Esta metodología destaca por su capacidad para identificar con precisión la polaridad del discurso económico y reflejar su evolución en el tiempo. Los resultados permiten detectar patrones y anomalías en la dinámica económica, constituyendo una herramienta útil para el análisis financiero, la formulación de políticas y la anticipación de cambios en indicadores clave (Shapiro & Sudhof, 2022). La Figura 4.1 resume el flujo general del proceso y la relación entre las etapas descritas.

4.0.1. Construcción de los datasets

El conjunto de datos utilizado se compone de tres principales fuentes de información, por un lado los documentos que emanen de las noticias económicas, por otro lado los informes del Banxico finalmente las estadísticas oficiales del INEGI.

4.0.1.1. Noticias económicas

Las noticias económicas se obtienen del repositorio de noticias que compila el Instituto de Investigaciones Económicas de la UNAM y publicadas a diario en el portal de la Universidad[1]. La ventaja del repositorio de noticias de la UNAM es que ofrece un resumen de las noticias pasadas donde no es posible rescatar el contenido de la fuente original.

Compilamos las noticias diarias publicadas desde el 25 de abril de 2006 hasta el 31 de diciembre de 2024 usando un crawler. Este último fue diseñado para extraer el contenido de las páginas web y almacenar los documentos en un formato estructurado usando como caraterísticas la fecha de publicación del documento y el título. Los documentos comprenden en general

[1]UNAM - Instituto de Investigaciones Económicas

noticias de los siguientes periódicos (se enlistan por orden alfabético):

- El Economista
- El Financiero
- El País
- Europa Press
- Excélsior
- Financial Times
- LaJornada
- Milenio
- New York Times
- Reforma

Es importante señalar que el número de periódicos incluidos no refleja la totalidad de los medios presentes en el ecosistema informativo, ya que muchas notas son reproducidas o citadas por los diarios enlistados. Asimismo, el corpus puede mostrar una sobrerrepresentación de ciertos medios, dado que algunos periódicos se especializan en temas económicos específicos, mientras que otros publican de manera más general. Esta distribución también depende de la disponibilidad de los artículos en los portales digitales y de las políticas editoriales de publicación de cada medio.

4.0.1.2. Informes del Banco de México

Los informes del Banco de México son documentos oficiales que contienen información relevante sobre la política monetaria y la situación económica

del país. Estos informes son publicados periódicamente y están disponibles en el sitio web del Banco de México[2]. Se componen de una serie de documentos que incluyen:

- Informe trimestral sobre la inflación
- Anuncios de decisiones de la política monetaria
- Minutas de las reuniones de la Junta de Gobierno
- Reportes de estabilidad financiera

Todos estos documentos se agregan usando la fecha de publicación para formar nuestro corpus. La importancia de estos informes radica en que contienen análisis detallados sobre la inflación, el crecimiento económico y otros indicadores clave que son fundamentales para la toma de decisiones económicas.

Asimismo, estos documentos reúnen los argumentos y fundamentos que sustentan las decisiones de política monetaria, lo que facilita una comprensión más profunda del contexto económico e institucional en el que dichas decisiones se formulan.

4.0.1.3. Estadísticas del INEGI

El Instituto Nacional de Estadística y Geografía (INEGI)[3] es la entidad responsable de producir las estadísticas oficiales en México. La institución proporciona una amplia gama de indicadores económicos, demográficos y sociales que son fundamentales para informar a la población. En nuestro caso, utilizamos las estadísticas del INPC como guía para evaluar las polaridades que emanen de los documentos.

[2]Banco de México
[3]INEGI

El INPC, compilado mensualmente, refleja la variación mes con mes de los precios de un conjunto de bienes y servicios que suelen consumir en promedio los hogares en México.

4.0.2. El corpus

4.0.2.1. Web scraping

El corpus se construyó a partir de los documentos extraídos de las fuentes anteriormente mencionadas. Primero, mediante BeautifulSoup[4] y Selenium[5] se extrajo el contenido de los documentos en formato HTML, de manera estructurada, organizándolo por fecha de publicación se almacenó en texto plano en un archivo CSV.

El periódico LaJornada ofrece de manera estructurada contenido de sus noticias desde 1996, lo que permite rescatar información desde 1999 hasta la fecha.

4.0.2.2. Extracción de las decisiones de Banxico

El procesamiento automatizado de las decisiones del Banco de México parte de la extracción de texto desde documentos en formato PDF. Para ello, se emplean herramientas especializadas que permiten convertir el contenido de los archivos manteniendo su estructura original, facilitando su posterior análisis vease 4.2.

El flujo general del procesamiento se compone de las siguientes etapas:

1. Lectura de documentos PDF almacenados en un directorio específico.

2. Extracción y unificación del contenido textual de todas las páginas en bloques procesables.

[4]BeautifulSoup
[5]Selenium

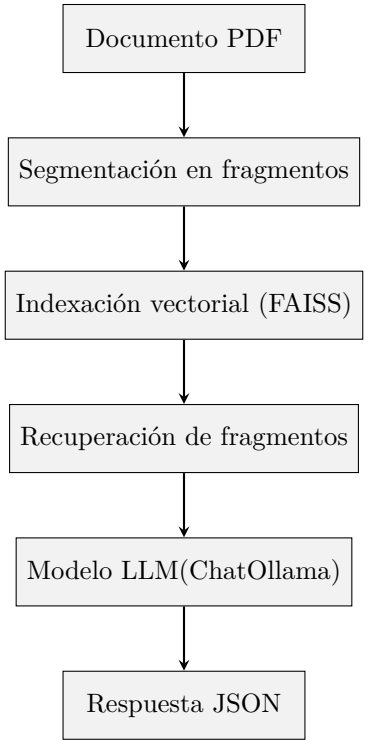

Figura 4.2: Pipeline del método la extracción de decisiones monetarias.

3. Segmentación del texto en fragmentos semánticos adecuados para la indexación y recuperación.

Para estructurar y extraer información clave de cada documento, se utiliza un LLM (Brown et al., 2020). Mediante un sistema de instrucción en español (véase Cuadro 4.0.2.2), el modelo identifica y extrae dos elementos principales:

- Fecha de la reunión de la Junta de Gobierno o publicación del documento.

- Resumen conciso de la decisión tomada por la Junta de Gobierno en materia de política monetaria.

Prompt: Extracción de las decisiones de política monetaria

Eres un asistente experto en analizar decisiones de política monetaria del Banco Central de México.

1. Tu tarea es extraer exclusivamente la **fecha de publicación** y la **decisión tomada** por la Junta de Gobierno respecto a la inflación, tasas de interés y otras medidas monetarias.
2. Si hay opiniones divergentes, inclúyelas en la respuesta.
3. Si no se menciona una decisión explícita, responde con `"No encontrada"`.
4. Regresa la respuesta en formato JSON.

Este proceso garantiza respuestas estructuradas en formato JSON, facilitando la integración y análisis.

4.0.2.3. Extracción automatizada usando LLM y Reglas

A pesar de la alta precision de los modelos de lenguaje, se identificaron casos en los que no lograron extraer la fecha o el resumen. Para mitigar esta limitación, se implementó un post-procesamiento mediante expresiones regulares (Regex) que permite detectar y corregir fechas no capturadas.

Los datos extraídos se almacenan en archivos JSON, cada uno asociado a un documento específico. La estructura básica de estos archivos incluye:

- La `fecha` de publicación del documento, corresponde a la fecha de la decisión del Banco.
- El `resumen` de la decisión monetaria.

Para asegurar la coherencia interna del corpus, se implementó un procedimiento de depuración y estandarización que comprendió:

Input: Ruta PDF *doc_path*
Output: Fecha y decisión monetaria en JSON

Paso *1: Cargar PDF*
  if *doc_path no existe* **then**
    return *Lista vacía*
  data ← `PDFMinerLoader`(*doc_path*).`load()`
  data ← `PyMuPDFLoader`(*doc_path*).`load()` **return** *Lista vacía*
Paso *2: Preprocesar y segmentar*
  if *data vacío* **then**
    return *Lista vacía*
  full_text ← concatenar `page_content` de *data*;
  full_text ← limpiar saltos y espacios dobles;
  chunks ← `RecursiveCharacterTextSplitter`(*chunk_size* = 2000, *chunk_overlap* = 300, *separators* = ["\n\n", "\n", " ", ""]).`split_text`(*full_text*);
  chunk_docs ← crear `Document` por chunk;
  return *chunk_docs*
Paso *3: Embeddings y FAISS*
  if *chunk_docs vacío* **then**
    return *None*
  embedding_model ← "sentence-transformers/all-MiniLM-L6-v2";
  relevant_chunks ← filtrar *chunk_docs* con "decisión de política monetaria" (case-insensitive);
  if *relevant_chunks vacío* **then**
    relevant_chunks ← *chunk_docs*
  vector_store ← `FAISS.from_documents`(*relevant_chunks*, *embedding_model*);
  return *vector_store*
Paso *4: Recuperar y extraer info*
  if *vector_store es None* **then**
    return *{fecha: "No encontrada", decisión: "No encontrada"}*
  llm ← `ChatOllama`(*modelo* = *llm_model*, *dispositivo* = *device*);
  query ← "Decisión política monetaria Banco de México en este documento";
  retriever ← *vector_store.as_retriever*(*k* = 5);
  relevant_docs ← *retriever.get_relevant_documents*(*query*);
  if *relevant_docs vacío* **then**
    return *{fecha: "No encontrada", decisión: "No encontrada"}*
  combined_text ← concatenar `page_content` de *relevant_docs*;
  mensaje ← construir prompt LLM para extraer fecha y decisión JSON;
  respuesta ← `ollama.chat`(*modelo* = *llm_model*, *mensajes* = *mensaje*);
  resultado ← `parsear JSON`(*respuesta*)
  resultado ← {*fecha* : ‘‘*Nodisponible*′′, *decisin* : ‘‘*Nodisponible*′′} **return** *resultado*
Ejecución principal
  data ← Paso 1(*doc_path*); *chunks* ← Paso 2(*data*); *vector_db* ← Paso 3(*chunks*); *info* ← Paso 4(*vector_db*); `imprimir`(info)

Algorithm 1: Proceso de extracción de decisiones de política monetaria del Banco de México

1. **Normalización de espacios en blanco.** Se suprimieron duplicidades de espacio y se corrigieron espacios indebidos adyacentes a signos de puntuación, además de homogeneizar saltos de línea. Con ello se garantizó una estructura oracional uniforme sin alterar el contenido.

2. **Regularización de la puntuación.** Se colapsaron repeticiones innecesarias (p. ej., secuencias de comas o puntos consecutivos) y se aplicaron reglas consistentes de espaciado tras signos. Este ajuste preservó los casos en que el punto funciona como separador decimal en expresiones numéricas.

3. **Depuración de redundancias textuales.** Se identificaron y retiraron enunciados duplicados o casi duplicados, tanto dentro de un mismo documento como entre documentos. Para ello se combinó coincidencia exacta con verificación de similitud aproximada, con el fin de minimizar sesgos por repetición sin perder cobertura informativa.

4.0.2.4. Recuperación de información de Banxico

Para facilitar la recuperación eficiente de información económica relevante, se desarrolló un índice semántico basado en técnicas de recuperación aumentada por generación (RAG, por sus siglas en inglés):

1. **Indexación:** se aplica al corpus limpio un modelo de `sentence-transformers` para generar los embeddings (Araci, 2019; Reimers & Gurevych, 2019).

2. **Almacenamiento vectorial:** estos embeddings se almacenan en un motor de bases de datos (FAISS) para realizar la búsqueda (J. Johnson et al., 2019).

3. **Recuperación:** cada fragmento de texto se representó mediante un vector de incrustación almacenado en FAISS; para cada consulta, se calculó su incrustación y se recuperaron desde dicho índice las oraciones más similares según su proximidad vectorial.

4. **Generación estructurada:** cada oración recuperada se reescribió mediante un modelo de lenguaje (`LLaMA`), asegurando el formato requerido y corrección gramatical.

El procesamiento automatizado de documentos del Banco de México combinó técnicas avanzadas de extracción, PLN y almacenamiento estructurado para obtener datos confiables y accesibles. La integración de modelos generativos con búsqueda vectorial (RAG) y técnicas de limpieza mejoró significativamente la precisión y utilidad de la información extraída.

4.0.2.5. Integración del corpus

El proceso de limpieza de los documentos se realizó utilizando librerías básicas de NLP como es el **NLTK**[6], **Spacy**[7] y **Regex**[8]. En ocasiones, se requiere recuperar la fecha de las ligas o directomente dentro del texto.

Los contenidos suelen tener etiquetas de HTML, anuncios, imágenes y otros elementos que no son relevantes para el análisis. Por lo que se procedió a eliminar estos elementos y dejar solo el texto lo más limpio posible.

La fase de limpieza de texto consta de los siguientes pasos:

- **Eliminación de ruido:** se descarta toda la información no relevante (publicidad, menús, pies de página, avisos legales, menús de navegación, secciones de contacto, etc.).

[6]NLTK
[7]Spacy
[8]Regex

- **Extracción de contenido económico:** se filtran únicamente las frases o párrafos que contienen datos cuantitativos, hechos, análisis y declaraciones económicas.

- **Corrección ortográfica y gramatical:** se revisan y corrigen errores de escritura, concordancia y sintaxis.

- **Normalización de puntuación y espacios:** se unifican signos redundantes, se elimina espaciado incorrecto y se asegura la correcta segmentación de oraciones, preservando números decimales.

- **Reagrupación y coherencia:** se mantiene el orden lógico de la información y se agrupan elementos afines, retornando `UNKNOWN` cuando la corrección no es posible.

4.0.2.6. Classificación de contenidos económicos con base en LLM

La extracción de contenidos se hace mediante el uso de los modelos de lenguaje. Esto debido a la alta precisión de dichos modelos para clasificar y predecir los subsiguientes tokens de una serie.

Para eso se usó la siguiente instrucción:

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"`.

A continuación se muestra el algoritmo utilizado para completar la limpieza del corpus. La idea es hacer uso de varios llms (Llama 3.3, Mistral y DeepSeek) donde se busca mantener la integridad del corpus. El texto de salida está basado en un sistema de puntuación que refleja la coherencia semántica con el texto original.

Input: Texto original `text`

Output: Texto limpio seleccionado

Paso *1: Cargar prompt de limpieza*
  prompt ← `load_prompt_template()`;

Paso *2: Refinar en paralelo con cada LLM*
  responses ← {};
  foreach $m \in$ {`"llama3.3"`, `"mistral"`, `"deepseek-r1"`} **do**
    responses[m] ← `call_llm`(m, *prompt*, *text*);

Paso *3: Evaluar cada respuesta*
  scored ← {};
  foreach (*model*, *candidate*) ***en*** *responses* **do**
    sim ← `measure_cosine_similarity`(*text*, *candidate*);
    cleaness ← `measure_cleanliness`(*candidate*);
    scored[*model*] ← $0.6 \times$ *sim* $+ 0.4 \times$ *cleaness*;

Paso *4: Seleccionar la mejor respuesta*
  best_model ← $\arg\max_m$ *scored*[m];
  return *responses*[*best_model*];

Algorithm 2: Proceso de limpieza usando LLMs

En conclusión, la Tabla 4.1 ilustra un ejemplo representativo del proceso de limpieza y extracción aplicado al corpus. En la primera columna se presenta la fecha de la publicación, mientras que la segunda columna contiene el contenido original extraído directamente de la fuente.

Este texto incluye información de naturaleza heterogénea, como datos económicos relevantes, anuncios institucionales y menciones a eventos administrativos o de servicios públicos, entre otros.

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.

Cuadro 4.1: Ejemplo de salida del proceso de limpieza

El campo “texto limpio” muestra el resultado del proceso de depuración. Aquí se observa una reducción significativa del ruido informativo,

eliminando detalles irrelevantes para el análisis económico, tales como instrucciones operativas de servicios públicos o resoluciones judiciales ajenas a la economía. Se prioriza la preservación de información económica clave, como indicadores de consumo, movimientos de reservas internacionales, importaciones y apreciaciones de las autoridades financieras.

Además, se realiza una normalización del lenguaje, corrigiendo inconsistencias gramaticales y mejorando la coherencia discursiva. El resultado es un texto más conciso, enfocado en los elementos económicos centrales y estructurado para facilitar tanto el análisis humano como el procesamiento automatizado posterior.

Este proceso no solo contribuye a la homogeneización del corpus, sino que también incrementa la calidad y relevancia de los datos que alimentarán los modelos de análisis de sentimiento, extracción de tópicos o cualquier otra tarea de procesamiento de lenguaje natural.

Así, se sientan las bases para obtener inferencias robustas y comparables a lo largo de las distintas observaciones temporales y documentales.

4.0.3. Extracción de tópicos (LDA + LLM)

La propuesta metodológica consta de dos grandes fases secuenciales que combinan métodos de representación semántica, agrupamiento y modelado de tópicos para aislar únicamente aquellas oraciones pertenecientes al dominio económico.

A continuación se presenta el esquema general (Figura 4.3) y su descripción detallada.

4.0.3.1. Generación de clusteres económicos

Dado un corpus de oraciones $\mathcal{S} = \{s_i\}_{i=1}^{N}$, se aplica un encoder semántico $f : \mathcal{S} \to \mathbb{R}^d$ (por ejemplo, Sentence-BERT Reimers y Gurevych, 2019) para

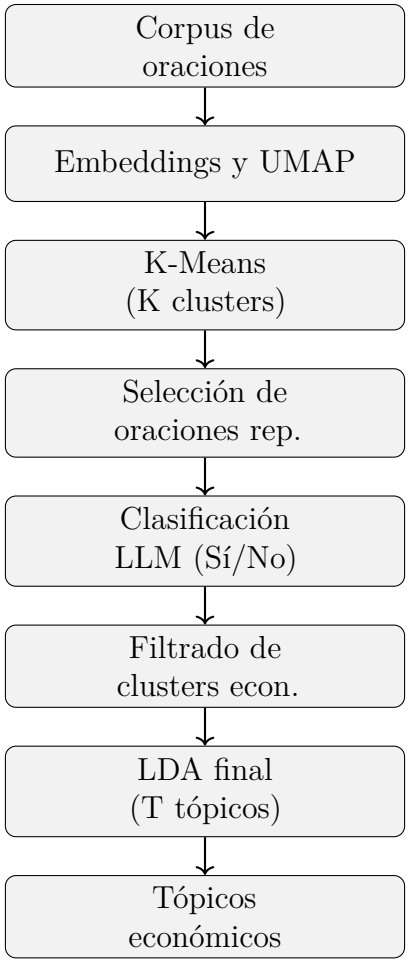

Figura 4.3: Extracción de tópicos económicos basado en LDA y LLM

obtener vectores $z_i = f(s_i)$.

A continuación, se reduce la dimensionalidad con UMAP McInnes et al., 2018 preservando la estructura de vecindad:

$$\{y_i\}_{i=1}^{N} = \text{UMAP}(\{z_i\}_{i=1}^{N}), \quad y_i \in \mathbb{R}^2. \tag{4.1}$$

4.0.3.2. Agrupamiento de oraciones

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

$$\{\mu_k\}_{k=1}^{K} = \arg\min_{\{\mu_k\}} \sum_{i=1}^{N} \min_{1\leq k\leq K} \|y_i - \mu_k\|^2, \tag{4.2}$$

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

4.0.3.3. Selección de oraciones representativas

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

$$\text{rep}_k = \underset{i:\, c(i)=k}{\arg\min\,\text{k}_\text{r}} \|y_i - \mu_k\|. \tag{4.3}$$

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

4.0.3.4. Filtrado de clusteres económicos

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

$$\rho_k = \frac{1}{|C_k|} \sum_{i:\, c(i)=k} \nVdash_{econ}(s_i), \tag{4.4}$$

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

4.0.3.5. Modelado temático con LDA

Sobre el subconjunto filtrado se entrena un modelo LDA Blei et al., 2003. El proceso generativo asume:

$$\theta_d \sim \text{Dir}(\alpha), \tag{4.5}$$

$$\phi_k \sim \text{Dir}(\beta), \tag{4.6}$$

$$z_{dn} \sim \text{Multinomial}(\theta_d), \tag{4.7}$$

$$w_{dn} \sim \text{Multinomial}(\phi_{z_{dn}}). \tag{4.8}$$

Los hiperparámetros α, β y el número de tópicos T se seleccionan mediante métricas de coherencia Greene et al., 2015.

A continuación se resumen los pasos clave del proceso:

1. Pre-filtrado: K-Means sobre *embeddings* para descartar clústeres con baja proporción de oraciones económicas.

2. Estimación LDA: entrenamiento sólo en el subconjunto resultante.

3. Selección de T: maximización de coherencia en validación.

4.0.4. Clasificación temática posterior a LDA

LDA identifica patrones de coocurrencia léxica en textos económicos, pero no asigna de forma directa esos patrones a categorías macroeconómicas. Para cerrar esa brecha, se aplica una capa de *clasificación ligera* sobre las oraciones candidatas, que mapea cada fragmento a una categoría única y utilizable en el análisis.

Eres un experto en economía. Tu tarea es clasificar los siguientes tópicos en una de las categorías macroeconómicas.
Tópico: <TÓPICO EXTRAÍDO POR LDA>
Categorías:

- Inflación
- Índice de Confianza del Consumidor
- Política Monetaria
- Consumo Privado
- Gasto Público
- Crecimiento Económico
- Comercio Exterior
- Mercados Financieros
- Empleo y Mercado Laboral
- Energía y Commodities
- Inversión
- Deuda y Crédito
- Otro (especificar)

Reglas:

- Elige - **Inflación** solo si aparecen cambios de precios, nivel de precios, poder adquisitivo o tipos de inflación.
- Elige - **Índice de Confianza del Consumidor** si hay menciones explícitas a sentimiento/expectativas de los consumidores.
- Si hay varias candidatas, prioriza: Inflación → Política Monetaria → Consumo Privado → Crecimiento Económico → resto.
- Devuelve **únicamente** el nombre exacto de la categoría, sin explicación adicional.

Este prompt cumple varias funciones críticas:

- Alineamiento semántico: transforma el output de LDA, consistente en distribuciones de probabilidad sobre temas, en etiquetas categóricas precisas, reduciendo la ambigüedad de los tópicos identificados P. Liu, Yuan et al., 2021.

- Filtro de interés: al centrar la clasificación explícita en Inflation o Consumer Confidence Index, se garantiza que las frases seleccionadas posteriores al LDA cumplan con los criterios de análisis de sentimiento sobre inflación Wei et al., 2022b.

- Consistencia metodológica: la lista exhaustiva de categorías y las instrucciones condicionales estandarizan el proceso, facilitando la reproducibilidad y la comparabilidad de resultados entre distintos lotes de noticias Pang y Lee, 2008.

- Preparación para análisis de sentimiento: al disponer de frases claramente clasificadas en temas de inflación, la etapa subsiguiente de análisis de sentimiento presenta mayor precisión y menor ruido, según demuestra la literatura en minería de opiniones Cambria et al., 2016.

4.1. Inferencia de polaridad

Una vez extraídas las oraciones relacionadas con la inflación, el siguiente paso es inferir su polaridad (positiva, negativa o neutral) en relación con el sentimiento inflacionario. Para ello, se emplea un enfoque híbrido que combina modelos de lenguaje (LLM) con un sistema basado en reglas léxicas.

4.1.1. Reglas

Basado en el trabajo de Brill (1995), Gorodnichenko et al. (2021) y A. Shah et al. (2023), se construyó una matriz de conceptos que permite inferir el sentimiento inflacionario de las frases extraídas. La matriz considera la relación entre los factores explicativos de la inflación y los verbos que definen el sentido de la inflación.

Partiendo de la definición de la inflación, como un aumento de los precios, se selecciona todos los sustantivos, adjetivos y verbos que explican la inflación y su relación con la actividad económica.

Panel A1	**Panel B1**
inflación, precios, índice de precios al consumidor, expectativa de inflación, tasa de interés, tasa bancaria, precios, inflación subyacente, actividad económica, desempleo, presiones inflacionarias, dólar, tipo de cambio	empleo, crecimiento económico, productividad, tipo de cambio, déficit, demanda, mercado laboral, política monetaria, peso, inversión, consumo
Panel A2	**Panel B2**
disminuir, reducir, bajar, caer, pausar, lento, declinar, someter, retroceder, desacelerar, despreciar, carecer, merma, menguar	apreciar, aliviar, facilitar, aumentar, incrementar, expandir, mejorar, repuntar, alto, elevar, subir, ascender, superar, ganar, acelerar
Panel C	
no estaban, no estaba, no va, no, no es, no era, no ha sido, no será, mantener sostener, sin señales, sin evidencias	

Cuadro 4.2: Matriz de reglas para inferir polaridad de inflación

Este enfoque de regla opera sobre listas de términos predefinidas normalizando cada oración a minúsculas sin acentos (`unidecode`). Se definen cuatro conjuntos léxicos y un conjunto de reversores:

- A_1: *sustantivos/constructos macroeconómicos* centrados en inflación, precios, tasas de interés, dólar, actividad, desempleo.
- B_1: *conceptos macroeconómicos* de empleo, crecimiento, productividad, tipo de cambio, demanda, política monetaria, etc.
- A_2: *verbos/tendencia de caída o contención* ("disminuir", "bajar", "pausa", "estable", "declinar", "desaceleración", …).
- B_2: *verbos/tendencia de aumento o expansión* ("aumentar", "incrementar", "expansión", "mejorar", "repunte", …).
- C: *reversores* y expresiones de mantenimiento ("no", "no va", "no es", "mantener", "sostiene", …).

Variante 1 (co-ocurrencia, lógica booleana) La versión comentada en el código clasifica la oración s como positiva si aparece una *co-ocurrencia* de tópico con dirección congruente:

$$\text{pos}(s) = \big[(A_1 \wedge A_2) \ \vee \ (B_1 \wedge B_2)\big]; \quad \text{neg}(s) = \big[(A_1 \wedge B_2) \ \vee \ (A_2 \wedge B_1)\big].$$

Si se detecta algún elemento de C, la señal se invierte ($y \leftarrow -y$). Esta lógica exige *dos* piezas: un término económico y un verbo direccional, reduciendo falsos positivos por coincidencias aisladas.

Variante 2 (activa en el código: voto por conteo) La implementación efectiva abandona la co-ocurrencia estricta y suma coincidencias tipo

bolsa de palabras:

$$\text{pos\_count}(s) = \sum_{w \in A_1} \mathbf{1}\{w \in s\} + \sum_{w \in A_2} \mathbf{1}\{w \in s\}, \quad \text{neg\_count}(s) = \sum_{w \in B_1} \mathbf{1}\{w \in s\}$$

La etiqueta escalar preliminar es

$$y(s) = \operatorname{sign}\big(\text{pos\_count}(s) - \text{neg\_count}(s)\big) \in \{-1, 0, 1\},$$

y se invierte si aparece algún $c \in C$ ($y \leftarrow -y$). Además, se calcula una *polaridad continua* normalizada por longitud:

$$\text{pol}(s) = \frac{\text{pos\_count}(s) - \text{neg\_count}(s)}{\#\text{palabras}(s)},$$

que se reporta junto con la etiqueta triclase (`positive`/`negative`/`neutral`).

4.1.2. Lexicón

El método basado en léxicon implementado utiliza un diccionario predefinido de palabras y frases con puntuación de polaridad para calcular el sentimiento económico de cada oración. Este enfoque sigue los principios de los sistemas clásicos de análisis de sentimiento, donde la suma ponderada de términos positivos y negativos determina la orientación global del texto (Taboada et al., 2011).

El procedimiento se desarrolla de la siguiente manera:

1. **Carga y preprocesamiento del léxicon:** se importa el archivo de léxicon y se normaliza el inventario sustituyendo guiones bajos por espacios, aplicando `unidecode` y minúsculas. Conforme al archivo fuente, las etiquetas con valor `2` se reasignan a -1 (negativas) y las `1` se mantienen como positivas; de este modo se evita ambigüedad y se deja el diccionario binario y consistente con la implementación.

2. **Tokenización y lematización:** cada oración se procesa con `spaCy` (`es_core_news_sm`), deshabilitando `parser` y `ner` para eficiencia. Se eliminan signos de puntuación y se trabaja con lemas normalizados (`unidecode`+`lower`) para maximizar coincidencias con el léxicon.

3. **Cálculo de polaridad con manejo de negaciones:** se recorre la secuencia de tokens; por cada coincidencia con el léxicon se suma $+1$ si el término es positivo y -1 si es negativo. Las palabras de negación (*no, ni, nunca, tampoco*, etc.) activan un *toggle* que invierte el signo de las coincidencias subsiguientes hasta una nueva negación. La polaridad continua de la oración se define como

$$\text{polaridad\_lexicon} = \frac{\#\text{positivos} - \#\text{negativos}}{\#\text{tokens}},$$

lo que corrige por longitud y permite comparabilidad entre oraciones.

4. **Asignación de etiqueta:** según el signo de polaridad_lexicon, la oración se clasifica como `positive`, `negative` o `neutral`.

5. **Salida estructurada:** las columnas `lexicon_polarity` (continua) y `lexicon_label` (triclase) se integran al `DataFrame` original para análisis posteriores y evaluación cruzada con otros métodos.

Este enfoque es transparente y trazable: cada decisión puede explicarse por coincidencias léxicas lematizadas y reglas de negación simples. Entre sus límites están el alcance global de la negación (no acotado por ventana) y la ausencia de desambiguación semántica; ambos pueden mitigarse con reglas de alcance (k-ventanas) o ponderaciones por término/fuente.

Cuadro 4.3: Distribución del léxicon utilizado (positivos = 1, negativos = −1)

Clase	Conteo	Proporción
Positivos (= 1)	8,146	40.5%
Negativos (= −1)	11,959	59.5%
Total	20,105	100%

4.1.3. Transformer

4.1.3.1. Pysentimiento

El etiquetado mediante **pysentimiento** se realizó usando el modelo `pysentimiento-robe` un modelo transformer basado en la arquitectura RoBERTa y entrenado específicamente para español sobre grandes corpus de redes sociales, noticias y foros (Pérez et al., 2021).

Este modelo destaca por su robustez en tareas de análisis de sentimiento y emociones en español, adaptándose a diferentes registros lingüísticos y expresiones idiomáticas frecuentes en el ámbito hispanohablante.

La secuencia de procesamiento incluye:

1. **Normalización y tokenización:** preparación del texto y segmentación en tokens conforme al tokenizador RoBERTa.

2. **Inferencia con modelo especializado:** clasificación de sentimiento mediante `pysentimiento-robertuito-base-uncased`, que asigna probabilidades a las clases `POS`, `NEG`, y `NEU`.

3. **Conversión y almacenamiento:** selección de la clase más probable y registro de la probabilidad asociada como polaridad continua.

4. **Integración de resultados:** adición de las columnas de sentimiento y polaridad al DataFrame original.

El uso de este modelo asegura una clasificación eficiente y confiable para textos en español, aunque puede presentar ligeras limitaciones en textos altamente técnicos o financieros.

4.1.3.2. BERT

Se utilizó **`nlptown-bert-base-multilingual-uncased-sentiment`**, un BERT multilingüe ajustado para clasificación de sentimiento en varios idiomas. Preentrenado sobre grandes corpus y afinado con reseñas, identifica polaridad en textos breves, incluido español, aprovechando la arquitectura Transformer de BERT (Devlin et al., 2019b).

El procedimiento se explica mediante los siguientes pasos:

1. *Tokenización*: Tokenización con el *tokenizer* del modelo.

2. *Inferencia*: asignación de una etiqueta ordinal de sentimiento (de muy negativo a muy positivo) por oración.

3. *Polaridad*: proyección de las probabilidades a una métrica escalar para análisis comparables.

El modelo es versátil y robusto en entornos multilingües; no obstante, puede perder matices económicos específicos frente a modelos especializados.

4.1.3.3. FinBERT

Se empleó el modelo **`bardsai-finance-sentiment-es-base`**, una variante de BERT adaptada al dominio financiero en español. Entrenado con noticias, reportes y textos bancarios, capta patrones léxicos y contextuales propios del discurso económico y de política monetaria (Araci, 2019).

A continuación se describen los pasos:

1. *Tokenización*: tokenización con el *tokenizer* del modelo.

2. *Inferencia*: clasificación de cada oración en `positive`, `negative` o `neutral` con probabilidades por clase.

3. *Polaridad*: conversión a métrica escalar (p. ej., $p(\text{pos}) - p(\text{neg})$).

4. *Salida*: incorporación de etiqueta y polaridad al `DataFrame`.

La especialización sectorial mejora la sensibilidad a términos técnicos y señales de mercado, proporcionando estimaciones más consistentes en textos económicos.

4.2. Enfoque semi-supervisado de inferencia de polaridades

Las polaridades inferidas anteriormente se basan en un enfoque completamente no supervisado. Si bien ello evita depender de etiquetas predefinidas, también puede introducir sesgos debido al desconocimiento de la distribución a priori de las clases. Para mitigar estos posibles sesgos, se propone una adecuación en la metodología que permite corregir el sesgo extraendo un subconjunto de observaciones cuyas polaridades se deducen a partir de la comparación con la serie de inflación.

Sea $\mathcal{D} = \{(x_i, s_i, t_i)\}_{i=1}^{N}$ un corpus de N oraciones económicas x_i, con etiqueta de fuente s_i (p. ej., `Regla`, `LDA`) e índice temporal t_i. El objetivo es inferir el sentimiento latente $y_i^* \in \{-1, 0, 1\}$ (negativo, neutro, positivo) o una polaridad continua $y_i^* \in \mathbb{R}$, con foco en fenómenos económicos (inflación, choques financieros, etc.).

A diferencia de dominios generales, aquí predominan escasez de anotación ($|\mathcal{L}| \ll |\mathcal{D}|$), lenguaje especializado y complejidad multilingüe. Las etiquetas disponibles $\tilde{y}_i$ provienen mayormente de supervisión débil (reglas,

léxicos, predicciones *zero-shot* de LLMs):

$$\tilde{y}_i = \mathcal{A}(x_i), \quad \mathcal{A} \in \{\text{reglas}, \text{léxicon}, \text{LLM}, \text{LLM financiero}\},$$

por lo que son ruidosas: $\mathbb{P}(\tilde{y}_i \neq y_i^*) > 0$.

En paralelo al esquema no supervisado, se implementa un refinamiento basado en LLMs para atenuar sesgos y ruido de la supervisión débil. Los LLMs, preentrenados en grandes corpus, capturan matices económicos en entornos de baja anotación Araci, 2019; Z. Zhang et al., 2022, pero sus salidas pueden estar descalibradas Ye y Shah, 2025. Para corregirlo, se usa *refinamiento iterativo* con *pseudoetiquetas*: un modelo f_θ se ajusta a predecir $\tilde{y}_i$ y luego se recalibra con el subconjunto anotado $\mathcal{L}$:

$$\hat{y}_i = f_\theta(x_i \mid \tilde{y}_i), \qquad \theta^* = \arg\min_\theta \sum_{i \in \mathcal{L}} \ell\big(g(f_\theta(x_i)),\, y_i\big),$$

donde $g(\cdot)$ es una capa de calibración (p. ej., Platt/isotónica) aprendida con las pocas etiquetas oro y_i. Así, la primera etapa aprovecha señal masiva (débil) y la segunda alinea niveles y umbrales al estándar humano, reduciendo el sesgo del método no supervisado.

4.2.1. Modelo de refinamiento supervisado débil (por método y fuente)

Se utiliza un *encoder* Transformer multilingüe (`bert-base-multilingual-cased`) con una cabeza de regresión de una sola salida (`num_labels`=1). El tokenizador segmenta cada oración x_i en subpalabras y el *encoder* produce una representación contextual; la cabeza lineal proyecta dicha representación a un escalar $\hat{y}_i \in \mathbb{R}$ que interpreta la polaridad continua.

El entrenamiento es de tipo *supervisión débil*: para cada combinación

[`método`, `fuente`] se ajusta un modelo a predecir las pseudoetiquetas continuas $\tilde{y}_i$ (obtenidas de `rule`, `lexicon`, `bert`, `finance`, `pysentimiento`) sobre el subconjunto de entrenamiento de esa fuente. La función de pérdida es el error cuadrático medio (MSE):

$$\min_{\theta} \frac{1}{|\mathcal{D}_{m,s}^{\text{train}}|} \sum_{i \in \mathcal{D}_{m,s}^{\text{train}}} \big(f_\theta(x_i) - \tilde{y}_i\big)^2,$$

donde f_θ es el modelo BERT+regresión. Esta formulación busca *calibrar* y suavizar el ruido de la señal débil preservando regularidades semánticas del dominio.

El entrenamiento se hace *por bloque* de método m y fuente s (p. ej., Regla vs LDA vs LDA+LLM), lo que induce modelos especializados $f_\theta^{(m,s)}$ que capturan contextos léxicos y de estilo propias de cada combinación. Para inferencia, se excluyen los textos usados en entrenamiento dentro de ese bloque, mitigando *leakage*.

Se emplea AdamW con tasa de aprendizaje 2×10^{-5}, batch (lotes) de 8 y entrenamiento por 5 épocas. En cada época se itera sobre `DataLoader`, se calcula la pérdida por `AutoModelForSequenceClassification` (modo `regression`) y se actualizan los parámetros. Se registra la pérdida media por época para monitorear convergencia.

El modelo produce una *polaridad corregida* continua $\hat{y}_i$ para los textos de inferencia del mismo $[m, s]$. Para facilitar lectura, se deriva una etiqueta de 3 clases mediante umbrales simétricos con tolerancia ε:

$$\texttt{label}(\hat{y}_i) = \begin{cases} \text{“positive”}, & \hat{y}_i > \varepsilon, \\ \text{“negative”}, & \hat{y}_i < -\varepsilon, \\ \text{“neutral”}, & \text{en otro caso.} \end{cases}$$

Las principales ventajas de este refinamiento son las siguientes:

- Aprender por $[m, s]$ actúa como *adaptación de dominio* fina, evitando que un único modelo mezcle distribuciones heterogéneas.
- El objetivo cuadrático sobre pseudoetiquetas reduce varianza y atenúa extremos espurios típicos de reglas/léxicos, induciendo una señal más estable.
- La proyección escalar preserva ordenaciones (útil para series temporales y agregados) y facilita posterior *calibración* con un pequeño conjunto con etiquetas disponibles.

4.3. Descomposición estacional de los datos observados

Sea $\{x_t\}_{t=1}^{N}$ la serie temporal de inflación mensual. Bajo el modelo aditivo, se asume que cada observación se descompone en tres componentes:

$$x_t \;=\; T_t \;+\; S_t \;+\; R_t,$$

donde

- T_t es la *tendencia* o evolución de largo plazo,
- S_t es la *estacionalidad* de periodo m (en nuestro caso $m = 12$ meses),
- R_t es el *residuo* o componente irregular.

Para extraer T_t, se aplica una media móvil centrada de ventana m:

$$\hat{T}_t \;=\; \frac{1}{m} \sum_{i=-k}^{k} x_{t+i} \quad \text{con} \quad k = \frac{m-1}{2}.$$

Este suavizado elimina las oscilaciones de frecuencia estacional y revela la evolución subyacente de la serie.

Una vez obtenida $\hat{T}_t$, la serie sin tendencia es:

$$x_t^{(d)} \;=\; x_t \;-\; \hat{T}_t.$$

A continuación, para cada posición estacional $r = 1, \ldots, m$ se calcula el promedio de las observaciones detrendizadas que caen en ese periodo del ciclo:

$$\hat{S}_r \;=\; \frac{1}{n_r} \sum_{t \equiv r \pmod{m}} \left(x_t^{(d)}\right),$$

donde n_r es el número de datos en la posición r. La secuencia estacional completa se define entonces por

$$\hat{S}_t = \hat{S}_{((t-1) \bmod m)+1}.$$

Finalmente, el residuo se obtiene restando tendencia y estacionalidad de la serie original:

$$\hat{R}_t \;=\; x_t \;-\; \hat{T}_t \;-\; \hat{S}_t.$$

La serie $\{\hat{T}_t\}$ representa la evolución de largo plazo de la inflación, libre de variaciones estacionales y ruido, y es la que se emplea para comparar con medidas externas (por ejemplo, polaridad de noticias). Esta aproximación es especialmente adecuada cuando se desea aislar tendencias subyacentes en series con marcada estacionalidad periódica.

Capítulo 5

Resultados

En este capítulo presentamos los principales resultados de nuestra propuesta metodológica. Para eso, empezaremos con una breve descripción de nuestro corpus, la distribución en el tiempo, del vocabulario, entre otras estadísticas necesarias para su entendimiento.

Luego, pasaremos al análisis de extracción y el proceso de etiquetado para finalemente comparar los resultados con nuestra propuesta de mejora.

5.1. Análisis del corpus

El corpus esta conformado por n noticias compiladas entre 1999 y 2024. El conjunto consta de $\approx 277,789$ documentos publicados entre el 2 de octubre de 1999 y el 31 de diciembre de 2024.

Después de realizar la limpieza, se eliminaron los documentos que no contenían información relevante o que eran demasiado cortos.

Cuadro 5.1: Resumen descriptivo de longitud de textos

Métrica	n	media	d.e.	min	50%	max
Número de palabras		42	32.43	0.0	34	384.0

Después de realizar la limpieza, se obtuvo un corpus de aproximadamente 277,000 documentos de noticias económicas. En la siguiente gráfica Figura 5.1, se muestra la distribución por año, donde la cantidad de noticias varia considerablemente a lo largo de los años, con un aumento significativo

en los últimos años.

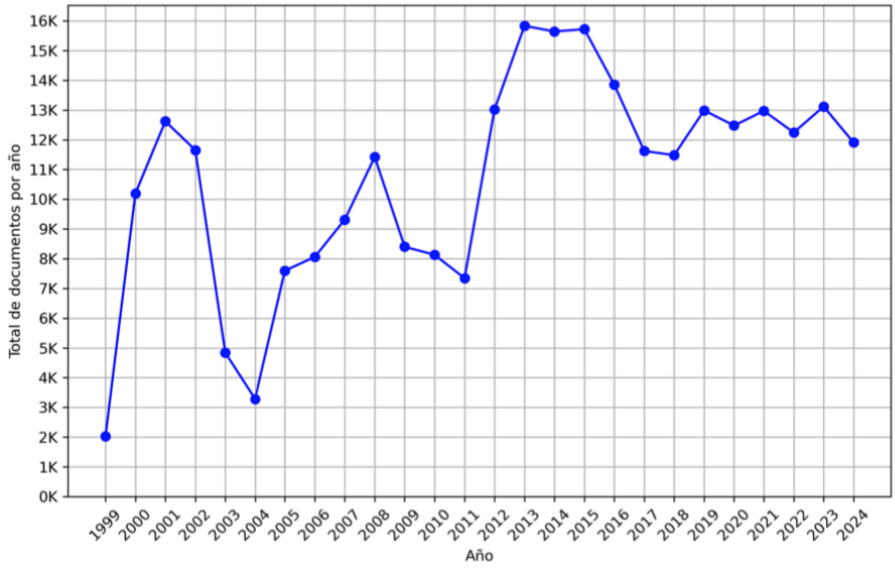

Figura 5.1: Distribución de noticias por año

A continuación, se presente la distribución de las noticias con base en las técnicas de extracción de tópico: Palabras clave, LDA y LDA+LLM pero antes revisamos los conceptos usados tanto para palabras clave como LDA.

5.2. Extracción de términos económicos (palabras clave)

Para la extracción se hizo uso de 2 conjuntos de conceptos: i) *términos directos* y ii) *factores relacionados*. El primero agrupa expresiones que denotan explícitamente inflación; el segundo, elementos que, sin mencionar la inflación, inciden en su dinámica.

Cuadro 5.2: Palabras clave directamente asociadas a la inflación

Términos	**Términos**	**Términos**
inflación	encarecimiento	INPC
presión inflacionaria	carestía	IPC
alza de precios	variación de precios	niveles de inflación
subida de precios	cambios en el índice de precios	nivel inflacionario
aumento de precios	índice nacional de precios al consumidor	baja de precios
incremento de precios	deflación	caída de precios

5.2.0.1. Factores relacionados con la inflación

Se agruparon en categorías aquellos factores que, sin mencionar la inflación de forma explícita, pueden incidir en su evolución:

Oferta y costos: costos de producción, insumos importados, materias primas, costos logísticos, problemas de suministro, cuellos de botella, cadena de suministro, precios de alimentos, precios energéticos, precios de combustibles, precios de gas, precios de electricidad, precios agropecuarios, aumento del costo de transporte.

Demanda interna: consumo privado, aumento del consumo, demanda interna, demanda agregada, expansión del gasto, crédito al consumo, recuperación económica.

Políticas públicas: política monetaria, tasas de interés, incremento de tasas, ajuste de tasas, política fiscal, aumento del gasto público, recorte presupuestal, apoyos sociales, transferencias monetarias, programas sociales.

Mercado laboral: salarios, negociación salarial, aumento salarial, salario mínimo, mercado laboral, presiones salariales.

Factores externos: tipo de cambio, depreciación del peso, apreciación del peso, precio del petróleo, conflictos internacionales, crisis energética, problemas logísticos globales, inflación importada.

Comercio e inversión: importaciones, exportaciones, déficit comercial, superávit comercial, costos de importación, tratado comercial, nearshoring.

Otros: expectativas de inflación, choques de oferta, choques de demanda, volatilidad económica, costo de vida, impacto inflacionario.

5.3. Extracción de tópicos económicos (LDA y LDA+LLM)

La aplicación de LDA permite identificar tópicos coherentes y relevantes para el análisis económico en noticias. La distribución de palabras en los tópicos refleja diferentes dimensiones del análisis económico: financiero, laboral, monetario, político y comercial.

Esto facilita la extracción automatizada de frases y conceptos económicos que son fundamentales para posteriores análisis de sentimiento, tendencias inflacionarias o monitoreo económico en tiempo real.

Cuadro 5.3: Tópicos y sus palabras principales con pesos, agrupados en bloques de 5.

Tópicos 1–5				
Tópico 1	Tópico 2	Tópico 3	Tópico 4	Tópico 5
dolar (3.5)	mexico (2.8)	ser (5.2)	mexico (2.5)	trabajo (1.1)
mercado (2.1)	unidos (2.7)	tener (0.9)	ser (2.1)	semana (1.0)
unidad (1.9)	producto (1.4)	mexico (0.8)	latino (1.6)	haber (0.9)
bolsa (1.6)	industria (1.3)	producto (0.8)	país (1.5)	nuevo (0.8)
mexicano (1.6)	ser (1.2)	haber (0.7)	inversion (1.5)	empresa (0.8)
peso (1.4)	comercial (1.1)	mundo (0.7)	remesa (1.4)	nacional (0.8)
punto (1.3)	exportación (1.0)	desarrollo (0.6)	america (1.3)	energia (0.7)
valor (1.2)	importación (1.0)	mayor (0.6)	extranjero (1.3)	federal (0.6)
acción (1.2)	vehiculo (0.9)	alimento (0.6)	flujo (1.2)	pasado (0.6)
moneda (1.1)	automotriz (0.9)	nuevo (0.6)	capital (1.0)	gobierno (0.6)
Tópicos 6–10				
Tópico 6	Tópico 7	Tópico 8	Tópico 9	Tópico 10
ser (2.8)	trabajador (3.1)	banco (5.0)	mes (2.8)	precio (3.2)
haber (2.2)	trabajo (2.9)	tasa (2.9)	anual (1.7)	mercado (1.6)
tener (1.5)	ser (2.9)	central (2.8)	tasa (1.6)	unidos (1.6)
empresa (1.3)	empleo (2.5)	inflacion (2.4)	trimestre (1.4)	petroleo (1.6)
presidente (1.2)	laboral (2.4)	pandemia (1.8)	ser (1.4)	haber (1.5)
poder (1.1)	persona (2.1)	inter (1.5)	aumento (1.3)	mundial (1.2)
decir (1.0)	salario (1.9)	covid (1.4)	primero (1.3)	demanda (1.1)
hacer (1.0)	haber (1.6)	haber (1.4)	dato (1.2)	decir (1.0)
gobierno (0.9)	empresa (1.0)	reserva (1.3)	pasado (1.2)	riesgo (0.9)
si (0.8)	tener (1.0)	ser (1.2)	punto (1.2)	economia (0.9)
Tópicos 11–15				
Tópico 11	Tópico 12	Tópico 13	Tópico 14	Tópico 15
millón (10.8)	ser (2.3)	ser (2.5)	ser (2.2)	comercio (4.4)
pesos (4.2)	reforma (1.3)	haber (2.3)	tasa (1.9)	mujer (2.0)
dolar (3.9)	fiscal (1.2)	crecimiento (2.3)	politica (1.4)	servicio (1.6)
ser (2.1)	gobierno (1.1)	economico (2.3)	haber (1.4)	libre (1.4)
total (2.0)	tener (1.0)	sector (1.8)	monetario (1.3)	tratado (1.2)
ingreso (0.9)	desarrollo (0.9)	economia (1.7)	dia (1.1)	mercado (1.2)
monto (0.9)	publico (0.9)	actividad (1.3)	gobierno (1.1)	ser (1.1)
gasto (0.8)	deber (0.8)	nivel (0.7)	dar (0.7)	laboral (1.0)
billón (0.8)	haber (0.9)	mercado (0.8)	inter (0.9)	hombre (1.1)

La tabla 5.3 presenta los 15 tópicos más relevantes extraídos mediante LDA a partir del corpus de noticias económicas conformado. Cada tópico se describe mediante las palabras clave más representativas junto con sus pesos relativos, que reflejan la importancia o probabilidad de aparición de cada término dentro del conjunto.

El modelo LDA identifica grupos temáticos latentes en el conjunto de documentos, agrupando palabras que coocurren frecuentemente y que, en conjunto, representan conceptos o aspectos específicos del dominio económico. Los pesos asociados indican la relevancia relativa de cada palabra

dentro del tópico, facilitando la interpretación semántica. A continuación, se analiza cada tópico para ver su relación con el objeto de estudio:

- *Tópicos 1–5.* En este bloque predominan referencias a mercados financieros y a la economía en sentido amplio. El tópico 1 se vincula nítidamente con el mercado cambiario, aparecen *dólar*, *mercado*, *bolsa*, *peso*. El tópico 2 concentra términos sobre la estructura productiva de la economía mexicana (*México*, *producto*, *industria*, *exportación*). Los tópicos 3 y 4 combinan verbos muy frecuentes (*ser*, *tener*) con marcadores geográficos (*México*, *país*, *latino*), lo que sugiere un análisis descriptivo de fenómenos económicos. El tópico 5 reúne vocablos asociados al empleo y a la actividad empresarial (*trabajo*, *empresa*, *energía*), apuntando al mercado laboral y a sectores productivos.

- *Tópicos 6–10.* Aquí se acentúan tópicos relacionados con el empleo, banca e inflación. El tópico 7 destaca *trabajadores*, *salarios* y *empleo*, reflejando preocupaciones sobre el mercado laborales y aspectos sociales. El tópico 8 remite al ámbito financiero-monetario (*banco*, *tasa*, *inflación*, *reserva*), acercándose a la política monetaria y a las condiciones macroeconómicas. El tópico 10 incorpora términos sobre precios, mercados internacionales y energía, enfatizando en *commodities* y sus efectos en la economía.

- *tópicos 11 a 15:* Muestran temas diversos pero ligados a aspectos económicas, políticas públicas y comercio exterior en general. Por ejemplo, el tópico 11 concentra conceptos monetarios como *millón*, *pesos*, y *dólar*. El tópico 12 y 13 incluyen palabras relacionadas con reformas fiscales, crecimiento económico y sectores específicos, apuntando a análisis macroeconómicos y estructurales. El tópico 15 destaca términos de comercio y relaciones internacionales, como *comercio*,

tratado y *libre*, sugiriendo enfoques en tratados comerciales y políticas externas.

Finalmente, después de filtar el corpus, se revisa la distribución de las frases por año.

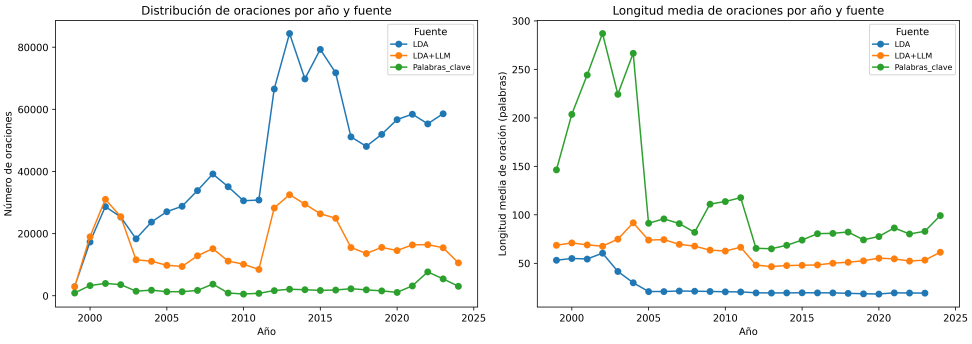

Figura 5.2: Distribución de las noticias recuperadas por método de extracción de tópico

Como se puede observar en la Figura 5.2, por cada método de selección de tópicos contamos con la siguiente distribución:

- **Palabras clave.** Corresponde al conjunto más pequeño del corpus. Su volumen de oraciones es reducido y relativamente estable a lo largo del tiempo, con presencia más marcada en los años iniciales y un repunte moderado en los años recientes. Esto es consistente con un esquema de selección más conservador, enfocado en fragmentos claramente económicos y con alta precisión temática.

- **LDA.** Es la fuente cuantitativamente dominante: a partir de mediados de los 2000 el número de oraciones se incrementa de forma notable y alcanza sus máximos entre 2012 y 2017. Esta expansión refleja la mayor capacidad del enfoque puramente estadístico para extraer masivamente oraciones asociadas a los tópicos económicos, aunque con menor control sobre la pertinencia fina de cada segmento.

- **LDA+LLM.** Presenta un volumen intermedio entre el método basado en regla y LDA. Se observan incrementos importantes en los años en que se aplica la combinación de *topic modeling* con filtrado mediante LLM, lo que sugiere que la etapa de refinamiento reduce parte del ruido inherente a LDA pero mantiene una cobertura temporal amplia.

En cuanto a la **longitud media de las oraciones**, la figura del panel derecho muestra que:

- Las oraciones de `palabras clave` son sustancialmente más largas, especialmente en los primeros años, con promedios cercanos a 250–300 palabras antes de 2004 y estabilizándose después alrededor de 80–120 palabras. Esto indica textos más discursivos y densos en términos de información económica por oración.

- La fuente `LDA` produce las oraciones más cortas del corpus. A partir de 2003 la longitud media cae a alrededor de 15–20 palabras y se mantiene prácticamente constante, lo que refleja una segmentación más agresiva y posiblemente la inclusión de fragmentos más genéricos o contextuales.

- Las oraciones de `LDA+LLM` se sitúan en una posición intermedia: son más largas que las de LDA pero más compactas que las de palabras clave, con una ligera tendencia descendente en el tiempo. Esto es coherente con un proceso en el que el LLM filtra los segmentos más irrelevantes, pero conservando unidades de análisis relativamente breves.

En conjunto, estos resultados muestran que no sólo importa el *método de extracción y segmentación de texto* (Palabras clave, LDA, LDA+LLM),

sino también el *enfoque con el que se construyen las series de polaridad* (no supervisado vs. semi-supervisado). En el enfoque no supervisado, las polaridades se derivan directamente de la aplicación de los modelos de sentimiento sobre las oraciones seleccionadas por cada técnica de tópicos, lo que incorpora de manera casi mecánica la heterogeneidad en longitud, contenido y ruido temático descrita en la sección anterior. En cambio, el enfoque semi-supervisado introduce una capa adicional de corrección y filtrado sobre las polaridades, guiada por criterios económicos y de coherencia temporal, que atenúa parte de esa heterogeneidad.

El resultado es que algunos modelos (como BERT y, en menor medida, Finance y PySentimiento) generan señales más suaves y consistentes tras la corrección, mientras que otros (como el léxico) amplifican el contraste entre meses con contenido valorativo explícito. Así, la combinación entre técnica de selección de oraciones y estrategia de rotulado (no supervisada vs. semi-supervisada) condiciona la forma final de las series de polaridad que se utilizarán para inferir la dinámica económica.

5.4. Inferencia de polaridad basada en enfoque semi-supervisado

A partir del contexto previo —donde se documentó la estructura del corpus por método de selección de oraciones y la distribución de polaridades agregadas— el siguiente paso consiste en construir conjuntos de entrenamiento para el modelo **BERT+Regresión lineal** a partir de la coincidencia en términos de tendencia entre las series de polaridad y la inflación observada de INEGI.

Las siguientes 2 figuras resumen esta coincidencia en forma de mapa de calor: para cada mes y cada combinación de método de polaridad y esquema

de selección de oraciones se representa la dirección de la tendencia (subida, caída o estabilidad). Los meses en los que el color de una fila de polaridad coincide con el de la fila del INPC constituyen candidatos naturales para el conjunto de entrenamiento, pues reflejan episodios en los que el modelo de lenguaje "lee" la misma dirección del ciclo inflacionario que los datos oficiales.

5.4.1. Concordancia basada en base en palabras clave/expertos

En la siguiente gráfica (véase Figura 5.3), asociada al esquema de selección por palabra clave o expertos, se observa un patrón relativamente fragmentado en el tiempo: la serie del INPC alterna episodios de alzas y bajas marcadas, mientras que las filas de los modelos de polaridad muestran bloques de coincidencia con distinta intensidad según el método. En este contexto, el conjunto de entrenamiento derivado del enfoque de palabras clave tiende a ser:

- **Más reducido en número de meses**, dado que la selección de oraciones vía juicio de experto opera sobre un subconjunto pequeño y muy depurado del corpus.

- **De alta precisión temática**, ya que las oraciones fueron previamente filtradas para contener conceptos y verbos económicos relevantes; cuando la polaridad coincide con el INPC, la señal suele estar sustentada en noticias explícitamente vinculadas a precios, salarios o política monetaria.

- **Heterogéneo entre métodos de polaridad**: algunos modelos (por ejemplo, Finance o PySentimiento) muestran bloques de acuerdo más

extensos con el INPC que otros (como el léxico), de modo que el solapamiento efectivo entre modelos al construir etiquetas “confiables” es limitado.

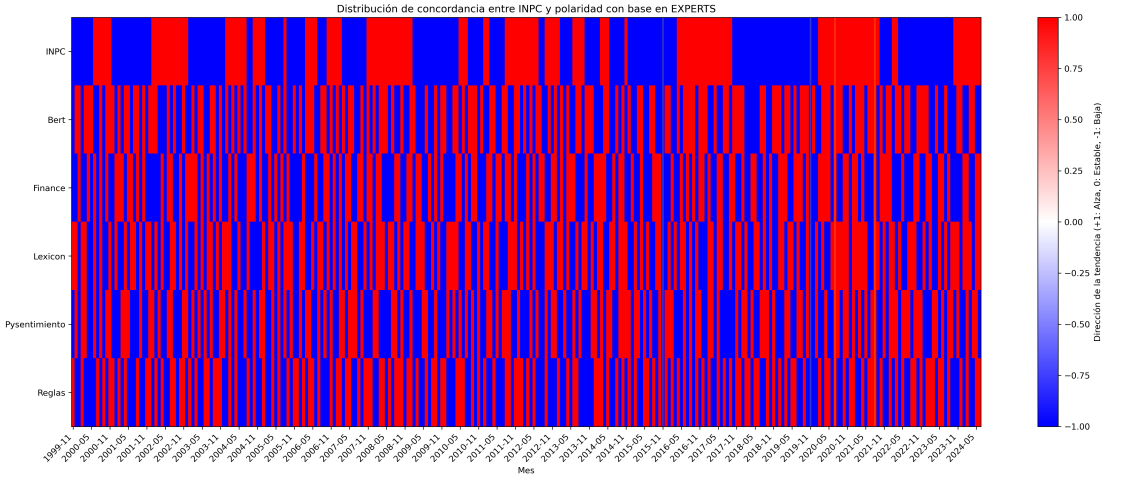

Figura 5.3: Representación de coincidencia entre INPC y polaridad con base método de extracción de oraciones: palabras clave/expertos

En términos de entrenamiento, este primer conjunto aporta observaciones relativamente escasas pero muy informativas, adecuadas para anclar el modelo BERT+Regresión lineal en episodios donde existe un consenso claro entre texto y inflación observada.

5.4.2. Concordancia basada en LDA.

Aquí la gráfica muestra lo siguiente:

- **Mayor densidad de episodios** en los que al menos uno de los métodos de polaridad coincide con la dirección del INPC, simplemente porque el número de oraciones y la cobertura temporal de LDA son mucho más amplios.
- **Patrones de acuerdo más persistentes** en ciertos periodos (por ejemplo, fases prolongadas de inflación alta o baja), donde varias

filas de polaridad comparten el mismo color que el INPC, generando columnas verticales relativamente homogéneas.

- **Mayor ruido temático**: al provenir de tópicos estimados de forma puramente estadística, algunas coincidencias pueden estar impulsadas por noticias menos específicas sobre inflación (por ejemplo, actividad económica general o mercados financieros), lo que introduce variabilidad adicional en la calidad de las etiquetas.

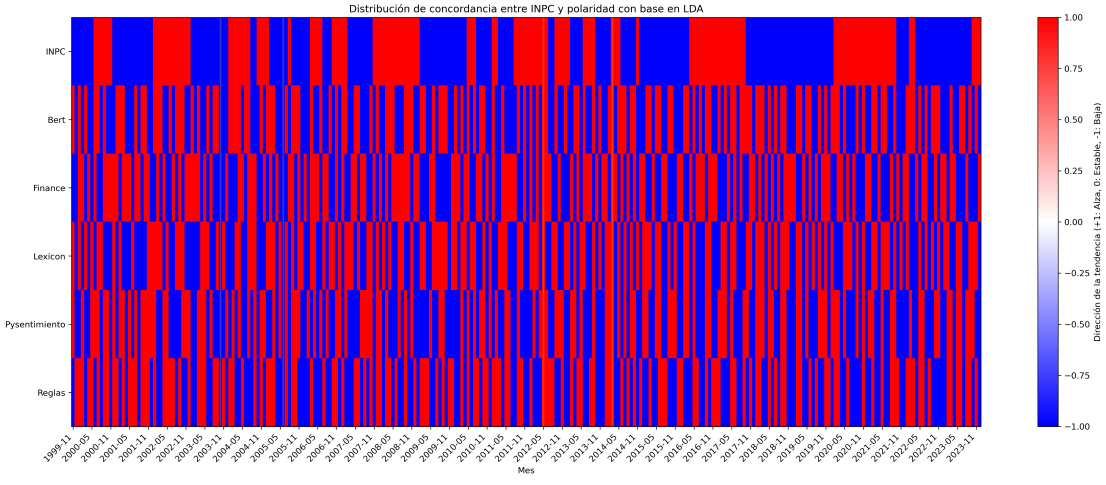

Figura 5.4: Representación de coincidencia entre INPC y polaridad con base método de extracción de oraciones: LDA

Este segundo conjunto de entrenamiento es, por tanto, mucho más voluminoso que el anterior y cubre con mayor continuidad el horizonte 1999–2024, pero a costa de incorporar más heterogeneidad en el tipo de información económica que respalda cada coincidencia texto–inflación.

5.4.3. Concordancia basada en LDA+LLM.

En está gráfica se ilustra el esquema híbrido LDA+LLM, en el que las oraciones seleccionadas por tópicos se someten a un filtrado adicional con

un modelo de lenguaje. Visualmente, el mapa combina rasgos de los dos casos anteriores:

- La **cobertura temporal** sigue siendo amplia, similar a LDA, pero con una ligera reducción en los episodios aislados de acuerdo, resultado del filtrado de oraciones menos relevantes desde el punto de vista económico.

- Los **bloques de coincidencia sostenida** entre las filas de polaridad y el INPC se vuelven más nítidos en ciertos intervalos, lo que sugiere que el LLM ayuda a concentrar el conjunto de entrenamiento en meses donde la información textual es más claramente alineable con el comportamiento inflacionario.

- La **coherencia entre métodos de polaridad** mejora moderadamente: aunque persisten discrepancias entre modelos (por ejemplo, entre el enfoque léxico y Finance), se observan más columnas en las que varios métodos comparten el mismo signo que el INPC, lo que facilita construir etiquetas de "alto consenso" para el BERT+Regresión lineal.

En consecuencia, el conjunto de entrenamiento LDA+LLM se sitúa en un punto intermedio: más grande y diverso que el derivado de palabra clave, pero más depurado y temáticamente focalizado que el obtenido con LDA puro. Este balance resulta especialmente valioso para el modelo BERT+Regresión lineal, pues proporciona suficientes observaciones para aprender patrones temporales, sin perder la conexión económica explícita entre noticias y dinámica inflacionaria.

En resumen, los tres conjuntos de entrenamiento se diferencian menos por el algoritmo de polaridad en sí mismo y más por la interacción entre

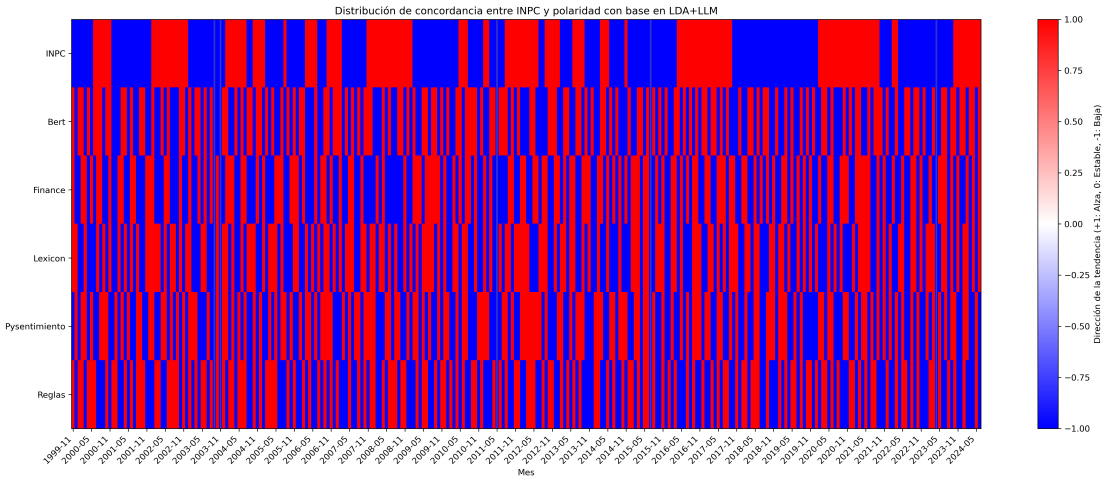

Figura 5.5: Representación de coincidencia entre INPC y polaridad con base método de extracción de oraciones: palabras LDA+LLM

dicho algoritmo y el *pipeline de selección de oraciones*. El conjunto de palabra clave ofrece pocos meses pero de alta calidad; LDA aporta volumen y cobertura, a costa de un mayor ruido; y LDA+LLM construye un compromiso entre ambos extremos, generando un insumo más equilibrado para el entrenamiento supervisado del modelo BERT+Regresión lineal.

5.5. Conjunto de entrenamiento para enfoque semi-supervisado

La primera serie de gráficas muestra, para cada mes, la dirección de la variación del índice de precios al consumidor y de los distintos indicadores de polaridad derivados del corpus de noticias económicas. Cada trayectoria se representa como una señal discreta que toma valores en $\{-1, 0, +1\}$: el valor $+1$ indica una tendencia al alza (presiones inflacionarias positivas), el valor -1 una tendencia a la baja (alivio de presiones inflacionarias) y el valor 0 corresponde a un tramo prácticamente estable, en el que la variación mensual se mantiene dentro de una banda de tolerancia previamente

fijada. En el eje vertical se grafica esta codificación discreta de direcciones, mientras que en el eje horizontal se representan los meses del periodo de análisis.

Las franjas sombreadas resaltan los meses en los que existe *concordancia* entre la dirección del índice de precios y la señal agregada de los modelos de polaridad, es decir, aquellos periodos en los que la mayoría de los indicadores textuales apuntan en el mismo sentido que la inflación observada.[1]

La lectura de estas figuras permite: (i) verificar si las medidas de polaridad extraídas de las noticias logran capturar *puntos de giro* en la dinámica inflacionaria (cambios de signo en la tasa de variación), (ii) evaluar la persistencia de episodios de desacuerdo entre texto y datos oficiales (tramos en los que los signos se mantienen sistemáticamente opuestos) y (iii) identificar posibles sesgos estructurales de los modelos, como la tendencia a evitar el estado neutro (0) o a sobrerreaccionar en una sola dirección.

5.6. Inferencia de polaridad basada en BERT + Regresión lineal

El modelo propuesto estima la polaridad a nivel oración como una variable continua, combinando representaciones semánticas profundas de BERT multilingüe (bert-base-multilingual-uncased) considerando de manera explícita el tiempo de ocurrencia del evento. Cada oración se tokeniza hasta un máximo de 512 tokens de acuerdo a la arquitectura Transformer (Devlin et al., 2019b; Vaswani, Shazeer, Parmar et al., 2017b); se utiliza el vector

[1] De manera formal, puede definirse una regla de concordancia a partir de la comparación entre la señal discreta de la inflación y la de cada modelo en el mismo mes, y considerar que hay acuerdo cuando el número de coincidencias supera un umbral mínimo K (por ejemplo, mayoría simple).

[CLS] ($\mathbf{h}_{[\text{CLS}]} \in \mathbb{R}^{d_B}$) como resumen contextual del enunciado.

Por otro lado, el vector de tiempo $\mathbf{t} \in \mathbb{R}^3$ se proyecta a un espacio latente de dimensión $d_H = 256$ mediante una capa densa con activación ReLU, $\mathbf{z}_t = \text{ReLU}(\mathbf{W}_t\mathbf{t} + \mathbf{b}_t)$. La combinación de información lingüística y temporal se realiza por concatenación $[\mathbf{h}_{[\text{CLS}]}; \mathbf{z}_t]$, seguida de una capa lineal que devuelve la predicción escalar $\hat{y}$ de polaridad.

Esta parametrización es consistente con la arquitectura BERT que captura dependencias semánticas y sintácticas de largo alcance, mientras que la variable de tiempo ayudan a modelar y captar la estacionalidad informativa que afectan la distribución de señales en noticias económicas. La implementación se apoya en Transformers/PyTorch (Wolf et al., 2020).

5.6.1. Entrenamiento del modelo

El modelo se entrena de forma independiente para cada método de polaridad (`Regla`, `Lexicón`, `Bert` y otros). Para cada objetivo, se utiliza una partición 80/20 (semilla = 42) y se optimiza una pérdida MSE con **AdamW** ($\eta = 2 \times 10^{-5}$, *weight decay* por defecto) durante ~ 5 épocas y *batch* = 8 (Loshchilov & Hutter, 2019).

5.6.2. Evaluación del modelo

En el conjunto de prueba se informan MAE, RMSE y R^2, junto con una gráfica Predicho–vs–Real por objetivo, lo que permite evaluar sesgos y dispersión del error en torno a la diagonal. Además, se exportan archivos `CSV` con los pares $\{\texttt{true}, \texttt{pred}\}$ para auditoría y posterior agregación. Con el modelo calibrado, se ejecuta inferencia masiva sobre el corpus completo para corregir etiquetas ruidosas y completar valores faltantes; las series resultantes se agregan por ventana de tiempo y fuente, y se contrastan con la inflación observada (INEGI) para cuantificar capacidad de señal.

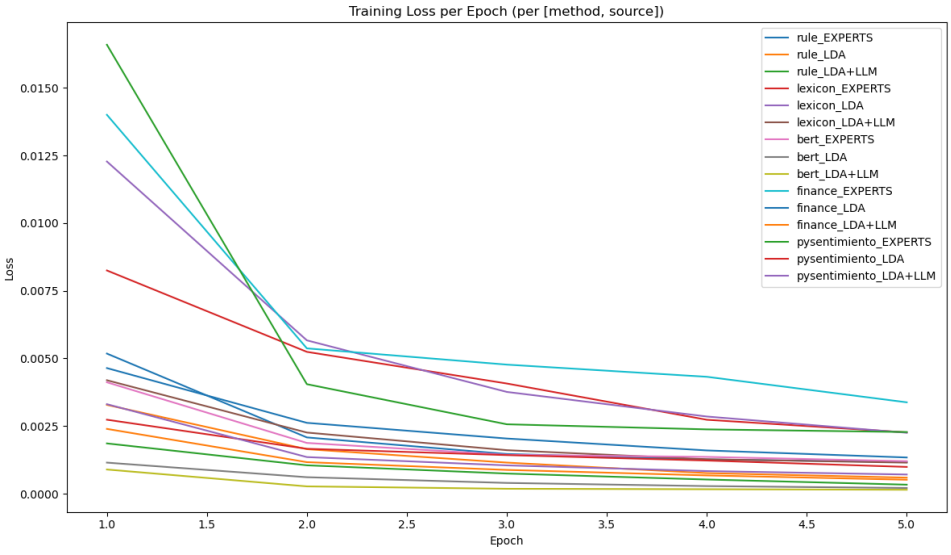

Figura 5.6: Resultado del entrenamiento

En la Figura 5.6 se muestra la evolución de la pérdida (MSE) por época para cada combinación [`método de extracción`, `polaridad`]. Se observan tres hechos robustos.

- Primero, convergencia rápida: la mayor reducción de pérdida ocurre entre las épocas 1 y 2, con rendimientos decrecientes a partir de la época 3.

- Segundo, dependencia de la fuente: los modelos entrenados con señales provenientes de `palabras clave` y `LDA+LLM` alcanzan pérdidas finales sistemáticamente menores que sus contrapartes con LDA puro, lo que sugiere menor ruido de etiquetado y mejor alineación semántica.

- Tercero, heterogeneidad por método: algunos métodos (p. ej., variantes financieras o léxicas) conservan pérdidas más altas durante todo el entrenamiento, indicio de desajuste entre la señal débil y la representación del modelo o de mayor varianza en las pseudoetiquetas.

5.7. Análisis descriptivo de las polaridades

En la sección anterior se mostró que la estructura del corpus varía de forma importante según el método de selección de oraciones: el enfoque basado en palabras clave trabaja sobre un subconjunto reducido pero muy denso en contenido económico, mientras que los métodos automáticos (LDA y LDA+LLM) generan un volumen mucho mayor de oraciones, generalmente más cortas y con mayor heterogeneidad temática.

Sobre este trasfondo, la Tabla 5.4 presenta un resumen de las *polaridades mensuales* por método de análisis de sentimiento, distinguiendo entre el enfoque **No supervisado** (polaridad clásico) y el **Semi-supervisado** (polaridades corregidas).

En términos generales, todos los métodos reportan polaridades medias positivas en ambos enfoques, lo que indica que, en promedio, el tono de las noticias económicas tiende a ser ligeramente optimista o, al menos, no marcadamente negativo.

No obstante, la magnitud de esa positividad varía de manera sistemática entre modelos: el método de Regla y el enfoque Léxico producen valores medios muy cercanos a cero (entre 0.02 y 0.07), coherentes con una detección más conservadora del sentimiento; BERT se sitúa en un nivel intermedio (en torno a 0.23–0.24), mientras que Finance y PySentimiento concentran las polaridades más altas (≈0.43–0.44 y 0.56–0.59), reflejando una mayor sensibilidad a vocabulario financiero y afectivo explícito.

En todos los casos la dispersión es moderada, con desviaciones estándar reducidas y rangos intercuartílicos relativamente estrechos, lo que sugiere una señal agregada bastante estable a lo largo del tiempo.

La comparación entre el enfoque **No supervisado** y el **Semi-supervisado** revela que la corrección en el método de inferencia de la polaridad no alte-

Cuadro 5.4: Distribución de la polaridad por método y enfoque de inferencia

Enfoque	Método	n	Media	Desv. est.	Mínimo	Q_{25}	Mediana	Q_{75}	Máximo
No supervisado	BERT	884	0.236	0.011	0.222	0.226	0.232	0.246	0.288
No supervisado	Finance	884	0.426	0.030	0.359	0.388	0.441	0.449	0.478
No supervisado	Léxico	884	0.065	0.027	-0.021	0.044	0.059	0.090	0.126
No supervisado	PySentimiento	884	0.565	0.025	0.534	0.546	0.550	0.595	0.627
No supervisado	Regla	884	0.022	0.032	-0.004	0.000	0.001	0.063	0.086
Semi-supervisado	BERT	884	0.234	0.006	0.224	0.226	0.236	0.238	0.258
Semi-supervisado	Finance	884	0.437	0.030	0.385	0.409	0.426	0.472	0.507
Semi-supervisado	Léxico	884	0.072	0.033	0.015	0.043	0.067	0.102	0.142
Semi-supervisado	PySentimiento	884	0.593	0.037	0.553	0.565	0.571	0.640	0.657
Semi-supervisado	Regla	884	0.020	0.031	-0.007	-0.001	-0.000	0.058	0.083

ra de forma drástica el nivel medio de polaridad, pero sí introduce ajustes finos y cambios en la variabilidad.

- En **BERT**, la media prácticamente se mantiene constante, mientras que la desviación estándar cae a la mitad, lo que indica una señal más homogénea tras eliminar oraciones ambiguas o marginalmente económicas.

- **Finance** y **PySentimiento** muestran ligeros incrementos en la polaridad media al pasar al corpus semi-supervisado, consistentes con un filtrado que refuerza la presencia de oraciones con contenido económico y afectivo más marcado.

- En el caso del modelo **Léxico**, la media aumenta levemente y la dispersión se amplía, reflejando una mayor concentración de vocabulario explícitamente valorativo (positivo y negativo).

- Finalmente, el modelo de **Regla** apenas se ve afectado por la corrección: sus estadísticas casi no cambian entre enfoques, en línea con el hecho de que opera sobre un subconjunto pequeño y relativamente homogéneo de oraciones ya altamente depuradas.

En conjunto, estos resultados confirman que la forma en que se construye el corpus interactúa con las características de cada modelo de polaridad.

Los métodos aplicados sobre un conjunto más restringido y denso en información económica generan polaridades de menor magnitud pero muy estables, mientras que los modelos diseñados para capturar señales financieras o afectivas tienden a amplificar las variaciones del ciclo económico y responden con mayor intensidad a la depuración de las oraciones.

5.7.1. Distribución de polaridad: método corregido (semi-supervisado) vs. no supervisado

La comparación entre ambas versiones muestra tres efectos consistentes. Primero, menor dispersión tras la corrección: en `LEXICON`, `FINANCE` y `PYSENTIMIENTO` el rango intercuartílico se reduce y desaparecen muchos extremos, señal de menor ruido en la señal de polaridad. Segundo, desplazamientos de nivel por familia: en `BERT` la mediana cae y queda en una banda más estrecha; en `FINANCE` y `PYSENTIMIENTO` la mediana sube levemente y la variabilidad disminuye; en `REGLA` y `LEXICON` los cambios de nivel son modestos, pero las distribuciones se compactan. Tercero, jerarquía de magnitudes más nítida: se mantiene el orden `PYSENTIMIENTO` > `FINANCE` > `BERT` > `LEXICON`/`REGLA`, con menos solapamiento entre cajas.

Por método de extracción, la corrección *homogeneiza* sin colapsar: *LDA* queda con las cajas más compactas; *LDA+LLM* conserva sensibilidad pero sin extremos (el alza es visible sobre todo en `FINANCE` y `PYSENTIMIENTO`); y *extracción de términos* reduce de forma marcada su dispersión. El cuadro general es de señales más estables y comparables entre modelos, condición útil cuando la polaridad se emplea como indicador para seguir la dinámica inflacionaria.

En suma, el paso al esquema corregido aporta: (i) reducción de varianza, (ii) recentrado de medianas (a la baja en `BERT`, al alza moderada en `FINANCE`/`PYSENTIMIENTO`) y (iii) mejor separación entre familias y fuentes.

Estos efectos son coherentes con una etapa de depuración guiada por información macro que filtra ruido y calibra amplitudes sin borrar estructura (Arazo et al., 2019; Devlin et al., 2019b; Vaswani, Shazeer, Parmar et al., 2017b; Zhu, 2009).

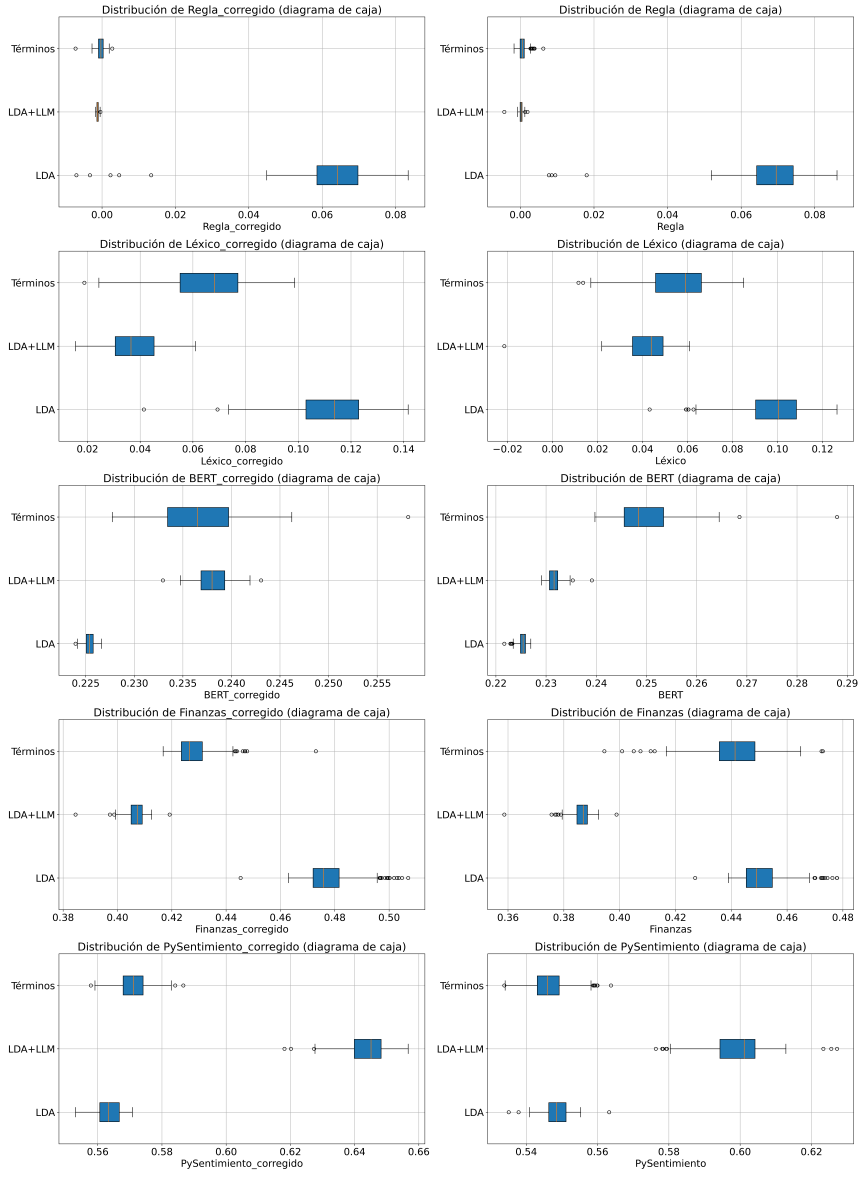

Figura 5.7: Distribución de polaridad por método y fuente de extracción (versiones no supervisada y corregida).

5.8. Análisis de la inflación (INEGI)

A continuación se analiza la tendencia mensual del INPC *subyacente* y *no subyacente* que publica el INEGI como medición de la inflación. La serie subyacente deja fuera precios muy volátiles (energéticos, algunos alimentos y tarifas administradas) y suele usarse como una aproximación a la tendencia "de fondo" de la inflación. La no subyacente, en cambio, recoge justamente esos componentes más inestables.

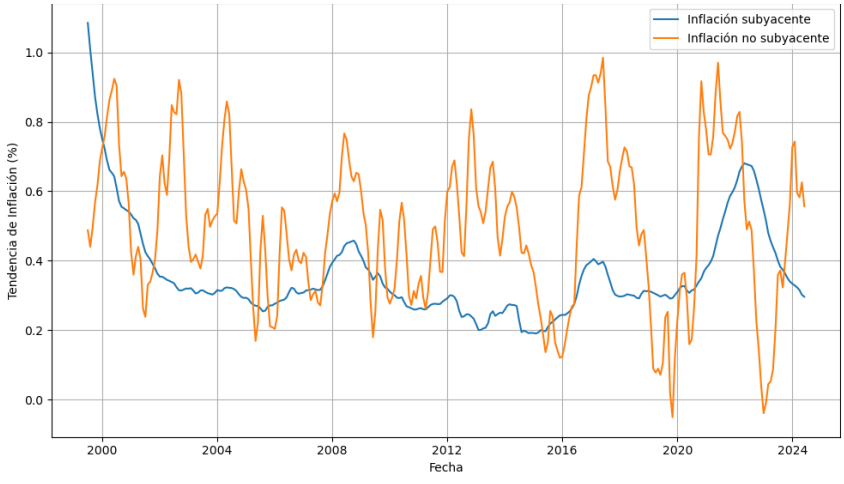

Figura 5.8: Tendencia de la inflación mensual (INEGI)

Como se puede observar en la gráfica anterior 5.8 la distribución de la línea azul (subyacente), al inicio del periodo se observa una inflación alta y va cayendo rápido a principios de los 2000, hasta quedar en niveles mensuales cercanos a 0.3 %. A partir de ahí la serie se mueve poco: hay pequeños vaivenes, pero durante varios años se mantiene más o menos en una banda estrecha, lo que coincide con la etapa de desinflación y el esfuerzo del Banco de México por anclar la inflación alrededor de su meta. Después de 2016, y sobre todo tras la pandemia, la subyacente vuelve a subir de forma gradual, reflejando el episodio inflacionario más reciente, y luego empieza a moderarse otra vez al final de la muestra.

La línea naranja (no subyacente) cuenta una historia distinta. La serie es mucho más nerviosa: aparecen picos y caídas bruscas a lo largo de todo el periodo, con meses en los que la inflación se dispara por arriba de 0.8–1 % y otros en los que prácticamente desaparece o incluso es negativa. Esos saltos están asociados a choques temporales en precios agropecuarios, energéticos o ajustes de tarifas reguladas, que el INEGI agrupa en la parte no subyacente. Varias veces las dos curvas se separan claramente: la no subyacente sube o baja con fuerza mientras la subyacente se mantiene relativamente estable. Esto confirma que la subyacente refleja mejor la trayectoria persistente del proceso inflacionario, mientras que la no subyacente está dominada por shocks de corto plazo que añaden ruido al indicador general.

5.8.1. Análisis comparativo entre polaridad e inflación observada

5.8.1.1. Comparacion basada en extracción de términos/expertos

La comparación entre el INPC no subyacente y las series de polaridad revela que la etapa de corrección (semi–supervisada) mejora de forma sistemática la lectura cíclica respecto a las versiones no supervisadas. En términos generales, las señales corregidas reducen el componente de alta frecuencia, estabilizan los niveles y alinean mejor los cambios de tendencia del INPC en episodios relevantes de la muestra (recesión de 2009, choque pandémico de 2020 y repunte inflacionario de 2021–2022). Esta ganancia en estabilidad hace más nítidas la comparación (ver 5.9).

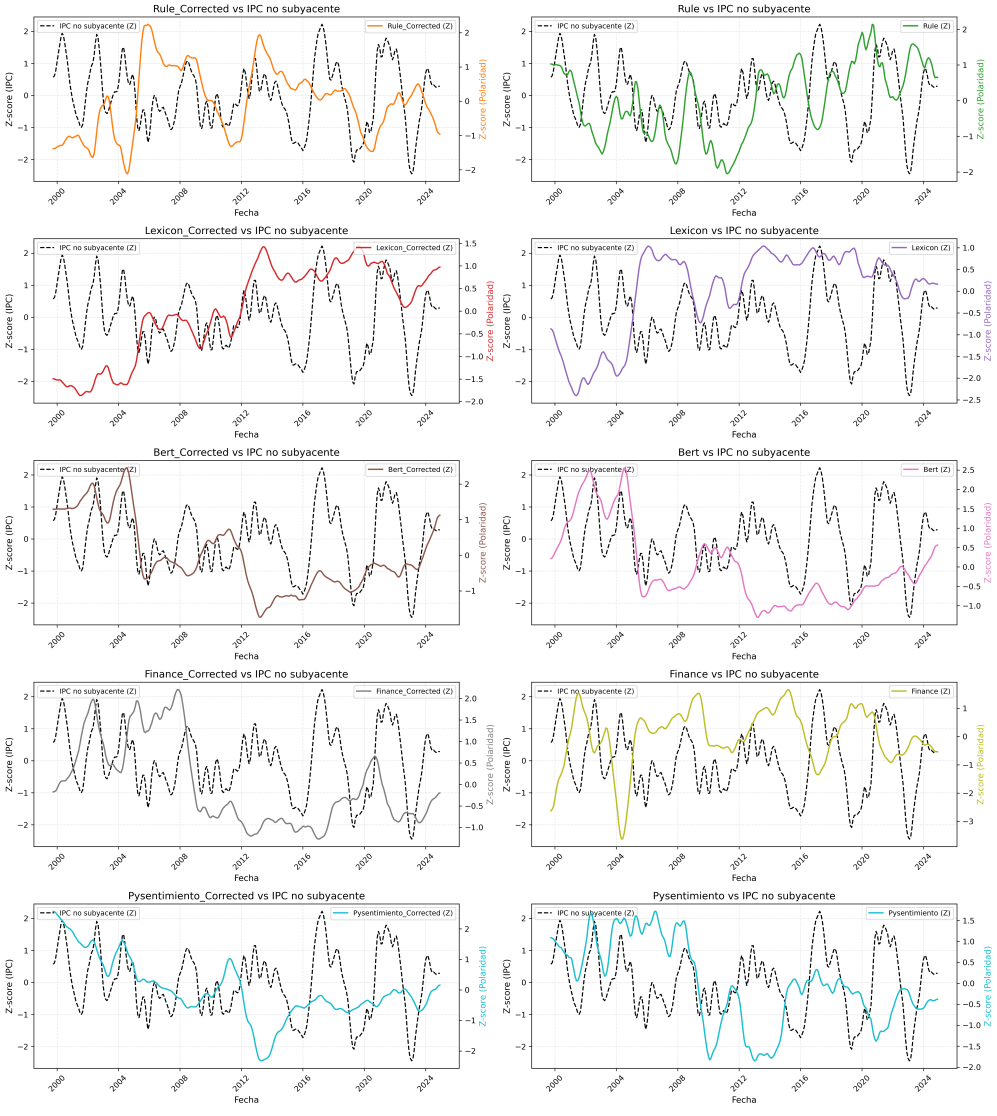

Figura 5.9: Comparación de los métodos de polaridad corregidos vs no supervisados para extracción de términos

En conjunto, las correcciones introducen un patrón más sobrio y útil para seguimiento: la señal basada en reglas actúa como un filtro de baja varianza que acompaña el ciclo con rezagos cortos y sin sobresaltos. El enfoque basado en léxicon, una vez corregido, conserva bien el nivel pero sacrifica los cambios observados entre 2013–2015; su fortaleza radica en anclar tendencias de mediano plazo más que en señalar cambios de ten-

dencias. En ausencia de corrección, reproduce mejor la forma del ciclo en 2008–2012 y 2020–2022, aunque con amplitudes sobredimensionadas.

Por su parte, el modelo BERT corrige hacia una trayectoria contenida y estable que marca con claridad las caídas de 2009 y 2020 y los rebotes de 2017 y 2021, típicamente con uno o dos periodos de rezago; la versión no supervisada, en cambio, muestra respuestas más débiles y tramos de baja frecuencia mal definidos al inicio de la muestra. Para el método de Transformer especializado en finanzas, la corrección preserva la capacidad de reacción pero acota picos: los cambios de 2017 y 2021–2022 aparecen con adelantos cortos y la contracción de 2009 queda nítida, mientras que el enfoque no supervisado tiende a sobrerresponder en fases intermedias. Finalmente, `pysentimiento` reduce la saturación en niveles altos sin perder lectura cíclica: los mínimos de 2009 y 2020 y el repunte 2021–2022 se delinean con mayor coherencia que en su contraparte no supervisada, que suele amplificar choques.

A partir de esta evidencia, las series corregidas comparten menor varianza, menos cambios de alta frecuencia y mejor sincronía con la serie del INPC. Ello habilita una arquitectura parsimoniosa para uso operativo: tomar `BERT_corregido` como ancla estable; emplear `FINANCE_corregido` y `PYSENTIMIENTO_corregido` como detectores de cambio con *anticipaciones breves*; reservar `REGLA_corregido` como capa de suavizado; y considerar `LEXICON` como señal de nivel cuya utilidad para el *timing* es acotada tras la corrección. Esta combinación separa mejor movimientos transitorios de cambios en el ciclo y reduce la probabilidad de falsas alarmas ante choques de corta duración.

5.8.1.2. Comparación basada en LDA

En términos de desempeño de cada familia de métodos, `REGLA` bajo *LDA* sigue el ciclo con fluctuaciones moderadas alrededor de los mínimos de 2009 y 2020 y del repunte de 2021–2022, mientras que la versión no supervisada preserva el ciclo con un cambio más marcado; con selección experta, la corrección reduce aún más el ruido de alta frecuencia y atenúa sobre–reacciones tempranas. `LEXICON` conserva el trazo del ciclo con amplitud acotada en *LDA*; sin corrección, replica mejor los vaivenes de 2008–2012 y 2020–2022 a costa de sobrepasos, y en experto su aporte es principalmente de nivel más que de *timing*.

`BERT` corregido identifica con precisión las caídas de 2009 y 2020 y el rebote 2021–2022 con rezagos cortos y baja varianza; el no supervisado responde débilmente a choques, y en experto la estabilidad se mantiene aunque con segmentos iniciales más planos, por lo que `BERT_corregido` funciona como columna vertebral en ambos esquemas. La familia `FINANCE` preserva sensibilidad y elimina picos: los giros de 2017 y 2021–2022 se anticipan levemente y la caída de 2009 queda bien marcada; el no supervisado amplifica choques intermedios, y en experto la corrección controla las colas sin perder oportunidad, ofreciendo el mejor equilibrio entre *timing* y control de ruido. Finalmente, `PYSENTIMIENTO` corregido reduce saturación en niveles altos y mejora la coherencia cíclica (2009, 2020 y 2021–2022 quedan bien delineados); el no supervisado tiende a sobrerreaccionar, mientras que en experto se mantienen adelantos cortos en repuntes y se suprimen mesetas artificiales.

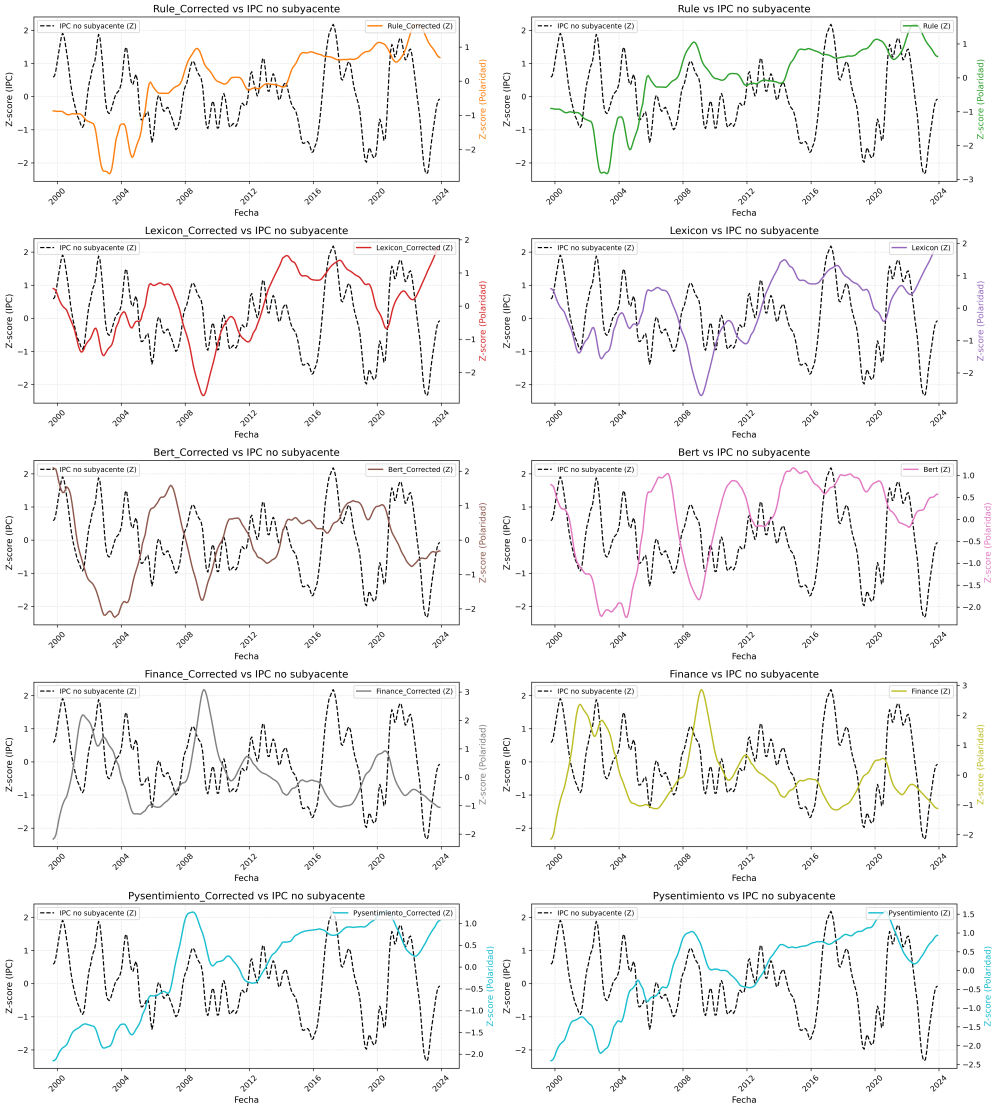

Figura 5.10: Comparación de los métodos de polaridad corregidos vs no supervisados para extracción LDA

En síntesis, con *LDA* las versiones corregidas de `BERT`, `FINANCE` y `PYSENTIMIENTO` ofrecen el mejor compromiso entre estabilidad y sensibilidad, con menor varianza y alineación clara con los puntos de giro del INPC; en selección experta, la corrección resulta especialmente valiosa para `REGLA` y `PYSENTIMIENTO`, al reducir ruido sin borrar ciclos, mientras `LEXICON` conserva utilidad como ancla de nivel pese a su *timing* limitado. Operativamente,

una combinación parsimoniosa consiste en anclar con `BERT_corregido`, sumar un detector de cambio entre `FINANCE_corregido` o `PYSENTIMIENTO_corregido`, y, dependiendo de la fuente, privilegiar con *LDA* la precisión cíclica de baja varianza o, en experto, añadir el suavizado de `REGLA_corregido` y dejar a `LEXICON` la fijación del nivel. Esta arquitectura discrimina mejor choques transitorios de giros de ciclo relevantes para la inflación.

5.8.1.3. Comparación basada en extracción en LDA+LLM

La extracción *LDA+LLM* aporta una base semántica más coherente y, tras la corrección semi–supervisada, genera trayectorias de polaridad con menor varianza idiosincrática y mejor sincronía con los giros del INPC no subyacente (2009, 2020 y 2021–2022). La Figura 5.11 ilustra esta mejora en la lectura cíclica frente a los esquemas no supervisados. En este marco, la señal basada en reglas actúa como un filtro de baja varianza que acompaña con rezagos cortos los valles de 2009 y 2020 y el tramo alcista de 2021–2022; sin corrección, persisten picos locales y cambios que no corresponden a cambios macroeconómicos sostenidos.

El enfoque léxico, una vez depurado, mantiene información de nivel y amortigua extremos, aunque atenúa la ondulación cíclica en la segunda mitad de la muestra; su contraparte no supervisada replica con mayor amplitud los vaivenes de 2008–2012 y 2020–2022, a costa de sobrepasos, por lo que su aporte es más bien como ancla de tendencia que como señal de *timing*. Con BERT, la corrección produce una serie contenida y claramente co-móvil con el INPC: identifica caídas y rebotes (2009, 2017, 2020, 2021–2022) con uno o dos periodos de rezago y sin sobrerreacciones, mientras que el modo no supervisado diluye la señal en tramos planos. Los métodos se benefician especialmente del acoplamiento *LDA+LLM*: preserva sensibilidad a los giros con colas acotadas, perfila bien la contracción de 2009

y el repunte de 2021–2022, y capta el choque de 2020 sin amplificarlo; el ajuste no supervisado, en cambio, exacerba movimientos intermedios.

Por su parte, `pysentimiento` corrige la saturación en niveles altos, perfila con claridad mínimos y ascensos (2009, 2020, 2021–2022) y mantiene adelantos cortos en fases de aceleración, mientras que su versión no supervisada tiende a sobrerresponder con mesetas y caídas abruptas de difícil interpretación.

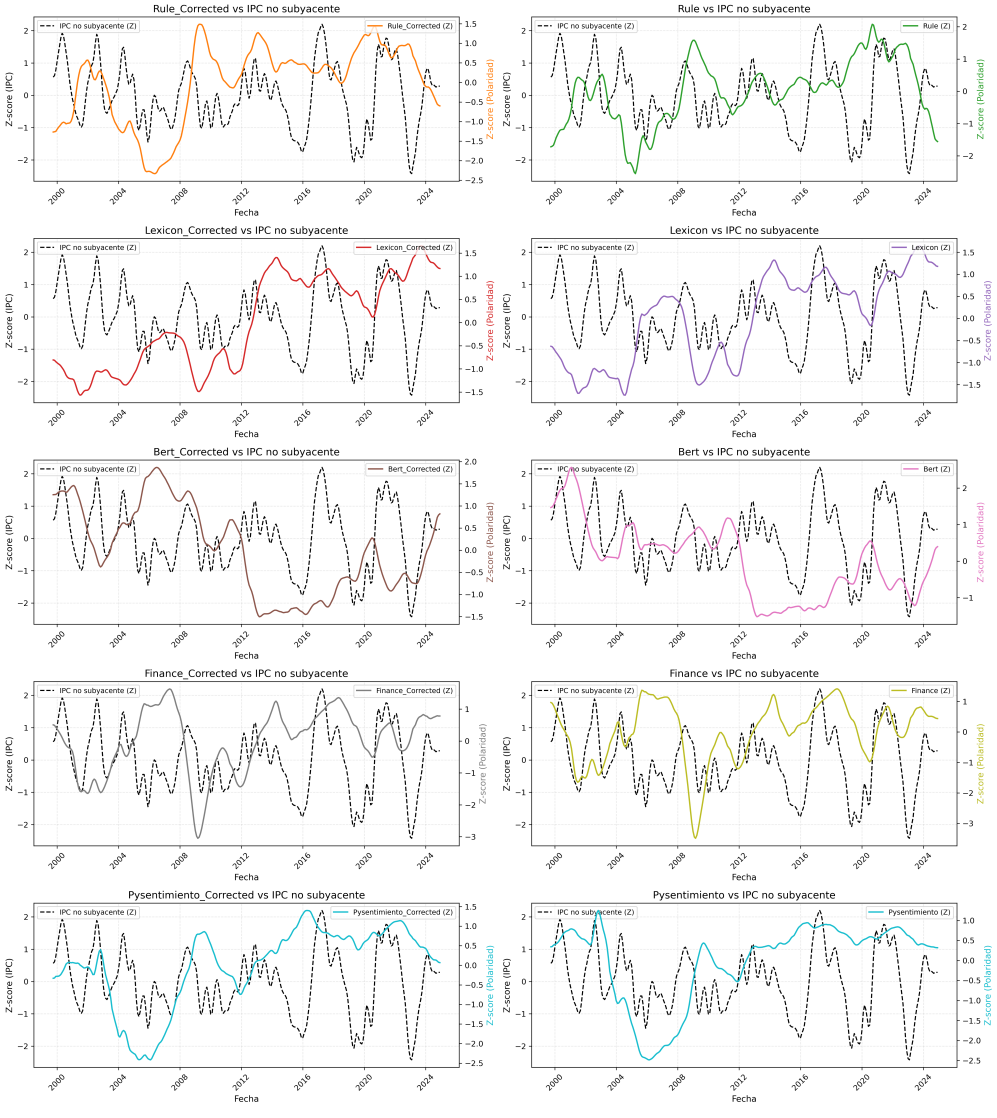

Figura 5.11: Comparación de los métodos de polaridad corregidos vs no supervisados para extracción LDA+LLM

En síntesis, bajo *LDA+LLM* las correcciones consolidan tres ventajas: reducción de varianza y de cambios de alta frecuencia, mayor alineación con la tendencia del INPC y una diferenciación funcional más nítida entre familias. Operativamente, una arquitectura parsimoniosa toma `BERT_corregido` como referencia estable, añade `FINANCE_corregido` y `PYSENTIMIENTO_corregido` como detectores de cambio de corto plazo, y reserva `REGLA_corregido` como capa de suavizado; `LEXICON` conserva valor como componente de nivel, si bien su utilidad para *timing* es acotada tras la depuración. Este arreglo reduce falsas alarmas y mejora la identificación de transiciones de régimen con relevancia macroeconómica.

5.9. Divergencia de Kullback–Leibler entre Inflación y Polaridades de Sentimiento

En este apartado describimos detalladamente el procedimiento para comparar la distribución de la inflación no subyacente con las distribuciones de distintas polaridades de sentimiento, utilizando la divergencia de Kullback–Leibler (KL). Incluimos desde las fórmulas de suavizado y normalización, pasando por la estimación de densidades, hasta la interpretación pormenorizada de los resultados.

5.9.1. Preprocesamiento de las series

5.9.1.1. Filtro de Savitzky–Golay

Para reducir el ruido de las series mensuales se aplica un filtro de Savitzky–Golay. Dado un vector de observaciones x_t, el valor suavizado $\widetilde{x}_t$ viene dado por

$$\widetilde{x}_t \;=\; \sum_{k=-m}^{m} c_k \, x_{t+k},$$

donde el número de puntos de la ventana es $w = 2m + 1$ y los coeficientes c_k provienen del ajuste de un polinomio de orden $p \leq m$ en dicha ventana. En nuestro caso usamos $w = 9$ y $p = 2$.

5.9.1.2. Normalización (Z-score)

Tras el suavizado, cada serie se normaliza mediante puntuación estándar (Z-score):

$$z_t \;=\; \frac{\widetilde{x}_t \;-\; \mu_{\widetilde{x}}}{\sigma_{\widetilde{x}}},$$

donde

$$\mu_{\widetilde{x}} \;=\; \frac{1}{N}\sum_{t=1}^{N}\widetilde{x}_t, \quad \sigma_{\widetilde{x}} \;=\; \sqrt{\frac{1}{N-1}\sum_{t=1}^{N}(\widetilde{x}_t - \mu_{\widetilde{x}})^2}.$$

Se aplica este proceso tanto a la inflación no subyacente (P) como a cada polaridad de sentimiento (Q).

5.9.2. Estimación de las distribuciones

Para comparar dos series normalizadas P y Q utilizamos histogramas de densidad con $B = 30$ intervalos equiprobables:

$$\hat{p}_i \;=\; \frac{\#\{P_t \in \text{bin } i\}}{N\,\Delta x}, \quad \hat{q}_i \;=\; \frac{\#\{Q_t \in \text{bin } i\}}{N\,\Delta x},$$

donde Δx es la amplitud del bin y N el número de observaciones. Para evitar ceros en las entradas, se añade un $\varepsilon = 10^{-8}$ a cada $\hat{p}_i$ y $\hat{q}_i$.

5.9.3. Divergencia de Kullback–Leibler

La divergencia de KL asimétrica se define como

$$D_{\mathrm{KL}}(P \parallel Q) \;=\; \sum_{i=1}^{B}\hat{p}_i\,\ln\!\Big(\tfrac{\hat{p}_i}{\hat{q}_i}\Big),$$

y su versión inversa

$$D_{\mathrm{KL}}(Q \parallel P) \;=\; \sum_{i=1}^{B} \hat{q}_i \,\ln\!\Big(\tfrac{\hat{q}_i}{\hat{p}_i}\Big).$$

Finalmente, la divergencia simétrica se calcula como

$$D_{\mathrm{KL}}^{\mathrm{sym}}(P,Q) \;=\; \tfrac{1}{2}\Big[D_{\mathrm{KL}}(P \parallel Q) \;+\; D_{\mathrm{KL}}(Q \parallel P)\Big].$$

La matriz de divergencia de Kullback–Leibler (KL) entre las distribuciones de polaridad y la del INPC no subyacente permite comparar la cercanía estadística de cada combinación *(extracción, método)*; valores menores indican mejor acoplamiento distributivo (Kullback & Leibler, 1951). Bajo **selección experta**, los desempeños más cercanos son `BERT` y `Pysentimiento_refined` (ambos KL ≈ 0.10), seguidos por `BERT_refined` (0.11) y `Finance_refined` (0.17), mientras `Lexicon` se distancia notablemente (0.75). Con **LDA**, domina la familia financiera: `Finance_refined` alcanza la menor divergencia de toda la tabla (0.05), muy cerca de `Finance` sin corrección (0.07); en contraste, `BERT` sin corrección (0.68) y `Pysentimiento_refined` (0.60) exhiben desacoples importantes. En **LDA+LLM**, la opción más cercana es `Regla` sin corrección (0.23), seguida por `BERT`/`BERT_refined` (0.30–0.30) y `Finance` (0.34); la mayor distancia aparece en `Pysentimiento` (0.85), que mejora con corrección pero permanece alta (0.54).

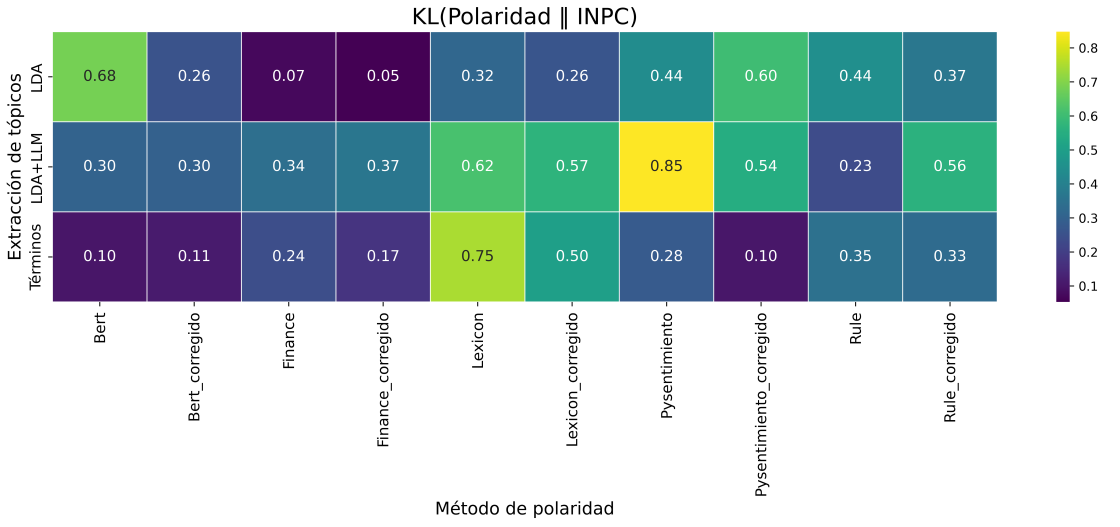

Figura 5.12: Matriz de divergencia KL entre la inflación no subyacente y las polaridades de sentimiento, por método y fuente. Los colores más claros indican menor divergencia (mejor acoplamiento).

Dos patrones emergen con claridad. Primero, la *corrección semi-supervisada* reduce la divergencia en varios casos clave (p. ej., `Finance` con LDA: 0.07 → 0.05; `Pysentimiento` con experto: 0.28 → 0.10), aunque no es universalmente beneficiosa (`Regla` con LDA+LLM: 0.23 → 0.56), lo que sugiere que un suavizado excesivo puede desalinear la distribución respecto al INPC. Segundo, la *dependencia de la fuente de extracción* es marcada: con LDA conviene la familia financiera, con selección experta destacan BERT y Pysentimiento corregido, y con LDA+LLM la regla sin corrección ofrece el mejor encaje.

5.10. Pruebas de causalidad de Granger entre polaridades y la inflación no subyacente

5.10.1. Prueba para método de extracción basado en términos

Nota de contexto. La matriz de valores-p corresponde al *método de extracción basado en términos* (selección experta por reglas/palabras clave). En este entorno, las series ya están enfocadas en vocabulario económico, por lo que la prueba de Granger (*polaridad* $\rightarrow$ *INPC*) evalúa si la señal textual *anticipa* movimientos del índice.

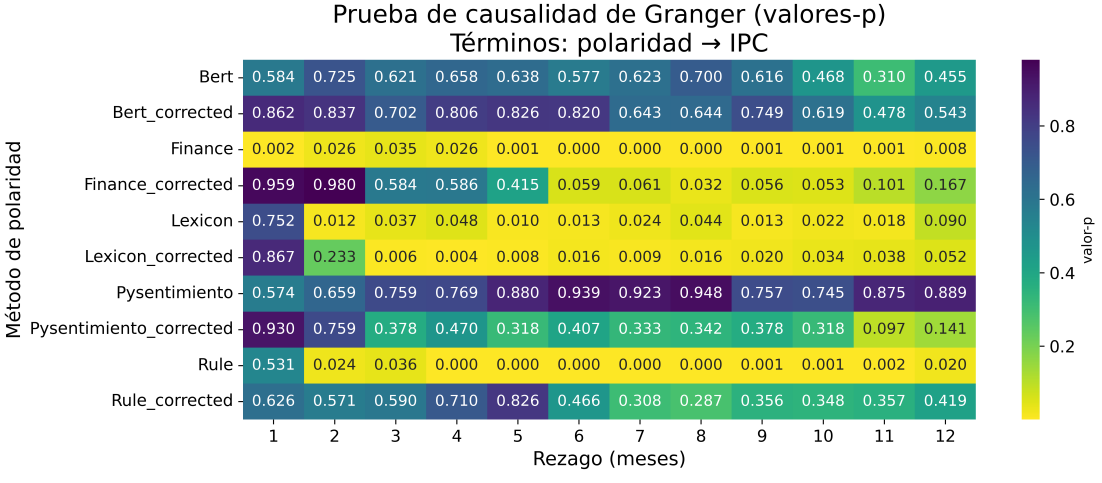

Figura 5.13: Mapa de calor de valores-p de las pruebas de causalidad

Hallazgos principales. (i) Las variantes *no corregidas* de `Regla`, `Lexicon` y `Finance` muestran evidencia robusta de precedencia temporal: `Regla` es significativa en un corredor amplio de rezagos (4–11 meses, $p \approx 0.000$–0.002), `Lexicon` concentra significancia al 5% en 2–9 meses ($p \approx 0.012$–0.044) y `Finance` es significativa prácticamente en todo el rango (1–12

meses, $p \approx 0.001$–0.035). (ii) La *corrección* tiende a atenuar la potencia predictiva: `Regla_coregido`, `Bert_coregido` y `Finance_coregido` elevan sus valores-p; una excepción parcial es `Lexicon_coregido`, que conserva tramos significativos (3–11 meses, $p \leq 0.04$). (iii) `BERT` (con y sin corrección) y `Pysentimiento` en crudo no alcanzan significancia sistemática, por lo que no evidencian causalidad de Granger en esta especificación.

Lectura sustantiva para extracción por términos. El filtrado experto ya concentra contenido económico; sobre esa base, las señales `Regla`, `Lexicon` y `Finance` *sin corrección* preservan variabilidad informativa que parece anteceder al INPC (especialmente a 4–10 meses). La etapa de corrección, útil para estabilizar niveles, puede sobre–suavizar y borrar parte del desfase líder.

5.10.2. Prueba para método de extracción LDA

La prueba de causalidad de Granger para la extracción por *LDA* no arroja evidencia estadística de precedencia temporal al 5% en ninguno de los métodos de polaridad: los valores-p se mantienen, en general, por encima de 0.10 a lo largo de los 12 rezagos. Las mínimas cercanías al umbral aparecen en horizontes muy cortos con `Bert` y `Bert_coregido` (rezago 1: $p = 0.103$ y 0.090), y en media duración con `Lexicon_coregido` (rezagos 6–10: $p \approx 0.124$–0.188), pero ninguna cae por debajo de 0.05. Las familias `Finance` y `Regla`, corregidas o no, muestran valores-p consistentemente altos ($\gtrsim 0.30$–0.95), lo que descarta señal líder robusta en esta especificación; `Pysentimiento` y su versión corregida se comportan de forma similar, con p elevados en casi todos los rezagos.

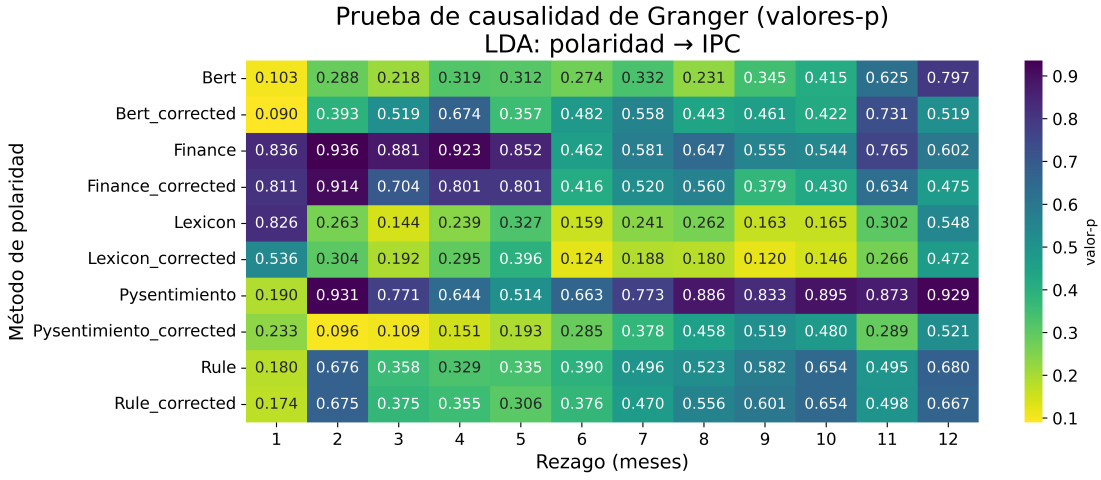

Figura 5.14: Mapa de calor de valores-p de las pruebas de causalidad para extracción LDA

Esta ausencia de significancia sugiere que, bajo *LDA*, las series de polaridad capturan sobre todo componentes coincidentes o de nivel, más que información adelantada respecto al INPC. En términos operativos, conviene utilizar estas señales como *monitores de estado* —por ejemplo, `Bert` o `Bert_coregido` para una lectura estable de corto plazo y `Lexicon_coregido` para anclar el nivel— y reservar la búsqueda de señal líder para otras fuentes de extracción (p. ej., términos expertos) donde la evidencia de Granger fue más favorable. Como precaución metodológica, la interpretación debe acompañarse de pruebas de estacionariedad, elección adecuada del número de rezagos y, de ser necesario, ajustes por multiplicidad (FDR), antes de trasladar estos resultados a un esquema de pronóstico.

5.10.3. Prueba para método de extracción LDA+LLM

La prueba de causalidad de Granger para la extracción *LDA+LLM* (polaridad → INPC) muestra un patrón concentrado: la evidencia de precedencia temporal se limita esencialmente a `Regla_corregido` y, de forma acotada, a `Pysentimiento_corregido`. En `Regla_corregido` los valores-p

caen por debajo del 5% de manera sostenida en rezagos medios y largos ($L = 6$: 0.022; 7: 0.045; 8: 0.018; 9: 0.014; 10: 0.023; 11: 0.042; 12: 0.028). `Pysentimiento_corregido` alcanza significancia sólo al final de la ventana ($L = 12$, $p = 0.027$; umbral cercano en $L = 10$, $p = 0.052$).

El resto de combinaciones no registra valores-p inferiores a 0.05: `Bert` y `Bert_corregido` permanecen elevados; `Finance` y `Finance_corregido` no resultan significativos en ningún lapso; `Lexicon` y `Lexicon_corregido` se mantienen sistemáticamente por encima del umbral.

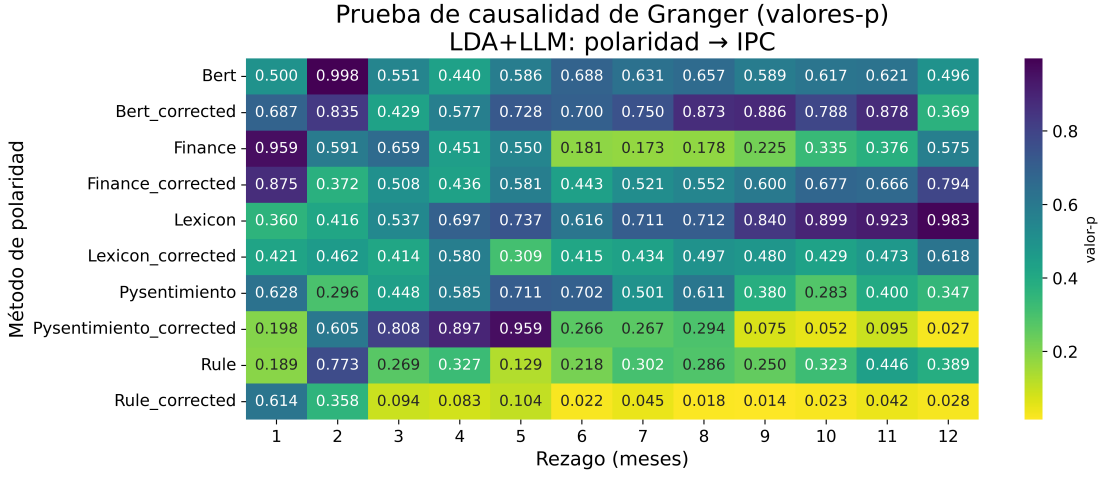

Figura 5.15: Mapa de calor de valores-p de las pruebas de causalidad para extracción LDA+LLM

Bajo *LDA+LLM*, la corrección semi–supervisada potencia a la familia basada en reglas: `Regla_coregido` emerge como *indicador líder* robusto a horizontes de 6–12 meses, mientras que el componente afectivo corregido aporta señal más tardía (12 meses). A diferencia del esquema basado en términos, aquí ni `Finance` ni `Lexicon` ofrecen evidencia de precedencia, lo que sugiere que la coherencia semántica de *LDA+LLM* favorece estructuras regulares (reglas) sobre señales léxicas o financieras para anticipar movimientos del índice.

En la canalización *LDA+LLM*, priorizar `Regla_coregido` como motor

de señal adelantada (6–12 meses) y utilizar `Pysentimiento_coregido` como confirmación en horizontes largos; reservar `BERT`, `Finance` y `Lexicon` como monitores coincidentes/estables más que como detectores de giro. *Caveat metodológico:* corroborar estacionariedad, la selección de rezagos y ajustar por multiplicidad (p. ej., FDR) antes de incorporar estas evidencias a un esquema de pronóstico.

Capítulo 6

Perspectiva y conclusiones

Este trabajo muestra que es factible construir un flujo cómputo–lingüístico capaz de extraer señales de polaridad económica en español y enlazarlas con la dinámica inflacionaria de México (1999–2024). La propuesta articula la extracción temática y depuración semisupervisada (LDA y LDA+LLM) con distintas familias de métodos de polaridad (reglas, léxicos, modelos preentrenados y variantes financieras), evaluadas frente a los componentes subyacente y no subyacente del INPC. En términos globales, la etapa de corrección reduce la varianza y atenúa el ruido de alta frecuencia sin borrar estructura útil; además, la *fuente de extracción* condiciona qué familia rinde mejor y permite configurar una arquitectura parsimoniosa para uso operativo que combina un ancla estable (p. ej., `BERT_corregido`), detectores de cambio (`FINANCE_corregido` y `PYSENTIMIENTO_corregido`) y un mecanismo de suavizado basado en reglas. Estas pautas son consistentes con la evidencia distribucional (divergencia KL) y con los contrastes temporales (causalidad de Granger) reportados en los capítulos empíricos.

La comparación fina por método y fuente matiza el diagnóstico. Las matrices KL indican que la corrección semisupervisada suele acercar la distribución de polaridades a la del INPC, pero el efecto depende del origen de las oraciones: con LDA, el componente financiero logra el encaje más estrecho; con selección experta destacan `BERT` y `PYSENTIMIENTO` tras la depuración; y con LDA+LLM, la señal basada en reglas *sin* corrección puede aproximarse más al índice, lo que advierte contra un suavizado excesivo en ciertos con-

textos. Las pruebas de Granger refuerzan la idea de funciones complementarias: en extracción por términos, `Regla`, `Lexicon` y `Finance` no corregidos exhiben precedencia temporal amplia; en LDA+LLM, `Regla_corregido` se consolida como indicador líder a horizontes de 6–12 meses, mientras el componente afectivo corregido opera como confirmador de más largo plazo. De aquí se desprende que conviene combinar señales estables con sensores sensibles, ajustando la mezcla a la fuente de extracción.

Estos resultados responden a las preguntas iniciales. Los textos económicos, tratados con herramientas adecuadas, permiten caracterizar tendencias y anticipar movimientos del índice sin depender de etiquetado profesional intensivo; además, se establecen procedimientos para comparar y alinear las inferencias con referentes oficiales (INEGI/Banxico), enmarcadas por la trayectoria del INPC subyacente y no subyacente—con episodios de desinflación, el choque pandémico y la posterior moderación, frente a la mayor volatilidad del componente no subyacente—que sirven de telón de fondo para evaluar las señales textuales.

Reafirmación de hipótesis. La evidencia respalda tres ideas centrales. Primero, la *representatividad* del texto: el corpus de noticias contiene información útil para inferir variables macroeconómicas, algo que se refleja en co–movilidad y precedencia temporal en subconjuntos bien definidos. Segundo, la *arquitectura híbrida*: combinar extracción temática (LDA/LDA+LLM) con corrección semisupervisada mejora la calidad de la señal frente a enfoques puramente no supervisados o exclusivamente léxicos, tanto por reducción de varianza como por mejor encaje distribucional. Tercero, la *complementariedad funcional*: distintas familias desempeñan roles diferenciados—ancla, detector y suavizador—y su desempeño depende del origen de las oraciones (términos, LDA o LDA+LLM).

Ubicación en la literatura. En el ámbito de PLN y modelos funda-

cionales, los resultados son coherentes con la capacidad de las arquitecturas Transformer para capturar regularidades semánticas relevantes (Devlin et al., 2019b; Vaswani, Shazeer, Parmar et al., 2017b; Wolf et al., 2020), pero añaden una contribución metodológica: la calidad de la señal no depende sólo del modelo, sino también de la ingeniería de extracción y del grado de corrección aplicado, y su evaluación debe combinar métricas distribucionales (KL) con contrastes temporales (Granger). En la línea de "texto como datos" aplicada a economía, el marco ayuda a explicar por qué algunos estudios hallan poder predictivo y otros sólo co–movilidad: lo determinante no es el "texto" en abstracto, sino *cómo* se construye y depura la señal; además, se ofrece una sintaxis operativa para ensamblar indicadores que mezclen componentes estables y sensibles. Finalmente, respecto a la comunicación de política y el contenido informativo de documentos, la combinación LDA+LLM sugiere que una base temática más coherente puede realzar la señal basada en reglas a horizontes de 6–12 meses y, al mismo tiempo, moderar sobre–reacciones de otras familias, ofreciendo una lectura unificada de qué componente del texto se alinea mejor con la dinámica inflacionaria según el tipo de extracción.

Limitaciones. La mejora por corrección no es monótona: un exceso de suavizado puede degradar la capacidad de *timing*. La señal depende de la cobertura del corpus y del *drift* semántico a lo largo del tiempo. Además, la interpretación de las pruebas temporales exige verificar estacionariedad, seleccionar rezagos de forma cuidadosa y, cuando corresponde, ajustar por multiplicidad.

Líneas de trabajo futuro. Un camino natural es integrar extracción y corrección en un esquema *end-to-end* con aprendizaje débil, explorando autoetiquetado y pérdidas robustas a ruido de etiquetas; ampliar el análisis a otros dominios (confianza del consumidor, empleo, actividad) y fuentes

adicionales (minutas de política, redes verificadas); combinar señales LDA y LLM en marcos de *nowcasting* y *forecasting* multivariado mediante ensamblados jerárquicos que ponderen familias por desempeño contextual; auditar sesgos y estabilidad temporal mediante pruebas de *drift* semántico; y liberar recursos reproducibles—corpus, código y modelos—para facilitar transferencia a entornos de política pública y analítica financiera.

En suma, la tesis tiende un puente operacional entre modelos de lenguaje de última generación (Devlin et al., 2019b; Vaswani, Shazeer, Parmar et al., 2017b; Wolf et al., 2020) y la inferencia macroeconómica en español, especificando *cuándo* y *por qué* distintas construcciones textuales informan sobre la trayectoria inflacionaria, y poniendo a disposición un conjunto de herramientas prácticas para monitoreo y pronóstico.

Bibliografía

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175-185.

Antweiler, W., & Frank, M. Z. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance*, *59*(3), 1259-1294. https://doi.org/https://doi.org/10.1111/j.1540-6261.2004.00662.x

Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063*. https://arxiv.org/abs/1908.10063

Arazo, E., Ortego, D., Álvarez, P. A., O'Connor, N. E., & McGuinness, K. (2019). Unsupervised Label Noise Modeling and Loss Correction. *arXiv preprint arXiv:1904.11238*. https://arxiv.org/abs/1904.11238

Arnold, D., Balkan, L., Meijer, S., Humphreys, R., & Sadler, L. (1994). Example-Based Machine Translation. *Proceedings of the 12th International Conference on Computational Linguistics (COLING '94)*, 110-114.

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 1027-1035.

Ash, E., & Hansen, S. (2023). Text Algorithms in Economics. *Annual Review of Economics*, *15*, 659-688. https://doi.org/10.1146/annurev-economics-082222-074352

Baker, S., Bloom, N., & Davis, S. (2015). *Measuring Economic Policy Uncertainty* (NBER Working Papers Nº 21633). National Bureau of Economic Research, Inc. https://EconPapers.repec.org/RePEc:nbr:nberwo:21633

Balle, B., Ranzato, M., & Synnaeve, G. (2016). Interpretation of random forests. *arXiv preprint arXiv:1610.09506*.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, *5*(2), 157-166.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful? *International conference on database theory*, 217-235.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Bommasani, R., Hudson, D. A., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, C. J., & Stone, R. A. (1984). *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, *21*(4), 543-565.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877-1901.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121-167.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, *31*(2), 102-107.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2017). Sentiment analysis: The state of the art and a comparative review. *Information Sciences*, ** 2017*, 128-157.

Chen, Y., Li, X., Xu, H., & Zhang, J. (2022). BERT4ECON: Pre-trained Language Models for Economic and Financial Text Mining. *Expert Systems with Applications*, *202*, 117177. https://doi.org/10.1016/j.eswa.2022.117177

Chen, Z., Chen, W., Smiley, C., Shah, S., Borova, I., Langdon, D., Moussa, R. N., Beane, M. I., Huang, T.-H. '., Routledge, B. R., & Wang, W. Y. (2021). FinQA: A Dataset of Numerical Reasoning over Financial Data. *ArXiv*, *abs/2109.00122*. https://api.semanticscholar.org/CorpusID:235399966

Cho, K. e. a. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP*.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Jude, C., Krueger, G., et al. (2022). PaLM: Scaling Language Modeling with Pathways. *arXiv preprint arXiv:2204.02311*.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493-2537.

Company, M.

bibinitperiod. (2021). AI in Banking: Can Banks Meet the Challenge?

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, *13*(1), 21-27. https://doi.org/10.1109/TIT.1967.1053964

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783-2792.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019a). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies*, 4171-4186.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019b). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL*, 4171-4186.

Dos Santos, C. N., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 69-78.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179-211.

Feldman, R., & Sanger, J. (2006). Introduction to Text Mining. En *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data* (pp. 1-18). Cambridge University Press.

for Economic Co-operation, O., & Development. (2021). Using Artificial Intelligence for Financial Fraud Detection.

for International Settlements, B. (2020). Artificial Intelligence and Big Data in Finance [BIS FSI Insights on Policy Implementation No. 35].

Fund, I. M. (2018). The Rise of Artificial Intelligence: Implications for the Financial Sector. *Finance & Development*. https://www.imf.org/en/Publications/fandd/issues/2018/06/artificial-intelligence-and-finance.

García-Díaz, J. A., García-Sánchez, F., & Valencia-García, R. (2023). Smart Analysis of Economics Sentiment in Spanish Based on Linguistic Features and Transformers. *IEEE Access*, *11*, 14211-14224. https://api.semanticscholar.org/CorpusID:256777196

Gorodnichenko, Y., Pham, T., & Talavera, O. (2021, marzo). *The Voice of Monetary Policy* (Working Paper Nº 28592). National Bureau of Economic Research. https://doi.org/10.3386/w28592

Greene, D., O'Callaghan, D., & Cunningham, P. (2015). How Many Topics? Stability Analysis for Topic Models. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 555-564.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Grishman, R. (1997). Information extraction: Techniques and challenges. En *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology* (pp. 10-27). Springer. https://doi.org/10.1007/978-3-642-59336-4_2

Gutiérrez Andrade, A., Osvaldo Zurita Moreno. (2006). Sobre la inflación. *PERSPECTIVAS*. https://www.redalyc.org/articulo.oa?id=425942413004

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd). Springer.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 328-339. https://doi.org/10.18653/v1/P18-1031

Hutchins, J. (1986). *Machine Translation: Past, Present, Future*. Ellis Horwood Limited.

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. *European Conference on Machine Learning*, 137-142.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs.

Johnson, S. C. (1967). Hierarchical Clustering Schemes. *Psychometrika*, *32*(3), 241-254. https://doi.org/10.1007/BF02289588

Jokubaitis, S., Celov, D., & Leipus, R. (2020). Sparse structures with LASSO through principal components: Forecasting GDP components in the short-run. *International Journal of Forecasting*. https://api.semanticscholar.org/CorpusID:225194400

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd). Prentice Hall.

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd) [Draft version available online]. https://web.stanford.edu/~jurafsky/slp3/

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 655-665.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746-1751.

Kipper, K., Palmer, M., & Bod, R. (2005). A Large-Scale Classification of English Verbs. *Language Resources and Evaluation*, *39*(4), 305-340.

Kochmar, E. (2022). *Getting Started with Natural Language Processing*. Manning. https://books.google.com.mx/books?id=I4yKEAAAQBAJ

Koppel, M., & Schler, J. (2004). Authorship Verification as a one-class classification problem. En *Proceedings of the Twenty-first International Conference on Machine Learning* (pp. 1-7, Vol. 1). https://doi.org/10.1145/1015330.1015448

Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*, *104*, 38-48. https://doi.org/10.1016/j.dss.2017.10.001

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79-86. https://doi.org/10.1214/aoms/1177729694

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *International Conference on Learning Representations (ICLR)*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4331-4345.

Liu, P., Wang, X., Chen, L., & Chen, Y. (2021). RoBERTa-Econ: Financial and Macroeconomic Text Analysis with Contextualized Embeddings. *Journal of Computational Finance*, *25*(3), 67-92. https://doi.org/10.21314/JCF.2021.410

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*. https://arxiv.org/abs/1907.11692

Lloyd, S. (1982). *Least squares quantization in PCM* (inf. téc.). IEEE.

Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. *7th International Conference on Learning Representations*.

Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35-65. https://doi.org/https://doi.org/10.1111/j.1540-6261.2010.01625.x

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.

Malo, P., Sinha, A., Korhonen, P. J., Wallenius, J., & Takala, P. (2013). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, *65*. https://api.semanticscholar.org/CorpusID:7700237

Mankiw, N. (2015). *Principles of Economics*. Cengage Learning. https://books.google.com.mx/books?id=c76FCwAAQBAJ

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. *AAAI-98 workshop on learning for text categorization*, *752*(1), 41-48.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations (ICLR)*.

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. *Interspeech*, 1045-1048.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. *IEEE Access*, *8*, 131662-131682. https://api.semanticscholar.org/CorpusID:220836326

Mlyahilu, J., Kim, Y., & Kim, J. (2019). Classification of 3D Film Patterns with Deep Learning [Full-text available on ResearchGate]. ––. https://www.researchgate.net/publication/338174593_Classification_of_3D_Film_Patterns_with_Deep_Learning.

Mulvey, J. M., Gu, J., Holen, M., & Nie, Y. (2022). Applications of Machine Learning in Wealth Management. *Journal of Investment Consulting*, *21*(1), 66-82. https://ssrn.com/abstract=4114934

Murtagh, F. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(1), 86-97.

of England, B. (2020). Machine Learning in UK Financial Services.

Otter, D., Medina, J. R., & Kalita, J. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 604-624.

Pan, R., García-Díaz, J. A., García-Sánchez, F., & Valencia-García, R. (2023). Evaluation of transformer models for financial targeted sentiment analysis in Spanish. *PeerJ Computer Science*, *9*. https://api.semanticscholar.org/CorpusID:258596166

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1-135.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543.

Pérez, J., Reverte, C., Puebla, G., Salinas, A., & Figueroa, A. (2021). Pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks. *arXiv preprint arXiv:2106.09462*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227-2237. https://doi.org/10.18653/v1/N18-1202

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners* (Technical Report). OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, *21*(140), 1-67.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982-3992. https://doi.org/10.18653/v1/D19-1410

Richardson, A., van Florenstein Mulder, T., & Vehbi, T. (2021). Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, *37*(2), 941-948. https://doi.org/https://doi.org/10.1016/j.ijforecast.2020.10.005

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence*, *3*(22), 41-46.

Sádaba-Campo, N., & Gómez-Moreno, H. (2025). Exploration of Generative Neural Networks for Police Facial Sketches. *Big Data and Cognitive Computing*, *9*(2). https://doi.org/10.3390/bdcc9020042

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513-523.

Sang, E. T. K., & Meulder, F. D. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Conference on Computational Natural Language Learning*. https://api.semanticscholar.org/CorpusID:2470716

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv preprint arXiv:1910.01108*.

Schölkopf, B., & Smola, A. J. (2002). Learning with kernels: Support vector machines, regularization, optimization, and beyond. *MIT press*.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

Sehrawat, S. (2019). Learning Word Embeddings from 10-K Filings for Financial NLP Tasks. *CompSciRN: Computer Principles (Topic)*. https://api.semanticscholar.org/CorpusID:214123409

Shah, A., Paturi, S., & Chava, S. (2023). Trillion Dollar Words: A New Financial Dataset, Task Market Analysis. https://arxiv.org/abs/2305.07972

Shah, R. S., Chawla, K., Eidnani, D., Shah, A., Du, W., Chava, S., Raman, N., Smiley, C., Chen, J., & Yang, D. (2022). When FLUE Meets FLANG: Benchmarks and Large Pretrained Language Model for Financial Domain. *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:253244049

Shapiro, A. H., & Sudhof, M. (2022). Measuring economic sentiment from news texts. *Economic Analysis Journal*, *35*, 114-130.

Silge, J., & Robinson, D. (2020). *Text Mining with R: A Tidy Approach*. "O'Reilly Media, Inc."

Silva, T. C., Moriya, K., & Veyrune, M. R. M. (2025, junio). *From Text to Quantified Insights: A Large-Scale LLM Analysis of Central Bank Communication* (IMF Working Papers Nº 2025/109). International Monetary Fund. https://doi.org/None

Somers, H. (2003). Review Article: Machine Translation. *Machine Translation*, *17*(2), 77-99.

Sousa, M. G., Sakiyama, K. M., de Souza Rodrigues, L., de Moraes, P. H., Fernandes, E. R., & Matsubara, E. T. (2019). BERT for Stock Market Sentiment Analysis. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 1597-1601. https://api.semanticscholar.org/CorpusID:211208214

Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing. https://books.google.com.mx/books?id=48RiDwAAQBAJ

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 3104-3112.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.

Tang, X., & Lei, N. (2023). Research on CPI Prediction Based on Natural Language Processing. https://arxiv.org/abs/2303.05666

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, *62*(3),

1139-1168. https://doi.org/https://doi.org/10.1111/j.1540-6261.2007.01232.x

Tetlock, P. C., Saar-Tsechansky, M., & Mackassy, S. (2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*, *63*(3), 1437-1467. https://doi.org/https://doi.org/10.1111/j.1540-6261.2008.01362.x

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017a). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017b). Attention Is All You Need. *Advances in Neural Information Processing Systems*, *30*.

Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F., & Tu, K. (2020). Automated Concatenation of Embeddings for Structured Prediction. *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:222290783

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236-244.

Wei, J., et al. (2022a). Emergent Abilities of Large Language Models. *arXiv preprint arXiv:2206.07682*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., et al. (2022b). Chain of Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*.

Weinberger, K. Q., & Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. *Advances in neural information processing systems*, 1473-1480.

Weizenbaum, J. (1966). ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, *9*(1), 36-45.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38-45.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A Large Language Model for Finance. *ArXiv*, *abs/2303.17564*. https://api.semanticscholar.org/CorpusID:257833842

Xie, Q., Han, W., Chen, Z., Xiang, R., Zhang, X., He, Y., Xiao, M., Li, D., Dai, Y., Feng, D., Xu, Y., Kang, H., Kuang, Z.-Z., Yuan, C., Yang, K., Luo, Z., Zhang, T., Liu, Z., Xiong, G., … Huang, J. (2024). FinBen: A Holistic Financial Benchmark for Large Language Models. *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:267760226

Yang, Y., Uy, M. C. S., & Huang, A. (2020). FinBERT: A Pretrained Language Model for Financial Communications. *ArXiv*, *abs/2006.08097*. https://api.semanticscholar.org/CorpusID:219687757

Ye, Y., & Shah, D. (2025). Calibrating Pre-trained Language Classifiers on LLM-generated Noisy Labels via Iterative Refinement. *arXiv preprint arXiv:2503.XXXX*.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, *13*(3), 55-75.

Zhang, X., Xiang, R., Yuan, C., Feng, D., Han, W., Lopez-Lira, A., Liu, X.-Y., Qiu, M., Ananiadou, S., Peng, M., Huang, J., & Xie, Q. (2024). Dólares or Dollars? Unraveling the Bilingual Prowess of Financial LLMs Between Spanish and English. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. https://api.semanticscholar.org/CorpusID:267637074

Zhang, Z., Robinson, D., & Tepper, J. (2022). Automatic Extraction of Economic Indicators from News using Large Language Models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8190-8202. https://doi.org/10.18653/v1/2022.emnlp-main.574

Zhao, L., Li, L., & Zheng, X. (2020). A BERT based Sentiment Analysis and Key Entity Detection Approach for Online Financial Texts. *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 1233-1238. https://api.semanticscholar.org/CorpusID:210699957

Zheng, Y., Xu, Z., & Xiao, A. (2023). Deep learning in economics: a systematic and critical review. *Artificial Intelligence Review*, *56*(9), 9497-9539. https://doi.org/10.1007/s10462-022-10272-8

Zhu, X. J. (2009). *Semi-Supervised Learning*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00196ED1V01Y200906AIM006

Puebla, México

a 10 de noviembre de 2025