Aplicación de Procesamiento del Lenguaje Natural (PLN) para la Predicción del Impacto de Noticias Colombianas en el Índice COLCAP (Bolsa de Colombia)

Por:
David Rico Gaviria

Trabajo de grado:
Maestría en Ingeniería y Analítica de Datos,
Facultad de Ciencias Naturales e Ingeniería,
Universidad Jorge Tadeo Lozano

Profesor:
Jorge Ivan Romero Gelvez

Tabla de contenido

1. Introducción

El estudio del comportamiento de los mercados financieros ha estado históricamente centrado en variables cuantitativas como indicadores macroeconómicos, tasas de interés o reportes corporativos. No obstante, en las últimas décadas ha surgido un creciente interés por comprender cómo los factores cualitativos, particularmente el lenguaje empleado en los medios de comunicación, influyen en la percepción del riesgo y contexto mercantil y, en consecuencia, en las decisiones de inversión (Tetlock, 2007; Loughran & McDonald, 2011). Esta corriente investigativa ha cobrado mayor relevancia con el desarrollo de técnicas de Procesamiento de Lenguaje Natural (PLN) y el acceso masivo a fuentes de noticias digitales.

En el contexto colombiano, el índice COLCAP representa un referente clave para evaluar el rendimiento de la Bolsa de Valores de Colombia (BVC). Aunque existen estudios centrados en predicción bursátil con base en variables económicas, la exploración del impacto del contenido noticioso sobre este índice ha sido escasa. Esta investigación busca cerrar esa brecha mediante el análisis sistemático de noticias económicas colombianas, utilizando herramientas de PLN para extraer señales semánticas y evaluar su relevancia predictiva sobre la dirección diaria del COLCAP.

A lo largo del trabajo se recopilaron más de 8.500 noticias del periodo 2013-2025, las cuales fueron sometidas a procesos de limpieza, análisis de sentimientos y generación de resúmenes automáticos utilizando modelos de lenguaje. El corpus resultante permitió construir dos versiones diferenciadas del dataset: una basada en el contenido completo de las noticias y otra construida a partir de sus resúmenes generados con técnicas de NLP. Ambas versiones fueron enriquecidas con métricas como polaridad, subjetividad, frecuencia de términos, temáticas y codificaciones léxicas.

El objetivo central del estudio fue evaluar si la inclusión de variables lingüísticas derivadas del texto noticioso podía mejorar el poder predictivo de modelos de clasificación orientados a anticipar la dirección del índice bursátil. Para ello se diseñaron múltiples configuraciones de redes neuronales multicapa (MLP), explorando más de 700 combinaciones de hiperparámetros mediante búsqueda en malla. Se analizaron configuraciones con distintas arquitecturas, niveles de limpieza semántica, reducción dimensional (PCA) y regularización (dropout).

Los resultados revelan que los modelos entrenados sobre los resúmenes noticiosos superan consistentemente a aquellos basados en el corpus completo, tanto en

precisión como en estabilidad. El mejor modelo con resúmenes alcanzó un accuracy del 61.36% y un f1_score de 0.612, superando a su contraparte en el corpus original, que logró un accuracy de 60.13% y f1_score de 0.599. Este hallazgo sugiere que los resúmenes actúan como filtros semánticos efectivos, eliminando ruido textual sin perder información relevante para la predicción.

Adicionalmente, se observó que la inclusión indiscriminada de variables lingüísticas puede deteriorar el rendimiento del modelo si no se acompaña de técnicas de reducción o selección contextual. A medida que se incrementó el porcentaje de vocabulario lingüístico incorporado, las métricas de desempeño tendieron a disminuir, salvo en configuraciones donde se aplicó reducción dimensional agresiva vía PCA. Esta relación no lineal entre riqueza léxica y capacidad predictiva resalta la complejidad del lenguaje económico y la necesidad de enfoques adaptativos.

El análisis temático y temporal del corpus permitió también identificar patrones relevantes en la cobertura noticiosa. Términos y expresiones como "peace_talk", "tax", "mining" y "investment" mostraron asociaciones diferenciales con días de alza o baja del mercado, aunque su frecuencia y significancia variaron fuertemente con el tiempo. Esta volatilidad semántica refuerza la hipótesis de que el valor informativo del lenguaje no es estático, sino altamente dependiente del contexto político y económico.

En síntesis, este trabajo demuestra que el lenguaje de las noticias contiene señales útiles para anticipar movimientos del mercado, pero su aprovechamiento requiere un tratamiento cuidadoso. Los resultados obtenidos validan la utilidad de los resúmenes automáticos como insumo semántico y abren nuevas líneas de investigación en la intersección entre finanzas, PLN y aprendizaje automático, especialmente en contextos de mercados emergentes como el colombiano.

2. Marco Teórico

2.1. Fundamentos del Mercado de Valores Colombiano (BVC)

La Bolsa de Valores de Colombia (BVC) es la principal plataforma de negociación de activos financieros en el país. Su funcionamiento refleja el comportamiento de la economía nacional a través de instrumentos como acciones, bonos y derivados. Su sensibilidad a variables macroeconómicas (PIB, inflación, tasas de interés, tipo de cambio), políticas (reformas, elecciones, estabilidad institucional) y sociales (tendencias de consumo, cambio demográfico) la convierte en un objeto ideal para estudios de predicción cuantitativa (Ocampo, 2007; Mishkin, 2018).

Dentro de estos factores, la información mediática —en particular las noticias— representa un componente altamente influyente sobre la percepción del mercado. Noticias relacionadas con la política fiscal, estabilidad monetaria, decisiones del Banco de la República o eventos geopolíticos pueden detonar reacciones inmediatas en los precios de los activos (Tetlock, 2007; Loughran & McDonald, 2011). Esta influencia justifica el interés por cuantificar el impacto de las noticias utilizando herramientas computacionales modernas como el Procesamiento del Lenguaje Natural (PLN).

2.2. Fundamentos del Procesamiento del Lenguaje Natural (PLN)

El PLN es una subdisciplina de la inteligencia artificial cuyo objetivo es habilitar a las máquinas para analizar, comprender y generar lenguaje humano. Su aplicación en el análisis financiero permite interpretar grandes volúmenes de contenido textual como informes, noticias y opiniones de mercado.

Existen distintos niveles de análisis:

- Sintáctico: estructura gramatical y relaciones entre palabras.
- Semántico: interpretación del significado de frases y palabras en contexto.
- Pragmático: análisis de la intención comunicativa y su dependencia del entorno discursivo.

Para propósitos financieros, el PLN se ha utilizado en tareas como clasificación de documentos, extracción de entidades (empresas, sectores, eventos), análisis de tópicos, y detección de polaridad emocional.

2.3. Análisis de Sentimiento en Finanzas

El análisis de sentimiento consiste en cuantificar el tono emocional de un texto. En finanzas, se ha vinculado el sentimiento negativo en medios con caídas bursátiles y aumentos en la volatilidad del mercado (Tetlock, 2007; Loughran & McDonald, 2011).

Se clasifican tres enfoques principales:

- Basados en léxicos: utilizan diccionarios como VADER o Loughran-McDonald, asignando polaridad a cada palabra.
- Basados en modelos supervisados: entrenan clasificadores con ejemplos etiquetados, utilizando algoritmos como Random Forest, SVM o redes neuronales densas.
- Híbridos: combinan ambos enfoques para balancear simplicidad y adaptabilidad.

En esta investigación, se optó por un enfoque léxico ligero usando TextBlob, con polaridad continua entre -1 y 1 y subjetividad entre 0 y 1, para evaluar el tono general de las noticias económicas. Se examinó el impacto del preprocesamiento (lematización) sobre la calidad de las métricas, dado que ciertas formas gramaticales pueden contener matices emocionales que se pierden al normalizar.

2.4. Representación Léxica y Variables Derivadas del Texto

Para integrar la información textual al modelo predictivo, se construyeron representaciones vectoriales del contenido de noticias a partir de:

- N-gramas frecuentes y relevantes, extraídos tras un filtrado de bajo soporte estadístico.
- Frecuencia relativa de aparición de términos clave, normalizada por el número de noticias diarias.
- Categorización semántica mediante embeddings preentrenados (GloVe) y asignación temática con base en similaridad coseno con categorías económicas predefinidas.
- Agregaciones por fecha, colapsando los datos textuales en métricas numéricas alineadas temporalmente con las series de precios.

Esta estrategia permite representar de manera cuantitativa el lenguaje económico en un formato compatible con modelos supervisados.

2.5. Aplicaciones de PLN en predicción bursátil

Diversos estudios han aplicado PLN para predecir movimientos del mercado:

- Bollen et al. (2011) demostraron que el sentimiento en redes sociales podía anticipar el comportamiento del Dow Jones.
- Loughran & McDonald (2011) diseñaron un léxico financiero especializado, evidenciando que el uso de palabras específicas tenía correlación con retornos negativos.
- Awajan et al. (2021) integraron modelos de PLN y deep learning para predecir variaciones en precios accionarios con éxito moderado.

Sin embargo, la mayoría de estas investigaciones se enfocan en mercados desarrollados. La presente tesis contribuye a este campo al aplicar estas técnicas en un entorno emergente (Colombia), explorando sus limitaciones, validez empírica y potencial de aplicación.

3 Estado del Arte

3.1. Impacto de Noticias en el Mercado de Valores:

La comprensión y la predicción del impacto de las noticias en el mercado de valores ha sido un tema de interés persistente para académicos, profesionales de las finanzas e inversores. La hipótesis del mercado eficiente (EMH) sugiere que los precios de los activos reflejan toda la información disponible, incluyendo las noticias, lo que implica que solo las noticias inesperadas (sorpresas) deberían influir en los precios (Fama, E.F., 1970.). Sin embargo, las limitaciones de los inversores, los sesgos y las fricciones del mercado a menudo conducen a reacciones exageradas o insuficientes a las noticias, creando oportunidades para estrategias de negociación informadas (Barberis, N., & Thaler, R., 2003.).

3.2. Enfoques Tradicionales:

El análisis del impacto de las noticias en los mercados financieros se ha basado en métodos cualitativos, como el análisis basico, la opinión de expertos y el seguimiento de eventos (Reilly, F.K., & Brown, K.C., 2012). El análisis fundamental examina los estados financieros de las empresas, las condiciones de la industria y los factores macroeconómicos para evaluar el valor intrínseco de las acciones. El seguimiento de eventos implica la identificación manual de eventos noticiosos relevantes y el análisis de su impacto en los precios de las acciones. Aunque estos métodos pueden proporcionar información valiosa, son subjetivos, requieren mucho tiempo y son difíciles de escalar para analizar grandes volúmenes de noticias.

3.3. Previsión de Variaciones en la Curva de Precios (Forecasting):

Los modelos econométricos, como los modelos de series temporales (ARIMA, GARCH) y los modelos de vectores autorregresivos (VAR), se han utilizado para pronosticar las variaciones en los precios de las acciones basándose en datos históricos (Box, G.E.P., Jenkins, G.M., Reinsel, G.C., & Ljung, G.M., 2015). Estos modelos pueden incorporar variables exógenas, como indicadores macroeconómicos y noticias, para mejorar la precisión de las predicciones. Sin embargo, estos modelos a menudo asumen relaciones lineales y estacionarias, lo que puede no ser apropiado para los mercados financieros, que son dinámicos y no lineales.

3.4. Introducción del Aprendizaje Automático (Machine Learning) y Aprendizaje Profundo (Deep Learning):

Con el auge del procesamiento del lenguaje natural (PLN) y el análisis de sentimiento, se ha vuelto posible analizar grandes cantidades de datos de noticias de manera objetiva y eficiente y utilizar esta información para predecir el impacto de las noticias en el mercado de valores (Nassirtoussi, A.K., Aghabozorgi, S., Wah, T.Y., & Ngo, D.C.L., 2014.). Las técnicas de aprendizaje automático, como las máquinas de vectores de soporte (SVM), los árboles de decisión y las redes neuronales, se han utilizado para clasificar el sentimiento de las noticias y predecir los movimientos de los precios de las acciones (Mittermayer, M.A., 2019). El aprendizaje profundo, en particular las redes neuronales recurrentes (RNN) y los transformadores, ha demostrado ser muy eficaz para capturar las dependencias a largo plazo y las relaciones no lineales en los datos de noticias, mejorando la precisión de las predicciones (Hagenau, M., Wohlfahrt, R., & Knappe, R., 2013.).

Ejemplos de Estudios sobre la Interpretación de Noticias y la Variación del Mercado de Valores:

- Bollen et al. (2011) analizaron el sentimiento de los tweets y encontraron que el "estado de ánimo" de Twitter podía predecir los movimientos del mercado de valores.
- Loughran and McDonald (2011) desarrollaron un diccionario de análisis de texto específico para finanzas y demostraron que el uso de este diccionario mejoraba la precisión del análisis de sentimiento en el contexto del mercado de valores.
- Tetlock (2007) examinó el impacto del pesimismo de los medios en los precios de las acciones y encontró que un mayor pesimismo en los medios conducía a menores rendimientos de las acciones.
- Li et al. (2014) utilizaron el análisis de sentimiento de las noticias financieras para predecir la volatilidad del mercado de valores y encontraron que el sentimiento negativo estaba asociado con una mayor volatilidad.
- Awajan et al. (2021) usaron análisis de sentimiento de noticias y deep learning para predecir el mercado de valores.

3.5. Trabajos Similares para el Caso de la Bolsa Colombiana:

La investigación sobre el impacto de las noticias en la Bolsa de Colombia (BVC) es limitada en comparación con los mercados más desarrollados. Sin embargo, algunos estudios han comenzado a explorar esta área:

Vargas et al. (2023) investigaron la influencia de noticias en el mercado accionario colombiano mediante técnicas de procesamiento del lenguaje natural y aprendizaje automático, utilizando clasificadores como Naive Bayes y SVM para medir únicamente la polaridad de las noticias y su relación con los movimientos del índice COLCAP.

3.6. Direcciones Futuras:

La investigación futura en esta área debería centrarse en los siguientes aspectos clave para mejorar la precisión, aplicabilidad y robustez del sistema predictivo basado en noticias:

1. Mejoras en el análisis de sentimiento: Desarrollar modelos más sofisticados que capturen matices semánticos, ironía, sarcasmo y subjetividad, elementos clave en el lenguaje financiero y en la interpretación de noticias económicas. Explorar enfoques híbridos que combinen modelos lexicográficos, redes neuronales y técnicas de representación contextual como BERT o FinBERT.
2. Expansión de fuentes de datos: Incorporar una variedad más amplia de fuentes de información, incluyendo redes sociales (Twitter, Reddit, LinkedIn), blogs financieros, foros de discusión y reportes de analistas.
3. Aplicación de modelos avanzados de aprendizaje profundo: Explorar arquitecturas más complejas, como transformers multimodales y modelos autoregresivos, que permitan capturar dependencias a largo plazo, mejorarando la capacidad predictiva mediante el análisis de secuencias temporales.
4. Segmentación sectorial y diversificación de activos: Investigar el impacto de las noticias en distintos sectores de la economía y en diversos tipos de activos.
5. Integración con estrategias de inversión automatizadas: Desarrollar estrategias de trading basadas en el análisis de noticias, evaluando su efectividad en escenarios simulados. Esto permitiría convertir los modelos predictivos en herramientas prácticas para gestión de portafolios.
6. Análisis de impacto de noticias falsas y manipulación informativa: Investigar cómo la difusión de fake news afectan el mercado. Implementar mecanismos para detectar y filtrar información poco confiable, asegurando la calidad de los datos utilizados en los modelos.
7. Adaptación a otros mercados emergentes: Extender la investigación a mercados financieros de otros países en América Latina, identificando

patrones comunes y diferencias en la manera en que la información influye en las bolsas de valores.

4. Planteamiento del Problema

El mercado de valores es un sistema altamente dinámico influenciado por múltiples factores económicos, políticos y sociales. En Colombia, donde la Bolsa de Valores de Colombia (BVC) opera en un entorno caracterizado por volatilidad e incertidumbre, la capacidad de anticipar sus movimientos representa un desafío crucial para inversionistas, analistas financieros y entes reguladores. Dentro de estos factores, las noticias juegan un papel determinante en la formación de expectativas sobre el mercado, ya que pueden afectar de manera significativa la percepción y toma de decisiones de los agentes económicos (Bollen, Mao & Zeng, 2011; Tetlock, 2007). Sin embargo, el impacto de las noticias en la BVC no se ha modelado con precisión, lo que deja un vacío en la capacidad predictiva para mercados emergentes, específicamente el colombiano.

El análisis tradicional de noticias en el ámbito financiero ha sido principalmente cualitativo, basado en la interpretación de expertos y en el análisis básico de los mercados. Sin embargo, estos enfoques presentan limitaciones significativas: son subjetivos, no escalables y no permiten el procesamiento eficiente de grandes volúmenes de información en tiempo real. Con el auge del Procesamiento del Lenguaje Natural (PLN) y el aprendizaje profundo (“deep learning”), se ha demostrado que es posible analizar automáticamente noticias y evaluar su impacto en los mercados financieros con un alto grado de precisión (Loughran & McDonald, 2011; Nassirtoussi et al., 2014). Sin embargo, la mayoría de los modelos existentes han sido diseñados para mercados financieros de países desarrollados y entrenados en inglés, lo que limita su aplicabilidad en el contexto colombiano, donde la semántica y estructura del lenguaje pueden ser muy diferentes. Debido a la dependencia de idioma, se adicional la necesidad de creación de un conjunto de datos en idioma inglés sobre noticias para Colombia.

Ante esta problemática surge la necesidad de desarrollar un modelo basado en Deep learning que permita predecir si la BVC subirá o bajara. Para lograrlo, es fundamental diseñar un sistema que aplique análisis de sentimiento, categorización temática, análisis de frecuencia de palabras y flujos de lenguaje, generando matrices de datos que alimenten modelos predictivos.

Este proyecto busca llenar este vacío investigativo y tecnológico, proporcionando una herramienta innovadora que contribuya a la toma de decisiones en el mercado bursátil colombiano mediante la integración de técnicas avanzadas de PLN y modelos predictivos. Con esto, se pretende mejorar la capacidad de anticipación de los agentes económicos, optimizando estrategias de inversión y gestión de riesgo en la BVC.

5. Objetivos

5.1. Objetivo General

Evaluar el impacto predictivo de distintas representaciones lingüísticas de noticias económicas sobre la dirección diaria del índice COLCAP, comparando el desempeño entre modelos entrenados con el corpus completo y con resúmenes automáticos, e incorporando variables semánticas como polaridad, subjetividad y tópicos temáticos.

5.2. Objetivos Específicos

1. Construir un conjunto de datos consolidado de noticias económicas relevantes para el contexto colombiano en inglés, alineado temporalmente con la evolución del índice COLCAP.
2. Generar representaciones semánticas a partir de las noticias, incluyendo análisis de polaridad, subjetividad, lematización temática y resúmenes automáticos mediante PLN.
3. Diseñar y entrenar modelos predictivos basados en redes neuronales multicapa, aplicando búsqueda de hiperparámetros para evaluar distintas configuraciones arquitectónicas.
4. Comparar el desempeño de los modelos entrenados con corpus completo frente a aquellos basados en resúmenes, considerando diferentes niveles de inclusión de variables lingüísticas.
5. Analizar el impacto de los hiperparámetros, la complejidad semántica y la reducción dimensional en el rendimiento de los modelos, con el fin de identificar configuraciones óptimas.
6. Reflexionar críticamente sobre los desafíos metodológicos del uso de lenguaje natural en predicción financiera y proponer líneas de investigación futuras.

6. Metodología

El enfoque adoptado se basa en la metodología CRISP-DM (‘*Cross-Industry Estándar Process for Data Mining’*). Se plantean las siguientes fases claves, comenzando con la comprensión del negocio y datos, seguida de la preparación de los datos, modelado, evaluación, despliegue y, finalmente, documentación del proceso y comunicación de resultados.

Fase 1: Comprensión del Negocio y datos.

- **A.** Revisión de literatura sobre sistemas de predicción de la valoración de la bolsa en general
- **B.** Entrevistas con expertos financieros para el entendimiento de las particularidades de la bolsa colombiana y su actual entendimiento.
- **C.** Recopilación de conjuntos de datos de la valoración de BVC, así como la obtención de un set de datos de noticias del contexto colombiano indexadas por fecha de publicación.

Fase 2: Preparación de los datos.

- **A.** Limpieza y transformación de los datos para tratar posibles problemas de calidad de la información, principalmente una validación de la extracción de noticias y su contenido.
- **B.** normalización y estandarización de los datos de la bolsa y su valoración en una curva que contenga los cambios de la bolsa.
- **C.** División de los conjuntos de datos en entrenamiento, validación y prueba, evitando clases sobrecargadas.

Fase 3: Modelado.

- **A.** Entrenamiento y comparación de algoritmos de análisis de sentimiento sobre las noticias.
- **B.** Entrenamiento y comparación de algoritmos de flujo de lenguaje y categorización semántica de las noticias.
- **C.** Entrenamiento y comparación de algoritmos para la predicción de impacto de información lingüística sobre la bolsa de valores colombiana.

Fase 4: Evaluación.

- **A.** Evaluación de los modelos desarrollados sobre los datos de prueba y validación.

B. Análisis de métricas de rendimiento, precisión y especificidad.

Fase 5: Despliegue.

A. Implementación del sistema de tratamiento de la información con el mejor desempeño.

B. Implementación del modelo de predicción sobre el impacto de las noticias sobre la valoración de BVC.

Fase 6: Documentación y comunicación de Resultados.

A. Elaboración de informes detallados sobre el aprendizaje obtenido en el desarrollo del proyecto.

B. Presentación de los resultados obtenidos a las partes interesadas.

Fase	Semana	1-2	3-4	5-6	6-7	7-8	9-10	11-12	13-14	15-16	17-18	19-20	21-22	23-24	25-26	27-28	29-30
1A																	
1B																	
1C																	
2A																	
2B																	
2C																	
3A																	
3B																	
3C																	
4A																	
4B																	
5A																	
5B																	
6A																	
6B																	

Tabla 1 referencia a las fases planteadas en la metodología, creación propia.

7. Presupuesto.

Se plantea el uso de una máquina virtual dedicada de 16 GB de memoria RAM y 100 GB de almacenamiento (recurso físico). Similarmente, el recurso humano necesario contempla la dedicación de un profesional en aprendizaje de maquina e ingeniería de datos.

Presupuesto para el desarrollo del proyecto:

Semanas	Recurso	Físico (hr/w)	Humano (hr/w)
1-2		0	15
3-4		0	15
5-6		40	15
7-8		40	15
9-10		40	10
11-12		40	10
13-14		40	10
15-16		40	10
17-18		40	10
19-20		40	15
21-22		40	15
23-24		40	15
25-26		40	15
27-28		40	15
29-30		0	20
Horas Totales		960	410
Costo Hora (USD)		0,25	15
Costo Total (USD)		240	6150

Tabla 2 referencia al presupuesto del proyecto.

8. Descripción del Dataset y Preprocesamiento

Con el fin de estudiar el impacto del lenguaje noticioso en la predicción de la fluctuación del índice COLCAP, se desarrollaron y analizaron dos versiones diferenciadas del dataset, cada una estructurada en torno a un enfoque distinto de representación lingüística del contenido de las noticias. Esta estrategia dual responde al objetivo metodológico de evaluar no solo el valor predictivo del lenguaje financiero, sino también cómo diferentes niveles de procesamiento semántico pueden alterar la señal extraída del texto.

El primer dataset está basado en el análisis directo del corpus textual completo de cada noticia. Este enfoque parte de la premisa de que todo el contenido, aunque redundante o periférico, puede contener señales distribuidas relevantes que afectan indirectamente el comportamiento del mercado. Para este conjunto se aplicaron técnicas tradicionales de Procesamiento de Lenguaje Natural (PLN) como lematización, eliminación de stopwords y extracción de frecuencias de términos (n-gramas), generando variables cuantitativas que reflejan el uso léxico en su forma más amplia.

El segundo dataset, en cambio, se apoya en un preprocesamiento más profundo del contenido mediante un modelo de resumen automático basado en aprendizaje profundo. La hipótesis que motiva este enfoque es que la compresión semántica del texto, centrada en su contenido informativo esencial, podría resaltar de forma más clara los patrones lingüísticos que realmente están vinculados con los movimientos del mercado, minimizando el “ruido” característico de los artículos periodísticos, como frases de contexto, citas indirectas o redacción editorial redundante.

Es importante destacar que, sobre este corpus resumido, se aplicaron las mismas técnicas de análisis lingüístico y extracción de variables que en el dataset basado en el corpus completo. Esto incluye la generación de n-gramas, conteo de términos, análisis de polaridad, subjetividad, categorización temática, y demás indicadores semánticos utilizados en el estudio. Al mantener constantes estas metodologías de análisis, se garantiza la comparabilidad entre ambos enfoques, permitiendo una evaluación directa del impacto del resumen sobre la calidad de las variables lingüísticas y, por ende, sobre el desempeño de los modelos predictivos entrenados posteriormente.

Este procedimiento de resumen automático no solo mejora la eficiencia computacional al reducir la longitud del texto, sino que además podría facilitar que

los modelos posteriores (como los clasificadores) identifiquen con mayor precisión señales semánticas relevantes. Esta estrategia también simula el comportamiento de lectura de un analista humano, quien rara vez lee todo el cuerpo noticioso y en cambio se enfoca en fragmentos clave con información relevante.

Otra justificación importante de comparar ambos enfoques es que permite evaluar hasta qué punto los métodos de PLN avanzados, como el resumen extractivo, pueden preprocesar el lenguaje natural de manera más eficaz que técnicas estadísticas tradicionales. En contextos financieros, donde la relación entre texto y mercado es altamente no lineal y contextual, esta evaluación se vuelve crítica.

Además, la comparación ayuda a determinar si el contenido redundante en los textos noticiosos introduce sesgos o simplemente dispersa la señal predictiva. Al contrastar ambos enfoques, también se espera identificar qué tipo de representación lingüística favorece modelos más precisos.

8.1. Obtención del corpus noticioso

El conjunto de datos principal usado en esta investigación fue construido mediante técnicas de web scraping, aplicadas a medios de comunicación digitales colombianos en inglés, usados por inversores como fuente de información. Se recopilaron noticias relacionadas con economía, política, mercados y coyuntura nacional. El proceso de extracción automatizada permitió capturar de forma estructurada los titulares, fechas de publicación, enlaces, como el cuerpo completo del texto de cada noticia.

En total, se recolectaron 12,152 noticias abarcando el periodo comprendido entre 2008-02-14 y 2025-02-27, lo cual proporciona una cobertura temporal alineada con las series históricas del índice COLCAP, con un promedio de noticias por día de 2.88. Esta sincronización es esencial para correlacionar las noticias con el comportamiento diario del mercado bursátil colombiano.

Descripción	Corpus Noticioso	Resumen Noticioso
Total, de noticias	12152	12152
Mínimo temporal	2008-02-14	2008-02-14
Máximo temporal	2025-02-27	2025-02-27
Promedio de Noticias por dia	2.88	2.88
Promedio de cantidad de caracteres del contenido	2427.31	297.56
Promedio de cantidad de palabras del contenido	1591.05	46.20

Tabla 3 Descripción de dataset de noticias.

8.2. Obtención del resumen del corpus noticioso

Con el propósito de evaluar el efecto de una representación condensada del lenguaje noticioso sobre la capacidad predictiva del modelo, se construyó una segunda versión del dataset a partir de resúmenes automáticos de cada artículo. Este enfoque busca reducir el ruido textual característico de los medios escritos, como información redundante, contextual o narrativa, y concentrarse en el contenido central de cada noticia, el cual se presume más alineado con el potencial efecto sobre el mercado financiero.

Para este proceso, se implementó un método de resumen extractivo basado en el modelo pre-entrenado BART (*Bidirectional and Auto-Regressive Transformers*), en su versión optimizada para tareas de resumen: facebook/bart-large-cnn. Este modelo fue elegido por su capacidad comprobada para generar resúmenes coherentes, informativos y gramaticalmente correctos, habiendo sido entrenado sobre un corpus masivo de noticias (CNN/DailyMail), lo que lo convierte en una opción ideal para tareas de compresión semántica de contenido periodístico.

Desde el punto de vista técnico, el procedimiento de resumen se estructuró en los siguientes pasos:

1. Carga del modelo y tokenizador: se utilizó BartTokenizer y el pipeline de HuggingFace summarization para encapsular el proceso de generación de resumen.
2. Segmentación del texto original: dado que muchos artículos exceden el límite de 1024 tokens del modelo, el texto completo se dividió en fragmentos de hasta 1024 palabras, preservando integridad semántica.
3. Generación de resumen por fragmento: cada segmento fue resumido individualmente, aplicando beam search con 4 haces (num_beams=4), un máximo de 200 tokens por resumen (max_length=200) y penalización de longitud (length_penalty=2.0) para evitar resultados demasiado breves o vagos.
4. Ensamblaje del resumen final: los resúmenes individuales de cada fragmento se concatenaron para formar un único texto representativo del contenido esencial de la noticia original.

Esta implementación permitió abordar la limitación de entrada del modelo y asegurar que incluso los artículos extensos pudieran ser resumidos sin truncamientos arbitrarios. La estrategia de “resumir por bloques” garantiza que se mantenga el contexto temático dentro de cada segmento, y al mismo tiempo ofrece una visión condensada y continua del contenido relevante.

Desde una perspectiva metodológica, el uso de resúmenes ofrece múltiples beneficios. Primero, reduce la dimensionalidad del texto, lo que impacta positivamente en la eficiencia computacional del pipeline de extracción de variables. Segundo, focaliza el análisis semántico en el contenido potencialmente más informativo, filtrando frases ornamentales o estructurales comunes en el periodismo, pero irrelevantes para el mercado. Finalmente, permite comparar el poder explicativo del lenguaje original vs. el lenguaje resumido, brindando evidencia empírica sobre si el PLN avanzado puede mejorar la calidad de las señales semánticas extraídas.

8.3. Preprocesamiento del texto

El contenido textual de cada noticia fue sometido a un proceso riguroso de limpieza y normalización con el fin de garantizar una representación lingüística uniforme, reducida en ruido y estructuralmente adecuada para su análisis computacional. Este proceso se aplicó de forma paralela e independiente a los dos conjuntos de datos construidos: (1) el corpus noticioso completo, y (2) el resumen automático generado a partir del mismo contenido.

8.3.1. Pasos comunes de preprocesamiento

En ambos casos, se aplicaron los siguientes procedimientos:

- Conversión a minúsculas para unificar las variantes de mayúsculas y minúsculas de los términos.
- Eliminación de signos de puntuación, números y caracteres especiales, que no aportan significado semántico útil para el análisis.
- Tokenización utilizando nltk.word_tokenize(), que divide el texto en unidades lingüísticas mínimas o *tokens*.
- Lematización para transformar cada palabra a su forma base o canónica. Este paso, realizado con modelos lingüísticos de spaCy, se prefirió frente al stemming por su capacidad para preservar el contexto gramatical y mejorar la coherencia semántica.

Este pipeline permitió reducir la dimensionalidad del texto, normalizar expresiones lingüísticas y mejorar la calidad de las representaciones utilizadas para extraer variables semánticas tales como conteo de términos, polaridad, subjetividad y temas dominantes.

8.3.2. Comparación entre corpus completo y resumen

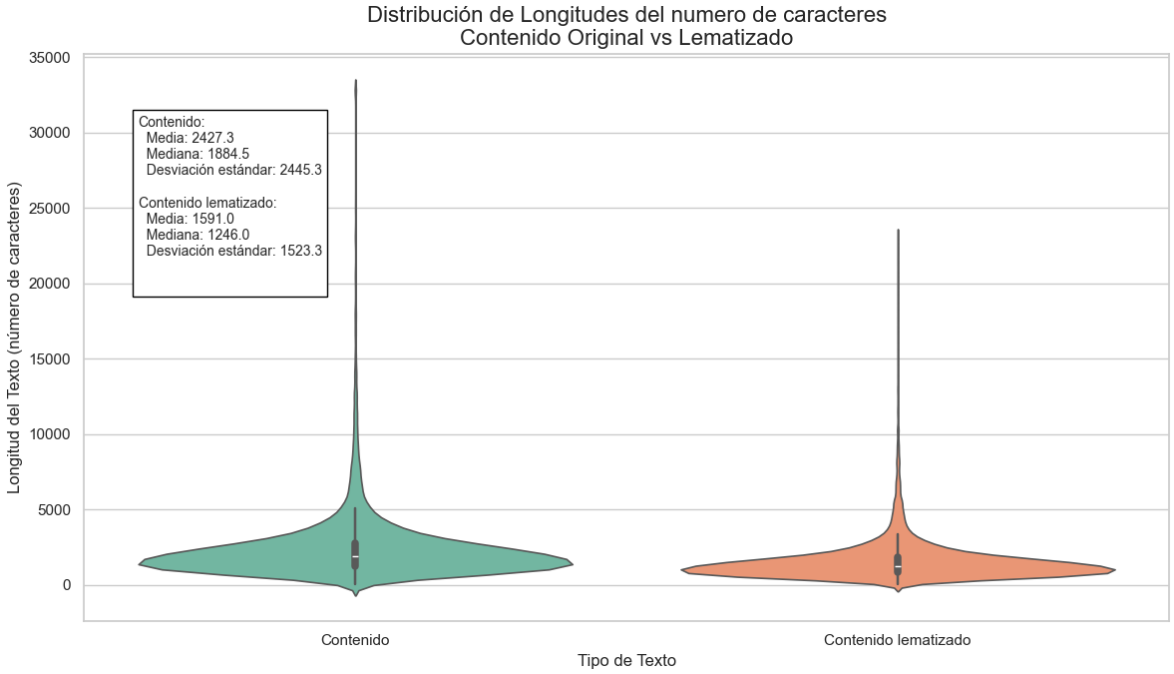

Ilustración 1 Grafica de violín para la distribución de longitudes del numero de caracteres del contenido original vs lematizado.

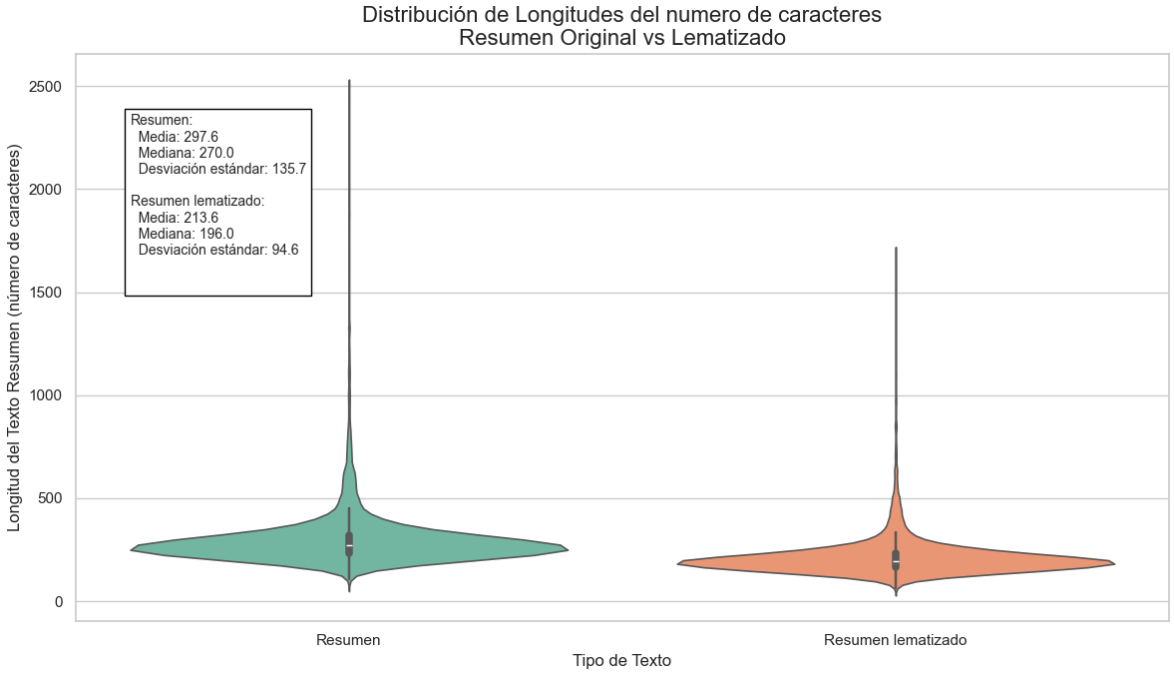

Ilustración 2 Grafica de violín para la distribución de longitudes del numero de caracteres del contenido resumido original vs lematizado

Como parte del diseño experimental, se realizó una comparación cuantitativa entre ambos tipos de texto procesado. En las Ilustración 1 y 2, se presenta la reducción promedio en la longitud del texto (en caracteres) antes y después de aplicar los pasos anteriores. En el caso del corpus completo, la longitud media por noticia pasó de 2.427 a 1.591 caracteres tras la limpieza, lo que representa una disminución de 836 caracteres. En contraste, el corpus de resúmenes, al ser generado por modelos NLP optimizados para condensar la información, presentó una longitud promedio aún más reducida desde el inicio, con valores cercanos a los 200 caracteres en promedio post-limpieza.

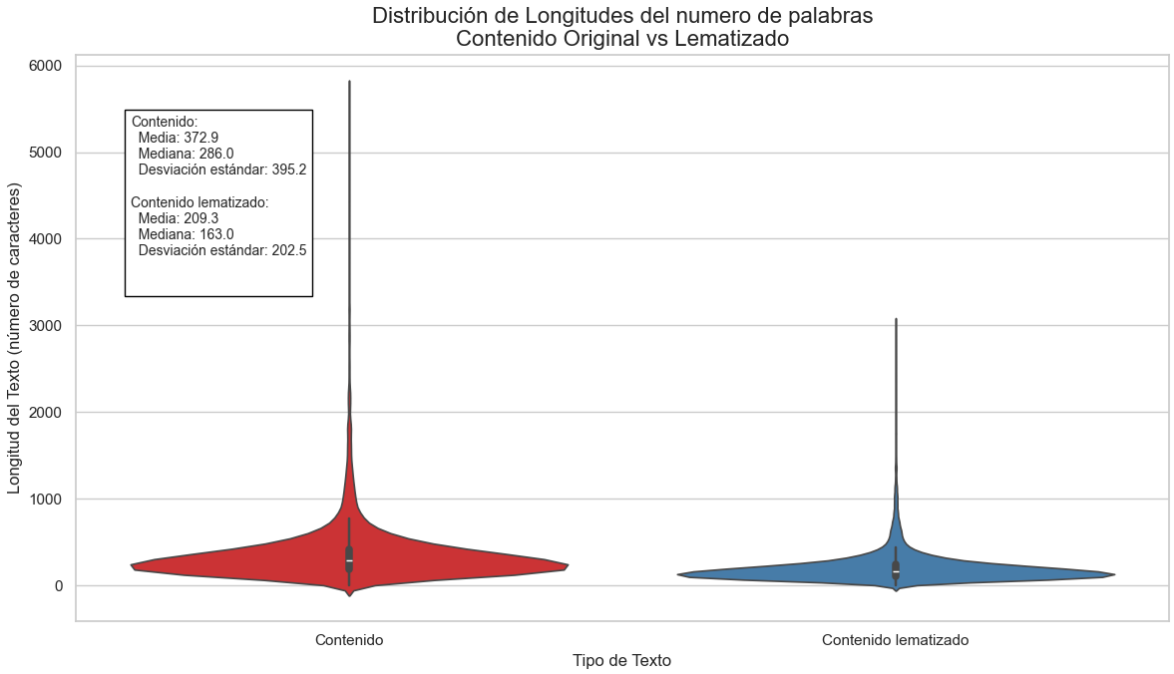

Ilustración 3 Grafico de violín para la distribución de longitudes del numero de palabras para el contenido original vs lematizado.

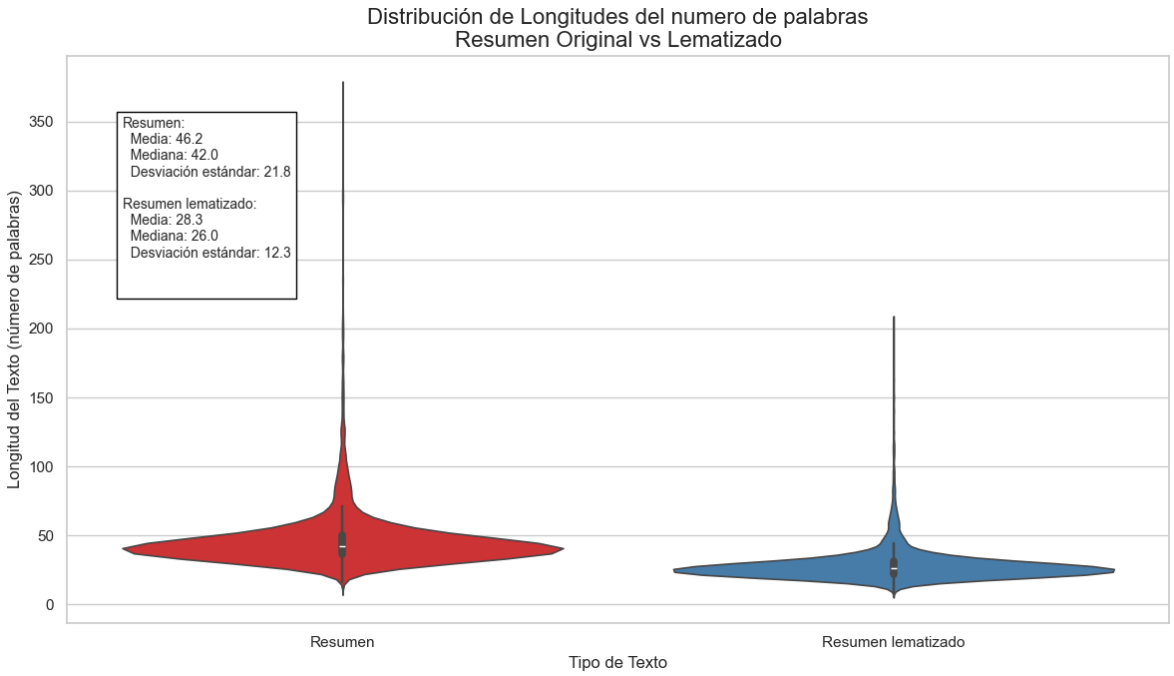

Ilustración 4 Grafico de violín para la distribución de longitudes del número de palabras en el resumen original vs lematizado.

De forma similar, en las Ilustraciones 3 y 4 se observa el impacto del preprocesamiento sobre la cantidad de palabras. Para el corpus completo, la media descendió de 373 a 209 palabras por noticia tras el procesamiento. En el caso de los resúmenes, la longitud promedio fue de apenas 46 palabras, lo que pone en evidencia su nivel de compresión semántica.

Estas diferencias no solo son cuantitativas, sino también cualitativas. En el corpus original, el lenguaje incluye un mayor número de elementos redundantes, conectores narrativos y referencias contextuales. En cambio, los resúmenes presentan una concentración más alta de términos claves, lo que se traduce en una menor dispersión en las distribuciones léxicas (como se evidencia en los gráficos de densidad). Esta menor dispersión facilita el entrenamiento de modelos estadísticos y de aprendizaje profundo, al reducir el espacio semántico de representación.

8.3.3. Implicaciones para el modelado

Ambos enfoques ofrecen ventajas y limitaciones. El corpus completo, por su extensión y riqueza contextual, puede capturar señales semánticas más complejas y matizadas, pero también introduce más ruido léxico. Por su parte, el resumen automático, aunque más breve, concentra información potencialmente más relevante para el análisis predictivo.

Al aplicar el mismo pipeline de extracción de variables a ambos conjuntos, se garantiza una comparación justa y controlada entre representaciones lingüísticas.

Esto permite evaluar no solo el impacto del contenido textual sobre los modelos predictivos, sino también la eficacia de los resúmenes automáticos como herramienta de preprocesamiento avanzado en tareas de PLN aplicadas a finanzas.

En suma, esta sección demuestra cómo el preprocesamiento léxico, tanto en el corpus completo como en su versión resumida, no solo reduce la complejidad superficial del texto, sino que actúa como un filtro estructural que potencia el análisis semántico posterior ya sea para extracción de tópicos, modelado de sentimiento o clasificación supervisada.

8.4. Construcción del diccionario de términos

Una vez procesado el contenido textual de ambos conjuntos de datos —el corpus completo de noticias y su correspondiente versión resumida generada mediante modelos de PLN— se procedió a la construcción de diccionarios de términos relevantes que permitieran representar cuantitativamente el lenguaje predominante en cada enfoque. Esta fase tuvo como objetivo central transformar el lenguaje natural en estructuras numéricas interpretables que pudieran alimentar los modelos de análisis semántico y predicción financiera desarrollados posteriormente.

Para ello, se empleó un enfoque clásico basado en n-gramas, mediante la implementación del algoritmo CountVectorizer. Esta herramienta permite convertir texto libre en una matriz de ocurrencias, contabilizando la frecuencia de aparición de unigramas y bigramas dentro de cada documento. En ambos datasets (corpus y resumen), se construyó una matriz independiente, manteniendo los mismos parámetros de vectorización para garantizar una comparación válida.

Como medida de depuración semántica, se estableció un umbral mínimo de frecuencia de aparición igual a dos. Esto permitió filtrar palabras espurias, errores de tokenización y términos utilizados de forma excepcional, preservando únicamente aquellas expresiones recurrentes que pudieran tener un valor informativo significativo. Esta decisión técnica es especialmente importante en el corpus completo, donde la longitud de los textos puede introducir términos irrelevantes, y en el resumen, donde se busca conservar únicamente la señal dominante.

El resultado fue un diccionario de términos filtrado para cada dataset, el cual contiene únicamente las palabras con uso repetido, permitiendo así una representación robusta del vocabulario informativo de las noticias analizadas. Además, para cada enfoque (corpus y resumen), se generaron dos versiones del diccionario: una basada en el contenido original sin normalización, y otra a partir del contenido lematizado, lo cual permite analizar el impacto de la normalización léxica en la estructura semántica del texto.

Como parte del análisis exploratorio, se construyeron nubes de palabras para los cuatro escenarios generados:

1. Corpus original
2. Corpus lematizado
3. Resumen original
4. Resumen lematizado

Cada una de estas nubes representa las 500 palabras más frecuentes dentro de su respectivo conjunto, lo que permite observar visualmente las diferencias léxicas y conceptuales entre ambas representaciones. En las Ilustraciones 5 y 6, correspondientes al corpus completo, se destacan términos como *Colombia*, *government*, *peace*, *farc*, *economy* y *president*, reflejando temas estructurales del discurso económico y político del país. En cambio, en las Ilustraciones 7 y 8, que representan los resúmenes, se observa una concentración semántica más enfocada, con predominancia de términos de impacto directo como *market*, *tax*, *strike*, *inflation*, *policy*, lo que respalda la hipótesis de que los resúmenes resaltan la información central y reducen la dispersión temática. Palabras como *"company"*, *"deal"*, *"rating"*, *"rebel"*, *"conflict"* son más visibles, reflejando una condensación semántica que privilegia actores y eventos clave.

La lematización, aplicada en ambos casos, contribuyó a reducir la redundancia morfológica del vocabulario. Por ejemplo, las formas *talk*, *talked* y *talking* se consolidaron en un solo término raíz (*talk*), lo cual mejora la eficiencia del análisis estadístico posterior y facilita una interpretación más clara de los temas dominantes.

En conclusión, la construcción de estos diccionarios constituye un insumo esencial para el modelado lingüístico y la extracción de variables textuales. Al aplicar este procedimiento tanto al corpus original como a su versión resumida, se estableció una base coherente para comparar la riqueza léxica, la estructura semántica y el valor predictivo de cada enfoque. Las nubes de palabras generadas permiten validar visualmente la calidad del vocabulario extraído y ofrecen un primer acercamiento a los temas predominantes que podrían estar vinculados con la fluctuación del mercado colombiano.

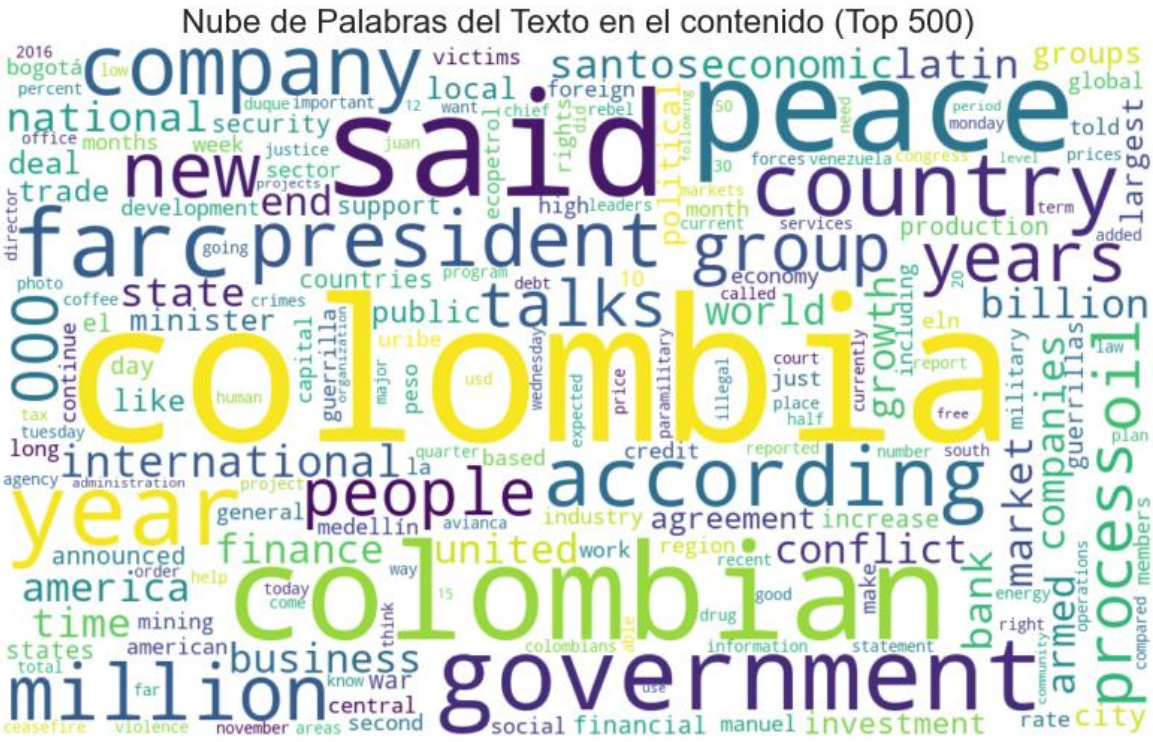

Ilustración 5 Grafico de nube de palabras del texto en el contenido.

Ilustración 6 Grafico de nube de palabras del texto en el contenido lematizado.

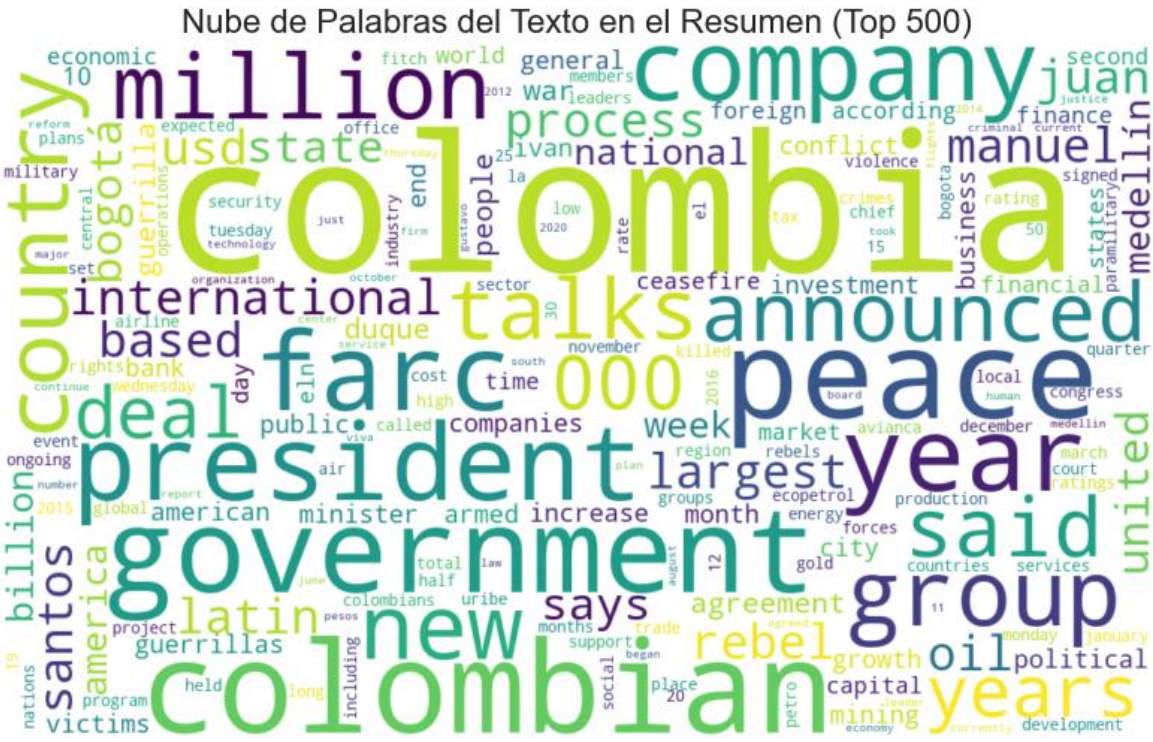

Ilustración 7 Grafico de nube de palabras del texto en el resumen.

Ilustración 8 Grafico de nube de palabras del texto en el resumen lematizado.

Aspecto de Evaluación	*Corpus Original*	*Corpus Lematizado*	*Resumen Original*	*Resumen Lematizado*
Redundancia morfológica	Alta	Baja	Media	Mínima
Palabras funcionales (ruido)	Alta	Media	Baja	Mínima
Presencia de actores clave	Alta	Alta	Muy alta	Muy alta
Representación de temas	Difusa	Enfocada	Clara	Muy clara
Densidad semántica por término	Baja	Media-alta	Alta	Muy alta

Tabla 4 Resumen de aspectos de evaluación para los datasets de corpus original, corpus lematizado, resumen original, y resumen lematizado.

8.5. Conteo de secuencias léxicas dentro de ventanas móviles

Con el objetivo de identificar patrones lingüísticos recurrentes más allá de palabras aisladas, se diseñó un procedimiento para calcular la frecuencia de aparición de **combinaciones de palabras** (o secuencias léxicas) dentro de **ventanas móviles de longitud variable**. Este tipo de análisis permite capturar estructuras composicionales del lenguaje que reflejan ideas, conceptos o narrativas más complejas, como *reforma tributaria*, *riesgo país* o *crecimiento económico*. Dichas expresiones compuestas resultan particularmente útiles para detectar patrones semánticos vinculados con eventos financieros o económicos específicos.

Este análisis fue aplicado de manera **simultánea e independiente** tanto al **corpus completo original**, como a los **resúmenes automáticos** generados mediante modelos de lenguaje. La finalidad fue evaluar si el resumen, al comprimir semánticamente el contenido, potencia la aparición de secuencias relevantes o, por el contrario, pierde información de contexto útil.

El enfoque técnico consiste en recorrer el texto ya procesado (normalizado y lematizado), y para cada ventana de tamaño definido (window_size), generar todas las posibles combinaciones de palabras:

- Cuando allow_gap = False, el análisis considera únicamente **secuencias contiguas**. Por ejemplo, con window_size = 2, se generan n-gramas como *reforma_tributaria*, *sector_bancario* o *presidente_santos*. Estas combinaciones reflejan co-localizaciones sintácticamente compactas.

- Si allow_gap = True, el algoritmo permite **secuencias no adyacentes**, respetando un rango de distancia (rank_window). Esto habilita la detección de asociaciones más flexibles, como *riesgo_inversión* dentro de la frase “alto riesgo de inversión extranjera”. Esta modalidad es particularmente útil para identificar relaciones implícitas en lenguaje más narrativo.
- Se implementó además una lógica para **evitar repeticiones dentro de la misma ventana**, y se parametrizó la opción de considerar el **orden de aparición** como significativo o no. Esto permite capturar expresiones tanto orientadas a la sintaxis como a la semántica conceptual.

Para asegurar que las secuencias extraídas fueran representativas, se estableció un **umbral mínimo de frecuencia** de aparición de 10 veces a lo largo del corpus (o del conjunto de resúmenes). Solo las combinaciones que superaron este umbral fueron retenidas como secuencias relevantes, generando así un **diccionario de secuencias léxicas frecuentes**, acompañado de sus respectivas frecuencias absolutas.

Este procedimiento fue aplicado de forma paralela en los dos datasets. En el **corpus completo**, se obtuvo una alta diversidad de secuencias, muchas de las cuales resultaban contextualmente relevantes, pero también acompañadas de ruido semántico y combinaciones accidentales derivadas de la mayor longitud de los textos. En contraste, el **dataset de resúmenes** presentó una menor cantidad total de secuencias distintas, pero una **mayor concentración en combinaciones conceptualmente significativas**, como peace talk, peace deal, war crime, rebel group , y minimum wage; estos conjuntos de palabras reflejan el contexto noticioso de Colombia, como un mercado golpeado por conflictos e inestabilidad interna.

Este resultado sugiere que el preprocesamiento basado en resumen automático no solo reduce la longitud del texto, sino que **amplifica la frecuencia relativa de estructuras lingüísticas relevantes**, al eliminar fragmentos periféricos o irrelevantes para la información central. Así, el resumen no actúa únicamente como una reducción superficial del corpus, sino como un **filtro semántico que prioriza las unidades de significado clave**.

En conjunto, este tipo de análisis de secuencias léxicas contribuye a construir vocabularios compuestos, útiles para análisis posteriores de tópicos, sentimientos o identificación de eventos. Además, al comparar la densidad, variedad y significado de las secuencias extraídas del corpus y del resumen, se obtiene evidencia sobre la capacidad del resumen automático para **concentrar información predictiva útil en un espacio semántico reducido**.

Esta representación más rica y contextualizada del lenguaje, especialmente en su forma resumida, tiene un alto potencial para alimentar modelos de predicción y clasificación, mejorando tanto la interpretabilidad como la eficiencia de los algoritmos de PLN aplicados a noticias económicas.

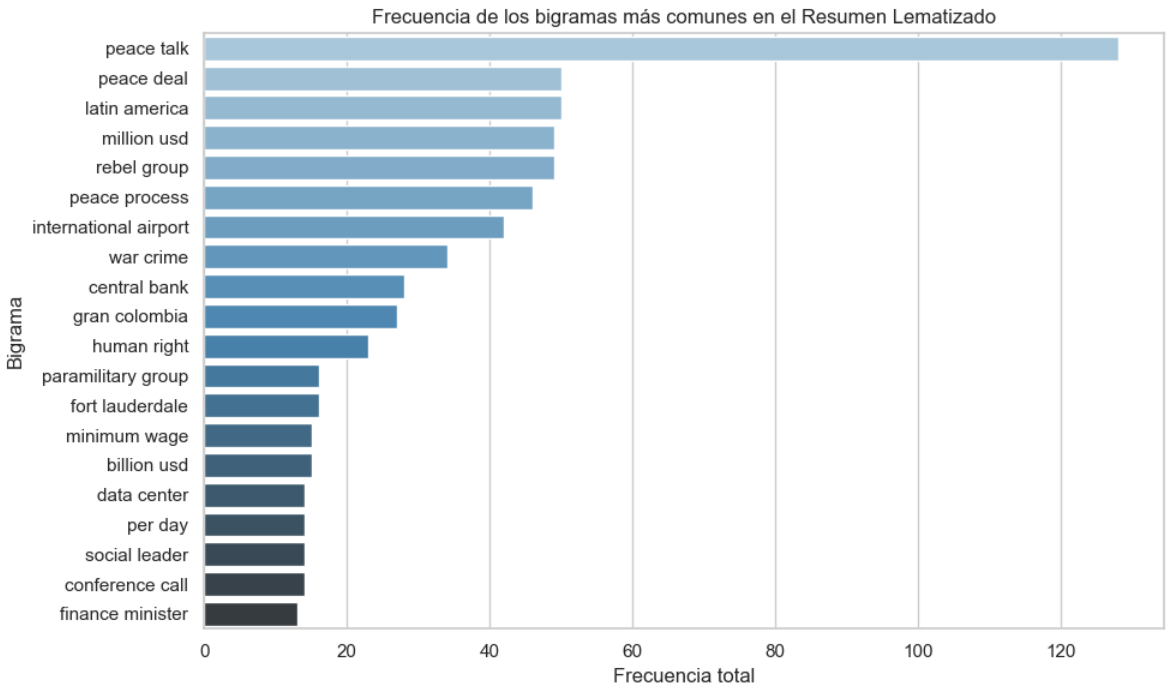

Ilustración 9 Grafico de Barras para la frecuencia de los bigramas mas comunes en el resumen lematizado.

La Ilustración 9 se muestran los **20 bigramas más frecuentes extraídos del conjunto de datos lematizado generado a partir de los resúmenes automáticos de noticias**. Entre las combinaciones más comunes se destacan expresiones directamente asociadas con temas de alta relevancia económica y política en el contexto colombiano, como *peace talk*, *peace deal*, *rebel group* y *peace process*, lo cual sugiere que el conflicto armado y los procesos de negociación continúan siendo ejes discursivos centrales en la cobertura noticiosa. También aparecen bigramas vinculados al entorno económico y financiero, como *million usd*, *central bank*, *billion usd*, *minimum wage* y *finance minister*, lo que confirma la presencia de señales explícitas relacionadas con variables macroeconómicas y decisiones de política fiscal o monetaria. La inclusión de combinaciones como *gran colombia*, *international airport* y *data center* denota la coexistencia de temas regionales y corporativos dentro del resumen noticioso. En conjunto, esta distribución evidencia que el resumen automático no solo conserva, sino que **amplifica las unidades de significado más relevantes** para el análisis económico, resaltando aquellas que tienen alta probabilidad de estar relacionadas con eventos de impacto en el mercado financiero.

8.6. Clasificación temática de vocabulario económico mediante embeddings semánticos

Con el fin de identificar los tópicos dominantes dentro del lenguaje utilizado en las noticias económicas analizadas, se implementó un procedimiento basado en modelos de lenguaje distribuido (*word embeddings*), utilizando como referencia el modelo pre-entrenado GloVe (*Global Vectors for Word Representation*), entrenado sobre el corpus de Wikipedia y Gigaword.

El enfoque parte de la hipótesis de que los vectores semánticos de palabras relacionadas conceptualmente tienden a ubicarse en regiones cercanas dentro del espacio vectorial. Aprovechando esta propiedad, se definieron manualmente 19 categorías temáticas macroeconómicas y financieras detalladas en la Tabla 5, tales como *política monetaria*, *sector bancario*, *riesgo geopolítico*, *cadenas de suministro*, entre otras. Cada categoría fue representada por una lista de palabras clave (semillas) específicas del dominio.

Categoría temática	*Palabras clave (semillas)*
macroeconomic_indicators	*inflation, gdp, cpi, ppi, deflation, recession*
monetary_policy	*interest rate, fed, tightening, qe, quantitative easing, rate hike*
fiscal_policy	*stimulus, tax cut, budget deficit, government spending, debt ceiling*
earnings_reports	*earnings, profit, revenue, guidance, eps, forecast*
mergers_acquisitions	*acquisition, merger, takeover, buyout, deal, joint venture*
market_sentiment	*bullish, bearish, sentiment, volatility, greed, fear*
regulatory_action	*regulation, sec, fine, lawsuit, antitrust, compliance*
geopolitical_risk	*war, conflict, sanctions, embargo, diplomacy, instability*
commodity_prices	*oil, gold, crude, wheat, energy, commodity*
central_banks	*federal reserve, ecb, boj, interest, minutes, policy*
labor_market	*employment, unemployment, jobs report, wage growth, payroll*
consumer_demand	*sales, consumer, spending, retail, demand, holiday season*
technology_news	*ai, chip, innovation, software, product launch, data breach*
supply_chain	*shortage, logistics, disruption, port, shipping, delay*
currency_fluctuations	*usd, euro, exchange rate, forex, devaluation, currency war*
banking_sector	bank, credit, liquidity, default, deposit, run on the bank

crypto_market	bitcoin, crypto, ethereum, blockchain, token, stablecoin
ratings_outlook	downgrade, upgrade, rating, moody's, s&p, outlook
environmental_events	hurricane, climate, flood, natural disaster, wildfire
insider_activity	insider buy, insider sell, stock options, executive trade, form 4

Tabla 5 resumen de las categorías semánticas evaluadas y sus palabras semilla.

Para cada categoría, se construyó un vector promedio de sus palabras semilla, lo que dio lugar a un centro semántico representativo del tema. Posteriormente, para cada palabra del vocabulario extraído de las noticias, se calculó su similaridad coseno con cada una de estas categorías. La palabra fue asignada a la categoría más cercana según dicha métrica.

Finalmente, se contabilizó el número de palabras asignadas a cada categoría, y se determinó la categoría temática predominante en el vocabulario de interés. Este procedimiento permitió realizar un mapeo temático del lenguaje utilizado en las noticias, evidenciando tendencias de contenido como predominancia de términos relacionados con política fiscal, volatilidad de mercado o fluctuaciones cambiarias, según el contexto.

Este método ofrece una aproximación flexible y semánticamente robusta para el análisis de tópicos, sin necesidad de entrenar clasificadores supervisados ni de aplicar modelos probabilísticos como LDA. Su base matemática está respaldada por técnicas de representación distribuida del significado léxico y por el uso de similaridad vectorial como criterio de agrupamiento.

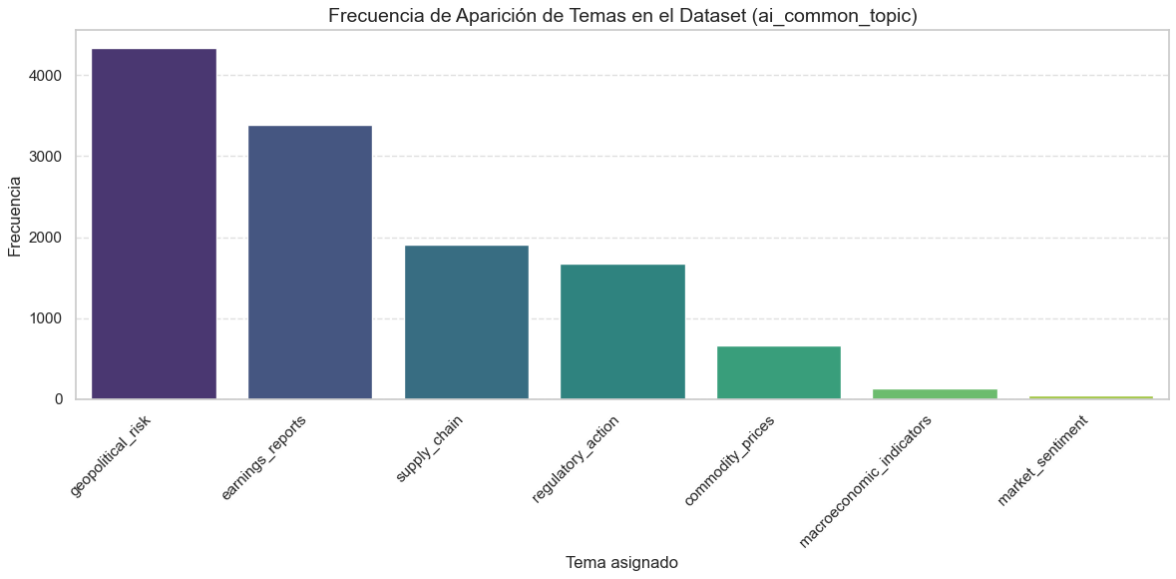

Ilustración 10 Grafico de Barras para la frecuencia de aparición de temas en el dataset.

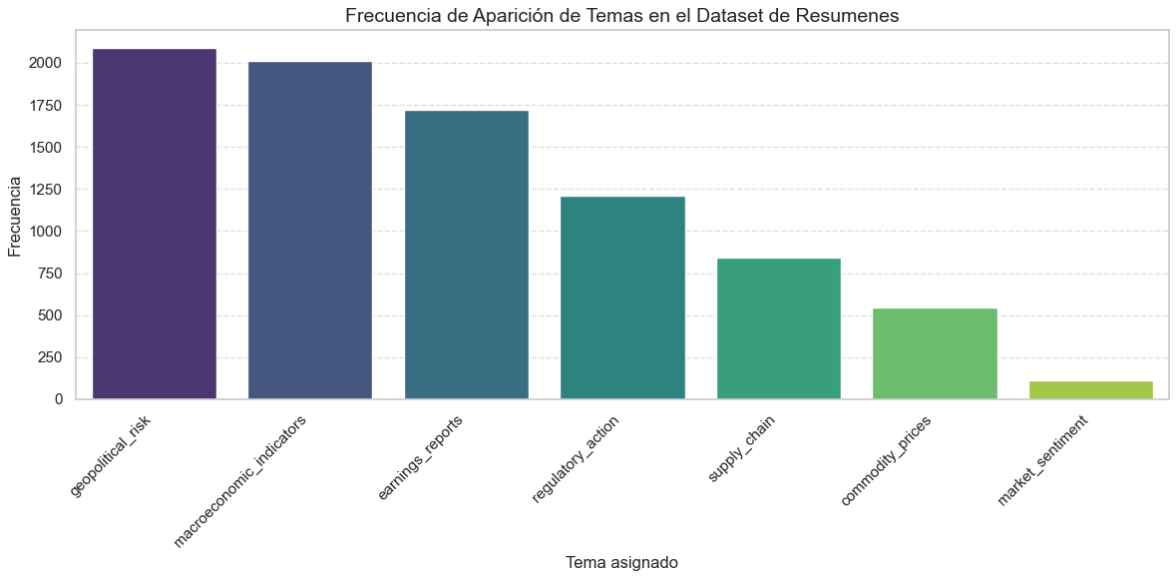

Ilustración 11 Grafico de Barras para la frecuencia de aparición de temas en el dataset de resúmenes automáticos.

Las Ilustraciones 10 y 11 se presenta una comparación directa entre la frecuencia de aparición de temas asignados en dos versiones del dataset: **el corpus completo de noticias** (Ilustración 10) y los **resúmenes automáticos generados a partir del mismo contenido** (Ilustración 11). Ambas visualizaciones reflejan la cantidad total de veces que cada noticia fue asignada a una de las categorías temáticas predefinidas, utilizando un modelo de embeddings semánticos y asignación por similitud coseno.

En ambas representaciones, el tema dominante es claramente **geopolitical_risk**, lo que sugiere una fuerte presencia de contenido vinculado a conflictos, inestabilidad regional y relaciones internacionales, tanto en las noticias originales como en sus resúmenes. Esta consistencia refuerza la importancia de este tópico como eje estructurante del discurso informativo en Colombia durante el periodo analizado.

Sin embargo, a partir del segundo y tercer puesto se evidencian diferencias significativas. En el **corpus completo**, el segundo tema más frecuente es **earnings_reports**, seguido por **supply_chain**. En cambio, en el **resumen automático**, la categoría **macroeconomic_indicators** toma el segundo lugar, desplazando a earnings_reports al tercer puesto. Este cambio sugiere que el proceso de resumen **resalta de forma más explícita variables macroeconómicas** como inflación, PIB o tasas de interés, probablemente porque estos términos suelen estar más concentrados en los fragmentos clave de las noticias.

Otra diferencia importante se encuentra en la categoría **market_sentiment**, que en ambos casos ocupa el último lugar, pero con una frecuencia mucho menor en el corpus completo que en los resúmenes. Esto indica que el resumen podría estar **filtrando contenido más interpretativo o emocional**, el cual queda diluido en el cuerpo completo del artículo.

La categoría **regulatory_action** aparece con similar relevancia en ambas versiones, lo cual sugiere que temas como legislación, regulación de mercado y decisiones de política pública son tratados de forma transversal en el contenido noticioso, y no se ven ni reforzados ni diluidos por el resumen.

En términos generales, el resumen automático tiende a **concentrar las apariciones en menos temas**, lo cual se refleja en la mayor altura relativa de los primeros tres tópicos. Este comportamiento es coherente con el objetivo del modelo de resumen: **reducir longitud textual mientras retiene lo más relevante**. Así, se logra una **mayor densidad temática por documento**, sacrificando parte de la diversidad que puede estar presente en el corpus completo.

Este análisis comparativo apoya la hipótesis de que el resumen automático no solo actúa como un proceso de compresión textual, sino como un **filtro semántico que potencia los tópicos de mayor relevancia informativa o económica**. Desde el punto de vista del modelado predictivo, esto puede traducirse en una mejor discriminación de clases si las categorías temáticas más frecuentes coinciden con señales latentes asociadas a movimientos de mercado.

En conclusión, aunque ambos enfoques capturan una estructura temática similar en términos generales, el resumen automático demuestra una mayor **capacidad de concentración semántica y priorización temática**, lo que lo convierte en un recurso valioso para tareas posteriores de análisis y predicción. Este hallazgo justifica su inclusión como segunda fuente textual en el diseño experimental de este estudio.

8.7. Análisis de sentimiento mediante procesamiento por lotes

El análisis de sentimiento fue aplicado sobre cada noticia del corpus con el propósito de cuantificar de forma objetiva el tono emocional de los textos. Para ello se empleó la biblioteca TextBlob, que permite obtener dos métricas principales:

- Polaridad: una medida continua en el rango [-1.0, 1.0], donde -1 representa un sentimiento negativo extremo y +1 un sentimiento positivo.
- Subjetividad: una medida entre 0 y 1 que indica cuán subjetivo (opinativo) o objetivo es el texto.

Dado que algunas noticias presentan una longitud considerable, se implementó un procedimiento de análisis por lotes (batching): el texto se divide en segmentos de hasta 10.000 palabras, que se procesan de forma independiente. Al finalizar, se calcula la media aritmética de la polaridad y la subjetividad obtenidas para cada lote, generando así un valor representativo para el contenido.

8.7.1. Comparación entre textos lematizado y no lematizado

Este análisis se realizó tanto sobre el corpus sin lematizar como sobre el corpus lematizado, con el objetivo de evaluar el impacto del preprocesamiento en el sentimiento capturado.

La hipótesis que motivó este análisis dual fue que la lematización, aunque beneficiosa para la normalización semántica y reducción de la dimensionalidad, podría ocasionar pérdida de información emocional, especialmente en adjetivos, verbos y expresiones coloquiales cuya forma original transmite una carga afectiva mayor que su lematización.

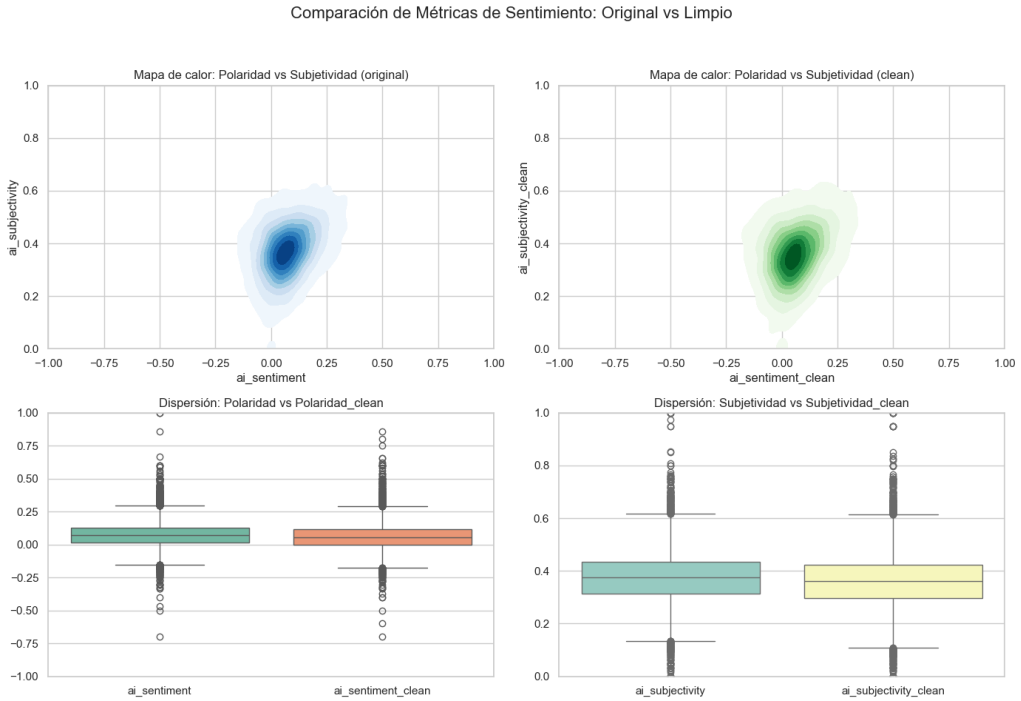

Ilustración 12 Grafico combinado de comparación de métricas de sentimiento original vs limpio, evaluando el espacio lineal entre la subjetividad y polaridad.

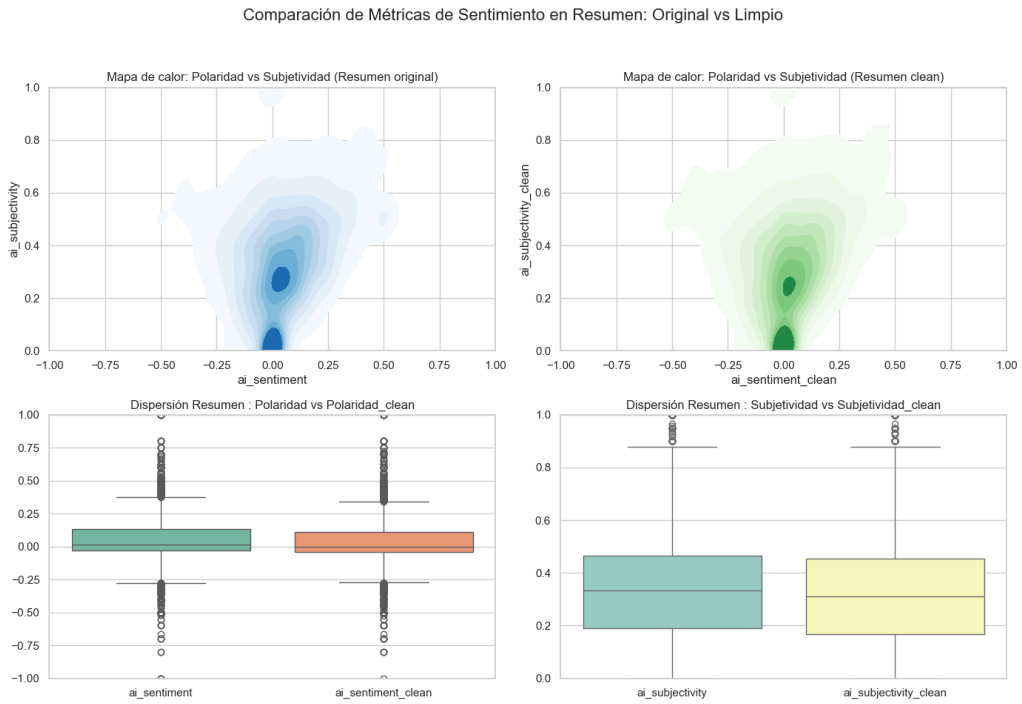

Ilustración 13 Grafico combinado de comparación de métricas de sentimiento para resumen automático original vs limpio, evaluando el espacio lineal entre la subjetividad y polaridad.

Conclusiones comparativas y proyección al modelado

1. **El resumen automático tiende a exhibir mayor dispersión emocional**, con polaridades y subjetividades más pronunciadas. Esto puede ser beneficioso para capturar reacciones del mercado sensibles al tono noticioso.
2. **El corpus completo es más neutro y conservador** en sus métricas de sentimiento, pero su volumen y ruido pueden dificultar la extracción de señales relevantes.
3. En ambos casos, la **lematización mejora la compactación de las métricas**, reduciendo outliers y dispersión, lo cual se espera que **mejore la estabilidad y generalización** de los modelos predictivos.
4. La versión lematizada del resumen podría ofrecer el **mejor balance entre expresividad y control léxico**, funcionando como una representación semántica óptima para tareas de clasificación.
5. Estas observaciones respaldan la decisión metodológica de evaluar múltiples representaciones textuales (corpus vs resumen, limpio vs original), ya que permiten **identificar la versión que ofrece mejor calidad de señal para predecir la fluctuación del mercado**.

8.8. Expansión de representaciones léxicas: selección y codificación de n-gramas frecuentes

Con el objetivo de incorporar las representaciones léxicas más informativas al modelo predictivo, se diseñó un procedimiento para seleccionar, filtrar y expandir columnas que contienen diccionarios de términos (por ejemplo, n-gramas o tópicos extraídos de noticias). Este procedimiento consta de dos fases principales:

8.8.1. Codificación expandida del diccionario

Una vez definidos los términos extraídos de los textos lematizados, se transformó cada fila en una representación vectorial explícita, en la cual cada uno de los términos seleccionados ocupa una columna distinta. El valor asociado a cada término corresponde a su frecuencia original (presente en el diccionario del documento), o cero si no aparece.

Este proceso de expansión controlada cumples múltiples objetivos:

1. Convierte estructuras no tabulares (diccionarios) en formatos compatibles con modelos supervisados.
2. Facilita la interpretabilidad del modelo, al mantener los nombres de los términos como identificadores de columnas.
3. Permite reducir el impacto del sesgo por términos raros o idiosincráticos, al aplicar un umbral de aparición mínimo mediante la selección de términos frecuentes.

El resultado es un conjunto de variables numéricas escasas (*sparse features*), etiquetadas como nombre_columna_término, las cuales pueden integrarse directamente en matrices de entrenamiento, visualizaciones exploratorias o procesos de selección de variables.

Este tipo de codificación es fundamental en tareas de modelado basadas en texto, especialmente cuando se desea preservar la semántica explícita de términos sin recurrir a representaciones densas como embeddings.

8.9. Agregación temporal de variables por fecha

Dado que el objetivo del estudio es analizar la influencia de las noticias sobre la dirección diaria del índice bursátil COLCAP, fue necesario transformar el conjunto de datos en una estructura agregada por fecha, en la que cada fila represente un día específico del calendario y resuma todas las observaciones (noticias) correspondientes a esa jornada.

Para ello se implementó una función de agregación personalizada que permite:

- Sumar valores numéricos acumulables, como conteos de aparición de términos o intensidades de sentimiento.
- Promediar variables continuas, como polaridad o subjetividad media de las noticias.
- Contar el número total de noticias publicadas en cada fecha (variable record_count).
- Construir diccionarios de frecuencia por valor categórico, por ejemplo, contabilizando cuántas veces aparece cada etiqueta temática (o palabra clave) por día.

Este procedimiento es altamente flexible, permitiendo excluir columnas irrelevantes antes del agrupamiento (exclude_columns), definir de forma explícita qué variables se deben sumar o promediar (sum_columns, avg_columns), y aplicar conteo de frecuencia sobre columnas categóricas (categorical_count_columns) que contengan temas, autores o etiquetas predeterminadas.

La salida del proceso es un Dataframe a nivel diario, en el cual se sintetiza toda la información textual y cuantitativa disponible en las noticias de esa fecha. Esta estructura sirve como insumo directo para:

- Generar visualizaciones exploratorias a lo largo del tiempo.
- Construir modelos supervisados alineados con las series del índice COLCAP.
- Realizar análisis de correlación entre lenguaje mediático y comportamiento de mercado.

Este enfoque de colapsamiento por fecha constituye una etapa clave en la integración de múltiples fuentes de datos heterogéneas (texto, sentimiento, tópicos, frecuencia), permitiendo su alineación temporal con variables financieras cuantificables.

8.10. Filtrado de variables con baja frecuencia informativa

Previo a la etapa de modelado, se aplicó un procedimiento de filtrado para eliminar variables con escaso poder explicativo. En particular, se identificaron y descartaron aquellas columnas derivadas del análisis léxico cuya suma total a lo largo del periodo analizado fuese menor a un umbral mínimo.

Este umbral (por defecto, 10) se seleccionó empíricamente para conservar únicamente los términos que aparecieron con suficiente frecuencia en el corpus de noticias. Las variables eliminadas representan palabras o combinaciones de palabras cuya ocurrencia fue marginal, lo que sugiere una contribución poco significativa al aprendizaje del modelo y una posible introducción de ruido o sobreajuste si se mantuvieran.

Esta técnica se alinea con prácticas estándar en procesamiento de texto, donde se busca balancear la riqueza semántica del vocabulario con la necesidad de reducir la dimensionalidad y mejorar la estabilidad de los modelos predictivos. En esencia, actúa como un filtro de baja varianza para variables léxicas.

La implementación garantiza que el filtrado se aplique de forma dinámica sobre las columnas generadas por la expansión de diccionarios (ai_word_*), lo que facilita la integración directa con la matriz de características construida en fases anteriores del pipeline de procesamiento.

8.11. Conformación del conjunto de entrenamiento y definición del objetivo

Una vez completadas las etapas de limpieza, normalización, agregación y enriquecimiento de variables, se procedió a construir el conjunto final de entrenamiento para los modelos de predicción.

8.11.1. Variable objetivo (y)

Se definió como variable dependiente (y) la columna fluctuación (*fluctuation*), que representa la dirección diaria del índice COLCAP en una escala ordinal o categórica (por ejemplo, subida, bajada o sin cambio). En esta versión del análisis, se consideró la predicción contemporánea, es decir, basada en los datos del mismo día.

8.11.2. Selección de variables predictoras (X)

Se seleccionaron todas las variables relevantes construidas durante el pipeline de ingeniería de características, excluyendo aquellas que no aportaban valor predictivo directo o cuya inclusión implicaría fuga de información:

Se eliminaron variables como la fecha original (date), y variables redundantes como day_of_year (ya representada en formato sinusoidal).

Las variables resultantes incluyen:

- Estadísticas móviles y rezagos (*lags*) de *fluctuación* y precio de inicio (*begin*).

- Variables estacionales derivadas del calendario.
- Variables normalizadas provenientes del análisis de texto (*ai_word_*).*
- Agregados ponderados como fluctuation_weighted_avg_5.

8.11.3. Filtrado final y consistencia

Se eliminaron todas las filas donde el objetivo (y) era nulo, lo cual típicamente ocurre en las últimas observaciones cuando se utilizan lags (rezagos) o desplazamientos. Con esto se garantizó la consistencia temporal entre las variables predictoras y la etiqueta correspondiente.

El resultado final es una matriz de entrenamiento (X) y una serie objetivo (y) completamente alineadas, estructuradas y listas para ser usadas en la etapa de modelado supervisado. Esta estructura permitió evaluar diferentes algoritmos de clasificación, manteniendo un contexto económico-temporal robusto y evitando problemas comunes como el leakage de información.

9. Resultados

Esta sección presenta los resultados experimentales del proyecto, centrados exclusivamente en los modelos y análisis implementados en el entorno de trabajo. Se abordan dos líneas principales: el análisis descriptivo del lenguaje económico y el rendimiento predictivo de los modelos evaluados sobre el comportamiento del índice COLCAP.

9.1. Análisis del lenguaje económico y sentimiento

Se realizó un análisis exploratorio del contenido textual del corpus y resumen de noticias, aplicando técnicas como conteo de n-gramas, análisis de sentimiento y comparaciones temáticas entre días positivos y negativos del COLCAP.

9.1.1. Distribución de sentimientos

El análisis de sentimiento de las noticias financieras se llevó a cabo utilizando un modelo basado en **TextBlob**, que permite calcular dos métricas fundamentales: **polaridad** (rango de -1 a 1, indicando negatividad o positividad del texto) y **subjetividad** (rango de 0 a 1, indicando el nivel de juicio/opinión frente a afirmaciones objetivas). Este análisis se aplicó tanto al contenido noticioso completo como a los resúmenes automáticos, permitiendo evaluar si existen diferencias relevantes en la distribución del sentimiento según el tipo de representación textual.

9.1.1.1. Sentimiento en el texto completo

En el caso del **corpus original**, en la Ilustración 14 se evidencio que la polaridad media diaria osciló entre -0.35 y +0.41, con una fuerte concentración en torno a cero, lo que sugiere un tono predominantemente **neutral**. La subjetividad se mantuvo en valores medios, lo que refleja el carácter semi objetivo del lenguaje periodístico económico.

Cuando se analiza la distribución de polaridad y subjetividad segmentada por la dirección del mercado (variable fluctuación, codificada como -1 para caída y 1 para subida), se observa que **las medianas de ambas métricas son muy similares entre los dos grupos**. Esto indica que, en promedio, no hay grandes diferencias en el tono emocional de las noticias según la dirección del mercado.

Sin embargo, los días de **fluctuación positiva tienden a mostrar mayor dispersión en polaridad en la Ilustración 14**, lo que puede indicar una presencia más marcada de lenguaje emocional en esos casos. Igualmente, la subjetividad muestra una ligera elevación bajo condiciones de mercado alcista, lo cual sugiere que las noticias asociadas a subidas del índice COLCAP podrían contener **expresiones más matizadas o valorativas**.

En el análisis temporal en la Ilustración 15 (gráfico de promedios mensuales y trimestrales de polaridad y subjetividad), se observa que estas métricas mantienen **una evolución relativamente estable a lo largo del tiempo**, con picos ocasionales que coinciden con eventos económicos críticos. No obstante, la **curva de fluctuación del mercado (línea roja)** presenta una volatilidad significativamente mayor. Aunque en ciertos momentos (por ejemplo, durante 2020 y 2022) se identifican coincidencias entre caídas del mercado y picos negativos de polaridad, estas correlaciones no son consistentes ni suficientes para derivar una relación causal.

Particularmente, la polaridad (indicador del tono emocional del lenguaje) muestra una alta volatilidad en ciertos periodos. Se observa *que,* durante momentos de recuperación o estabilidad del índice, como en los años 2017 o en la segunda mitad de 2021, el lenguaje noticioso adoptó un tono relativamente más negativo o conservador, con valores de polaridad bajos o decrecientes. Esto puede resultar contraintuitivo, ya que se esperaría un aumento en el optimismo informativo en consonancia con el desempeño del mercado.

Por otro lado, picos de polaridad positiva, por ejemplo, durante el primer trimestre de 2020 y a mediados de 2022*,* coinciden con altas fluctuaciones del mercado, posiblemente asociadas a eventos disruptivos como la llegada del COVID-19 (marzo 2020), donde los mercados mostraron fuertes caídas y posterior recuperación, o con la transición presidencial en Colombia en 2022, que generó reacciones mixtas ante la llegada de un gobierno de orientación política distinta a la tradicional.

La subjetividad muestra también un comportamiento oscilante, pero tiende a decrecer levemente a partir de 2023. Esto podría interpretarse como una mayor objetividad percibida en las narrativas financieras, aunque dicha tendencia coincide con una recuperación relativa del COLCAP en el mismo periodo. No obstante, la divergencia entre la estabilidad del mercado y la menor subjetividad puede estar influenciada por un cambio en el tono editorial tras los primeros meses del nuevo gobierno, o por la priorización de temas estructurales sobre coyunturales en la prensa económica.

En conjunto, estos hallazgos sugieren que el sentimiento noticioso en Colombia presenta una correlación parcial con la fluctuación del índice bursátil, pero también parece estar influenciado por factores externos como el clima político, percepciones ideológicas y la línea editorial de los medios. Esto pone en evidencia el carácter potencialmente sesgado del lenguaje económico, lo que refuerza la necesidad de aplicar técnicas robustas de preprocesamiento y análisis contextual para separar señales reales del ruido mediático.

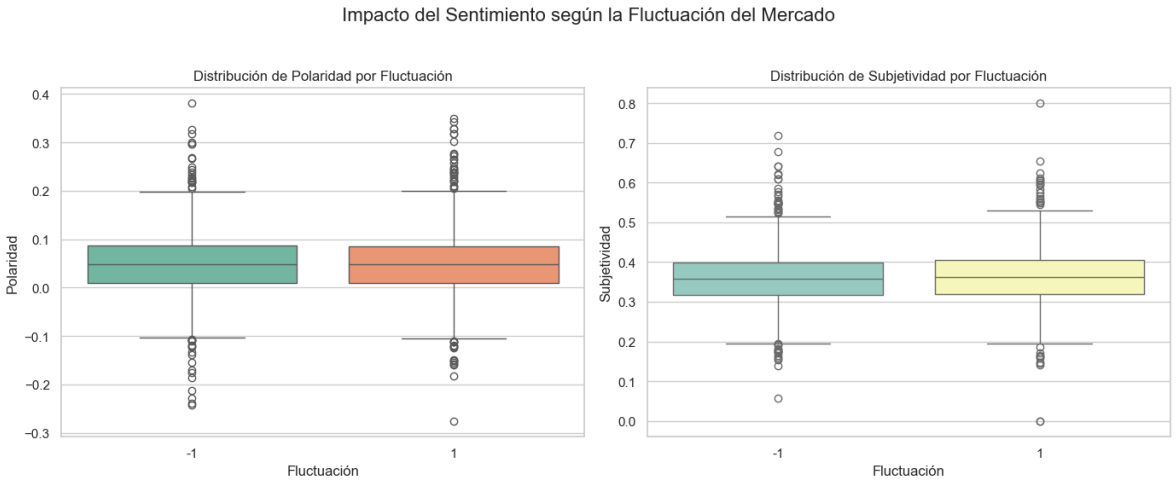

Ilustración 14 grafico de cajas para el impacto del sentimiento según la fluctuación del mercado.

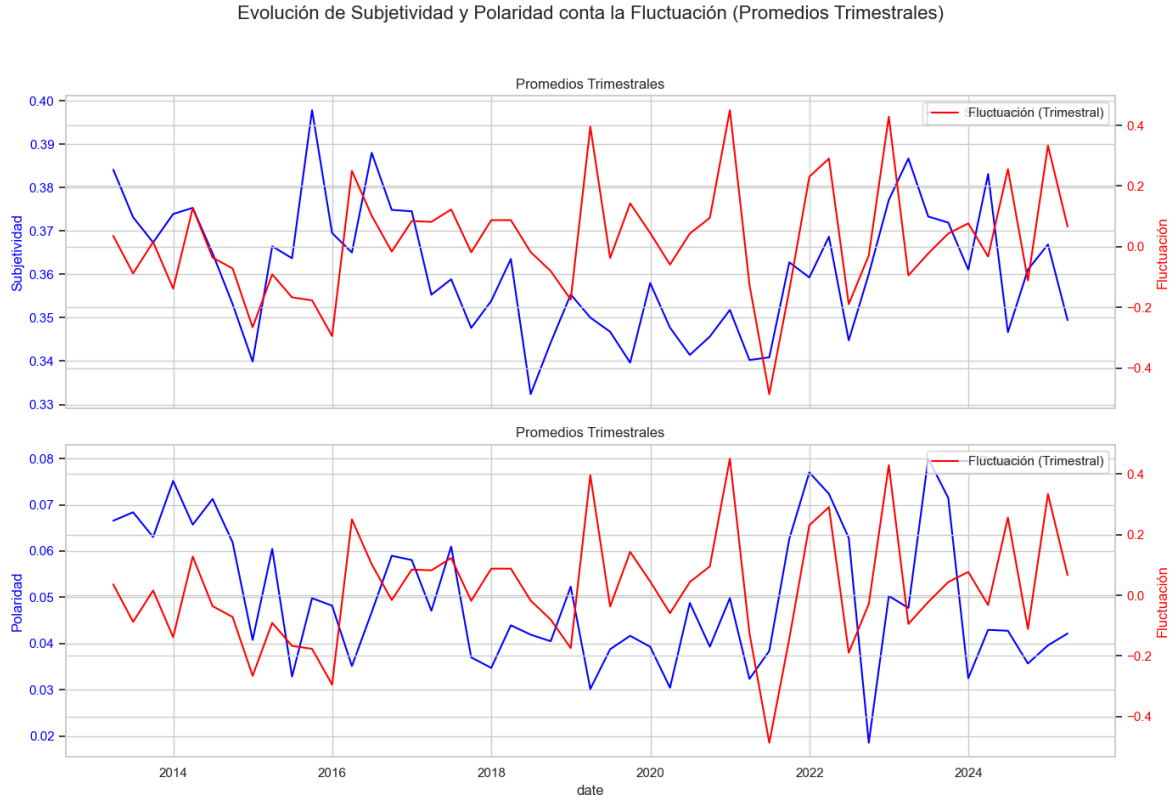

Ilustración 15 grafico de evolución de subjetividad y polaridad contra la fluctuación (promedios trimestrales.

9.1.1.2. Sentimiento en los resúmenes automáticos

El análisis de sentimiento aplicado a los resúmenes generados automáticamente mediante modelos NLP muestra patrones emocionales que, si bien se mantienen generalmente cercanos a la neutralidad, revelan **ciertas fluctuaciones estructurales a lo largo del tiempo** y **ligeras diferencias según el comportamiento del mercado**.

Las métricas emocionales agrupadas según la dirección del mercado (fluctuación = -1 para caída, 1 para subida). Los diagramas de caja de la Ilustración 16 nos revelan diferencias sutiles pero relevantes:

- **Polaridad**: Aunque la media y mediana se mantienen próximas a cero en ambos grupos, los días con **fluctuación positiva muestran una leve expansión hacia valores positivos extremos**, lo que sugiere mayor presencia de lenguaje favorable o valoraciones optimistas en las noticias que coinciden con alzas del mercado. Esta mayor dispersión puede amplificar la señal de eventos positivos en los modelos.
- **Subjetividad**: Las distribuciones en ambos grupos son bastante similares, con una leve tendencia a una mayor subjetividad durante días de subida. Esto podría estar asociado a un **incremento en la carga interpretativa o analítica** del lenguaje en contextos alcistas, cuando los medios tienden a explicar o justificar los movimientos del mercado con mayor énfasis narrativo.

Como se observa en la **Ilustración 17**, la subjetividad y polaridad presentan una evolución con **variabilidad contenida**, en contraste con la mayor volatilidad de la fluctuación del mercado (línea roja). En la parte superior, la subjetividad trimestral se mantiene en un rango estrecho entre 0.28 y 0.36, con ligeras caídas entre 2015 y 2021 y una recuperación leve posterior. No obstante, en años como 2020 y 2022 se aprecian **picos de volatilidad en la fluctuación del mercado** que coinciden con reducciones de subjetividad, lo que podría indicar un lenguaje más técnico o cauteloso durante periodos de incertidumbre económica, asociados a la llegada del COVID-19 y elecciones presidenciales respectivamente.

En la parte inferior, la polaridad trimestral se mueve alrededor de valores bajos (entre -0.04 y 0.06), sin una tendencia clara al alza o baja. Sin embargo, se identifican momentos en los que **la curva azul de polaridad refleja caídas notables cercanas a eventos de alta presión económica**, como la pandemia o la inflación post-COVID. Aunque estas coincidencias no establecen causalidad, sugieren que **el sentimiento textual agregado en los resúmenes podría reflejar el clima económico con cierto retraso o suavizado semántico**.

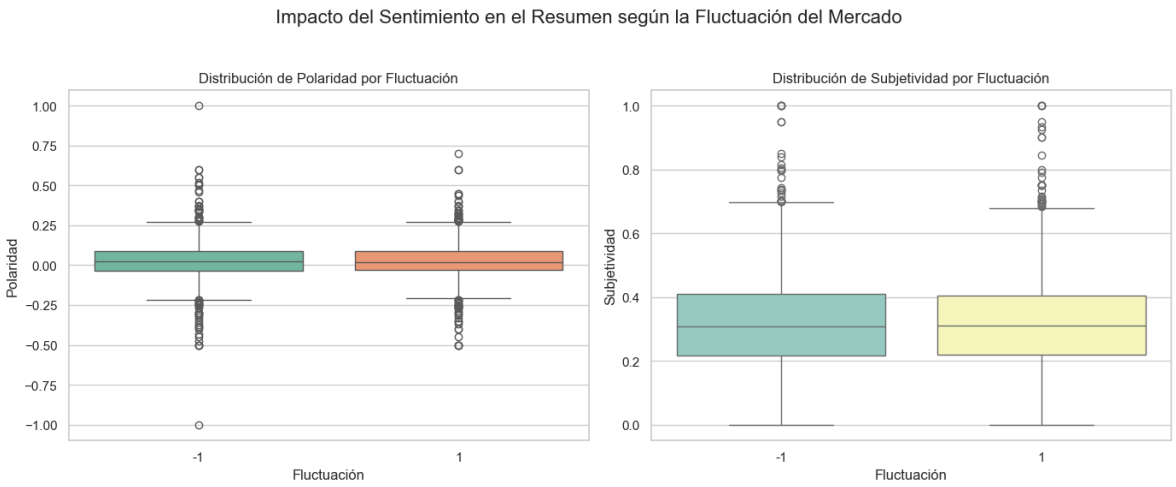

Ilustración 16 Grafico de cajas para el impacto del sentimiento en el resumen según la fluctuación del mercado.

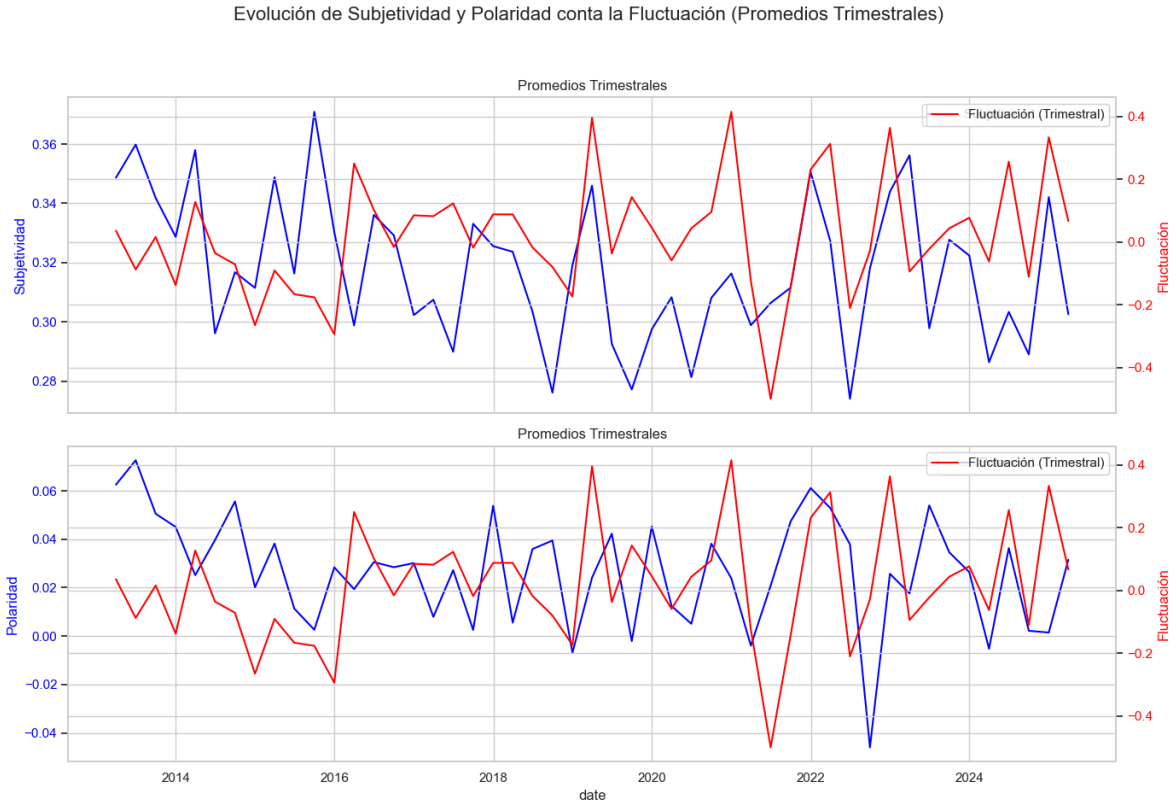

Ilustración 17 Grafico temporal para la evolución de la subjetividad y polaridad contra la fluctuación.

A diferencia del corpus completo, los resúmenes automáticos muestran una **mayor sensibilidad a la dinámica del mercado**, evidenciado en la Ilustración 17, especialmente en términos de polaridad. Esto es coherente con el hecho de que los modelos de resumen tienden a concentrar el contenido informativo más relevante,

lo que puede acentuar la aparición de señales semánticas ligadas a emociones positivas o negativas.

En términos de modelado predictivo, estas observaciones sugieren que el sentimiento extraído de resúmenes automáticos podría funcionar como un **indicador complementario útil**, siempre que se combine con otras variables (tópicos, frecuencia, entidades). Si bien la subjetividad y polaridad por sí solas no permiten predecir de manera determinística la dirección del mercado, su estabilidad, dispersión controlada y capacidad de reflejar cambios macroeconómicos críticos hacen de estas métricas una **fuente rica en objetividad de la información**.

9.1.2. Frecuencia de términos y diferencias por fluctuación:

Además del análisis de métricas emocionales como la polaridad y la subjetividad, resulta fundamental explorar la **frecuencia de aparición de términos específicos** en el contenido noticioso, y cómo esta frecuencia puede estar relacionada con la **dirección del mercado bursátil**. En esta sección se analizan los **términos léxicos más frecuentes**, tanto en su forma individual como en combinaciones (bigramas), y se examina cómo varía su presencia en días de **fluctuación positiva** frente a días de **fluctuación negativa** del índice COLCAP.

El objetivo es identificar si ciertos conceptos, eventos, entidades o narrativas aparecen **con mayor frecuencia en contextos alcistas o bajistas**, y si su comportamiento temporal ofrece indicios útiles para comprender las señales del mercado. Para ello, se presentan gráficos comparativos que contrastan la frecuencia de los términos entre los dos tipos de días, así como **series temporales** que permiten visualizar su evolución agregada a lo largo de los años.

Este análisis no solo enriquece la interpretación semántica del corpus, sino que también ofrece una base empírica para evaluar el **potencial predictivo de patrones lingüísticos recurrentes**, que pueden actuar como marcadores temáticos o contextuales asociados a cambios en el comportamiento del mercado financiero colombiano.

9.1.2.1. Analisis de aparicion de terminos en el corpus noticioso:

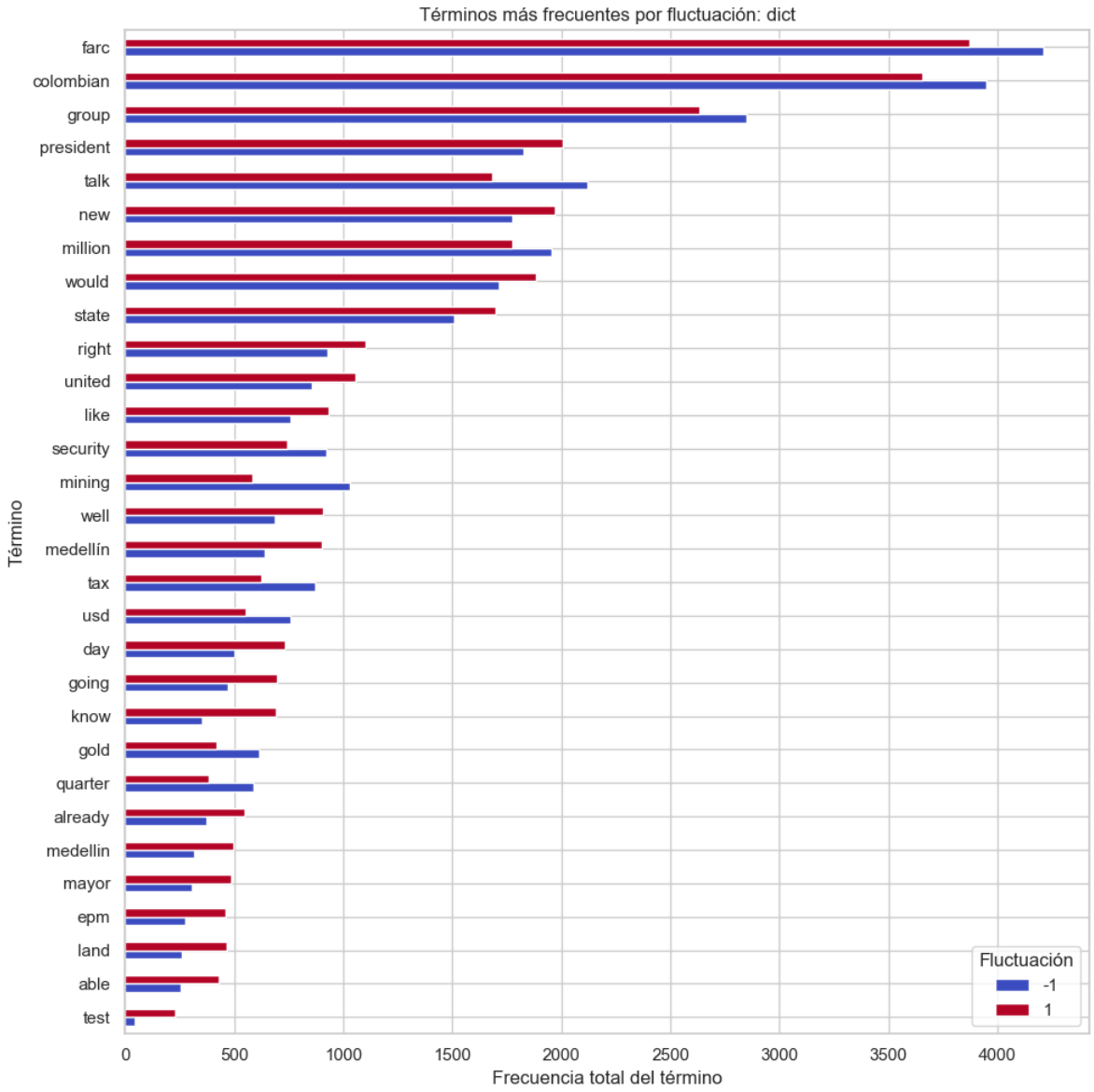

Ilustración 18 Grafico de barras combinadas para los términos más frecuentes por fluctuación.

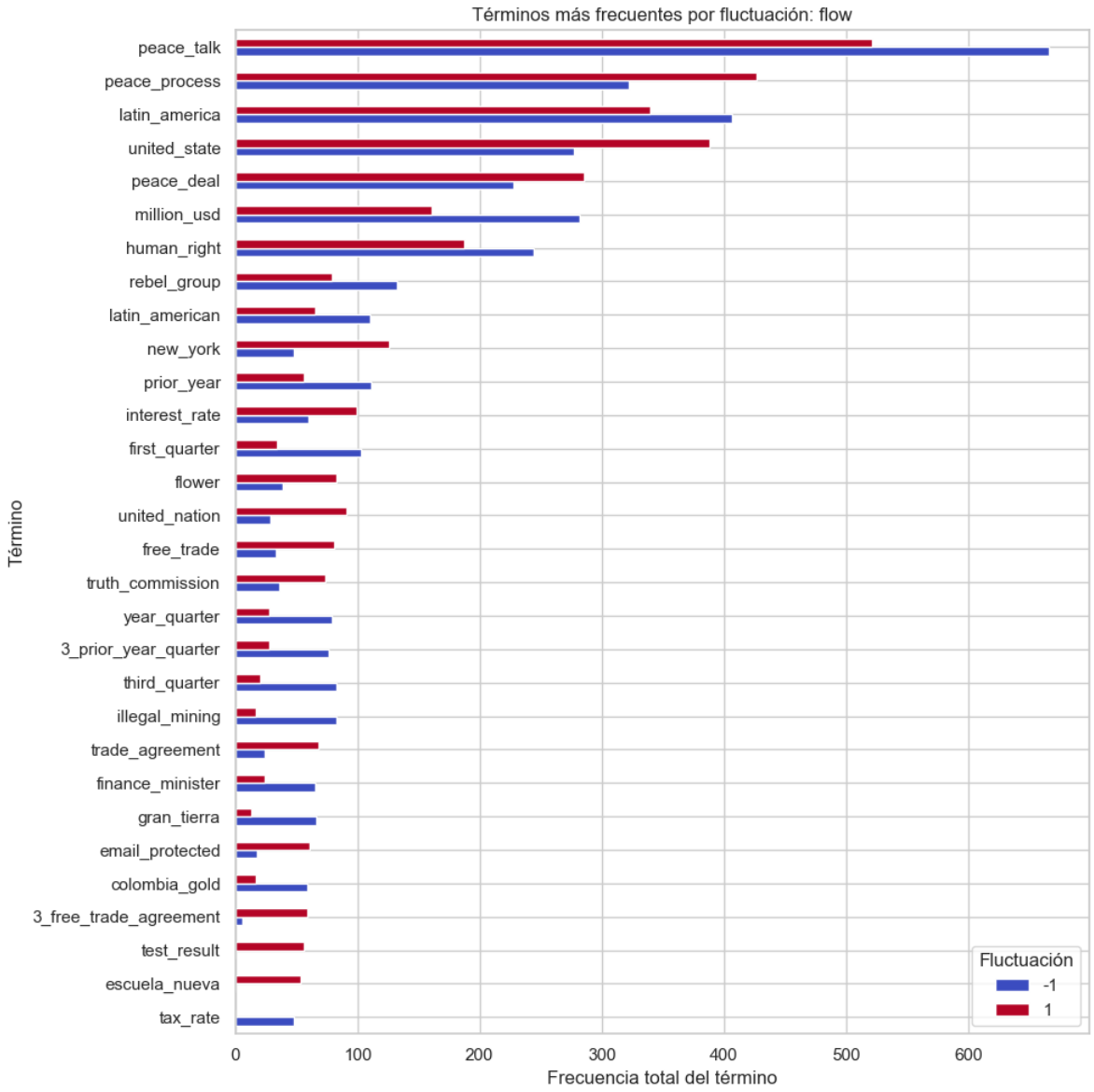

Ilustración 19 Grafico de barras combinadas para los términos más frecuentes por fluctuación, bigramas.

El análisis de los términos más frecuentes por tipo de día, considerando tanto palabras individuales (Ilustración 18) como expresiones compuestas (Ilustración 19), revela diferencias claras en el uso del lenguaje según la dirección del mercado.

En el gráfico correspondiente al grupo de palabras simples de la Ilustracion 18, se observan términos como **"farc"**, **"security"**, **"conflict"**, y **"group"**, los cuales aparecen con alta frecuencia durante días de **caída en el mercado** (fluctuation = -1). Esto sugiere una fuerte presencia de contenido asociado a temas de orden público, seguridad y política en los momentos donde el mercado experimenta retrocesos. Estos términos podrían estar relacionados con incertidumbre

sociopolítica, conflictos armados o crisis institucionales, aspectos históricamente vinculados a la percepción negativa del entorno económico.

Por otro lado, durante días de **subida del mercado** (fluctuation = 1), destacan términos como **"new"**, **"president"**, **"state"**, y **"epm"**, lo cual indica una orientación del contenido hacia temas financieros, empresariales o de crecimiento. Esta divergencia semántica sugiere que el lenguaje empleado en las noticias refleja, y posiblemente influencia, el estado emocional del mercado.

El gráfico del grupo de bigramas en la Ilustracion 19, que contempla n-gramas o frases compuestas, refuerza esta interpretación. En días negativos destacan expresiones como **"peace_talk "**, **"rebel_group"**, o **"human_right"**, mientras que en días positivos son más frecuentes expresiones como **"united_state"**, **"peace_deal"**, o **"peace_process"** Estas diferencias indican que el contexto temático de las noticias cambia no solo en contenido sino también en estructura lingüística cuando varía la dirección del mercado.

Los resultados evidencian que la fluctuación del mercado se asocia con diferencias significativas en la **frecuencia temática de los términos presentes en las noticias**. Los días de caída tienden a estar marcados por un léxico de conflicto, riesgo y tensión política, mientras que los días de subida presentan una narrativa más vinculada al crecimiento, la estabilidad institucional y la economía formal. Este hallazgo refuerza la utilidad del análisis de lenguaje como herramienta de diagnóstico contextual y como insumo potencial para modelos de predicción bursátil basados en noticias.

9.1.2.2. *Análisis temporal de aparición de términos en el corpus noticioso:*

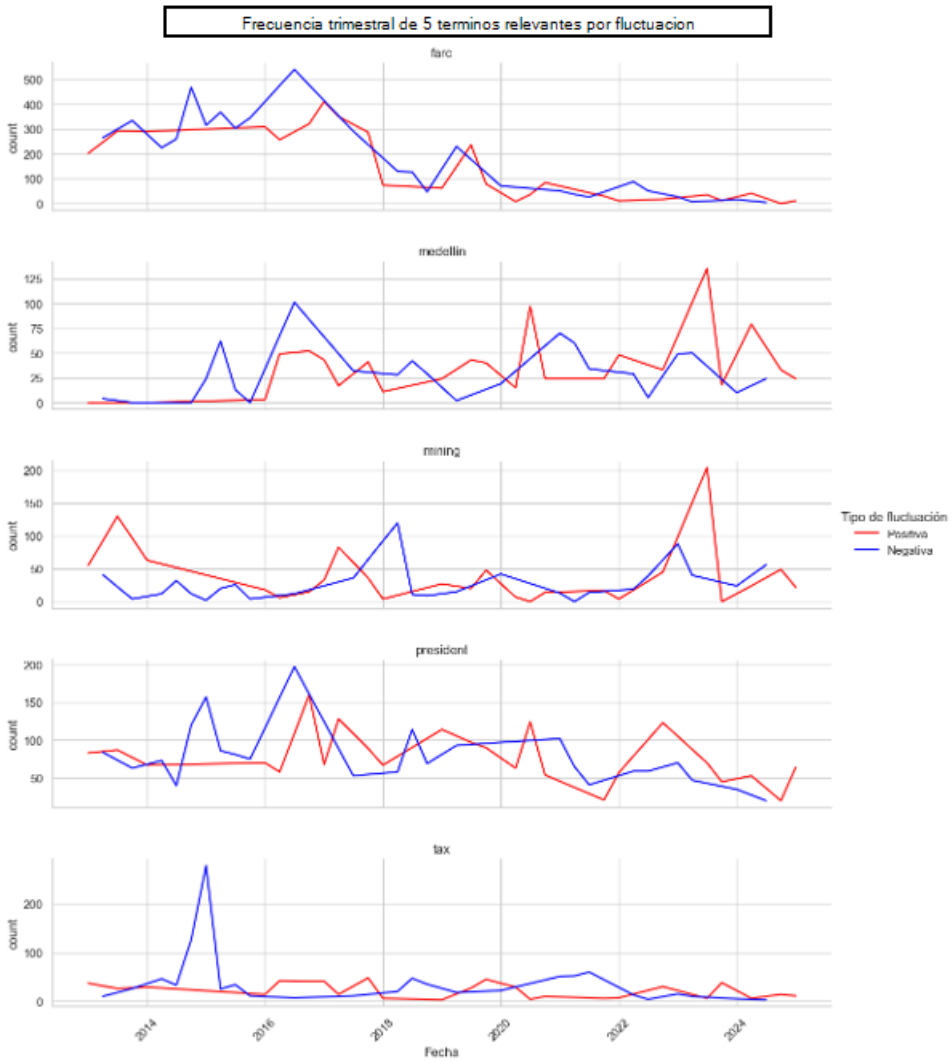

Ilustración 20 Grafico temporal para la frecuencia trimestral de 5 términos relevantes y tipo de fluctuación.

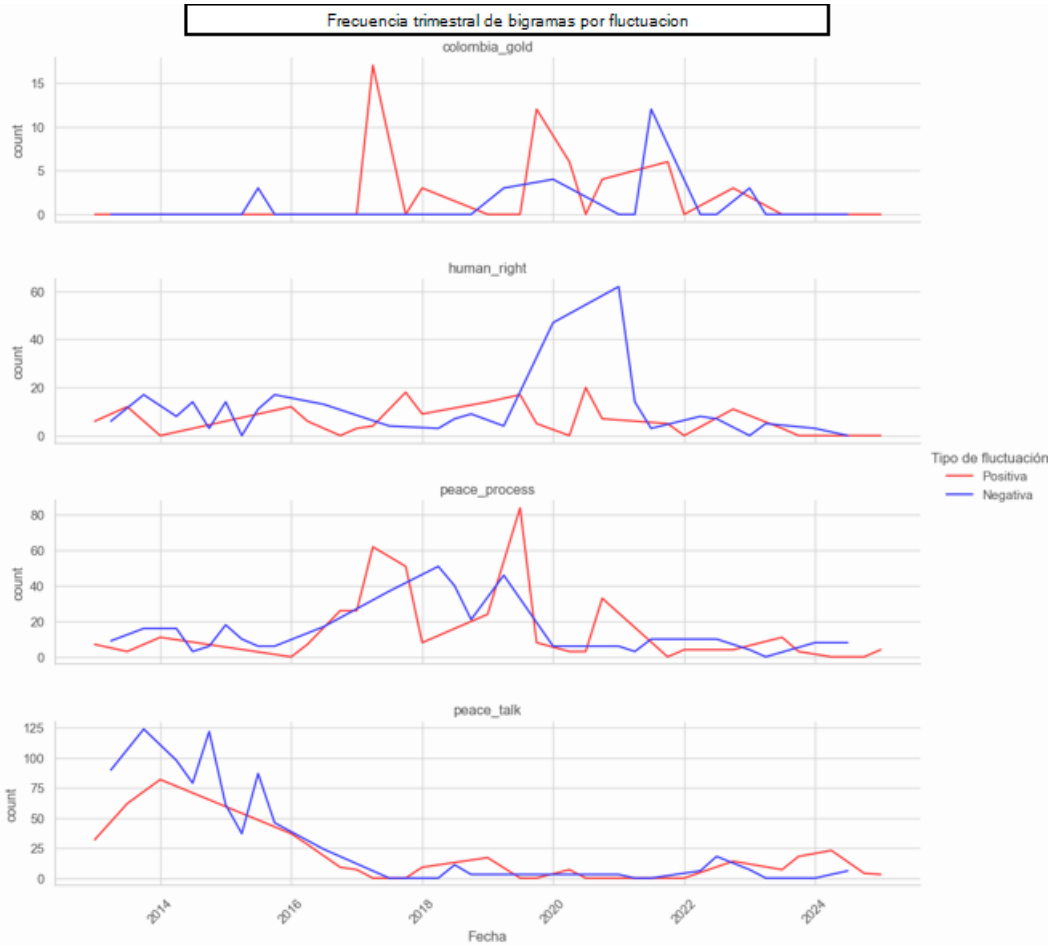

Ilustración 21 Grafico Temporal de frecuencia de bigramas por fluctuación..

Las Ilustraciones 20 y 21 muestran la evolución temporal de nueve términos altamente frecuentes en el corpus noticioso, segmentados según el tipo de fluctuación del mercado (**positiva** en rojo, **negativa** en azul). A partir del análisis de estos términos se pueden extraer las siguientes observaciones clave:

1. Temas de conflicto y paz (farc, peace_process, peace_talk, human_right)

- Alta frecuencia en días de fluctuación negativa, especialmente entre 2013 y 2016.
- Coinciden con el contexto del proceso de paz en Colombia, y reflejan cómo los temas de conflicto y derechos humanos **generan incertidumbre en el mercado**.
- Términos como peace_process y human_right muestran además una **disminución progresiva** tras la firma del acuerdo de paz, lo que evidencia una transición temática en la narrativa noticiosa y su impacto económico.

2. Temas económicos y fiscales (tax, company, mining, president)

- tax presenta picos que **coinciden con periodos de reformas fiscales** (como en 2016), altamente relacionados con fluctuaciones negativas, reflejando **temor regulatorio**.
- mining y company tienden a aparecer tanto en contextos positivos como negativos, pero con **mayor actividad en periodos volátiles**, lo que puede indicar su uso como marcadores de incertidumbre estructural.
- president es un término que fluctúa con ambos contextos, pero su patrón sugiere **relevancia en eventos políticos de alto impacto**, como elecciones o decisiones de política económica.

3. Recursos específicos y entidades (colombia_gold, medellín)

- colombia_gold y medellín aparecen como temas asociados a momentos puntuales de cobertura, principalmente en días de fluctuación negativa, lo que sugiere que las noticias relacionadas a empresas específicas o conflictos territoriales **generan ruido en el mercado local**.

9.1.2.3. *Análisis de aparición de términos en el resumen noticioso:*

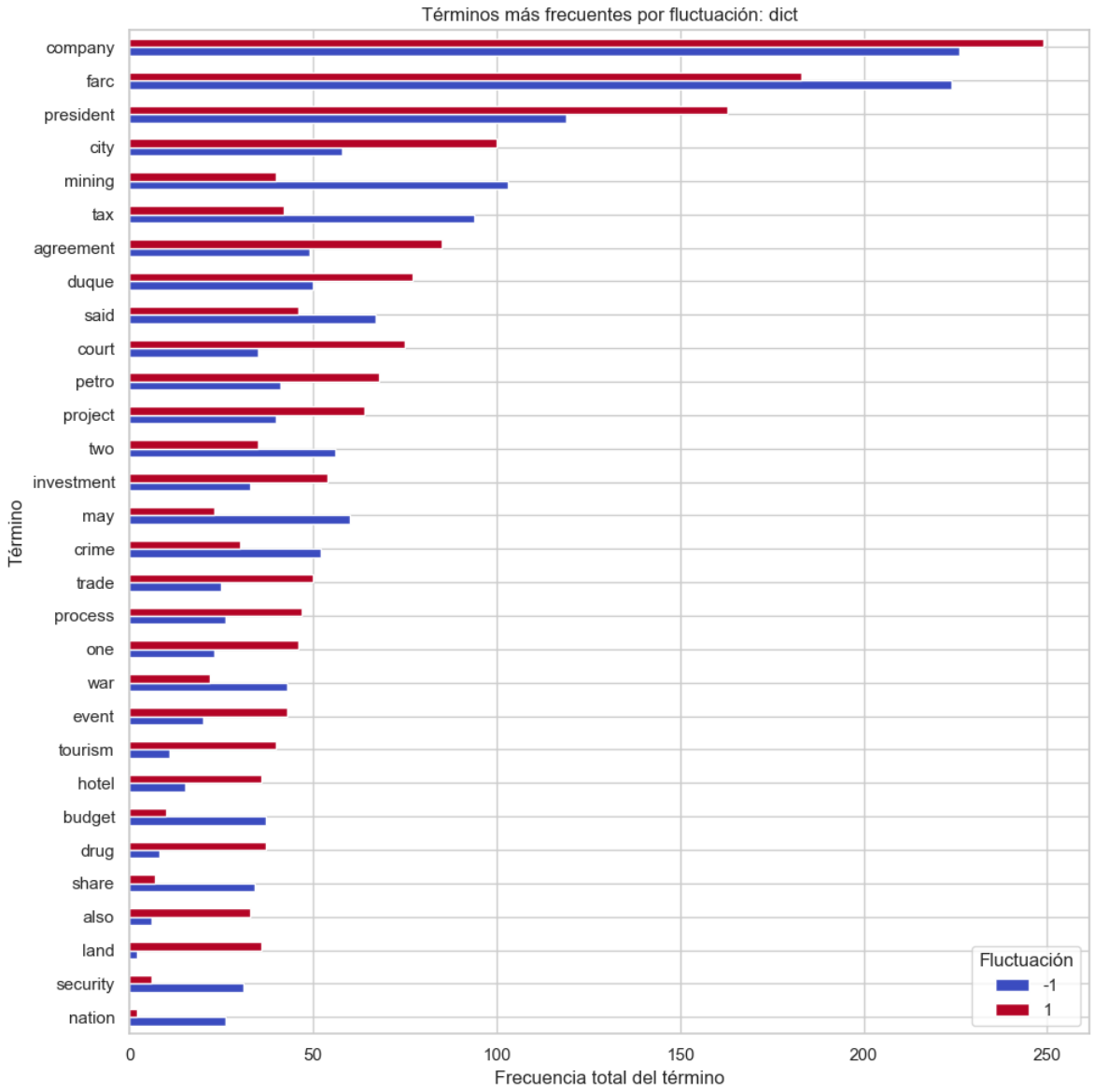

Ilustración 22 Grafico de barras compuesto para términos más frecuentes por fluctuación (resumen).

En el gráfico de la Ilustración 22 para los términos simples (unigramas), observamos que algunas palabras tienen **presencia relativamente equilibrada** entre subidas y bajadas. Sin embargo, otros términos muestran **preferencia notable por uno de los dos contextos**:

- Más frecuentes en días de caída (-1):
  - farc, mining, tax, crime, budget, security, share, budget, y said
  - Estos términos están fuertemente vinculados a **narrativas de riesgo, presión fiscal o inseguridad**, lo que podría reflejar la reacción del

mercado ante amenazas estructurales o eventos de alta sensibilidad social.

- Más frecuentes en días de subida (1):
    - Company, president, city, land, drug, hotel, tourism, y event
    - Aparecen términos asociados a **reactivación económica, inversión, o eventos públicos positivos**, lo que sugiere una mayor cobertura de noticias con carga constructiva o de recuperación.

Este contraste sugiere que **el contenido léxico reacciona a la dirección del mercado**, expresando más términos de conflicto, presión o controversia en contextos negativos, y más ideas asociadas a dinamismo o proyección en contextos positivos.

Realizando un comparativo de los términos de la Ilustración 18 y 22, evidenciamos que los términos en la ilustración 22 asociados a la información léxica proveniente del resumen del corpus noticioso, aumenta las señales lingüísticas de diferenciación en eventos de fluctuación para los términos, convalidando su aplicabilidad como una estrategia adecuada para esta iniciativa.

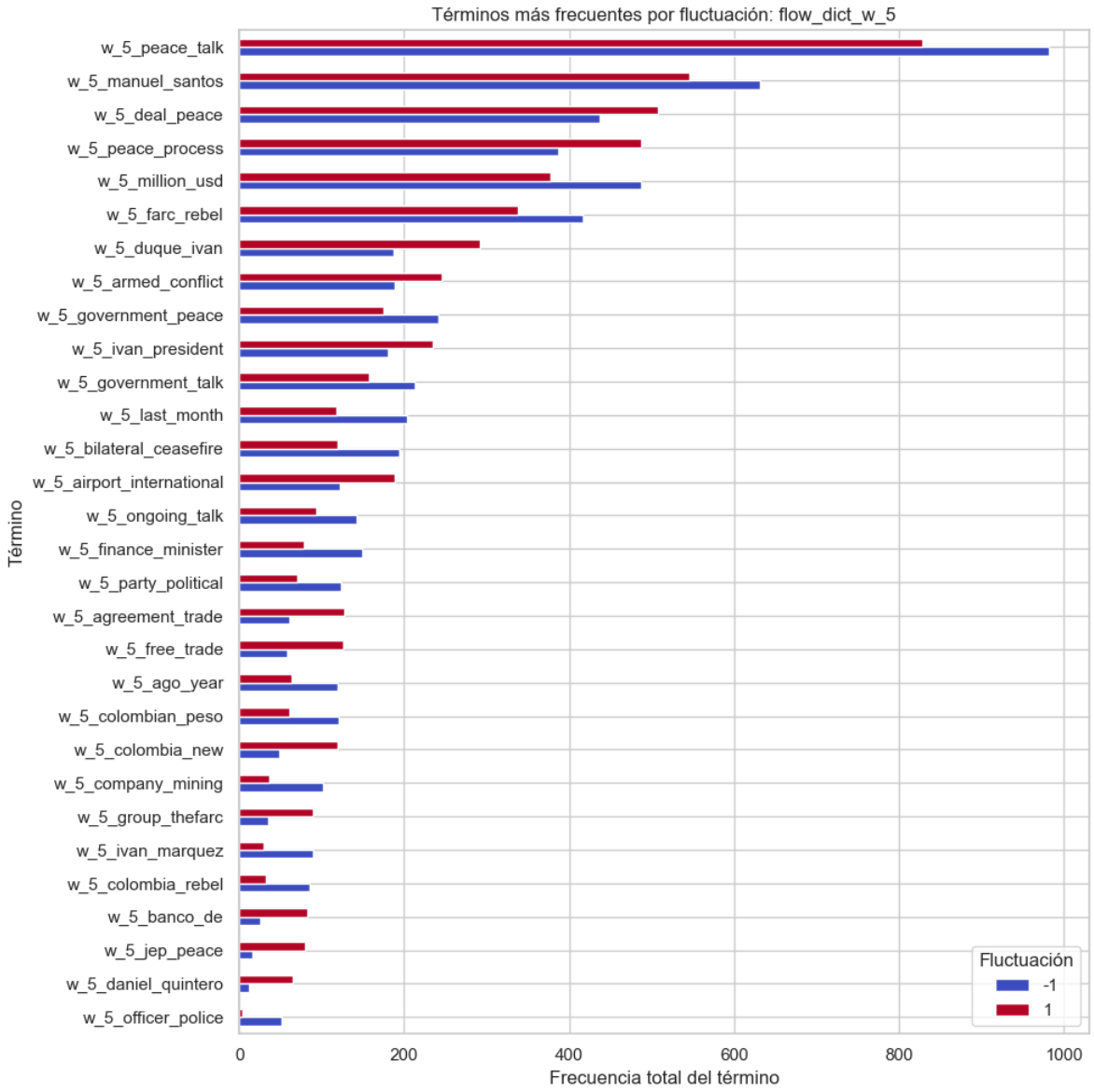

Ilustración 23 Grafico de barras combinadas para términos más frecuentes por fluctuación (resumen)

El gráfico de la Ilustracion 23 correspondiente al análisis de combinaciones frecuentes dentro de una **ventana móvil de 5 palabras** muestra resultados aún más reveladores, ya que las secuencias compuestas capturan **significados más estructurados** y relaciones semánticas más profundas.

- Las combinaciones con **mayor frecuencia absoluta** (como *peace_talk*, *manuel_santos*, *deal_peace*, *peace_process*) aparecen tanto en días positivos como negativos, pero con variaciones notables en proporción:
    - *peace_talk*: 982 (-1) vs 828 (1)
    - *deal_peace*: 437 (-1) vs 508 (1)
    - *jep_peace*: 16 (-1) vs 80 (1)

Esto sugiere que los **procesos de paz y las figuras políticas asociadas** son temas estructuralmente importantes, aunque su carga emocional y económica puede percibirse distinta según el tipo de noticia.

- Algunas combinaciones muestran una **clara inclinación hacia los días de alza**, como:
    - daniel_quintero, agreement_trade, free_trade, colombia_new
    - Estas expresiones están asociadas con **figuras públicas activas, acuerdos comerciales y dinámicas de apertura**, indicando un lenguaje más enfocado en progreso, relaciones institucionales y crecimiento.
- Otras combinaciones, como *company_mining*, *group_thefarc*, *colombia_rebel*, *armed_conflict*, dominan en días de caída, representando **narrativas de conflicto, explotación o polarización**.

9.1.2.4. análisis temporal de aparición de términos en el resumen noticioso:

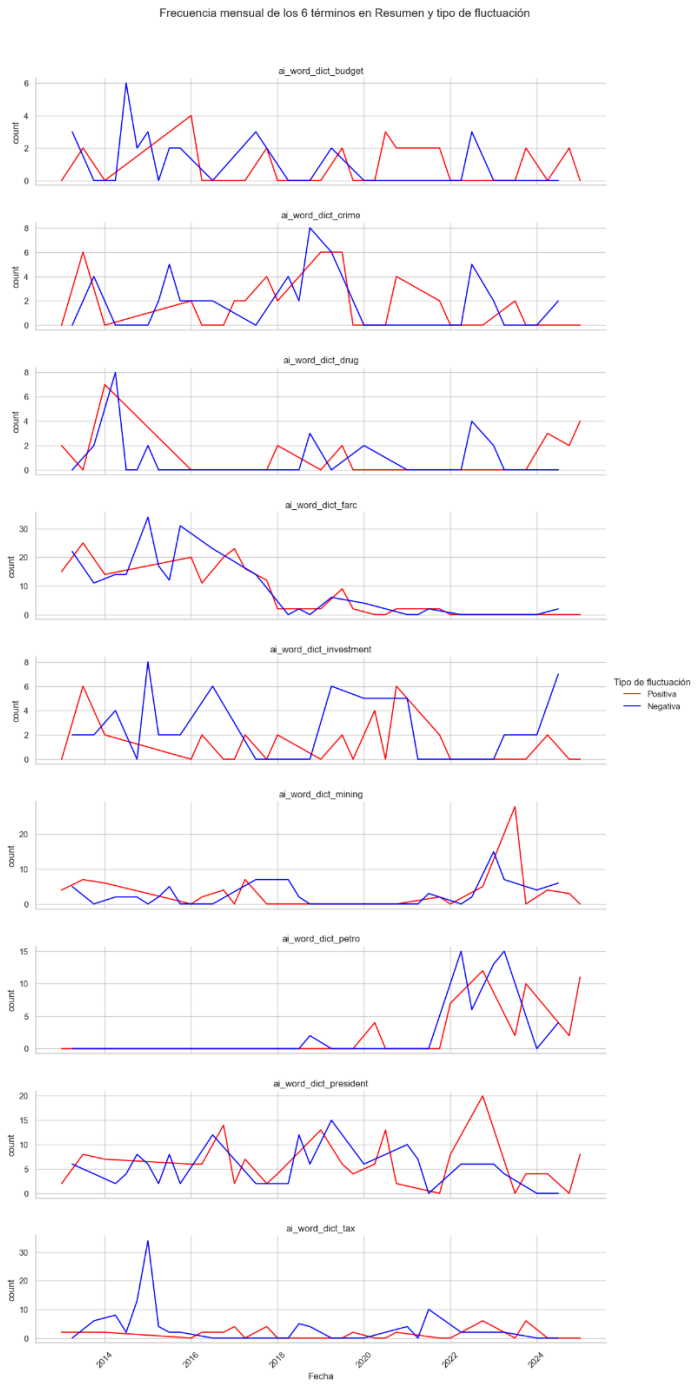

Ilustración 24 Grafico temporal de frecuencia de aparición de términos según fluctuación, donde el prefijo de los términos ai_word_dict, resalta que la información es proveniente de un preprocesamiento del corpus noticioso.

El gráfico presentado en la Ilustración 24 nos muestra la frecuencia mensual de aparición de diez términos clave dentro de los resúmenes automáticos de noticias económicas, segmentados por tipo de fluctuación del mercado (positiva en rojo, negativa en azul). Este análisis revela no solo la presencia o ausencia de ciertas

palabras en contextos de mercado específicos, sino también la **enorme variabilidad y dinamismo en su comportamiento temporal**, lo cual pone en evidencia uno de los principales desafíos del modelado lingüístico en entornos financieros: **la naturaleza cambiante del significado contextual de los términos**.

Los términos observados en el gráfico, como *farc*, *petro*, *mining*, *crime*, *tax*, o *investment*, **no mantienen una frecuencia estable a lo largo del tiempo**, sino que aparecen en picos abruptos y luego desaparecen por largos periodos. Esta característica evidencia que el lenguaje económico y político en la prensa **no responde a un patrón fijo**, sino que **fluctúa en función del contexto nacional**, coyunturas económicas, decisiones gubernamentales o crisis puntuales.

Por ejemplo:

- **farc** tuvo una alta frecuencia entre 2014 y 2017 (con especial presencia en días de caída), durante el proceso de paz con las FARC. A partir de 2018, su presencia cae drásticamente, volviéndose marginal en los resúmenes. Esto muestra cómo **una misma palabra puede perder relevancia semántica al cambiar el ciclo político o el enfoque mediático**.
- **petro** no aparece significativamente hasta después de 2021, coincidiendo con el ascenso de Gustavo Petro al poder. Su aparición está más relacionada con días de alza, posiblemente por coberturas optimistas asociadas a propuestas políticas o expectativas económicas. Esto refuerza la idea de que **la interpretación de un término cambia completamente según el momento histórico**.
- **tax** y **mining** exhiben picos abruptos en periodos específicos (por ejemplo, el término *tax* alcanza un valor inusualmente alto en 2015 y 2022), lo que sugiere que son **tópicos altamente sensibles a decisiones gubernamentales o reformas puntuales**. En estos casos, su aparición no está asociada a una narrativa estable, sino a eventos concretos que irrumpen en la agenda informativa.
- Términos como **investment**, **budget** o **crime** presentan una distribución aún más irregular, alternando entre momentos de alta y baja aparición sin una lógica temporal clara, lo que confirma que su relevancia textual está **altamente modulada por el entorno noticioso**.

Este comportamiento demuestra que el valor predictivo de un término no es inherente ni constante: **el mismo término puede ser irrelevante en un contexto, pero altamente informativo en otro**. En consecuencia, su interpretación no puede ser desvinculada del momento en el que aparece ni de la narrativa discursiva que lo enmarca.

Este fenómeno también introduce un alto grado de **ruido semántico** en los modelos de predicción basados en texto, ya que:

- Las señales lingüísticas no son estables ni estacionarias.
- La misma palabra puede tener connotaciones opuestas en distintos contextos.
- La mayoría de los términos informativos aparecen de forma **esporádica y concentrada**, lo que dificulta la generalización de patrones.

El gráfico confirma que la aparición de términos clave en los resúmenes de noticias es **altamente dinámica, impredecible y dependiente del contexto político, económico y mediático**. Este comportamiento reafirma que **el análisis semántico de noticias para propósitos predictivos requiere enfoques sensibles al tiempo y adaptativos al contenido**, como modelos con ventanas móviles, atención contextual o técnicas de seguimiento temporal. No es suficiente capturar la frecuencia de un término: es indispensable **comprender cuándo, por qué y cómo aparece en relación con el mercado**. Esta variabilidad constituye tanto un reto como una oportunidad para el desarrollo de modelos lingüísticos más robustos y sensibles al entorno.

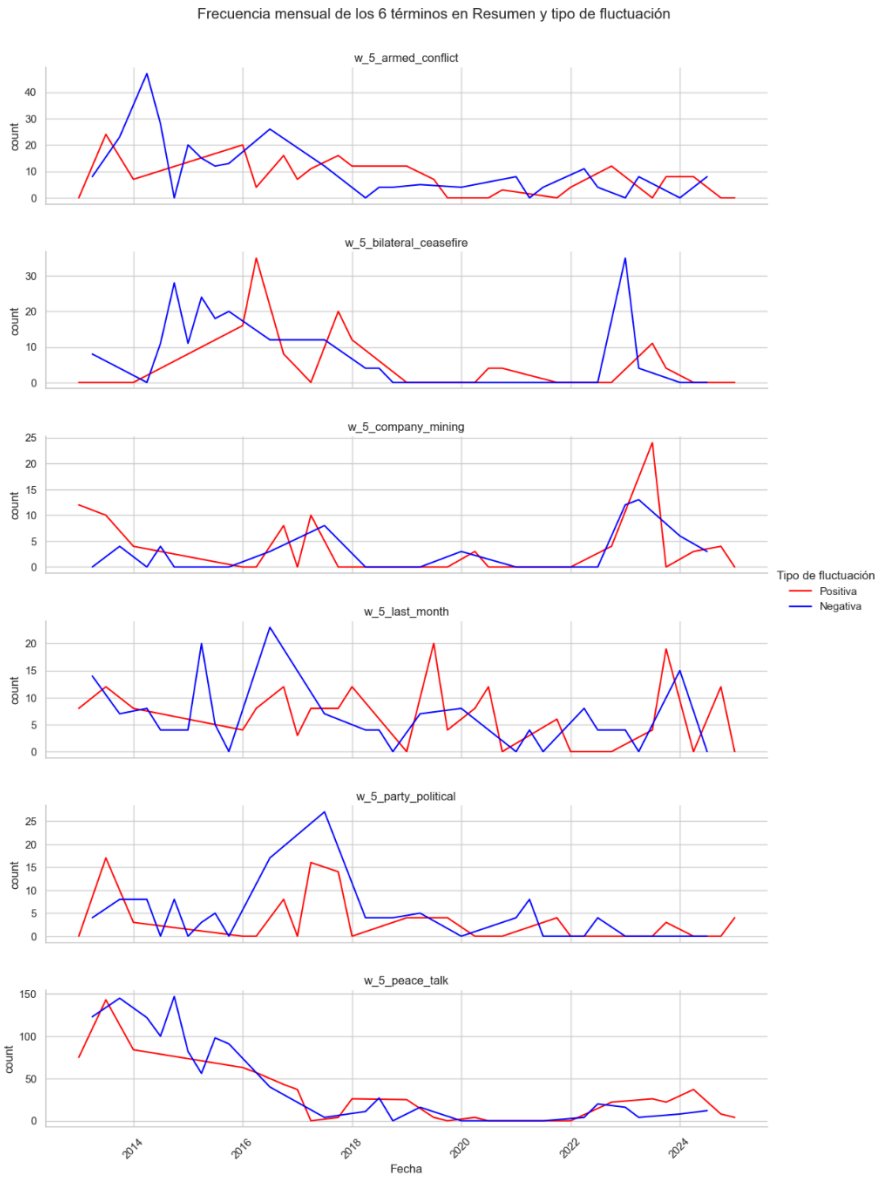

Ilustración 25 Grafico temporal para la aparición de bigramas según su fluctuación (resumen)

La Ilustracion 25 presenta la **evolución mensual** de seis expresiones compuestas (bigrams o frases de hasta cinco tokens) extraídas de los resúmenes automáticos del corpus noticioso. Estas secuencias léxicas fueron seleccionadas por su alta frecuencia y relevancia semántica, y se analizan diferenciando su aparición en días con **fluctuación positiva** del mercado (línea roja) y días con **fluctuación negativa** (línea azul).

Este análisis permite explorar no solo la **frecuencia de aparición de conceptos clave en la narrativa mediática**, sino también cómo su presencia se asocia con

distintos contextos económicos a lo largo del tiempo. A diferencia de los términos aislados, las expresiones compuestas aquí representadas —como *peace talk* o *company mining*— capturan **relaciones semánticas más precisas** y permiten una interpretación más contextualizada de los fenómenos descritos.

Observaciones clave por término

- **w_5_peace_talk**: muestra una clara concentración entre 2013 y 2017, coincidiendo con los años más activos del proceso de paz con las FARC. Su aparición es **mayor durante días de fluctuación negativa**, lo que podría sugerir que las noticias relacionadas con negociaciones de paz, aunque relevantes, no generan automáticamente confianza en los mercados financieros.
- **w_5_armed_conflict**: mantiene una presencia considerable en ambos tipos de días, con un ligero predominio en contextos de caída. La presencia sostenida de este término refleja su importancia como eje narrativo persistente en el entorno colombiano, aunque su intensidad disminuye progresivamente después de 2017.
- **w_5_bilateral_ceasefire**: al igual que *peace_talk*, presenta picos pronunciados entre 2014 y 2017, pero con una alternancia más balanceada entre días de subida y bajada. Esta ambigüedad semántica sugiere que **la narrativa del cese al fuego puede ser percibida como positiva o negativa dependiendo del contexto político en que se anuncie**.
- **w_5_company_mining**: muestra una aparición más tardía, con picos a partir de 2020. Su presencia se asocia más frecuentemente a días de caída, lo que podría interpretarse como una señal de incertidumbre regulatoria, conflictos territoriales o tensiones ambientales que afectan al sector extractivo.
- **w_5_last_month** y **w_5_party_political**: son expresiones más generales, asociadas a contexto temporal o dinámicas institucionales. A pesar de su carácter menos específico, muestran una **marcada oscilación en el tiempo**, lo que refuerza la idea de que incluso términos aparentemente neutros tienen **implicaciones distintas dependiendo del momento y la coyuntura informativa**.

Las expresiones compuestas analizadas presentan una frecuencia de aparición altamente variable a lo largo del tiempo, evidenciando su estrecha relación con ciclos discursivos condicionados por eventos históricos y contextos políticos específicos. Su vínculo con la dirección del mercado no es unívoco: términos como *peace talk* o *armed conflict* no generan respuestas bursátiles consistentes, ya que su impacto depende del momento y del contexto mediático en que se enuncian.

Esta variabilidad semántica demuestra que el significado de una expresión no es estático, sino que evoluciona y se resignifica según las circunstancias, representando un desafío considerable para los modelos de predicción basados en texto. En consecuencia, estos hallazgos enfatizan la importancia de utilizar representaciones lingüísticas dinámicas y contextuales, capaces de adaptarse temporalmente y de evitar suposiciones rígidas entre lenguaje y comportamiento del mercado.

En síntesis, la presencia oscilante y multivalente de estas expresiones en los resúmenes de noticias confirma que el lenguaje económico y político es profundamente contextual y volátil, lo que exige enfoques analíticos sofisticados para su interpretación y aplicación en sistemas predictivos de comportamiento financiero.

9.1.2.5. Análisis combinado de aparición de términos en el resumen y corpus noticioso:

El conjunto de análisis realizados a lo largo de esta sección proporciona evidencia robusta de que **el contenido lingüístico de las noticias financieras presenta una relación significativa con el comportamiento del mercado bursátil colombiano**. A través del examen de términos individuales, expresiones compuestas, patrones de frecuencia y evolución temporal, se identificaron diferencias consistentes entre el lenguaje empleado en días de alza y de baja del índice COLCAP, tanto en el corpus noticioso completo como en los resúmenes generados automáticamente.

En los días con **fluctuación negativa**, los términos más frecuentes tienden a asociarse con narrativas de **conflicto, riesgo institucional, inseguridad, presión fiscal y tensiones políticas**. Palabras como *farc*, *tax*, *crime*, *mining* o expresiones como *armed conflict*, *peace talk* y *company mining* son predominantes en estos contextos, reflejando la forma en que el mercado reacciona ante señales percibidas como amenazas a la estabilidad económica o social. Esto indica que el **lenguaje del conflicto y la incertidumbre tiene un correlato claro en el comportamiento bajista del mercado**.

Por el contrario, en los días con **fluctuación positiva**, se destaca un léxico enfocado en **temas económicos, institucionales y de crecimiento**, incluyendo términos como *company*, *city*, *agreement, investment*, y palabras compuestas como *deal peace*, *peace process, agreement trade*, o *free trade*. Esta narrativa favorable

sugiere una cobertura más orientada al progreso, la inversión o la estabilidad, elementos que probablemente contribuyen a generar confianza en los actores económicos.

No obstante, al extender el análisis hacia el **comportamiento temporal de estos términos**, se revela una dinámica más compleja y volátil. La aparición de palabras clave no sigue un patrón estacionario, sino que **responde a coyunturas políticas, económicas o mediáticas específicas**. Términos como *farc* o *peace talk*, que fueron centrales durante el proceso de paz (2013–2017), pierden relevancia en años posteriores. Por otro lado, términos emergentes como *petro*, *daniel quintero* o *investment* adquieren protagonismo en periodos recientes. Esta variabilidad semántica pone en evidencia que **el valor predictivo de un término no es estático ni universal**, sino profundamente dependiente del momento histórico y del contexto discursivo.

Asimismo, el análisis de expresiones compuestas refuerza la idea de que **la estructura lingüística del discurso cambia según el estado del mercado**. No solo varía el vocabulario, sino también las formas en que se articulan los conceptos, revelando una complejidad semántica que va más allá del conteo bruto de palabras.

En términos metodológicos, estos resultados resaltan que el modelado predictivo con base en lenguaje natural debe enfrentar desafíos significativos:

- Las señales lingüísticas son contextuales, no estacionarias y no lineales.
- El mismo término puede tener connotaciones opuestas en distintos años, actores o sectores.
- La mayoría de los términos relevantes aparecen de forma **puntual o concentrada**, lo que complica su generalización como variables robustas.

Por tanto, si bien **la incorporación de variables semánticas en modelos financieros es justificada y potencialmente poderosa**, su implementación requiere enfoques avanzados que consideren **la dimensión temporal, temática y contextual del lenguaje**, tales como ventanas móviles, codificación temporal de vocabulario, modelos de embeddings adaptativos o atención contextual.

Finalmente, estos hallazgos refuerzan la hipótesis fundamental de este trabajo: **el lenguaje noticioso refleja —y posiblemente condiciona— la percepción de riesgo e incertidumbre en los mercados financieros**. Aunque su efecto no es directamente determinista ni fácilmente cuantificable, constituye una fuente valiosa de información para enriquecer modelos predictivos y diagnósticos del comportamiento bursátil. La evidencia muestra que **los mercados no solo reaccionan a cifras y hechos, sino también a las narrativas que los envuelven**.

9.1.3. Frecuencia temática y diferencias por fluctuación

Con el objetivo de comprender cómo varía el enfoque temático de las noticias según el comportamiento del mercado, se analizó la columna ai_common_topic, la cual clasifica automáticamente cada noticia dentro de una de las categorías temáticas definidas previamente (*market_sentiment, regulatory_action, geopolitical_risk*, entre otras). Esta clasificación se derivó mediante un procedimiento basado en embeddings semánticos, que asigna cada término clave al centro temático más cercano mediante la métrica de similaridad coseno.

9.1.3.1. Frecuencia temática y diferencias por fluctuación para el corpus noticioso

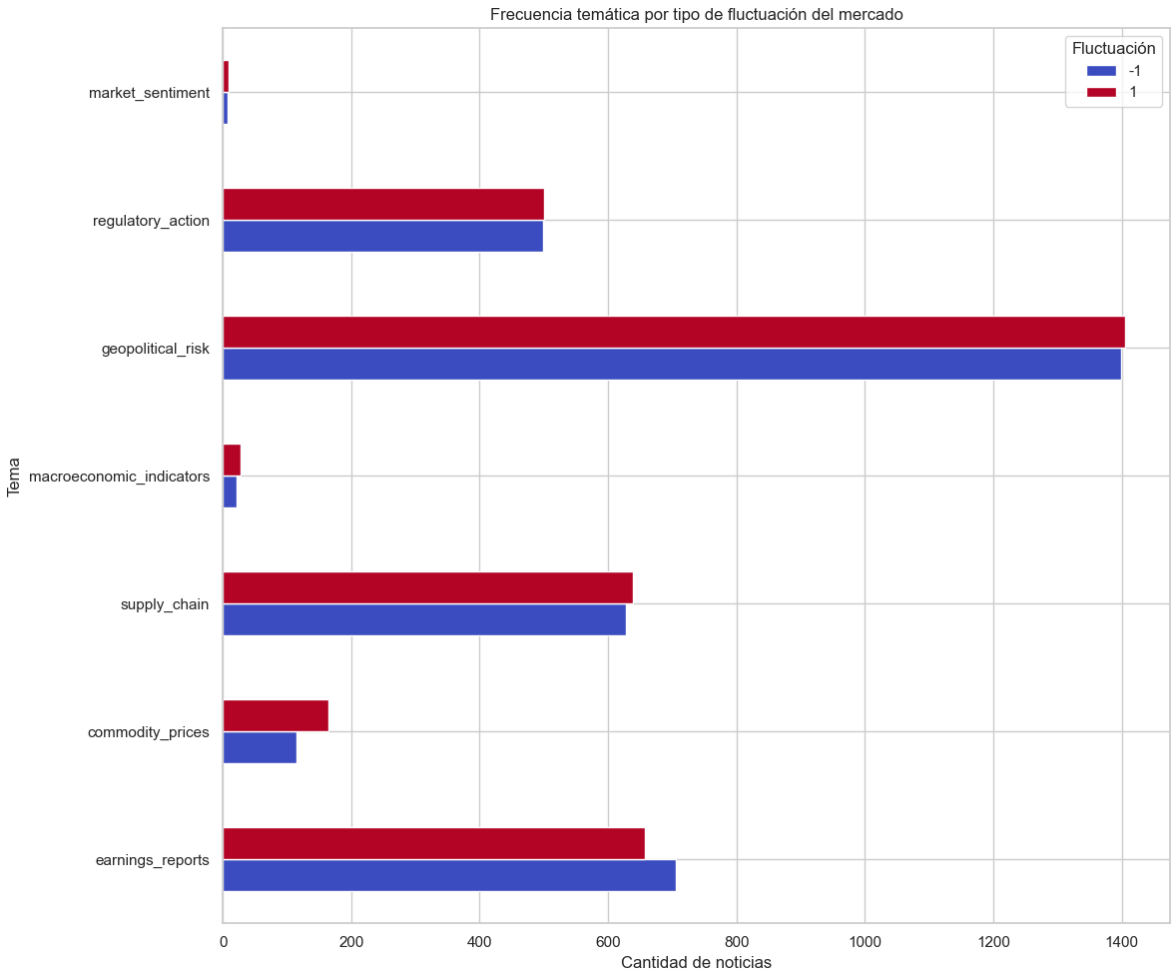

Ilustración 26 grafico de barras combinada para la frecuencia temática por tipo de fluctuación del mercado.

En la Ilustración 26 observamos que, independientemente de la dirección del mercado, existen temáticas que dominan el panorama informativo. No obstante, se evidencian **diferencias cuantitativas significativas entre los días positivos y negativos**. El tema más frecuente en ambos escenarios es **geopolitical_risk**, con más de 1.400 noticias asociadas tanto en días de alza como de baja. Esta constante sugiere que el entorno geopolítico es un eje transversal en la cobertura financiera, aunque su impacto puede variar según el contexto específico.

Temas como **regulatory_action** y **supply_chain** también presentan una alta frecuencia, con una **ligera predominancia en días de fluctuación positiva**. Esto podría interpretarse como una mayor cobertura de medidas de control, comercio y logística en contextos de recuperación o estabilidad. Por otro lado, el tema **earnings_reports** muestra una distribución relativamente equilibrada, aunque con una ligera inclinación hacia los días negativos, lo que refuerza la relación esperada entre **resultados financieros corporativos y desempeño del mercado, esto**

puede ser confirmado cuando se analizan n-gramas relacionados a reportes financieros como lo son first_quarter, year_quarter, y nombres de compañías colombianas como ecopetrol y epm.

En contraste, categorías como **market_sentiment** y **macroeconomic_indicators** tienen una presencia mucho menor, lo que puede deberse a que se expresan indirectamente en el texto o se solapan con otras categorías más explícitas como commodity_prices.

En conjunto, estos resultados evidencian que **la dimensión temática de las noticias no es neutral respecto al comportamiento del mercado**. La predominancia de ciertos tópicos en contextos positivos o negativos refuerza la utilidad de incorporar esta variable como insumo explicativo en modelos de predicción bursátil. Este tipo de análisis permite anticipar sesgos informativos y ajustar los modelos a los patrones discursivos más relevantes para el mercado.

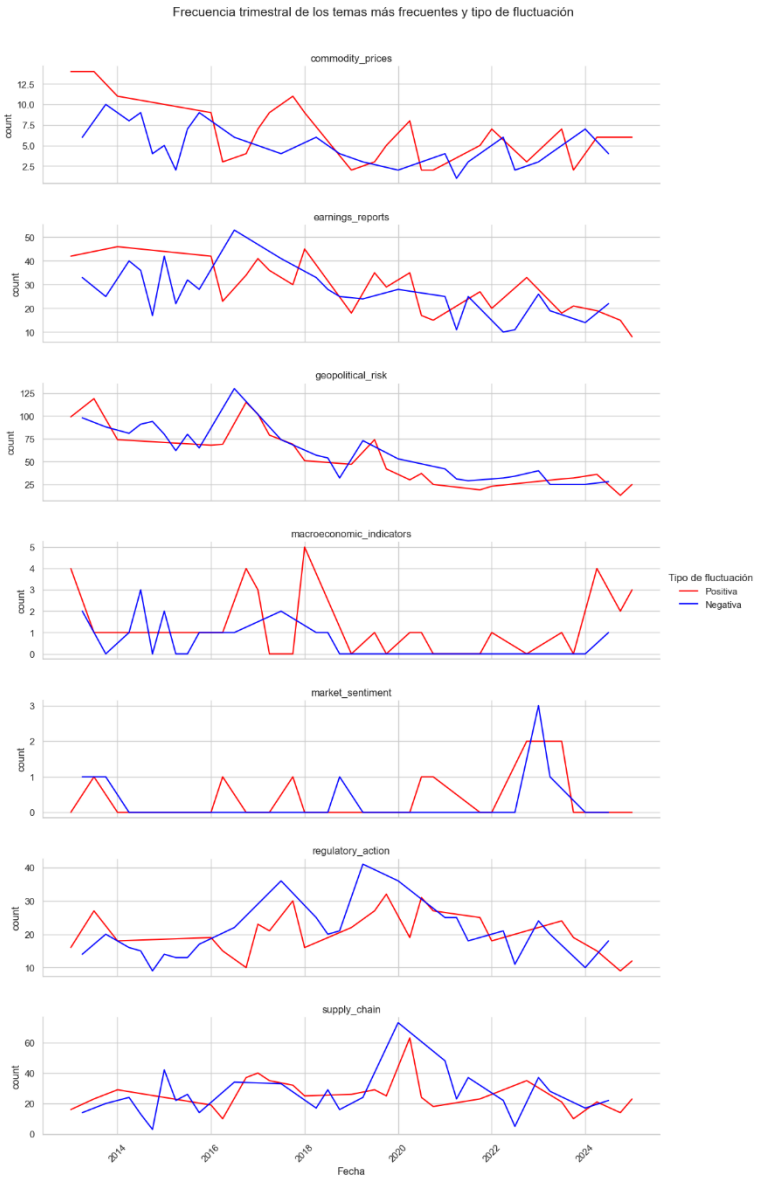

Ilustración 27 grafico temporal de la frecuencia de los temas mas frecuentes y tipo de fluctuación.

El análisis de los gráficos que se muestran en la Ilustración 27 reflejan que la **frecuencia temática de las noticias varía de forma consistente con la dirección del mercado**, reflejando que el contenido noticioso no es ajeno al contexto bursátil, sino que responde —y posiblemente influye— en la dinámica financiera. Temas como geopolitical_risk y supply_chain se consolidan como **indicadores de tensión**, predominando en contextos de mercado negativo. En cambio, categorías como earnings_reports y commodity_prices tienden a reforzar **narrativas optimistas**, siendo más frecuentes en trimestres con alzas.

Estos hallazgos validan la incorporación de variables temáticas en modelos predictivos de fluctuación del mercado, aportando una dimensión semántica que complementa los indicadores numéricos tradicionales. Además, permiten entender **cómo el discurso informativo se adapta a las condiciones económicas**, proporcionando una herramienta de análisis valiosa para estudios de finanzas conductuales y medios económicos.

9.1.3.2. Frecuencia temática y diferencias por fluctuación para los resúmenes noticiosos

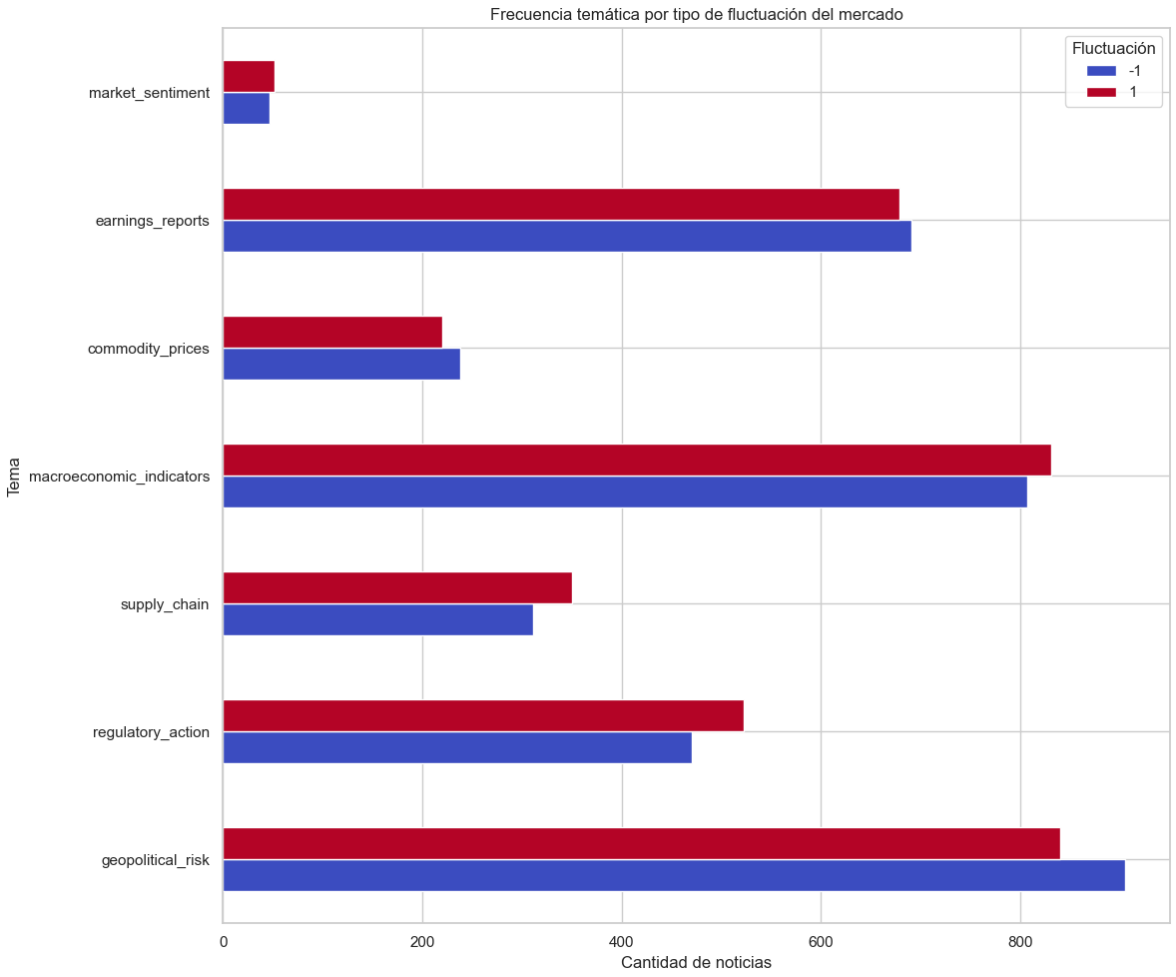

Ilustración 28 grafico de Barras combinado para la frecuencia temática por tipo de fluctuación del mercado (resumen)

El análisis temático aplicado a los resúmenes generados mediante modelos de lenguaje natural revela patrones informativos que reflejan diferencias significativas según la dirección del mercado. En la Ilustración 28 se observa la distribución de las categorías temáticas más frecuentes, diferenciando entre días de fluctuación positiva y negativa del índice COLCAP. Si bien existen tópicos que son consistentes en ambos escenarios, como *geopolitical_risk* y *macroeconomic_indicators*, también

emergen diferencias relevantes que permiten inferir una posible relación entre el contenido de los resúmenes y el comportamiento del mercado.

El tema *geopolitical_risk* lidera en frecuencia tanto en días positivos como negativos, lo cual sugiere que la incertidumbre geopolítica se mantiene como una constante narrativa en los medios, aunque su interpretación económica puede diferir según el contexto. Por su parte, *macroeconomic_indicators* y *earnings_reports* muestran una alta frecuencia y una distribución equilibrada entre los dos tipos de días, reflejando su rol estructural en la cobertura financiera. En contraste, temas como *commodity_prices* y *market_sentiment* presentan frecuencias notablemente menores, lo que puede deberse a su representación más implícita o a su integración dentro de narrativas más amplias.

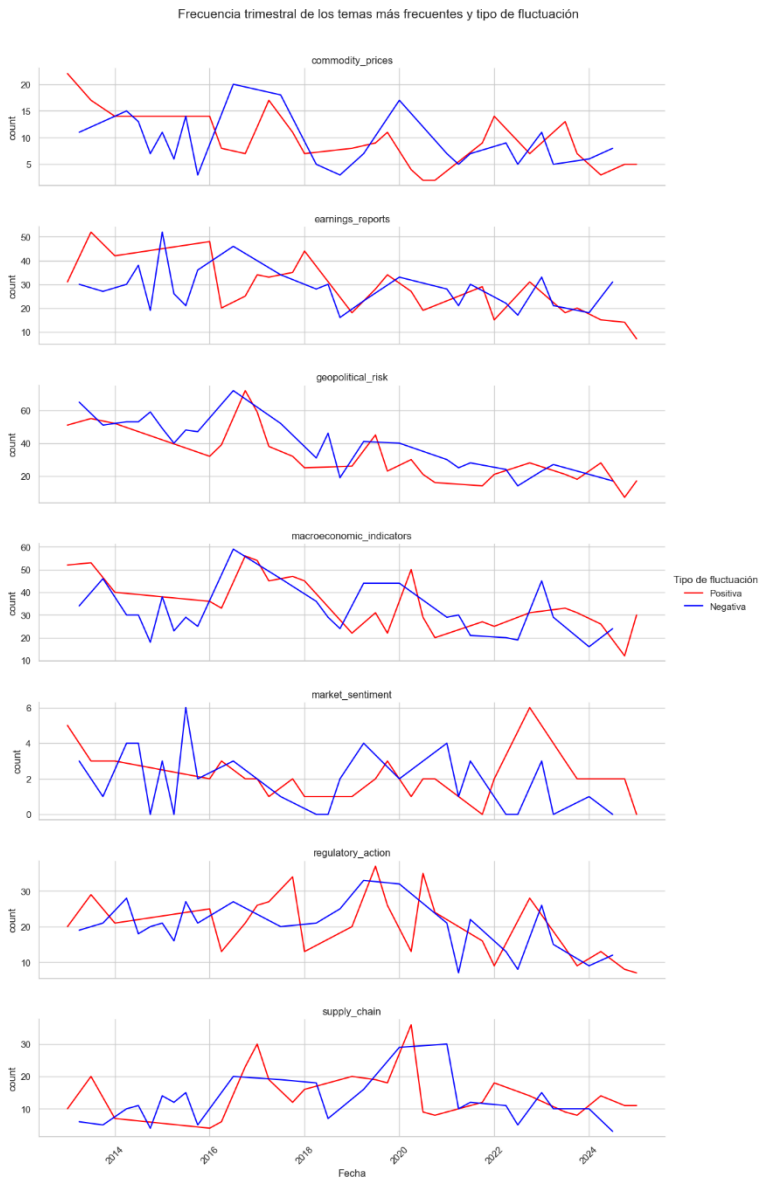

Ilustración 29 Grafico temporal para la frecuencia trimestral de los temas más frecuentes y tipo de fluctuación (resumen)

La Ilustración 29 profundiza este análisis mediante series temporales que muestran la evolución trimestral de cada categoría temática segmentada por tipo de fluctuación. En estas gráficas se evidencia el dinamismo de los tópicos a lo largo del tiempo: por ejemplo, *geopolitical_risk* presenta una disminución sostenida en su frecuencia desde 2015, lo que podría reflejar una normalización relativa del entorno internacional o una migración del enfoque noticioso hacia otras temáticas. *Regulatory_action* y *supply_chain* exhiben picos asociados a eventos específicos, como reformas regulatorias o disrupciones logísticas globales, mostrando mayor presencia en trimestres con alzas, lo que sugiere una cobertura orientada hacia medidas de control o recuperación.

Por el contrario, *earnings_reports* y *commodity_prices* muestran mayor volumen en trimestres de caída, posiblemente asociados a momentos de evaluación financiera o a fluctuaciones en los precios internacionales que generan sensibilidad en el mercado local. Esta variabilidad temática ofrece un reflejo del enfoque mediático cambiante, donde ciertas categorías emergen o disminuyen en relevancia según la coyuntura económica y política.

Estos resultados refuerzan la hipótesis de que el contenido temático de los resúmenes noticiosos no es neutro, sino que responde a la narrativa del momento y puede actuar como un indicador contextual del sentimiento económico general. La incorporación de estas categorías como variables explicativas en modelos predictivos no solo es viable, sino deseable, pues aportan una capa de significado que trasciende los indicadores cuantitativos tradicionales.

Además, se evidencia que la semántica de las noticias sintetizadas conserva elementos críticos del discurso económico, lo cual valida el uso de resúmenes automáticos como fuente informativa confiable para el análisis temático. Si bien se pierde parte de la riqueza contextual presente en los textos completos, se conserva una representación condensada de los tópicos más relevantes, facilitando el análisis de grandes volúmenes de información de forma eficiente y estructurada.

En conclusión, el análisis temático aplicado a los resúmenes permite identificar tendencias informativas coherentes con los ciclos de mercado, y destaca su potencial para enriquecer modelos de predicción con señales semánticas que capturan no solo el contenido informativo, sino también la intencionalidad narrativa de los medios financieros.

9.2. Modelado predictivo y evaluación

El núcleo metodológico de este trabajo se centró en construir y evaluar modelos de aprendizaje profundo capaces de predecir la dirección diaria del índice COLCAP, a partir de variables cuantitativas tradicionales y representaciones semánticas derivadas de noticias. El objetivo era determinar si las variables lingüísticas extraídas mediante técnicas de PLN aportaban valor predictivo y en qué condiciones arquitectónicas ese valor podía aprovecharse al máximo.

Para ello se definió una grilla de búsqueda (grid search) con múltiples combinaciones de arquitecturas de redes neuronales multicapa (MLP). Esta decisión se fundamentó en la necesidad de explorar modelos con diferentes niveles de profundidad y regularización, dada la alta dimensionalidad del dataset después de incorporar componentes semánticos. Se incluyeron configuraciones con diferentes tamaños de capa (layers), tasas de abandono (dropout), tamaños de lote (batch_size), niveles de reducción dimensional por PCA (pca_components), tasas de aprendizaje (learning_rate) y porcentajes de limpieza de variables lingüísticas (clean_percentile).

9.2.1. Modelos entrenados

El proceso de modelado predictivo fue respaldado por una estrategia de exploración intensiva basada en *grid search*, diseñada para capturar el efecto conjunto e individual de múltiples hiperparámetros críticos. En total, se evaluaron **720 combinaciones únicas para cada uno de los datasets (corpus completo y resumido)**, filtradas lógicamente para asegurar consistencia semántica entre las variables (por ejemplo, excluyendo configuraciones que aplican PCA cuando previamente se ha eliminado el 100% de las variables lingüísticas).

El espacio de búsqueda fue cuidadosamente definido para abarcar una amplia gama de arquitecturas y estrategias de regularización. Las variables evaluadas fueron:

- **Estructura de la red neuronal**: 6 configuraciones distintas de capas ocultas, desde redes compactas como [64, 32, 16] hasta modelos más profundos como [128, 128, 128, 64, 32], permitiendo capturar distintos niveles de complejidad no lineal en los datos.
- **Dropout**: 4 valores distintos (0.0, 0.1, 0.2, 0.4), aplicados como técnica de regularización para mitigar el sobreajuste, especialmente importante dado el alto número de variables semánticas.
- **Tamaño del lote**: fijo en 32, optimizado previamente como valor eficiente para el volumen de datos y arquitectura evaluada.
- **Reducción dimensional por PCA**: 3 configuraciones (sin PCA, varianza explicada del 90% y 98%), orientadas a reducir la dimensionalidad manteniendo la información más relevante del espacio semántico.
- **Porcentaje de limpieza de variables semánticas**: 5 niveles (0%, 50%, 90%, 98% y 100%), diseñados para evaluar el impacto de conservar o excluir términos léxicos menos frecuentes.
- **Tasa de aprendizaje**: fijada en 0.0005 con el optimizador *Adam*, elegida por su estabilidad y eficiencia en tareas con alto volumen de parámetros.
- **Condición de parada anticipada** (*early stopping*): establecida en 100 épocas para prevenir sobre-entrenamiento en configuraciones de alta capacidad.

Cada modelo fue entrenado con una duración máxima de 1000 épocas, utilizando validación cruzada y métricas de evaluación que incluyeron: *accuracy*, *f1_score*, *precision*, *recall* y *loss* (entendida como log-loss), esta última utilizada como criterio principal para evaluar la calibración probabilística del modelo.

Esta estrategia permitió no solo identificar las configuraciones más efectivas, sino también trazar patrones consistentes sobre cómo afectan los distintos hiperparámetros al rendimiento del modelo. El diseño de esta grilla permitió capturar

interacciones relevantes entre arquitectura, regularización y estrategia de representación semántica, constituyendo uno de los aportes metodológicos más significativos de esta investigación.

hiperparámetro	*Opciones evaluadas*
Estructura de capas (layer_options)	[256, 128, 128, 64] [128, 128, 64, 32] [128, 128, 128, 64, 32] [64, 32, 16] [64, 64, 32] [128, 64, 32]
Tasa de abandono (dropout_options)	0.0, 0.1, 0.2, 0.4
Tamaño de lote (batch_sizes)	32
Épocas máximas (epoch_options)	1000
Optimizador (optimizers)	'adam'
Componentes PCA (PCA_num)	0, 0.90, 0.98
Limpieza semántica (clean_percentiles)	0, 50, 90, 98, 100
Tasa de aprendizaje (learning_rates)	0.0005
Parada anticipada (early_stop)	100

Tabla 6 Resumen de la grilla de hiperparámetros de búsqueda para los modelos Deep Learning.

9.2.2. Desempeño de los modelos

La presente sección tiene como objetivo principal analizar el desempeño de los distintos modelos entrenados para predecir la fluctuación del índice COLCAP, utilizando tanto noticias completas como sus resúmenes automáticos. El enfoque de evaluación se basa en comparar diferentes configuraciones de redes neuronales profundas, variando parámetros arquitectónicos como el número de capas, la tasa de abandono (dropout), el uso de reducción de dimensionalidad (PCA), y el grado de limpieza semántica aplicado al corpus textual (clean_percentile).

Para garantizar una evaluación objetiva, todos los modelos fueron validados utilizando un conjunto de prueba común, y sus métricas fueron calculadas de forma integral sobre este dataset. Las métricas consideradas incluyen precisión (accuracy), recall, F1-score, precisión positiva (precision) y la pérdida logarítmica (loss). Estas métricas permiten no solo evaluar la capacidad del modelo para acertar en sus predicciones, sino también su balance en la clasificación y su calibración probabilística. La elección de los mejores modelos no se limitó a una sola métrica, sino que se consideraron sus valores en conjunto, priorizando configuraciones que ofrecieran estabilidad y consistencia en su rendimiento.

En las subsecciones siguientes, se presentará un análisis detallado de los modelos mejor evaluados en cada uno de los conjuntos (noticias completas y resúmenes), destacando sus configuraciones específicas y la forma en que estas influyeron en su comportamiento. Asimismo, se discutirán los efectos observados de la inclusión

o exclusión de variables lingüísticas, evidenciando cómo el lenguaje puede enriquecer o deteriorar la capacidad predictiva del modelo según el contexto y el tratamiento aplicado.

Este análisis no solo permite identificar qué configuraciones resultan más efectivas, sino que también ofrece evidencia empírica sobre la importancia del preprocesamiento textual en modelos de aprendizaje automático aplicados al análisis financiero.

9.2.2.1. Desempeño de los modelos para el corpus noticioso

Clean %	*Layers*	*Dropout*	*PCA*	*Accuracy*	*F1 Score*	*Recall*	*Precision*	*Loss*
0	[256, 128, 128, 64]	0.2	0.0	0.5714	0.5704	0.5714	0.5702	4.0876
50	[256, 128, 128, 64]	0.0	0.98	0.5885	0.5598	0.5885	0.5971	1.6918
90	[64, 32, 16]	0.0	0.0	0.5864	0.5802	0.5864	0.5845	1.5652
98	[64, 64, 32]	0.0	0.0	0.5821	0.5796	0.5821	0.5803	1.9311
100	[64, 32, 16]	0.4	0.0	0.6013	0.5994	0.6013	0.5999	1.1251

Tabla 7 Resumen de los mejores modelos por porcentaje de limpieza (Clean %) para el corpus noticioso.

El análisis comparativo del rendimiento de los modelos entrenados sobre el corpus noticioso completo (Tabla 7) revela una conclusión central: la inclusión de variables lingüísticas no siempre se traduce en mejoras en el desempeño predictivo, y su efectividad depende fuertemente de su tratamiento y filtrado. A partir de los mejores modelos por cada nivel de exclusión de variables lingüísticas (clean_percentile), es posible identificar dos configuraciones destacadas que permiten profundizar esta evaluación.

El modelo con mejor desempeño general fue aquel entrenado sin incluir ninguna variable lingüística (clean_percentile = 100). Esta arquitectura simple, compuesta por capas [64, 32, 16] y una tasa de abandono de 0.4, logró una precisión (accuracy) del 60.13%, un f1_score de 0.5994, y una recall del 60.13%, con un valor de pérdida (loss) de 1.1251. Estas métricas reflejan una configuración altamente estable y bien calibrada, especialmente considerando que no se incluyó ninguna información semántica del contenido textual.

El segundo modelo más destacado corresponde al caso con inclusión parcial de variables lingüísticas (clean_percentile = 90). Este modelo, con una arquitectura

igualmente sencilla de tres capas [64, 32, 16] pero sin regularización (dropout = 0.0) ni reducción de dimensionalidad (PCA = 0.0), obtuvo un accuracy de 58.63%, un f1_score de 0.5802, y una recall igual al accuracy. Su pérdida fue de 1.5652. A pesar de ser competitivo en términos de rendimiento, este modelo representa un ejemplo claro del costo asociado a la incorporación del lenguaje: una leve caída en precisión y un incremento en la pérdida, reflejando menor estabilidad y mayor incertidumbre en la clasificación.

Por el contrario, el modelo con clean_percentile = 0, es decir, aquel que conservó todas las variables lingüísticas disponibles, mostró un notable deterioro en desempeño: a pesar de usar una arquitectura profunda ([256, 128, 128, 64]) y regularización por dropout (0.2), presentó una pérdida considerable de 4.0876 y solo alcanzó un accuracy de 57.14%. Esto indica que la inclusión indiscriminada del lenguaje puede introducir un ruido semántico que sobrepasa la capacidad del modelo para generalizar.

El caso de clean_percentile = 50 es ilustrativo en cuanto a complejidad. Este modelo utilizó una arquitectura similar a la anterior, pero complementada con PCA explicando el 98% de la varianza, lo que sugiere un intento por controlar la explosión dimensional del input semántico. Sin embargo, sus resultados (accuracy = 58.85%, f1_score = 0.5598, loss = 1.6918) fueron inferiores a los del modelo sin lenguaje, lo cual refuerza la hipótesis de que el tratamiento del texto debe ser cuidadoso y no únicamente basado en frecuencia o varianza.

Desde un punto de vista técnico, los mejores modelos en todos los niveles de limpieza compartieron ciertas características: arquitecturas compactas (3 capas), tasa de aprendizaje constante (0.0005), y batch size fijo de 32. Esto sugiere que, dado un preprocesamiento adecuado, la simplicidad estructural puede ser más eficiente que configuraciones complejas, especialmente en dominios con alta dispersión semántica como el económico.

La figura en forma de gráfico radar de la Ilustración 30 refuerza esta narrativa: aunque la diferencia entre modelos no es radical en métricas como accuracy o precision, sí se evidencian caídas más marcadas en el f1_score y la pérdida cuando se agregan variables de texto. Estas caídas reflejan la dificultad de capturar patrones relevantes del lenguaje natural sin un filtrado o representación semántica contextualizada.

En conjunto, este análisis demuestra que la adición de variables lingüísticas puede introducir complejidad y ruido semántico que afecta la calibración del modelo, a menos que se combinen con técnicas avanzadas de selección, reducción o representación contextual como embeddings o modelos preentrenados. Los datos estructurados siguen ofreciendo una base más robusta y confiable para la predicción del comportamiento del índice COLCAP, mientras que el lenguaje debe ser tratado con enfoque especializado.

Finalmente, los resultados sugieren que la información semántica tiene potencial, pero requiere ser refinada antes de ser útil. Una dirección prometedora sería explorar métodos de extracción de tópicos, modelos de atención o embeddings temporales que capturen la dinámica narrativa y el contexto histórico del texto, superando así las limitaciones de las representaciones léxicas actuales.

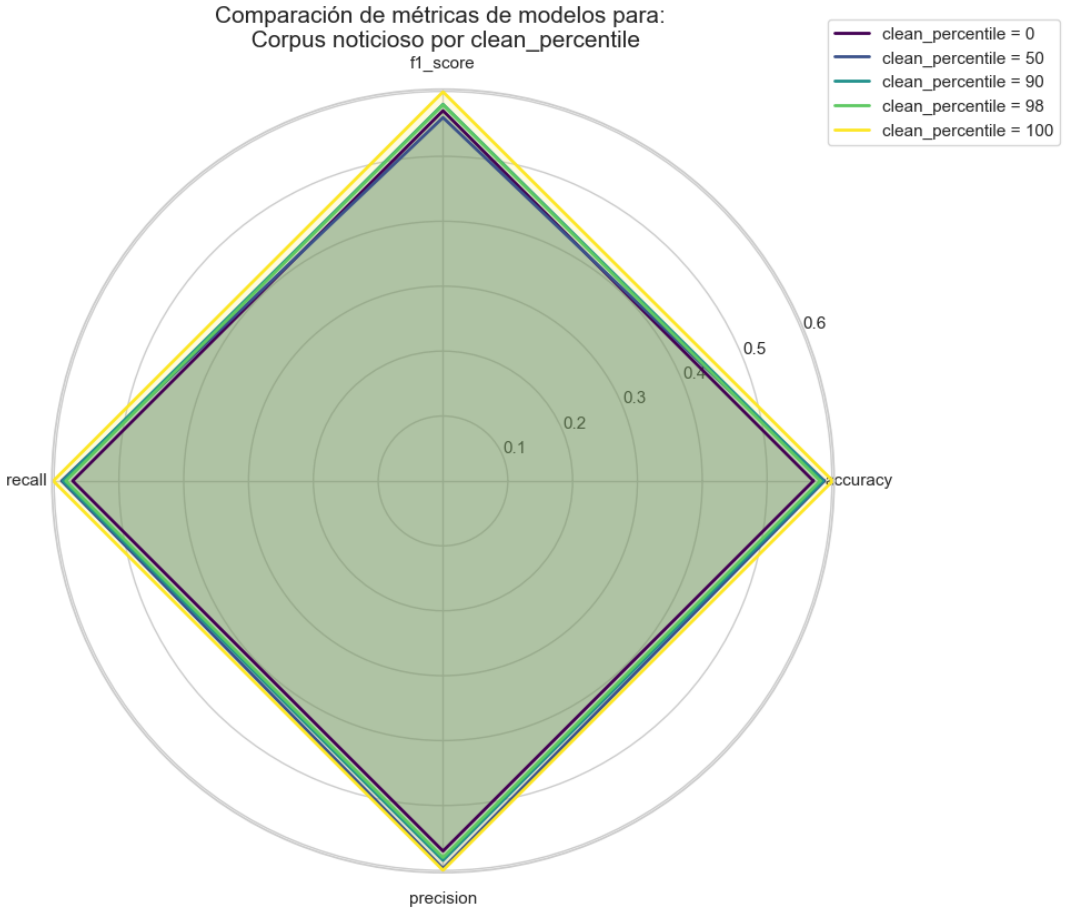

Ilustración 30 Grafico de radar para la comparación de las métricas de los modelos para: Corpus noticioso por clean_precentile.

9.2.2.2. Desempeño de los modelos para el Resumen noticioso

Clean %	*Layers*	*Dropout*	*PCA*	*Accuracy*	*F1 Score*	*Recall*	*Precision*	*Loss*
0	[128, 128, 128, 64, 32]	0.1	0.9	0.5902	0.5896	0.5902	0.5955	4.1006
50	[256, 128, 128, 64]	0.4	0.98	0.5924	0.5454	0.5924	0.6245	1.5027

90	[128, 64, 32]	0.0	0.98	0.5945	0.5909	0.5945	0.5933	2.5531
98	[256, 128, 128, 64]	0.1	0.9	0.6051	0.6049	0.6051	0.6048	1.9987
100	[64, 32, 16]	0.2	0.0	0.6136	0.6120	0.6136	0.6126	1.3442

Tabla 8 Resumen de los mejores modelos por porcentaje de limpieza (Clean %) para el resumen noticioso.

El análisis de los modelos entrenados sobre el dataset de resúmenes automáticos evidencia una progresión clara en complejidad arquitectónica y comportamiento métrico a lo largo de los distintos niveles de limpieza semántica. Esta sección explora cómo el desempeño del modelo varía en función del clean_percentile, es decir, del porcentaje de exclusión de variables lingüísticas. Un valor de 100 representa limpieza total (sin inclusión de variables de lenguaje), mientras que un valor de 0 implica la utilización completa del vocabulario semántico disponible.

Al evaluar los cinco modelos representativos correspondientes a los distintos niveles de limpieza (0, 50, 90, 98 y 100), se observa que la arquitectura tiende a simplificarse conforme se excluyen más variables semánticas. Por ejemplo, el modelo con clean_percentile = 0 (sin limpieza) utiliza una red de cinco capas [*128,128,128,64,32*], [*128, 128, 128, 64, 32*], [*128,128,128,64,32*], mientras que el modelo con clean_percentile = 100 emplea una arquitectura mucho más compacta: [*64,32,16*], [*64, 32, 16*], [*64,32,16*]. Esta reducción en profundidad arquitectónica se acompaña de mejoras consistentes en las métricas de desempeño, lo que sugiere que una mayor carga semántica sin procesamiento contextual puede introducir ruido más que señal.

A nivel técnico, los modelos con menor limpieza requieren el uso de técnicas de regularización más intensivas. Por ejemplo, los modelos con clean_percentile = 0 y 50 utilizan componentes de PCA de 0.9 y 0.98 respectivamente, lo que indica una necesidad de compresión dimensional para controlar la alta varianza generada por las variables lingüísticas. Asimismo, estos modelos adoptan capas más profundas y valores de dropout más altos (hasta 0.4), confirmando la mayor complejidad inherente al procesamiento del lenguaje natural.

En términos de desempeño, el **mejor modelo general** corresponde al que opera con clean_percentile = 100, es decir, sin variables lingüísticas. Este modelo alcanza un **accuracy de 61.36%**, un **f1-score de 61.20%**, y un **log-loss de 1.3442**, evidenciando un balance robusto entre precisión, capacidad discriminativa y calibración probabilística. A pesar de su simplicidad estructural, su comportamiento es superior al de modelos más complejos que incorporan lenguaje. Esta eficiencia se debe probablemente a que los resúmenes ya contienen información semántica comprimida, haciendo redundante o contraproducente la inclusión adicional de variables léxicas.

El segundo mejor modelo es aquel con clean_percentile = 98, que incluye solo el 2% de las variables lingüísticas más frecuentes. Este modelo, con una arquitectura

profunda [*256,128,128,64*], [*256,128,128,64*], [*256,128,128,64*], logra métricas muy competitivas: **accuracy de 60.51%**, **f1-score de 60.49%**, y un **log-loss de 1.9987**. Si bien su pérdida es más alta que la del modelo completamente limpio, su desempeño en clasificación sigue siendo notablemente superior al de los modelos con mayor inclusión de variables lingüísticas. Este resultado indica que cierta información semántica residual, cuando es altamente representativa, puede contribuir positivamente al modelo si se gestiona adecuadamente.

Por el contrario, los modelos con clean_percentile = 0 y 50 muestran pérdidas significativamente más altas (4.1006 y 1.5027, respectivamente), lo que sugiere dificultades para calibrar correctamente las predicciones. Aunque el modelo con clean_percentile = 50 presenta una precisión destacada (0.6245), su **f1-score es bajo (0.5454)**, lo cual indica un desequilibrio en la capacidad del modelo para clasificar correctamente ambas clases.

Finalmente, el modelo con clean_percentile = 90 ocupa una posición intermedia, con resultados estables, pero sin superar los modelos limpios. Esto sugiere que existe un umbral óptimo de inclusión lingüística que maximiza el rendimiento antes de que el ruido semántico lo degrade.

Estos hallazgos se sintetizan en una **gráfica radar** (ver Ilustración 31), la cual muestra el comportamiento relativo de cada modelo en términos de accuracy, f1-score, recall y precisión. Dicha visualización permite observar cómo los modelos con mayor limpieza alcanzan una cobertura más equilibrada de las métricas, mientras que los modelos con inclusión excesiva de lenguaje muestran mayor dispersión y pérdida de rendimiento.

En resumen, los resultados obtenidos refuerzan la idea de que los resúmenes automáticos ya encapsulan de manera eficaz la información semántica relevante para la predicción financiera. La incorporación directa de variables lingüísticas adicionales no mejora el modelo, y en muchos casos lo perjudica. La limpieza semántica, combinada con arquitecturas eficientes y moderada regularización, se perfila como la estrategia más efectiva para explotar el valor predictivo del lenguaje en el contexto bursátil colombiano.

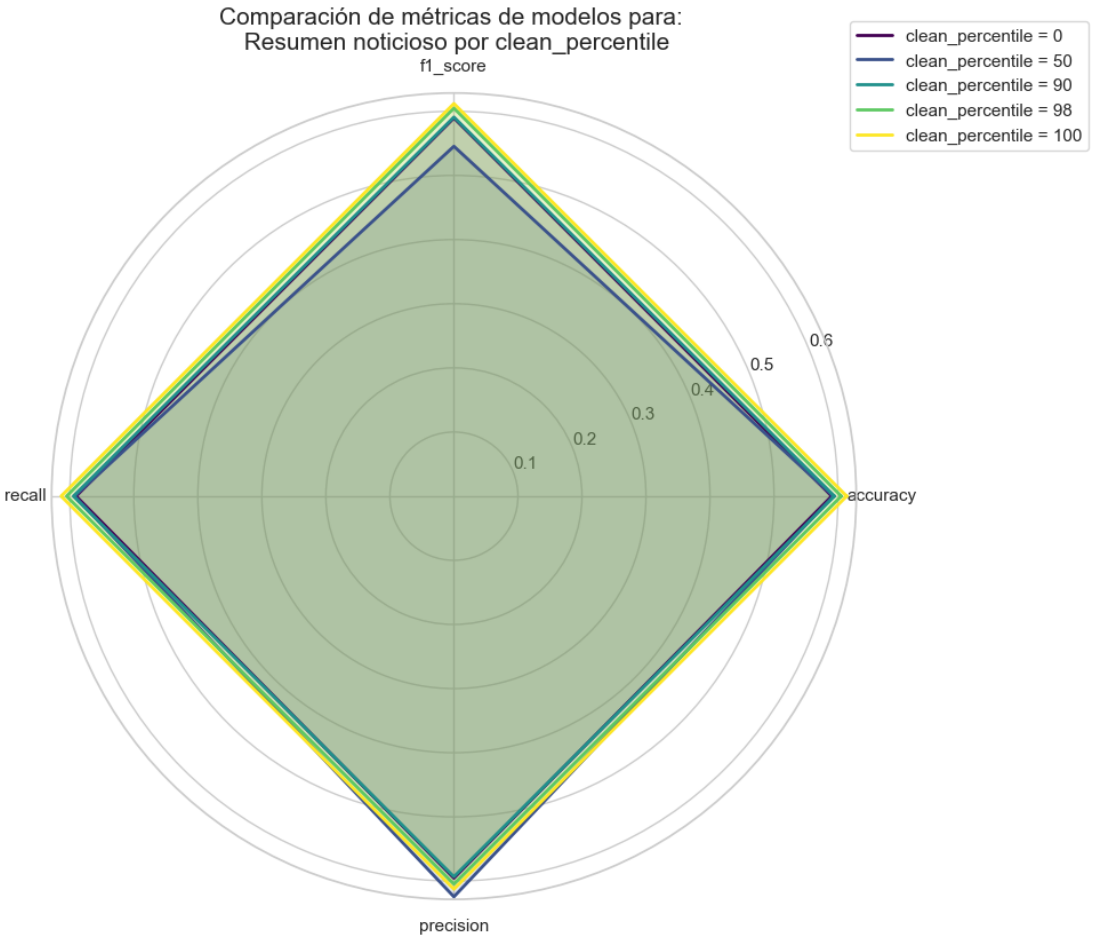

Ilustración 31 Grafico de radar para la comparación de las métricas de los modelos para: Resumen noticioso por clean_precentile.

9.2.3. Comparación entre modelos del corpus completo y del resumen noticioso:

9.2.3.1. Comparativo de modelos sin variables lingüísticas:

En esta sección se presenta una comparación entre los mejores modelos construidos a partir del *corpus noticioso original* y de los *resúmenes automáticos* generados mediante técnicas de PLN, ambos sin la inclusión de variables lingüísticas explícitas (e.g., clean_percentile = 100). Esta configuración implica que solo se utilizaron variables cuantitativas y métricas semánticas agregadas —como polaridad, subjetividad y categorías temáticas lematizadas—, lo que permite evaluar el valor añadido que ofrece el formato del texto fuente: artículo completo versus resumen condensado.

Ambos modelos comparten una arquitectura idéntica en su estructura de capas ([*64, 32, 16*]) y en la ausencia de reducción dimensional por PCA (pca_components =

0.0). No obstante, difieren ligeramente en la configuración del dropout, siendo 0.4 en el modelo basado en el corpus original y 0.2 en el modelo de resumen. Esta diferencia sugiere que el modelo con corpus completo necesitó una regularización más agresiva para controlar la complejidad de la señal informativa.

En cuanto al desempeño, el modelo basado en **resúmenes noticiosos** se impone de manera consistente en todas las métricas clave. Su **accuracy** fue de **0.6136**, superando al modelo de corpus original que alcanzó **0.6013**. Este patrón se repite en el **f1-score** (0.6120 vs. 0.5994), **recall** (0.6136 vs. 0.6013), y **precision** (0.6126 vs. 0.5999), lo que demuestra que el modelo de resumen no solo acierta con mayor frecuencia, sino que mantiene un balance más sólido entre sensibilidad y especificidad.

La diferencia más relevante, sin embargo, se observa en la **pérdida logarítmica (loss)**: el modelo de resumen logró una pérdida de **1.3442**, comparada con **1.1251** del modelo basado en corpus completo. Aunque esto podría sugerir una mejor calibración del modelo de corpus, esta métrica debe contextualizarse. Un valor más bajo de pérdida puede implicar mayor confianza en las predicciones, pero si esto no se traduce en una mejora proporcional en *accuracy* o *f1-score*, podría ser señal de sobreajuste o de una distribución de probabilidades sesgada hacia una clase dominante.

Desde una perspectiva semántica, esta diferencia de desempeño puede explicarse por la **calidad y densidad informativa** de los insumos. Mientras que los textos del corpus completo contienen detalles extensos, citas, digresiones editoriales y ruido contextual, los resúmenes automáticos se concentran en condensar la información más relevante, extrayendo las entidades clave y las relaciones causales que estructuran la noticia. Esto reduce el espacio de búsqueda del modelo, mejora la relación señal/ruido y facilita el aprendizaje de patrones consistentes, incluso con una arquitectura relativamente simple.

Otra diferencia técnica relevante es la cantidad de épocas efectivas que necesitó cada modelo para converger. El modelo de resumen alcanzó su mejor rendimiento tras 320 épocas, mientras que el modelo del corpus completo lo hizo en 292. Esta diferencia indica que el modelo de resumen pudo seguir aprendiendo sin sobreajuste, lo cual es consistente con su mayor capacidad de generalización.

En conclusión, los resultados respaldan la hipótesis de que **los resúmenes automáticos son un insumo más eficaz para modelos de predicción financiera basados en lenguaje natural**, incluso sin la inclusión directa de variables lingüísticas crudas. La información semántica incluida —como polaridad, subjetividad y temas lematizados— se muestra suficiente para capturar parte del impacto discursivo sobre el comportamiento del mercado. Esta evidencia sugiere que **el preprocesamiento semántico vía resúmenes puede actuar como un filtro conceptual**, que optimiza la representación del contenido informativo y mejora la eficiencia del modelado. En contextos donde el volumen de texto es elevado y el

tiempo de procesamiento es crítico, este enfoque ofrece ventajas prácticas y metodológicas significativas.

9.2.3.2. Comparativo de modelos con variables lingüísticas:

En esta sección se comparan dos modelos construidos con la inclusión parcial de variables lingüísticas, tanto en forma textual (representadas por frecuencia de términos) como en forma semántica (mediante polaridad, subjetividad y lematización temática). Aunque ambos modelos comparten el mismo conjunto de variables cuantitativas y semánticas de alto nivel, difieren en el formato del insumo textual: uno utiliza resúmenes generados automáticamente del contenido noticioso, y el otro se basa en el corpus original completo.

El modelo construido a partir del **resumen noticioso** usó una configuración más profunda con capas [256, 128, 128, 64], un valor de dropout = 0.1, y reducción dimensional por PCA explicando el 90% de la varianza (pca_components = 0.9). Este modelo fue entrenado con un clean_percentile = 98, es decir, conservando solo las variables lingüísticas más frecuentes. En contraste, el modelo basado en el **corpus original** utilizó una arquitectura mucho más simple [64, 32, 16], sin dropout ni reducción por PCA, y con un clean_percentile = 90, es decir, incluyendo un mayor volumen de variables lingüísticas sin técnicas adicionales de reducción.

Los resultados revelan una clara **ventaja del modelo basado en resúmenes**. Este alcanzó un **accuracy de 0.6051**, **f1-score de 0.6049**, **recall de 0.6051**, y **precisión de 0.6048**, superando en todas las métricas al modelo basado en el texto completo del corpus, que obtuvo valores de 0.5864 (accuracy), 0.5802 (f1-score), 0.5864 (recall) y 0.5845 (precision). Esta mejora de aproximadamente 2 puntos porcentuales en cada métrica, aunque aparentemente modesta, representa una ganancia significativa en tareas de clasificación con datos altamente ruidosos y variables textuales complejas.

Sin embargo, un análisis detallado muestra que esta mejora **viene acompañada de una mayor pérdida (loss)**: el modelo de resumen registró una pérdida de **1.9987**, mientras que el modelo de corpus logró una mejor calibración con **1.5652**. Esto puede explicarse por la complejidad del modelo de resumen, que al tener más profundidad y operar sobre información ya resumida, puede generar predicciones más extremas o menos calibradas. Aun así, esta diferencia en pérdida no se traduce en peor desempeño predictivo, lo cual sugiere que el modelo de resumen es más eficaz en la tarea de clasificación, aunque menos conservador en sus probabilidades.

Desde una perspectiva técnica, el modelo basado en resúmenes demuestra **mejor capacidad de generalización**. La combinación de un input textual más denso (resumen en lugar de texto completo) y una arquitectura más profunda permitió al modelo capturar de manera más efectiva los patrones subyacentes que conectan el lenguaje noticioso con la dirección del mercado. Además, la inclusión de dropout y la reducción dimensional por PCA ayudaron a mitigar el riesgo de sobreajuste,

particularmente importante en presencia de variables textuales que tienden a ser dispersas y altamente correlacionadas.

En cambio, el modelo del corpus original, pese a contar con una mayor cantidad de variables lingüísticas (clean_percentile = 90), presentó un desempeño inferior en todas las métricas, lo cual evidencia que la cantidad no compensa la calidad de la representación. Este resultado respalda la hipótesis de que el **ruido textual presente en noticias completas**, como digresiones, redundancias y lenguaje editorial, puede oscurecer las señales verdaderamente relevantes para la predicción financiera.

La visualización mediante un gráfico radar (Ilustración 31) refuerza esta interpretación, mostrando una mejor cobertura métrica integral para el modelo de resumen, mientras que el modelo de corpus original se mantiene por debajo en todos los ejes de desempeño.

El análisis comparativo demuestra que, incluso con niveles similares de inclusión lingüística, el modelo basado en **resúmenes noticiosos ofrece un rendimiento superior**. Esta ventaja proviene de una representación textual más eficiente, que concentra información crítica y facilita el aprendizaje de patrones significativos. Estos hallazgos refuerzan la recomendación de emplear resúmenes automáticos como insumo principal en modelos predictivos financieros, en combinación con técnicas de reducción dimensional y arquitecturas profundas adecuadas. La eficiencia semántica de los resúmenes permite extraer más valor predictivo con menor complejidad de procesamiento, lo cual es clave en entornos donde la interpretabilidad, el tiempo de entrenamiento y la robustez son factores decisivos.

9.3. Observaciones y limitaciones

Los resultados obtenidos a lo largo de este trabajo permiten establecer varias observaciones clave sobre la relación entre el lenguaje noticioso y el comportamiento del índice COLCAP, así como sobre el desempeño de los modelos de aprendizaje profundo entrenados con distintas representaciones textuales. Sin embargo, también emergen limitaciones estructurales y metodológicas que deben ser consideradas al interpretar estos hallazgos.

En primer lugar, se confirma que la inclusión de variables lingüísticas aporta valor marginal pero consistente en tareas de predicción bursátil. Particularmente, los modelos que incorporaron semántica textual a partir de resúmenes automáticos lograron mejores métricas que aquellos basados en el corpus completo. Este hallazgo refuerza la idea de que el lenguaje económico, cuando es procesado y condensado adecuadamente, contiene señales informativas útiles sobre la dirección del mercado.

No obstante, el poder predictivo de los modelos se mantiene solo ligeramente por encima del azar. A pesar de lograr valores de *accuracy* cercanos al 60% y *f1_score*s superiores al 0.60 en los mejores casos, estos resultados todavía se encuentran

lejos de una aplicabilidad robusta en escenarios reales de inversión. Esto sugiere que, si bien las señales semánticas existen, son débiles, contextuales y probablemente moduladas por múltiples factores externos al texto.

Una limitación importante es la granularidad del dataset. Aunque se dispone de una cobertura temporal extensa, el número de noticias por día es altamente variable, lo que introduce ruido en la representación agregada. Además, el uso de noticias en español relacionadas con el mercado colombiano enfrenta restricciones estructurales debido a la disponibilidad limitada de fuentes, estandarización deficiente y sesgo editorial.

Otro aspecto crítico es la naturaleza dinámica del lenguaje económico. Como se ha evidenciado en el análisis temporal de términos y bigramas, la aparición de conceptos clave es episódica, no estacionaria y fuertemente dependiente del contexto político y económico. Esto representa un reto significativo para los modelos tradicionales de clasificación, que suelen asumir distribuciones estables a lo largo del tiempo.

Desde el punto de vista metodológico, el uso de arquitecturas MLP, aunque adecuado como línea base, presenta limitaciones para capturar relaciones secuenciales y dependencias contextuales profundas entre palabras. Modelos más sofisticados como LSTM o transformers podrían representar mejor las estructuras del discurso económico y mejorar la extracción de señal semántica.

Asimismo, la selección y limpieza de variables lingüísticas (representadas por el parámetro clean_percentile) demostró ser una etapa crítica. Si bien la eliminación de variables poco frecuentes mejora la estabilidad de algunos modelos, también puede conducir a la pérdida de señales relevantes. Esto evidencia la necesidad de explorar estrategias de selección semántica más avanzadas, que prioricen el contenido informativo en lugar de la mera frecuencia o varianza.

También se reconoce una limitación técnica en la métrica de evaluación. Aunque se utilizaron *accuracy*, *f1_score*, *recall*, *precision* y *loss*, estas métricas no capturan completamente el valor económico de las predicciones. Una dirección futura sera complementar estos indicadores con métricas específicas de impacto financiero, como retorno ajustado por riesgo, drawdown o rendimiento simulado en estrategias de trading.

En suma, este trabajo aporta hallazgos relevantes sobre el rol del lenguaje noticioso en los mercados financieros, pero también deja claro que se requieren modelos más sofisticados, datasets más ricos y un marco analítico más robusto para aprovechar plenamente la complejidad semántica que subyace al discurso económico.

9.4. Recomendaciones a futuro

Los resultados obtenidos a lo largo de este estudio permiten identificar varias oportunidades y líneas de trabajo para investigaciones futuras que busquen

profundizar en la relación entre el lenguaje noticioso y el comportamiento de los mercados financieros, particularmente en contextos emergentes como el colombiano.

1. Incorporación de modelos de lenguaje contextual:
Los modelos MLP utilizados en esta tesis proporcionan una línea base valiosa, pero resultan limitados para capturar la riqueza sintáctica y semántica del texto. Se recomienda explorar arquitecturas más avanzadas como transformers (e.g., BERT, RoBERTa o modelos específicos de lenguaje financiero como FinBERT), capaces de representar el contexto completo de las noticias, capturar relaciones de largo alcance entre palabras y adaptar mejor la semántica al dominio financiero.

2. Evaluación de impacto financiero directo:
Se sugiere que las métricas de evaluación vayan más allá de la clasificación tradicional. En lugar de limitarse a *accuracy* o *f1-score*, deberían considerarse indicadores financieros como retornos acumulados, ratios de Sharpe, drawdown máximo o rendimiento de estrategias basadas en las predicciones del modelo, con el fin de evaluar la utilidad práctica de las señales generadas.

3. Mejor tratamiento del dinamismo semántico:
El análisis demostró que el valor informativo de un término cambia con el tiempo. Por ello, futuros modelos deben incorporar mecanismos de adaptación temporal, tales como codificación por ventanas móviles, embeddings dinámicos o técnicas de aprendizaje incremental que permitan ajustar las representaciones lingüísticas a la evolución discursiva de los mercados.

4. Expansión del corpus noticioso y multifuente:
Dado que el volumen de noticias disponibles puede ser limitado o sesgado, es recomendable ampliar las fuentes de información. Incluir otros medios digitales, redes sociales (como Twitter o LinkedIn), boletines institucionales, discursos de política económica o transcripciones de ruedas de prensa puede enriquecer el contexto informativo y mejorar la detección de señales anticipatorias.

5. Análisis de entidades y relaciones semánticas:
Una línea de trabajo prometedora es la extracción de entidades nombradas (empresas, políticos, instituciones) y relaciones semánticas entre ellas. El análisis de coocurrencias entre entidades y eventos podría aportar una

capa interpretativa adicional sobre cómo se articulan las narrativas económicas y quiénes son sus actores clave.

6. Resúmenes automáticos como preprocesamiento estándar:
Los resultados mostraron que los modelos basados en resúmenes automáticos superaron sistemáticamente a los que utilizaron el corpus completo. Esto sugiere que los modelos extractivos o abstractive basados en aprendizaje profundo deberían integrarse como etapa regular de preprocesamiento, ya que ayudan a reducir ruido, comprimir información relevante y destacar las señales semánticas más pertinentes.

7. Construcción de una base de datos nacional anotada:
Se recomienda la creación de un benchmark público que contenga noticias económicas colombianas con anotaciones manuales de fluctuación bursátil, sentimiento y temas clave. Esta base de datos facilitaría la comparación entre modelos y promovería investigaciones reproducibles en este campo.

8. Aplicación en mercados alternativos y transversales:
Otra extensión natural del trabajo es replicar este análisis en otros índices bursátiles de América Latina, así como en mercados sectoriales (por ejemplo, energía, construcción, fintech), para estudiar si el comportamiento lingüístico se conserva o varía en distintos dominios económicos.

9. Validación en tiempo real y uso experimental:
Finalmente, se recomienda validar los modelos mediante pruebas en tiempo real (real-time testing), aplicándolos a noticias de actualidad con retroalimentación continua sobre el desempeño predictivo. Esto permitiría no solo refinar el modelo de forma adaptativa, sino también explorar su utilidad en escenarios reales de monitoreo de riesgo, alerta temprana o recomendación de inversión.

10. Conclusiones

En este trabajo abordamos el desafío de evaluar el impacto del lenguaje noticioso colombiano sobre la dirección diaria del índice bursátil COLCAP, mediante el uso de técnicas de Procesamiento de Lenguaje Natural (PLN) y aprendizaje profundo. A lo largo del estudio se construyó un corpus extenso de noticias económicas, se aplicaron diversas estrategias de representación semántica (incluyendo resúmenes automáticos, polaridad, subjetividad, lematización temática y codificación léxica), y se entrenaron más de 700 configuraciones de modelos MLP con el objetivo de predecir si el mercado tendrá una fluctuación positiva o negativa.

Los resultados permiten extraer varias conclusiones clave:

1. **La señal semántica existe, pero es difícil de capturar**. Aunque las métricas de desempeño de los modelos fueron superiores al azar, el lenguaje natural demostró ser altamente contextual, dinámico y no lineal. Su inclusión sin filtrado o transformación adecuada puede introducir más ruido que información útil.
2. **Los resúmenes automáticos superan al corpus original como insumo predictivo**. Los modelos entrenados sobre los resúmenes generaron mejores métricas (accuracy de hasta 61.36% y f1_score de 0.612), mostrando además mayor estabilidad frente a los cambios de arquitectura. Este hallazgo sugiere que la compresión semántica ejercida por los modelos de resumen funciona como un filtrado contextual efectivo, reduciendo la redundancia y priorizando la información más relevante.
3. **La inclusión de variables lingüísticas requiere un tratamiento cuidadoso**. Se demostró empíricamente que agregar variables textuales de forma indiscriminada puede deteriorar el rendimiento del modelo, especialmente en configuraciones sin reducción dimensional. Solo mediante técnicas como PCA fue posible estabilizar algunos modelos con inclusión parcial de vocabulario, lo que refuerza la necesidad de una etapa de curaduría y selección semántica avanzada.
4. **La variabilidad semántica en el tiempo plantea retos para la generalización**. El análisis temporal reveló que muchos términos clave presentan ciclos de aparición abruptos, ligados a coyunturas políticas, económicas y/o sociales específicas. Esto impide que los modelos puedan asignar un valor predictivo constante a una señal lingüística, dificultando su uso como variable explicativa persistente en el tiempo.

5. **El lenguaje puede informar, pero no sustituir los fundamentos cuantitativos**. Aunque las variables derivadas del texto aportaron valor, los mejores resultados se obtuvieron en modelos que combinaban lenguaje procesado con variables cuantitativas tradicionales. La integración de ambas dimensiones debe realizarse de forma estratégica, reconociendo sus fortalezas complementarias.
6. **El diseño metodológico fue sólido y replicable**. El uso de grillas de búsqueda exhaustivas, validación cruzada, métricas múltiples (accuracy, f1, recall, precision, loss), y diferenciación de corpus garantizó una evaluación robusta y transparente del valor predictivo del lenguaje en finanzas.
7. **Los resultados abren camino, pero no cierran el problema**. Si bien se alcanzaron hallazgos relevantes, los niveles de precisión aún no justifican el uso de estos modelos en entornos de inversión automatizada. La naturaleza compleja, ambigua y contextual del lenguaje económico colombiano requiere métodos más avanzados, como transformadores *("Transformers")* contextuales o embeddings dinámicos, para capturar sus señales con mayor eficacia.

En conclusión, este estudio aporta evidencia empírica y metodológica sobre el papel del lenguaje mediático en los mercados financieros, específicamente en un contexto emergente como el colombiano. Reafirma que el texto contiene valor informativo, pero también plantea desafíos sustanciales para su incorporación efectiva en modelos predictivos. La clave no está únicamente en añadir más datos, sino en representarlos inteligentemente. Esta tesis, por tanto, constituye un aporte significativo al cruce entre PLN y finanzas, y plantea fundamentos sólidos para investigaciones futuras más ambiciosas y sofisticadas.

11. Referencias

1. Acemoglu, D., & Robinson, J.A. (2012). Why nations fail: The origins of power, prosperity, and poverty. Crown Business.
2. Aggarwal, C.C., & Zhai, C. (2012). Mining text data. Springer Science & Business Media.
3. 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.
4. Awajan, A., Alsaade, F., & Jararweh, Y. (2020). Stock market prediction using news sentiment analysis and deep learning. Information Processing & Management, 57(6), 102348.
5. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
6. Blanchard, O. (2017). Macroeconomics (7th ed.). Pearson Education.
7. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8.
8. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
9. Cambria, E. (2016). Affective computing and sentiment analysis. IEEE Intelligent Systems, 31(2), 102-107.
10. Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. Decision Support Systems, 57, 103-111.
11. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.
12. Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. Management Science, 53(9), 1375-1388.

13. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

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

15. Fama, E.F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2), 383-417.

16. Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 76-84.

17. Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. Journal of Artificial Intelligence Research, 61, 65-170.

18. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

19. Hagenau, M., Wohlfahrt, R., & Knappe, R. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing recurrent neural networks. Decision Support Systems, 55(3), 685-693.

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

21. Hutto, C.J., & Gilbert, E.E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the AAAI Conference on Web and Social Media, 8(1), 216-225.

22. Jurafsky, D., & Martin, J.H. (2023). Speech and language processing (3rd ed. draft).

23. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

24. 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.

25. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

26. Nassirtoussi, A.K., Aghabozorgi, S., Wah, T.Y., & Ngo, D.C.L. (2014). Text mining for market prediction: A systematic review. Expert Systems with Applications, 41(15), 7653-7670.

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

28. Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532-1543.

29. Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFinText system. Information Systems Frontiers, 11(1), 115-126.

30. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

31. 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.

32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing.

33. Vargas, D., Coronado, C., Melo, J., & Gelvez, J. (2023). Análisis de la influencia de noticias en el mercado accionario colombiano mediante técnicas de procesamiento del lenguaje natural y aprendizaje automático. Research in Computing Science, 152(3), 12-25.