Análisis comparativo multiclase de emociones de inversionistas y noticias financieras del mercado S&P 500 y Tesla durante el periodo de enero de 2022 – abril de 2026, mediante fine-tuning de un modelo de lenguaje preentrenado especializado en dominio financiero

Edison Fabián Moreno Hernández

Víctor Alfonso Rodríguez Guzmán

Universidad Ean

Facultad de Ingeniería

Programa de Maestría en Ciencia de Datos

Bogotá, Colombia

15/10/2026

Análisis comparativo multiclase de emociones de inversionistas y noticias financieras del mercado S&P 500 y Tesla durante el periodo de enero de 2022 – abril de 2026, mediante fine-tuning de un modelo de lenguaje preentrenado especializado en dominio financiero

Edison Fabián Moreno Hernández

Víctor Alfonso Rodríguez Guzmán

Trabajo de grado presentado como requisito para optar al título de:

Magíster en Ciencia de Datos

Director:

José Alexander Fuentes Montoya

Modalidad:

Monografía

Universidad Ean

Facultad de Ingeniería

Programa de Maestría en Ciencia de Datos

Bogotá, Colombia

15/10/2026

Nota de aceptación:

Firma del jurado

Firma del jurado

Firma del director del trabajo de grado

Bogotá, 15/10/2026

(Dedicatoria o frase. Página opcional)

Dedicatoria:

A nuestras familias, por acompañarnos en este camino de aprendizaje y crecimiento.

A quienes han contribuido en nuestra formación académica y profesional, inspirándonos a comprender que el conocimiento adquiere verdadero valor cuando se transforma en nuevas preguntas, aprendizajes y oportunidades.

Agradecimientos

Expresamos nuestro sincero agradecimiento a Miguel Ángel González Curbelo, tutor del seminario de investigación, por su orientación académica y metodológica durante la formulación y estructuración de este trabajo. Sus aportes, recomendaciones y conocimientos contribuyeron de manera significativa a fortalecer las bases de nuestra investigación y a orientar su desarrollo con rigurosidad.

Asimismo, agradecemos especialmente a José Alexander Fuentes Montoya, director del trabajo de grado, por su disposición, experiencia y permanente orientación durante las diferentes etapas de este proceso investigativo. Sus observaciones, sugerencias y aportes fueron fundamentales para enriquecer el análisis, fortalecer la calidad académica del trabajo y avanzar hacia su consolidación.

A la Universidad Ean y a la Maestría en Ciencia de Datos, expresamos nuestra gratitud por brindarnos los conocimientos, recursos y herramientas conceptuales, metodológicas y tecnológicas que hicieron posible el desarrollo de esta investigación y que han contribuido al fortalecimiento de nuestras competencias profesionales e investigativas.

A nuestras familias, agradecemos profundamente su comprensión, respaldo y motivación constante durante esta etapa académica. Su presencia ha sido fundamental para afrontar con compromiso y perseverancia los retos que implicó la realización de este trabajo.

Finalmente, extendemos nuestro agradecimiento a todas las personas que, de manera directa o indirecta, aportaron conocimientos, experiencias, orientaciones y perspectivas que contribuyeron al desarrollo y culminación de esta investigación.

Resumen

La presente investigación analiza comparativamente las categorías emocionales presentes en noticias financieras institucionales y comentarios de inversionistas retail relacionados con el índice S&P 500 y Tesla Inc. durante el periodo comprendido entre enero de 2022 y abril de 2026. Ante la necesidad de estudiar con mayor granularidad el discurso financiero, se adoptó un enfoque cuantitativo, descriptivo y comparativo. Se construyó un corpus inicial de 10.000 textos, del cual se conservaron 7.752 después del proceso de preprocesamiento. Mediante FinBERT base y reglas heurísticas de mapeo se generaron pseudoetiquetas correspondientes a cinco categorías emocionales: enojo, entusiasmo, miedo, neutralidad y tristeza. Posteriormente, FinBERT fue ajustado mediante fine-tuning y evaluado con métricas de clasificación multiclase. En el conjunto de prueba independiente de 1.163 textos, el modelo obtuvo un Macro-F1 de 0,5564 y una exactitud de 0,8813.

La comparación de las pseudoetiquetas emocionales entre fuentes retail e institucionales mostró una asociación estadísticamente significativa entre el tipo de fuente y la categoría emocional, con $\chi^2(4) = 126{,}422$ y $p < 0{,}0001$. No obstante, el V de Cramér de 0,128 indicó que la magnitud de la asociación fue limitada, mientras que la divergencia KL simétrica de 0,0356 se utilizó como medida descriptiva complementaria de separación entre las distribuciones. Para la contrastación de H2, las predicciones del modelo multiclase agrupadas en entusiasmo y no entusiasmo obtuvieron un Macro-F1 de 0,8624, frente a 0,8311 del clasificador binario independiente. Sin embargo, la diferencia no fue estadísticamente significativa (McNemar, $p = 0{,}8955$), por lo que H2 no fue respaldada.

Los resultados muestran que la clasificación multiclase permite caracterizar y comparar distribuciones emocionales con mayor granularidad dentro del corpus analizado, aunque

su interpretación está condicionada por el desbalance entre clases, la calidad de las pseudoetiquetas, la cobertura de las fuentes y el alcance lingüístico y temporal del estudio.

Palabras clave: análisis de emociones, textos financieros, procesamiento del lenguaje natural, clasificación multiclase, pseudoetiquetado, FinBERT, inversionistas retail.

Abstract

This study comparatively analyzes emotional categories in institutional financial news and retail investor comments related to the S&P 500 index and Tesla Inc. between January 2022 and April 2026. To examine financial discourse at a more granular level, a quantitative, descriptive, and comparative approach was adopted. An initial corpus of 10,000 texts was constructed, of which 7,752 were retained after preprocessing. FinBERT base predictions and heuristic mapping rules were used to generate pseudo-labels for five emotional categories: anger, enthusiasm, fear, neutrality, and sadness. FinBERT was subsequently fine-tuned and evaluated using multiclass classification metrics. On an independent test set of 1,163 texts, the model achieved a Macro-F1 score of 0.5564 and an accuracy of 0.8813.

The comparison of emotional pseudo-label distributions between retail and institutional sources showed a statistically significant association between source type and emotional category, $\chi^2(4) = 126.422$, $p < 0.0001$. However, Cramér's V of 0.128 indicated that the magnitude of the association was limited, while the symmetric Kullback–Leibler divergence of 0.0356 was used as a complementary descriptive measure of distributional separation. For H2, the predictions of the multiclass model grouped into enthusiasm and non-enthusiasm achieved a Macro-F1 score of 0.8624, compared with 0.8311 for the independently trained binary classifier. However, the difference was not statistically significant (McNemar, $p = 0.8955$), and H2 was therefore not supported.

The findings indicate that multiclass classification can characterize and compare emotional distributions at a finer level of granularity within the analyzed corpus. Nevertheless, the interpretation of the results is constrained by class imbalance, the quality of the pseudo-labels, source coverage, and the linguistic and temporal scope of the study.

Keywords: emotion analysis, financial texts, natural language processing, multiclass classification, pseudo-labeling, FinBERT, retail investors.

Contenido

Pág.

Lista de Figuras **10**

Lista de Tablas **11**

Introducción **13**

Contexto y antecedentes de la investigación *13*

Problema y vacío de investigación *18*

Delimitación y enfoque de la investigación *20*

Pregunta de investigación *20*

Estructura del documento *21*

Objetivos **22**

Objetivo general *22*

Objetivos específicos *22*

Justificación **23**

Marco Teórico **26**

Inteligencia artificial y transformación del análisis financiero *26*

Finanzas conductuales y la dimensión emocional del inversionista *29*

Procesamiento del lenguaje natural (PLN) en el dominio financiero *30*

Modelos transformer y fine-tuning para textos financieros *32*

Análisis de sentimiento y emociones multiclase en textos financieros *35*

Fuentes textuales financieras: noticias institucionales versus comunidades retail *37*

Modelo conceptual propuesto *41*

Estado del arte *46*

Hipótesis **50**

Variables **51**

Metodología **54**

Enfoque y alcance de la investigación *54*

Diseño y tipo de investigación *55*

Modelo conceptual propuesto *56*

Especificación estadística formal *57*

Contraste estadístico de H1 57

Divergencia de Kullback-Leibler 58

Evaluación del desempeño del modelo de clasificación 59

Macro-F1 61

Micro-F1 61

Weighted-F1 62

Coeficiente de correlación de Matthews 62

Contraste estadístico de H2 64

Intervalos de confianza mediante bootstrap 66

Integración de los procedimientos estadísticos *67*

Aplicación analítica *68*

Fases de la investigación *70*

Fase 1. Identificación y recolección del corpus 70

Fase 2. Preprocesamiento y depuración 72

Fase 3. Representación semántica de los textos financieros 75

Fase 4. Identificación y clasificación de emociones 77

Fase 5. Comparación de la distribución emocional entre fuentes 80

Fase 6. Evaluación del desempeño de los modelos 83

Manejo de desbalance de clases *85*

Riesgos metodológicos *87*

Contribución científica *91*

Trabajo de Campo 94

Recolección del corpus *94*

Preprocesamiento y depuración del corpus *96*

Distribución Emocional del Corpus Etiquetado 98

Análisis de resultados *101*

Entrenamiento del modelo: fine-tuning de FinBERT 101

Contrastación de la hipótesis H1: diferencias distribucionales entre fuentes 103

Hallazgo H1: distribuciones emocionales estadísticamente diferentes entre fuentes 105

Evaluación Final del Modelo y Contrastación de la Hipótesis H2 107

Hallazgo H2: comparación del clasificador binario independiente y el modelo multiclase agrupado 110

Contraste H2: comparación del desempeño de los modelos 111

Discusión de la hipótesis H2 113

Validación humana de las pseudoetiquetas 115

Resumen consolidado de resultados 118

Propuesta de solución a la problemática *122*

Situación actual 122

Oportunidades 127

Propuesta de solución al problema planteado 132

Discusión **139**

Limitaciones de la investigación *139*

Desbalance estructural del corpus *139*

Representatividad temporal del corpus *140*

Alcance geográfico y sesgo de fuente *142*

Reflexión final sobre las limitaciones *143*

Conclusiones y trabajo futuro **145**

Conclusiones *145*

Conclusión 1. Caracterización del corpus y comportamiento emocional 145

Conclusión 2. Desempeño del modelo FinBERT en la clasificación emocional multiclase 146

Conclusión 3. Diferencias entre las distribuciones emocionales de las fuentes 147

Conclusión 4. Diferencia de desempeño entre los enfoques multiclase y binario y relevancia de la granularidad emocional 148

Síntesis frente al objetivo general y al problema de investigación 149

Trabajo futuro *151*

Referencias **156**

Anexos **163**

Anexo A. Distribución de textos por año, fuente, tipo y activo del corpus final *163*

Anexo B. Distribución de eliminaciones por fuente *165*

Anexo C. Matriz de confusión del modelo FinBERT fine-tuned v2 ..... *166*

Anexo D. Guía de anotación *168*

Validación humana de las pseudoetiquetas 168

Diseño de la validación humana 168

Guía de anotación 170

Resultados de la validación 171

Comparación entre anotadores humanos y pseudoetiquetas 174

Interpretación y discusión de los resultados 174

Circularidad del procedimiento como limitación metodológica 175

Implicaciones para trabajos posteriores 176

Anexo E. Procedencia de los datos, evidencia de ejecución y trazabilidad metodológica *178*

Procedencia de los datos 178

Software y configuración computacional 178

Construcción de las categorías emocionales.....179

Listas léxicas y reglas de asignación de pseudoetiquetas.....179

Evidencias de ejecución y trazabilidad 180

Custodia del código fuente y alcance de la reproducibilidad 189

Lista de Figuras

Pág.

Figura 1 Modelo conceptual propuesto 41

Figura 2 Pipeline de investigación 44

Figura 3 Matriz de confusión del modelo FinBERT fine-tuned v2 166

Figura 4 Fase 1 - Recolección del corpus (Mínimo 5.000 textos por fuente, ene 2022 – abr 2026, S&P 500 y TSLA) 181

Figura 5 Fase 1 - Consolidando corpus final 182

Figura 6 Fase 2 – Preprocesamiento y depuración 182

Figura 7 Fase 2 – Validación por fuente y activo 183

Figura 8 Fase 3 – Representación semántica y etiquetado con FinBERT 183

Figura 9 Fase 3 – Distribución de emociones 184

Figura 10 Fase 4 – Fine-tuning de FinBERT 184

Figura 11 Fase 4 – Evaluación sobre set de validación 185

Figura 12 Fase 4 – v2 – Reentrenamiento con oversampling 185

Figura 13 Fase 4 – v2 – finbert fintuned 186

Figura 14 Fase 5 – Comparación de distribuciones emocionales 186

Figura 15 Fase 5 – Resultado Chi – cuadrado 186

Figura 16 Fase 6 – Evaluación final del modelo 187

Figura 17 Fase 6 – Métricas globales 187

Figura 18 Fase 7 – Crear etiquetas binarias 188

Figura 19 Fase 7 – Distribución binaria - entrenamiento 188

Lista de Tablas

Pág.

Tabla 1 Comparación de antecedentes sobre comportamiento financiero y análisis de información textual 16

Tabla 2 Comparativa inteligencia artificial y transformación del análisis financiero 28

Tabla 3 Operacionalización de las variables y del componente metodológico de clasificación 53

Tabla 4 Distribución del corpus por fuente y activo financiero 94

Tabla 5 Balance del corpus por tipo de fuente y activo financiero..... 95

Tabla 6 Resultados del proceso de preprocesamiento 96

Tabla 7 Distribución de emociones en el corpus etiquetado 98

Tabla 8 División estratificada del corpus: conjuntos de entrenamiento, validación y prueba 99

Tabla 9 Hiperparámetros configurados para el fine-tuning: comparación v1 vs. v2 ..101

Tabla 10 Evolución del entrenamiento por época: versión v2 (Modelo Final) 102

Tabla 11 Comparativa v1 vs. v2 103

Tabla 12 Distribución emocional comparada: fuentes retail vs. Institucionales 104

Tabla 13 Resultados de las pruebas estadísticas para la contrastación de H1.....104

Tabla 14 Métricas globales de evaluación del modelo FinBERT fine-tuned v2 107

Tabla 15 F1-Score por clase emocional en el conjunto de prueba.....108

Tabla 16 Comparación del desempeño de los modelos 111

Tabla 17 Evaluación de FinBERT v2 vs. frente a las anotaciones humanas.....117

Tabla 18 Resumen consolidado de resultados 119

Tabla 19 Distribución temporal del del corpus final por año.....141

Tabla 20 Distribución de textos por año, fuente, tipo y activo del corpus final.....163

Tabla 21 Distribución de eliminaciones por fuente.....165

Tabla 22 Distribución de las pseudoetiquetas en la muestra de validación.....169

Tabla 23 Definiciones operacionales para la anotación de las cinco emociones 170

Tabla 24 Resultados de la validación humana y comparación con las pseudoetiquetas 172

Tabla 25 Matriz de acuerdo entre los dos anotadores humanos.....173

Introducción

Contexto y antecedentes de la investigación

Históricamente, el análisis del mercado bursátil se ha apoyado principalmente en información cuantitativa, como precios de cierre, volúmenes de negociación e indicadores técnicos. Sin embargo, el desarrollo de las finanzas conductuales ha ampliado esta perspectiva al cuestionar los supuestos de racionalidad plena y reconocer que el comportamiento, las expectativas y los procesos de formación de creencias de los participantes pueden aportar elementos adicionales para comprender fenómenos financieros que no siempre quedan plenamente representados por enfoques exclusivamente cuantitativos (Shiller, 2003). En esta línea, Anufriev et al. (2022) mostraron experimentalmente que el horizonte de inversión y la forma en que los participantes extrapolan tendencias pasadas pueden influir en la estabilidad de los precios, favoreciendo una mayor volatilidad y la aparición de burbujas y caídas bajo horizontes más cortos. En este contexto, factores asociados con percepciones, expectativas y respuestas psicológicas adquieren relevancia para el análisis financiero.

Desde la perspectiva de las finanzas conductuales, Shiller (2003) cuestiona la suficiencia de los enfoques basados exclusivamente en la hipótesis de mercados eficientes y destaca la importancia de considerar factores psicológicos y comportamentales para comprender la dinámica de los mercados financieros. En consonancia con esta perspectiva, los resultados de la presente investigación muestran diferencias en la distribución de las pseudoetiquetas emocionales según el tipo de fuente: el entusiasmo representó el 9,3% de los textos retail y el 5,0% de los institucionales, mientras que la tristeza correspondió al 2,3% y al 5,5%, respectivamente. Aunque estos resultados no

permiten establecer relaciones causales ni verificar directamente los estados emocionales de los autores, ilustran la utilidad de comparar sistemáticamente el contenido emocional de distintas fuentes financieras para ampliar la caracterización de la información que circula en los mercados.

Con el avance del procesamiento del lenguaje natural (PLN), el análisis de información textual se ha convertido en una herramienta relevante para estudiar el sentimiento asociado con los mercados financieros. Bollen et al. (2011) analizaron cerca de 9,85 millones de publicaciones de Twitter, correspondientes al periodo comprendido entre febrero y diciembre de 2008, para estudiar la relación entre los estados de ánimo colectivos y el comportamiento del índice Dow Jones Industrial Average (DJIA). Los autores reportaron una precisión del 86,7% en la predicción de los cambios diarios al alza o a la baja del índice y una reducción superior al 6% en el error porcentual absoluto medio (MAPE) al incorporar determinadas dimensiones del estado de ánimo colectivo. Estos resultados ilustran el potencial analítico de la información textual para estudiar su relación con los indicadores financieros, aunque corresponden al contexto, los datos y los modelos evaluados por dichos autores y no implican que el mismo desempeño pueda generalizarse a otros mercados o periodos.

Li et al. (2014) analizaron el impacto del sentimiento extraído de noticias financieras sobre los retornos de las acciones, mientras que Kearney y Liu (2014) sistematizaron las principales fuentes, métodos de análisis de contenido y modelos empíricos utilizados en la literatura financiera. Asimismo, Nguyen et al. (2015) estudiaron sentimientos derivados de medios sociales para la predicción de movimientos bursátiles, Ruan et al. (2018) analizaron la relación entre sentimiento en Twitter y retornos anormales de empresas del S&P 500, y

Wang y Duxbury (2021) examinaron el sentimiento institucional en 38 mercados. Aunque estos antecedentes muestran que la información textual puede incorporarse a tareas predictivas, la presente investigación no busca modelar precios ni retornos, sino caracterizar y comparar distribuciones emocionales entre diferentes tipos de fuentes financieras.

En este estudio, la comparación estadística de las distribuciones emocionales se realizó sobre pseudoetiquetas generadas mediante FinBERT base y reglas heurísticas de clasificación emocional, mientras que el modelo FinBERT ajustado mediante fine-tuning fue entrenado y evaluado posteriormente como clasificador multiclase.

El desarrollo de arquitecturas basadas en *transformers* fortaleció esta línea de investigación al permitir representaciones contextuales más especializadas del lenguaje. En el dominio financiero, Araci (2019) introdujo FinBERT, un modelo basado en BERT adaptado a textos financieros, cuyos resultados mostraron mejoras frente a métodos de referencia en tareas de clasificación de sentimiento. Más recientemente, Sun et al. (2025) propusieron un enfoque basado en modelos de lenguaje preentrenados que incorpora conocimiento procedente de diccionarios financieros y características específicas orientadas, entre otros aspectos, a mejorar la identificación del sentimiento neutral. Estos desarrollos muestran una evolución hacia modelos cada vez más sensibles a las particularidades semánticas del lenguaje financiero.

Tabla 1 *Comparación de antecedentes sobre comportamiento financiero y análisis de información textual*

Autor(es)	Enfoque o información analizada	Resultado o aporte descrito en la Introducción	Relación con la investigación
Shiller (2003)	Finanzas conductuales y comportamiento de los participantes del mercado.	Cuestiona la suficiencia de la teoría de mercados eficientes para explicar determinados fenómenos y destaca la relevancia de factores psicológicos y comportamentales.	Fundamenta la importancia de estudiar dimensiones emocionales en el contexto financiero.
Anufriev et al. (2022)	Experimentos sobre horizontes de inversión y extrapolación de tendencias pasadas.	El horizonte de inversión y la extrapolación de tendencias pueden influir en la estabilidad de los precios y en la aparición de burbujas y caídas.	Aporta evidencia sobre la relevancia del comportamiento de los participantes para comprender la dinámica del mercado.
Bollen et al. (2011)	Estados de ánimo colectivos derivados de publicaciones en Twitter.	Encontraron que ciertas dimensiones del estado de ánimo aportaban información para predecir cambios diarios del Dow Jones Industrial Average.	Sustenta la utilidad potencial de analizar información textual para estudiar fenómenos financieros.
Li et al. (2014)	Sentimiento extraído de noticias financieras y retornos accionarios.	Examinaron la relación entre el sentimiento de las noticias y los retornos de las acciones.	Relaciona el contenido de fuentes institucionales con variables del mercado financiero.
Kearney y Liu (2014)	Revisión de fuentes, métodos de análisis de contenido y modelos empíricos.	Sistematizaron enfoques utilizados en la investigación del sentimiento financiero.	Proporciona antecedentes metodológicos para el análisis de textos financieros.
Nguyen et al. (2015)	Sentimiento extraído de medios sociales.	Estudiaron su utilidad para predecir movimientos bursátiles.	Se relaciona con el análisis de textos procedentes de fuentes de inversionistas.
Ruan et al. (2018)	Sentimiento en Twitter y retornos anormales de empresas del S&P 500.	Analizaron la relación entre las señales de sentimiento y los retornos anormales.	Aporta un antecedente sobre información textual asociada con empresas del mercado estadounidense.

Wang y Duxbury (2021)	Sentimiento institucional en 38 mercados.	Examinaron el sentimiento institucional en distintos mercados.	Se relaciona con la comparación entre fuentes institucionales y otras fuentes financieras.
Investigación actual	Comparación de cinco categorías emocionales en textos retail e institucionales asociados con el S&P 500 y Tesla.	Se obtuvo una diferencia estadísticamente significativa entre las distribuciones de las pseudoetiquetas emocionales de las dos fuentes: $(\chi^2)(4) = 126{,}422$, $p < 0{,}0001$; la divergencia KL simétrica fue 0,0356.	Integra la comparación de fuentes con una clasificación emocional de cinco categorías. Los resultados de H1 corresponden a pseudoetiquetas generadas con FinBERT base y reglas heurísticas, no a emociones verificadas

En conjunto, los antecedentes revisados muestran que el comportamiento de los participantes, el sentimiento presente en las noticias y las señales emocionales de los medios sociales constituyen dimensiones relevantes para el estudio de los mercados financieros. Sin embargo, los trabajos citados abordan fuentes, variables y objetivos analíticos diversos, por lo que sus resultados no son directamente equivalentes. En este contexto, la presente investigación aporta una comparación de las distribuciones de cinco pseudoetiquetas emocionales entre fuentes retail e institucionales asociadas con el S&P 500 y Tesla. La prueba de independencia identificó diferencias estadísticamente significativas entre ambas distribuciones; no obstante, este resultado debe interpretarse considerando que las etiquetas se generaron mediante FinBERT base y reglas heurísticas, y que la significación estadística no implica por sí misma un efecto de gran magnitud ni demuestra emociones reales de los autores. De esta manera, el estudio contribuye a caracterizar comparativamente el contenido textual financiero y a reconocer tanto su utilidad analítica como las limitaciones de su procedimiento de etiquetado.

Problema y vacío de investigación

A pesar de estos avances, persisten aspectos metodológicos que justifican profundizar en el análisis comparativo de emociones presentes en diferentes fuentes de información financiera. En primer lugar, la literatura ha estudiado noticias, mensajes publicados en internet y contenidos generados por usuarios en plataformas digitales (Kearney y Liu, 2014), pero resulta pertinente profundizar en comparaciones realizadas bajo un mismo esquema de clasificación emocional, sobre los mismos activos analizados y dentro de un marco metodológico común.

En segundo lugar, una parte importante del análisis de sentimiento financiero se ha desarrollado mediante esquemas de polaridad que clasifican los textos como positivos, negativos o neutrales. El análisis de sentimiento consiste en identificar la orientación valorativa general de un mensaje frente a un tema, activo o acontecimiento financiero. Aunque estas categorías permiten representar dicha orientación, el análisis de emociones específicas busca distinguir estados afectivos particulares, como el entusiasmo, el miedo o la tristeza, ofreciendo una caracterización más detallada del contenido textual. En esta línea, Lee et al. (2023) desarrollaron StockEmotions a partir de 10.000 comentarios en inglés recopilados de StockTwits y propusieron doce categorías emocionales de grano fino. Asimismo, García-Méndez et al. (2023) aplicaron el análisis de emociones basado en aspectos, un enfoque que identifica las emociones asociadas con elementos concretos mencionados en un mensaje financiero, como un activo o una oportunidad de inversión, con el propósito de distinguir señales de oportunidad y precaución. Por su parte, Sun et al. (2025) estudiaron métodos para mejorar el análisis de sentimiento financiero mediante modelos de lenguaje preentrenados, incorporando conocimiento de diccionarios

financieros y características orientadas al reconocimiento del contenido neutral. Este último se refiere a mensajes cuya orientación no se identifica claramente como positiva o negativa y que, por tanto, no deben forzarse a una de esas dos categorías. En conjunto, estos enfoques permiten una caracterización más detallada del lenguaje financiero; sin embargo, también plantean retos de clasificación, especialmente al diferenciar categorías emocionalmente próximas o distinguir el contenido neutral de expresiones con una orientación afectiva definida.

En tercer lugar, la reproducibilidad de los estudios basados en procesamiento de lenguaje natural (PLN) requiere documentar de manera clara las características del corpus, las fuentes utilizadas y los procedimientos de preparación de los datos. En esta investigación se trabajó con textos financieros en inglés asociados con el S&P 500 y Tesla Inc. (TSLA), provenientes de fuentes retail e institucionales. El corpus inicial estuvo conformado por 10.000 registros, distribuidos de manera equilibrada entre ambos tipos de fuente y entre los dos activos analizados. Posteriormente, los textos fueron sometidos a procedimientos de depuración y preprocesamiento, obteniéndose un corpus de 7.752 registros utilizado en las etapas posteriores de clasificación y análisis.

En cuarto lugar, resulta pertinente fortalecer la comparación estadística de las distribuciones emocionales obtenidas a partir de diferentes tipos de fuentes financieras. Además de evaluar el desempeño del modelo de clasificación, la identificación de diferencias entre las emociones presentes en contenidos retail e institucionales permite examinar si ambos tipos de discurso financiero presentan perfiles emocionales diferenciados. En esta investigación, dicha comparación se aborda mediante estadísticas

descriptivas, una prueba Chi-cuadrado de independencia y medidas descriptivas de divergencia distribucional.

Delimitación y enfoque de la investigación

De manera complementaria, el estudio evalúa si el desempeño del esquema multiclase difiere del obtenido mediante una formulación binaria, con el propósito de examinar la relación entre granularidad emocional y desempeño de clasificación.

Metodológicamente, la investigación distingue dos etapas complementarias. En una primera etapa, las cinco categorías emocionales fueron construidas como pseudoetiquetas a partir de las predicciones de FinBERT base y reglas heurísticas de mapeo; estas pseudoetiquetas se utilizaron para caracterizar el corpus y comparar estadísticamente las distribuciones entre fuentes retail e institucionales. En una segunda etapa, el corpus pseudoetiquetado fue utilizado para realizar el fine-tuning de FinBERT y evaluar el desempeño del clasificador multiclase.

Pregunta de investigación

A partir de estos antecedentes y del vacío de investigación identificado, se plantea la siguiente pregunta de investigación:

¿En qué medida difiere la distribución de las pseudoetiquetas emocionales multiclase identificadas en noticias institucionales y comentarios de inversionistas retail sobre el índice S&P 500 y Tesla Inc. durante el periodo de enero de 2022 a abril de 2026, a partir de un esquema de clasificación basado en FinBERT y reglas heurísticas de mapeo emocional?

Estructura del documento

El documento se organiza en los siguientes apartados. En primer lugar, se presentan los objetivos y la justificación de la investigación. Posteriormente, se desarrolla el marco teórico, que aborda los fundamentos de la inteligencia artificial aplicada al análisis financiero, las finanzas conductuales, el procesamiento del lenguaje natural, los modelos Transformer y la clasificación de emociones en textos financieros. A continuación, se expone el estado del arte y se formulan las hipótesis y las variables de investigación. En el capítulo de metodología se describen el enfoque, el diseño y el tipo de investigación, la especificación estadística, los procedimientos de evaluación del desempeño de los modelos y las fases de ejecución del estudio. Seguidamente, el capítulo de trabajo de campo presenta la conformación y depuración del corpus, la distribución emocional, el entrenamiento y la evaluación de FinBERT, los contrastes de hipótesis y la validación humana de las pseudoetiquetas. Después, se desarrolla la propuesta de solución a la problemática identificada. El apartado de discusión examina los resultados y las limitaciones de la investigación. Finalmente, se presentan las conclusiones y las líneas de trabajo futuro, seguidas de las referencias bibliográficas y los anexos que complementan la documentación del estudio.

Objetivos

Objetivo general

Analizar comparativamente las emociones financieras multiclase identificadas en noticias institucionales y comentarios de inversionistas retail relacionados con el índice S&P 500 y Tesla Inc. durante el periodo de enero de 2022 a abril de 2026, mediante la construcción de pseudoetiquetas emocionales basadas en FinBERT y reglas heurísticas, la clasificación mediante fine-tuning de FinBERT y la comparación estadística de sus distribuciones.

Objetivos específicos

- Caracterizar el corpus de textos financieros provenientes de fuentes retail e institucionales relacionadas con el S&P 500 y Tesla Inc., mediante estadísticas descriptivas de su composición y distribución, considerando el tipo de fuente, el activo financiero y la frecuencia y proporción de las cinco categorías emocionales —enojo, entusiasmo, miedo, neutralidad y tristeza— asignadas mediante pseudoetiquetado.
- Clasificar los textos financieros en cinco categorías emocionales —enojo, entusiasmo, miedo, neutralidad y tristeza— mediante un modelo FinBERT ajustado mediante fine-tuning, evaluar su desempeño con métricas de clasificación multiclase y analizar la variación de dichas métricas al agrupar las predicciones en una formulación binaria que distingue entre entusiasmo y el resto de las emociones.
- Comparar estadísticamente la distribución de las pseudoetiquetas emocionales entre fuentes retail e institucionales relacionadas con el S&P 500 y Tesla Inc., utilizando estadísticas descriptivas, la prueba Chi-cuadrado de independencia y medidas descriptivas de divergencia distribucional.

Justificación

El análisis de emociones multiclase en textos financieros constituye un campo relevante para la ciencia de datos debido al creciente volumen de información no estructurada generada tanto por medios especializados como por inversionistas en plataformas digitales. En este contexto, el procesamiento del lenguaje natural (PLN) permite transformar contenidos textuales en representaciones susceptibles de clasificación y comparación, facilitando el estudio de patrones emocionales presentes en diferentes fuentes de información financiera (Kearney y Liu, 2014; Araci, 2019).

Desde la perspectiva teórica, la investigación contribuye al estudio del componente emocional presente en el discurso financiero mediante un enfoque que supera la descripción exclusiva de polaridad positiva, negativa o neutral y se orienta hacia categorías emocionales más específicas. Este nivel de granularidad permite caracterizar con mayor detalle diferentes estados afectivos presentes en los textos financieros y examinar sus distribuciones entre contenidos generados por inversionistas retail y noticias institucionales (Lee et al., 2023; García-Méndez et al., 2023). Asimismo, el uso de modelos especializados en el dominio financiero resulta pertinente frente a las particularidades semánticas del lenguaje y a los desafíos asociados con categorías como la neutralidad (Araci, 2019; Sun et al., 2025). En esta línea, Jiang y Zeng (2023) emplearon FinBERT para extraer sentimiento financiero e integrarlo en un modelo LSTM orientado a la predicción de movimientos bursátiles, evidenciando la utilidad de los modelos de lenguaje especializados como insumo para tareas analíticas en finanzas.

Desde la perspectiva aplicada, la investigación caracteriza un corpus de textos financieros en inglés relacionados con el índice S&P 500 y Tesla Inc. (TSLA), recopilados durante el periodo comprendido entre enero de 2022 y abril de 2026. El corpus inicial estuvo conformado por 10.000 registros, distribuidos equitativamente entre fuentes retail e institucionales, con 5.000 textos por tipo de fuente. Después del proceso de depuración y preprocesamiento, se conservaron 7.752 textos, equivalentes al 77,5% del conjunto inicial, mientras que se excluyeron 2.248 registros (22,5%). Las fuentes retail estuvieron representadas principalmente por StockTwits y Twitter Financial, mientras que las institucionales incluyeron contenidos de Reuters, CNBC y otros medios financieros especializados. La caracterización considera, además, la distribución de los textos según el tipo de fuente y el activo financiero, así como las frecuencias y proporciones de las cinco categorías emocionales asignadas mediante pseudoetiquetado: enojo, entusiasmo, miedo, neutralidad y tristeza. Esta descripción permite contextualizar el alcance del corpus y establecer la base empírica para comparar las distribuciones emocionales de las fuentes estudiadas.

Desde la perspectiva metodológica, el estudio integra procedimientos de construcción y depuración de corpus, procesamiento del lenguaje natural (PLN), fine-tuning de un modelo especializado en dominio financiero, evaluación mediante métricas de clasificación multiclase y comparación estadística de distribuciones emocionales. Esta integración permite documentar los procedimientos aplicados al análisis de textos financieros y evaluar tanto el desempeño del clasificador como las diferencias entre las distribuciones emocionales de las fuentes retail e institucionales estudiadas. Además de su aporte metodológico, el trabajo aborda necesidades relevantes para el sector financiero y la

analítica de datos, como el procesamiento sistemático de información textual no estructurada, la comparación de contenidos provenientes de diferentes fuentes y la evaluación transparente del desempeño de modelos de clasificación. Estas capacidades pueden contribuir a que investigadores, analistas financieros y equipos de ciencia de datos dispongan de información documentada para explorar patrones en el discurso financiero. Asimismo, la documentación de las evidencias de ejecución, la trazabilidad de los procedimientos y la custodia del código fuente permiten delimitar el alcance de la reproducibilidad del estudio. No obstante, los resultados de esta investigación se circunscriben al corpus, al periodo y a las categorías emocionales estudiadas, por lo que su aplicación a otros contextos requiere validación adicional.

Marco Teórico

Inteligencia artificial y transformación del análisis financiero

La transformación digital de los mercados financieros ha ampliado el volumen y la diversidad de información disponible para el análisis. Junto con datos estructurados como precios, volúmenes de negociación, indicadores contables y variables macroeconómicas, actualmente se emplean fuentes no estructuradas como noticias, publicaciones en redes sociales y otros contenidos textuales relacionados con los mercados. En este contexto, las técnicas de inteligencia artificial, minería de texto y procesamiento del lenguaje natural (PLN) han permitido incorporar este tipo de información a diferentes tareas de análisis financiero (Nassirtoussi et al., 2014; Ashtiani y Raahemi, 2023).

Dentro de este desarrollo, el aprendizaje automático (machine learning) y el aprendizaje profundo (deep learning) proporcionan métodos para identificar patrones en conjuntos de datos complejos, mientras que el procesamiento del lenguaje natural (PLN) permite representar y analizar computacionalmente información expresada en lenguaje humano. Por ejemplo, ante una noticia financiera que indique que Tesla reportó un crecimiento de sus ventas superior al esperado, las técnicas de PLN permiten procesar el texto e identificar expresiones relevantes para su interpretación financiera. Posteriormente, un modelo de análisis de sentimiento puede clasificar la orientación general del mensaje, mientras que un modelo de clasificación emocional puede asignarle una categoría específica de acuerdo con las clases definidas para el estudio. De esta manera, un contenido textual puede transformarse en una representación estructurada susceptible de análisis estadístico y comparación con otros mensajes financieros. En la presente

investigación, este principio se aplica a textos provenientes de fuentes retail e institucionales relacionados con el S&P 500 y Tesla Inc., mediante el procesamiento textual y la clasificación en cinco categorías emocionales: enojo, entusiasmo, miedo, neutralidad y tristeza. Las categorías asignadas deben interpretarse como etiquetas generadas por el procedimiento de clasificación y no como una verificación directa de los estados emocionales reales de quienes escribieron los textos (Nassirtoussi et al., 2014; Rodríguez-Ibáñez et al., 2023).

Para sintetizar los antecedentes y facilitar la comparación de sus aportes, la Tabla 2 presenta el alcance, los datos cuantitativos, los enfoques metodológicos y los principales hallazgos de dos revisiones de literatura relacionadas con el análisis de sentimientos en redes sociales y la predicción de mercados financieros a partir de noticias. Asimismo, se incorpora la caracterización de la presente investigación para precisar su relación con los estudios previos. Mientras que las revisiones citadas examinan tendencias y métodos desarrollados en la literatura, este trabajo aplica un procedimiento de clasificación emocional multiclase a un corpus de textos financieros de fuentes retail e institucionales, con el propósito de comparar sus distribuciones de pseudoetiquetas emocionales.

Tabla 2 *Comparativa inteligencia artificial y transformación del análisis financiero*

Estudio	Datos y alcance de la revisión	Métodos o enfoque	Principales hallazgos y aportes
Rodríguez-Ibáñez et al. (2023)	Revisión sobre análisis de sentimientos en redes sociales. Los autores reportan que, entre 2008 y 2022, las publicaciones que incluyen el concepto de análisis de sentimientos en redes sociales crecieron a una tasa geométrica anual del 34%.	Revisión de técnicas, aplicaciones y tendencias del análisis de sentimientos, con especial atención a las redes sociales y Twitter.	Documenta la expansión del campo y sus aplicaciones en predicción de mercados financieros, analítica de clientes, marketing y otros dominios. También identifica oportunidades de investigación en técnicas y aplicaciones.
Ashtiani y Raahemi (2023)	Revisión sistemática de investigaciones publicadas entre 2015 y 2022. Se identificaron inicialmente 293 artículos y se seleccionaron 61 estudios para el análisis final, a partir de búsquedas en cinco bibliotecas digitales.	Revisión sistemática de estudios que utilizan minería de texto y aprendizaje automático para predecir mercados financieros a partir de noticias.	Identifica tendencias en la predicción financiera basada en noticias y el uso de redes neuronales y aprendizaje profundo. Señala oportunidades para aprovechar más las técnicas avanzadas de procesamiento del lenguaje natural.
Presente investigación	Corpus de 7.752 textos financieros en inglés, relacionados con el S&P 500 y Tesla Inc., correspondientes al periodo enero de 2022 a abril de 2026.	Preprocesamiento textual, asignación de pseudoetiquetas emocionales, ajuste fino de FinBERT, evaluación multiclase y comparación estadística de las distribuciones emocionales.	Compara cinco categorías emocionales —enojo, entusiasmo, miedo, neutralidad y tristeza— entre fuentes retail e institucionales. Las categorías corresponden a pseudoetiquetas del procedimiento aplicado, no a emociones verificadas directamente en cada texto.

Finanzas conductuales y la dimensión emocional del inversionista

La teoría de los mercados eficientes ha constituido uno de los principales marcos para explicar la formación de precios en los mercados financieros. Sin embargo, el desarrollo de las finanzas conductuales amplió esta perspectiva al documentar anomalías y comportamientos que no se explican completamente bajo supuestos de racionalidad plena. En este enfoque, los procesos de formación de expectativas, las creencias y determinados patrones de comportamiento de los participantes adquieren relevancia para comprender fenómenos observados en los mercados financieros (Shiller, 2003).

Desde esta perspectiva, el análisis de componentes psicológicos y emocionales resulta pertinente para estudiar el comportamiento de los participantes del mercado. Las respuestas frente a situaciones de incertidumbre, riesgo o expectativas favorables pueden expresarse también a través del lenguaje utilizado por inversionistas y otros actores financieros. El desarrollo de plataformas digitales y redes sociales ha generado, en este sentido, grandes volúmenes de contenido textual susceptible de ser analizado mediante técnicas computacionales para aproximarse a dimensiones de sentimiento y estado de ánimo colectivo.

En el ámbito del análisis textual, Bollen et al. (2011) examinaron si los estados de ánimo colectivos extraídos de publicaciones de Twitter estaban relacionados con el comportamiento del mercado bursátil. Los autores utilizaron OpinionFinder para medir polaridad positiva y negativa y Google-Profile of Mood States (GPOMS) para representar el estado de ánimo mediante seis dimensiones. Sus resultados mostraron que determinadas dimensiones del estado de ánimo aportaban información para mejorar la predicción de cambios diarios del Dow Jones Industrial Average, constituyendo un

antecedente relevante del uso de contenidos generados en redes sociales para el análisis financiero.

La evolución de esta línea de investigación ha llevado a considerar representaciones emocionales más específicas que la polaridad positiva o negativa. Ge et al. (2020) analizaron emociones expresadas en redes sociales durante el desplome del mercado bursátil chino de 2015 mediante un marco basado en cognición, valencia y activación. Sus resultados identificaron al miedo y al disgusto como emociones subcategóricas significativamente relacionadas con la cognición del mercado durante el episodio analizado. Estos hallazgos muestran que una representación emocional más detallada puede aportar información que no queda completamente reflejada mediante categorías generales de polaridad y sustentan la pertinencia de estudiar emociones diferenciadas en textos financieros.

Procesamiento del lenguaje natural (PLN) en el dominio financiero

El procesamiento del lenguaje natural (PLN) comprende un conjunto de métodos computacionales orientados a representar, analizar y extraer información a partir del lenguaje humano. En el ámbito financiero, estas técnicas permiten trabajar con fuentes textuales como noticias, informes corporativos, publicaciones en redes sociales y comentarios de inversionistas, convirtiendo información no estructurada en representaciones susceptibles de clasificación y análisis cuantitativo (Kearney y Liu, 2014).

Una dificultad particular del análisis textual financiero es que palabras con determinadas connotaciones en el lenguaje general pueden adquirir significados diferentes dentro del

dominio. Loughran y McDonald (2011), al estudiar reportes corporativos 10-K, mostraron que los diccionarios generales podían clasificar incorrectamente numerosos términos financieros y desarrollaron listas de palabras específicas para representar con mayor precisión el tono de estos documentos. Este trabajo constituye un antecedente fundamental de la adaptación de herramientas de análisis de sentimiento a las particularidades semánticas del lenguaje financiero.

La evolución de estos enfoques ha permitido pasar de métodos predominantemente léxicos hacia modelos capaces de representar relaciones contextuales y lingüísticas más complejas. Daudert (2021) propuso un enfoque de análisis financiero de sentimiento de grano fino que combina el contenido textual con información contextual y relacional asociada a los registros, con el propósito de capturar sentimientos implícitos que pueden resultar difíciles de identificar utilizando únicamente el texto aislado. Por su parte, Xiang et al. (2022) desarrollaron un modelo neural que incorpora información semántica y sintáctica para representar de manera explícita la relación entre un objetivo financiero y su contexto textual, mostrando la utilidad de incorporar estructura lingüística al análisis de sentimiento financiero.

El conocimiento específico del dominio también puede integrarse directamente con modelos de lenguaje. Lin y Liao (2024) propusieron un método basado en prompts enriquecidos con léxicos financieros para el análisis dimensional del sentimiento y distinguieron entre representaciones categóricas y dimensionales de los estados afectivos. De manera complementaria, Sheetal y Aithal (2025) combinaron el diccionario Loughran-McDonald con BERT para analizar titulares financieros, integrando información léxica específica del dominio con representaciones contextuales aprendidas por el modelo. En

conjunto, estos desarrollos muestran que la adaptación del vocabulario, las representaciones semánticas y el conocimiento especializado constituye un componente relevante para el análisis computacional del lenguaje financiero.

En el contexto de la presente investigación, estos antecedentes sustentan el uso de herramientas especializadas para procesar y clasificar textos financieros provenientes de fuentes con características discursivas diferentes. La necesidad de representar adecuadamente el significado contextual de los términos resulta particularmente relevante cuando se busca distinguir categorías emocionales específicas y comparar posteriormente sus distribuciones entre contenidos retail e institucionales.

Modelos transformer y fine-tuning para textos financieros

La arquitectura Transformer, propuesta por Vaswani et al. (2017), introdujo un enfoque para el procesamiento de secuencias basado principalmente en mecanismos de atención, prescindiendo de las estructuras recurrentes y convolucionales utilizadas tradicionalmente en este tipo de tareas. El mecanismo de autoatención permite modelar relaciones entre diferentes posiciones de una secuencia y facilita un mayor grado de paralelización durante el entrenamiento. Este desarrollo constituyó la base arquitectónica de numerosos modelos de lenguaje preentrenados posteriores.

Sobre esta arquitectura, Devlin et al. (2019) desarrollaron BERT (*Bidirectional Encoder Representations from Transformers*), un modelo de lenguaje preentrenado diseñado para aprender representaciones contextuales bidireccionales. Su entrenamiento original combinó tareas de predicción de palabras enmascaradas y predicción de la siguiente

oración. Posteriormente, el modelo puede ajustarse mediante fine-tuning a tareas específicas de procesamiento del lenguaje natural incorporando modificaciones relativamente pequeñas en la arquitectura de salida. Esta estrategia de preentrenamiento y ajuste fino estableció un marco ampliamente utilizado para transferir conocimiento lingüístico general hacia dominios y tareas particulares.

En el dominio financiero, Araci (2019) propuso FinBERT, un modelo basado en BERT adaptado específicamente al lenguaje financiero mediante entrenamiento adicional sobre un corpus del dominio y posterior fine-tuning para tareas de análisis de sentimiento. Los resultados mostraron mejoras frente a métodos de referencia en los conjuntos de datos evaluados, incluso utilizando cantidades relativamente reducidas de datos etiquetados. Este trabajo evidenció la utilidad de adaptar modelos de lenguaje preentrenados a las particularidades semánticas del dominio financiero.

Jiang y Zeng (2023) extendieron el uso de FinBERT al emplear las estimaciones de sentimiento obtenidas a partir de noticias financieras como insumo de un modelo LSTM orientado a la predicción de movimientos bursátiles. Sus resultados mostraron que el sentimiento podía aportar información útil dentro del modelo predictivo. Aunque dicha aplicación persigue un objetivo diferente al del presente estudio, constituye un antecedente relevante sobre la utilización de FinBERT como herramienta de extracción de información semántica en textos financieros.

Desarrollos posteriores han explorado distintas estrategias para incorporar información especializada y mejorar las representaciones del lenguaje financiero. Sheetal y Aithal

(2025) combinaron el diccionario financiero de Loughran-McDonald con BERT para el análisis de titulares de noticias financieras, integrando señales léxicas específicas del dominio con representaciones contextuales aprendidas por el modelo. Peng et al. (2024), por su parte, propusieron SuCroMoCo, un enfoque de aprendizaje contrastivo supervisado para alinear representaciones de textos financieros con ejemplos prototípicos de categorías de sentimiento, obteniendo mejoras frente a varios modelos de lenguaje evaluados.

La evolución reciente también incluye modelos de lenguaje de mayor escala. Mahdavi Ardekani et al. (2024) desarrollaron FinSentGPT, un modelo de análisis de sentimiento financiero basado en el ajuste de un modelo de lenguaje de gran escala, cuyo desempeño resultó competitivo frente a modelos financieros de referencia y métodos tradicionales de aprendizaje automático en los conjuntos estudiados. De manera complementaria, Sun et al. (2025) propusieron EnhancedFinSentiBERT, integrando preentrenamiento específico del dominio, conocimiento procedente de diccionarios y características orientadas a mejorar la identificación de contenido neutral. Sus resultados mostraron mejoras particularmente en el reconocimiento de la clase neutral.

En conjunto, estos desarrollos muestran una evolución desde arquitecturas generales de representación del lenguaje hacia modelos y estrategias específicamente adaptados al dominio financiero. Para la presente investigación, este antecedente sustenta la elección de FinBERT como modelo base y el uso de fine-tuning para adaptar sus representaciones a una tarea de clasificación emocional multiclase. La utilización de un modelo especializado resulta pertinente frente a las particularidades semánticas del lenguaje

financiero y a la necesidad de distinguir categorías emocionales con diferentes niveles de representación en el corpus.

Análisis de sentimiento y emociones multiclase en textos financieros

El análisis de sentimiento financiero puede abordarse mediante diferentes niveles de representación. Los esquemas categóricos más comunes clasifican los textos según polaridades generales como positiva, negativa o neutral. Sin embargo, una representación más detallada puede incorporar intensidad, aspectos específicos, dimensiones continuas o categorías emocionales discretas. Estas aproximaciones no son equivalentes: el sentimiento de grano fino puede referirse, por ejemplo, a la polaridad asociada con un objetivo específico o a una puntuación continua de intensidad, mientras que una clasificación emocional multiclase busca distinguir estados afectivos diferenciados como miedo, tristeza, enojo o entusiasmo.

Ge et al. (2020) analizaron emociones expresadas en redes sociales durante el desplome del mercado bursátil chino de 2015 mediante un enfoque que combinó dimensiones de valencia y activación con emociones subcategóricas. Sus resultados identificaron al miedo y al disgusto como emociones particularmente relacionadas con la cognición del mercado durante el episodio analizado. Este trabajo muestra que el análisis emocional puede incorporar dimensiones y categorías más específicas que la polaridad positiva o negativa.

El análisis dirigido a objetivos constituye otra forma de aumentar la granularidad. García-Méndez et al. (2023) propusieron un sistema de análisis emocional basado en aspectos

para mensajes financieros de Twitter capaz de identificar señales de oportunidad y precaución asociadas con activos específicos dentro de un mismo mensaje. De manera relacionada, Consoli et al. (2022) desarrollaron FiGAS, un enfoque no supervisado de análisis de sentimiento financiero de grano fino que identifica la polaridad asociada con temas específicos y asigna puntuaciones continuas entre −1 y +1. Lengkeek et al. (2023), por su parte, utilizaron modelos jerárquicos para predecir clases de aspectos financieros y su polaridad correspondiente. Estos estudios muestran que la granularidad puede incrementarse mediante el análisis de aspectos y objetivos concretos sin implicar necesariamente una clasificación de emociones discretas.

Otras investigaciones han profundizado en la representación contextual del sentimiento. Daudert (2021) propuso un enfoque de análisis financiero de sentimiento de grano fino que incorpora contenido textual e información contextual y relacional para facilitar la identificación de sentimientos implícitos. Xiang et al. (2022) desarrollaron un modelo neural que integra información semántica y sintáctica para inferir orientación alcista o bajista e intensidad de sentimiento asociada con entidades financieras específicas. Estos enfoques evidencian que el análisis de sentimiento de grano fino puede involucrar múltiples dimensiones de representación además de la simple polaridad global de un texto.

Dentro de los enfoques centrados específicamente en emociones discretas, Lee et al. (2023) introdujeron StockEmotions, un conjunto de datos compuesto por 10.000 comentarios financieros en inglés recopilados de StockTwits y organizados en doce categorías emocionales de grano fino inspiradas en las finanzas conductuales. A diferencia de los conjuntos tradicionales centrados exclusivamente en polaridad, este recurso incorpora categorías emocionales específicas asociadas con la experiencia del

inversionista. Su propuesta constituye un antecedente especialmente relevante para el análisis multiclase de emociones en contenidos generados por inversionistas.

En conjunto, estos trabajos muestran que el análisis del discurso financiero puede abordarse con distintos niveles de granularidad, desde categorías generales de polaridad hasta representaciones basadas en aspectos, dimensiones continuas y emociones discretas. Una mayor granularidad permite caracterizar con mayor detalle el contenido afectivo del texto, pero también incrementa la complejidad de la tarea de clasificación. Por ello, el desempeño de un esquema multiclase frente a formulaciones más simples debe evaluarse empíricamente y no asumirse como superior de manera previa. En la presente investigación, esta perspectiva sustenta la clasificación de los textos en cinco categorías emocionales —enojo, entusiasmo, miedo, neutralidad y tristeza— y la evaluación explícita del desempeño del modelo mediante métricas de clasificación multiclase.

Fuentes textuales financieras: noticias institucionales versus comunidades retail

La información financiera circula a través de diferentes tipos de fuentes textuales que presentan características comunicativas y discursivas particulares. Las fuentes institucionales, como agencias de noticias y medios financieros especializados, producen contenidos elaborados bajo procedimientos editoriales y orientados a comunicar información económica, corporativa y de mercado. Por su parte, las plataformas digitales y comunidades de inversionistas permiten la publicación directa de opiniones, expectativas y valoraciones generadas por usuarios. Estas diferencias en el origen, propósito y estilo de

comunicación justifican analizar por separado los contenidos institucionales y los generados por inversionistas retail.

Diversos estudios han analizado el potencial de las redes sociales como fuente de información financiera. Oliveira et al. (2017) evaluaron el valor predictivo de variables derivadas de Twitter para diferentes indicadores bursátiles, incluidos retornos, volatilidad, volumen de negociación e índices de sentimiento. Sus resultados mostraron que determinadas características extraídas de Twitter aportaban información útil para algunas tareas de predicción, particularmente para ciertos retornos bursátiles y medidas de sentimiento. Nisar y Yeung (2018), por su parte, analizaron la relación de corto plazo entre el sentimiento expresado en Twitter y los movimientos del FTSE 100 durante un evento político en el Reino Unido. Sus resultados mostraron asociaciones entre el estado de ánimo agregado de los usuarios y determinadas variaciones del mercado, aunque la evidencia estadística no fue concluyente para todas las relaciones evaluadas.

La especificidad lingüística de los contenidos retail también representa un desafío metodológico. Wilksch y Abramova (2023) construyeron un conjunto de 10.000 tweets financieros anotados manualmente y mostraron que distintos modelos de sentimiento entrenados en dominios relacionados presentaban dificultades al enfrentarse con el vocabulario especializado del discurso financiero en Twitter. Los autores desarrollaron posteriormente un modelo específico para esta tarea que superó a varios métodos de referencia, evidenciando la importancia de adaptar los modelos a las características lingüísticas de los contenidos financieros generados por usuarios.

Las noticias financieras institucionales también han sido utilizadas ampliamente como fuente para el análisis de sentimiento. Kelly y Ahmad (2018) estudiaron noticias específicas de distintos dominios financieros y analizaron su relación con activos de los mercados de acciones y petróleo, destacando la importancia de considerar el vocabulario y contexto propios del dominio. Audrino et al. (2020) combinaron medidas de sentimiento y atención provenientes de redes sociales, noticias y otras fuentes digitales y encontraron que algunas de estas variables aportaban mejoras estadísticamente significativas en la predicción de volatilidad, aunque con una magnitud económica limitada. Costola et al. (2023) analizaron 203.886 artículos relacionados con la COVID-19 publicados en MarketWatch, The New York Times y Reuters mediante un modelo BERT adaptado al dominio financiero, encontrando una relación estadísticamente significativa entre el sentimiento de las noticias y el comportamiento del S&P 500 durante el periodo estudiado.

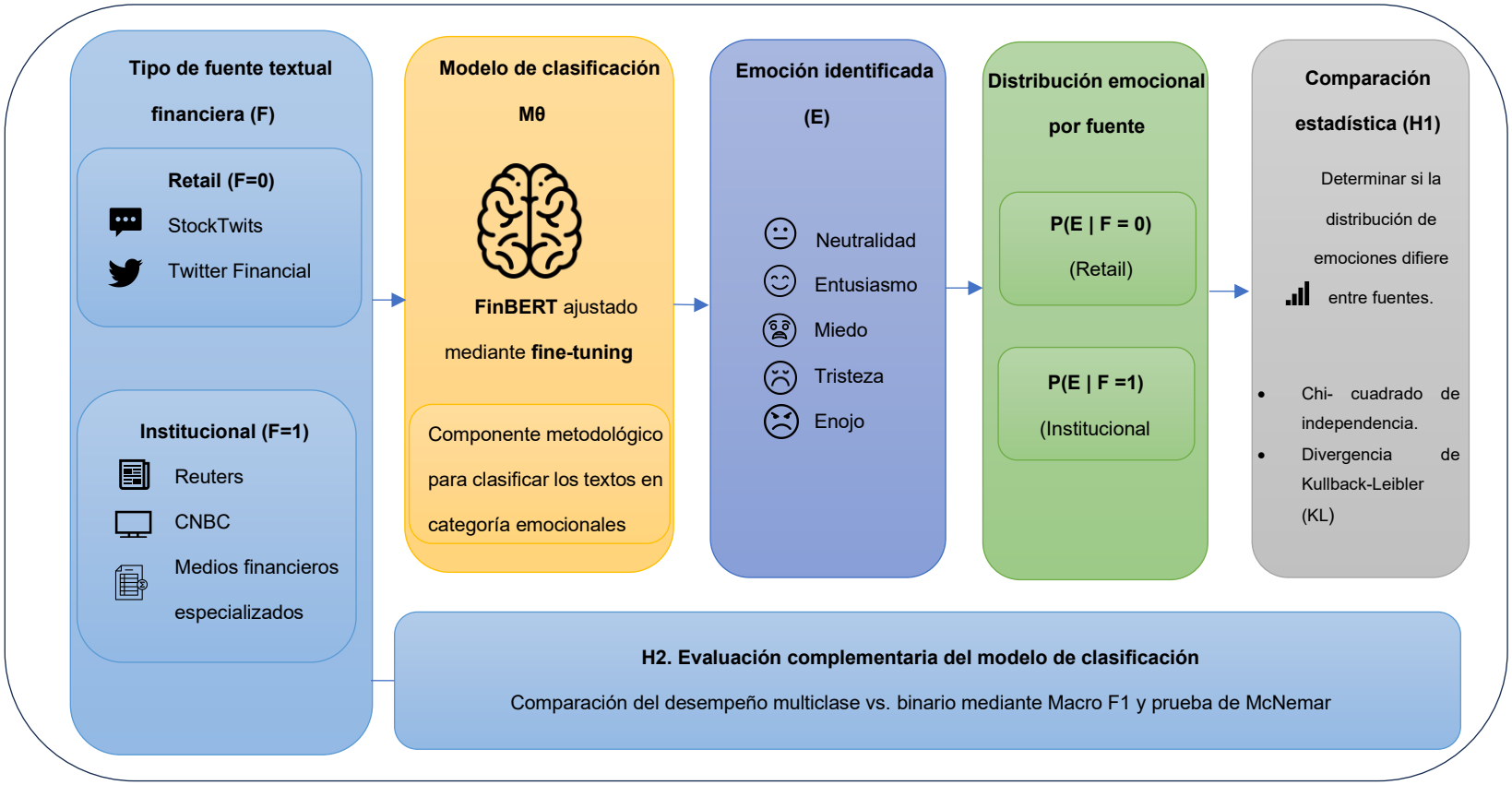
La diferenciación entre participantes retail e institucionales también encuentra respaldo en investigaciones que comparan directamente sus medidas de sentimiento. Cui et al. (2025), utilizando empresas pertenecientes al CSI 300 durante el periodo 2015–2023, encontraron asociaciones diferenciadas entre el sentimiento alcista de inversionistas retail e institucionales y los retornos bursátiles. Aunque su medida de sentimiento se construye a partir de flujos de negociación y no de información textual, sus resultados evidencian que el tipo de participante constituye una dimensión relevante para analizar el sentimiento financiero de manera diferenciada.

En conjunto, estos antecedentes muestran que tanto las noticias financieras como los contenidos generados en plataformas digitales pueden contener información susceptible

de análisis mediante técnicas de sentimiento y procesamiento del lenguaje natural (PLN), aunque presentan características diferentes según su origen y contexto comunicativo. En la presente investigación, esta distinción se operacionaliza mediante dos tipos de fuente: contenidos retail provenientes principalmente de StockTwits y Twitter Financial, y contenidos institucionales procedentes de Reuters, CNBC y otros medios financieros especializados. El propósito no es establecer relaciones causales entre estas fuentes y los precios bursátiles, sino comparar las distribuciones de las emociones identificadas mediante un mismo esquema de clasificación multiclase.

Modelo conceptual propuesto

Figura 1 *Modelo conceptual propuesto*



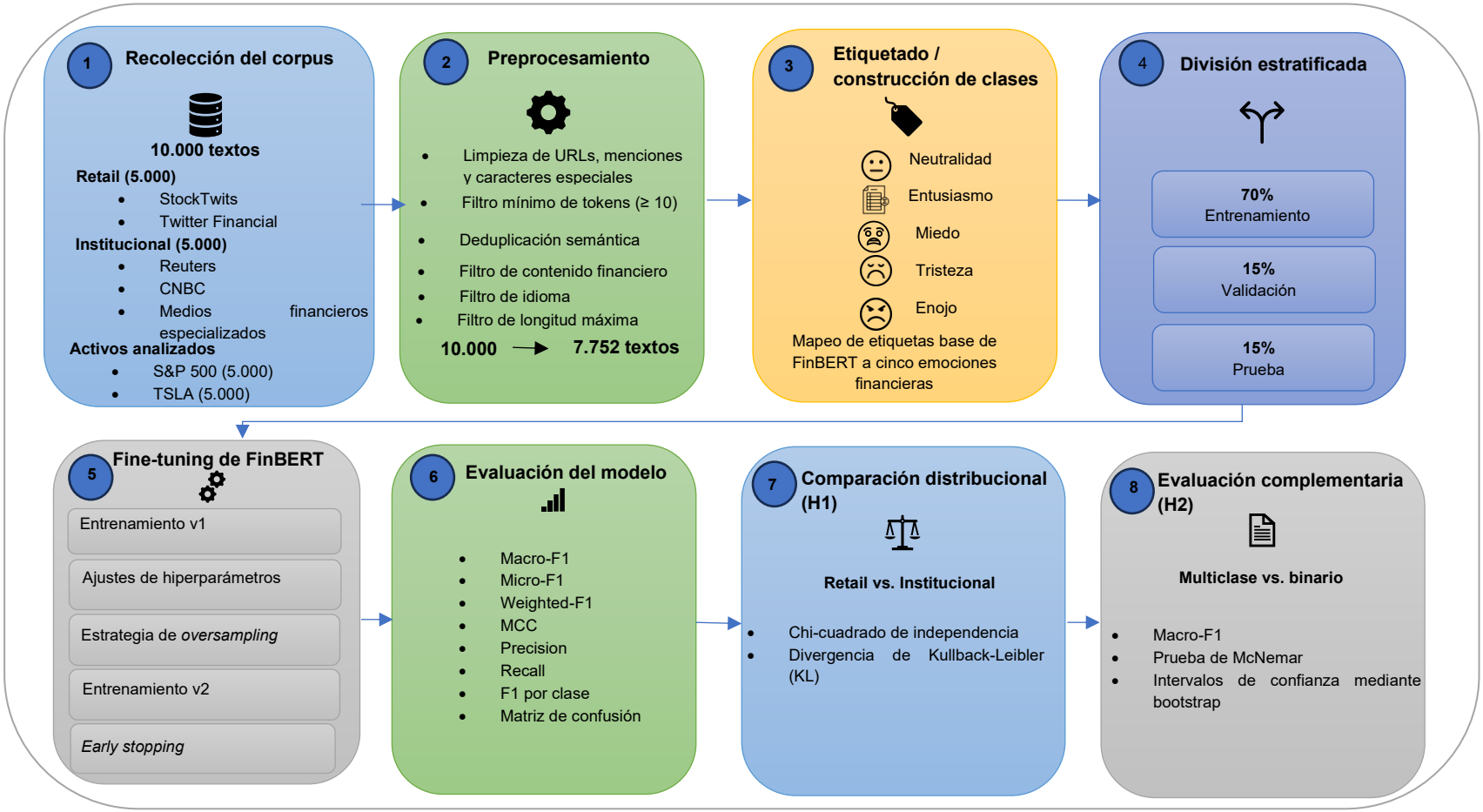
Nota. Elaboración propia

El modelo conceptual de esta investigación, presentado en la Figura 1, organiza los componentes analíticos utilizados para comparar las distribuciones de emociones identificadas en textos financieros procedentes de fuentes retail e institucionales. La variable de agrupación corresponde al tipo de fuente textual financiera, definida como retail ((F=0)) e institucional ((F=1)). Los textos se procesan mediante FinBERT ajustado por fine-tuning para su clasificación en cinco categorías emocionales: neutralidad, entusiasmo, miedo, tristeza y enojo. A partir de estas categorías, se comparan las distribuciones emocionales entre ambos grupos mediante pruebas estadísticas, con el fin de contrastar la hipótesis H1. De manera complementaria, se evalúa el desempeño de la clasificación multiclase frente a una formulación binaria, en el marco de la hipótesis H2. La figura sintetiza las relaciones analíticas examinadas en el estudio y no representa relaciones causales entre el tipo de fuente y las emociones identificadas.

La Figura 2 presenta la secuencia metodológica seguida en la investigación, desde la conformación y depuración del corpus hasta la evaluación del modelo de clasificación y los contrastes estadísticos asociados con las hipótesis H1 y H2. El procedimiento comprende la preparación de los textos financieros, la construcción de las categorías emocionales, la división del corpus en conjuntos de entrenamiento, validación y prueba, el ajuste fino de FinBERT y la evaluación de su desempeño. Asimismo, se distinguen dos componentes analíticos: la comparación de las distribuciones emocionales entre las fuentes retail e institucionales y la evaluación complementaria del desempeño de la clasificación multiclase frente a la formulación binaria. Esta representación resume las etapas del estudio; las decisiones de procesamiento, los criterios de clasificación y las métricas

empleadas se detallan en los apartados metodológicos y en el capítulo de trabajo de campo.

Figura 2 *Pipeline de investigación*



Nota. Elaboración propia

Los textos son procesados mediante un modelo de clasificación Mθ, correspondiente a FinBERT ajustado mediante fine-tuning. En este esquema, FinBERT no se considera una variable moderadora, sino el componente metodológico utilizado para transformar cada texto en una categoría emocional. La salida del modelo es una variable emocional E con cinco clases discretas: enojo, entusiasmo, miedo, neutralidad y tristeza.

La relación conceptual puede expresarse como:

$F \rightarrow M\theta(\text{texto}) \rightarrow E \rightarrow$ comparación de distribuciones emocionales

Este esquema implica que los textos provenientes de cada tipo de fuente son procesados bajo el mismo procedimiento de clasificación, permitiendo estimar las distribuciones condicionales $P(E \mid F = 0)$ para fuentes retail y $P(E \mid F = 1)$ para fuentes institucionales. Posteriormente, dichas distribuciones se comparan mediante estadísticas descriptivas y pruebas de comparación distribucional.

La contrastación principal corresponde a la hipótesis H1, según la cual la distribución de las cinco emociones difiere entre fuentes retail e institucionales. Esta comparación se realiza mediante la prueba Chi-cuadrado de independencia y la divergencia de Kullback-Leibler. De manera complementaria, la hipótesis H2 evalúa el desempeño de la clasificación multiclase frente a una formulación binaria mediante Macro-F1 y la prueba de McNemar, sin asumir previamente que el esquema multiclase presente un desempeño superior.

Estado del arte

La investigación sobre análisis de sentimiento y emociones en textos financieros ha experimentado una evolución metodológica desde enfoques basados principalmente en léxicos y modelos estadísticos hacia arquitecturas de aprendizaje profundo y modelos de lenguaje especializados en el dominio financiero. Para organizar esta evolución, la literatura revisada puede agruparse en cuatro etapas que reflejan cambios en las fuentes de información, las representaciones lingüísticas y las estrategias de modelado empleadas.

En una primera etapa, los estudios se concentraron en métodos léxicos, análisis de contenido y técnicas estadísticas o de aprendizaje automático convencional. Loughran y McDonald (2011) mostraron que los diccionarios generales podían interpretar de manera inadecuada numerosos términos utilizados en documentos financieros y desarrollaron listas de palabras específicas para este dominio. Posteriormente, Kearney y Liu (2014) sistematizaron las principales fuentes de información textual, métodos de análisis de contenido y modelos empleados en finanzas, mientras que Nassirtoussi et al. (2014) realizaron una revisión sistemática de la minería de texto aplicada a la predicción de mercados financieros, identificando fuentes, técnicas de representación y métodos de modelado utilizados en la literatura. Estos trabajos evidenciaron tempranamente la necesidad de adaptar las herramientas de análisis textual a las particularidades del lenguaje financiero.

En una segunda etapa, la investigación amplió las fuentes de información hacia redes sociales, microblogs y noticias digitales. Bollen et al. (2011) analizaron dimensiones

agregadas del estado de ánimo derivadas de Twitter y su relación con los movimientos del Dow Jones Industrial Average, convirtiéndose en un antecedente relevante del uso de contenido generado por usuarios en análisis financiero. Oliveira et al. (2017) y Nisar y Yeung (2018) exploraron posteriormente el valor informativo de Twitter para diferentes indicadores y movimientos bursátiles, mostrando tanto posibilidades como limitaciones de estas fuentes. Desde el ámbito de las noticias, Kelly y Ahmad (2018) analizaron el sentimiento específico del dominio y Audrino et al. (2020) estudiaron la contribución de medidas de sentimiento y atención procedentes de diferentes fuentes digitales al análisis de volatilidad. Paralelamente, Ge et al. (2020) avanzaron más allá de la polaridad general al estudiar dimensiones y categorías emocionales durante el desplome del mercado bursátil chino de 2015.

Una tercera etapa está marcada por una mayor especialización de los modelos y recursos lingüísticos para el dominio financiero. Araci (2019) propuso FinBERT, adaptando BERT al lenguaje financiero mediante entrenamiento adicional sobre un corpus especializado y posterior fine-tuning para tareas de sentimiento. En paralelo, Li et al. (2021) desarrollaron un léxico de sentimiento específico para el contexto financiero chino mediante técnicas de representación de palabras y modelos de aprendizaje profundo, aplicándolo posteriormente a la predicción de distress financiero. Daudert (2021) incorporó información textual, contextual y relacional al análisis de sentimiento financiero de grano fino, mientras que Xiang et al. (2022) integraron información semántica y sintáctica en un modelo neural dirigido a entidades financieras. Consoli et al. (2022), mediante FiGAS, avanzaron en el análisis de sentimiento de grano fino basado en aspectos. En conjunto, estos trabajos muestran una transición desde recursos léxicos especializados hacia

modelos capaces de representar contexto, relaciones y estructura lingüística de manera más compleja.

La etapa más reciente amplía estas líneas mediante modelos de lenguaje preentrenados, aprendizaje contrastivo, modelos de gran escala y representaciones afectivas más granulares. Wilksch y Abramova (2023) desarrollaron un modelo especializado para sentimiento en tweets financieros, mientras que Lengkeek et al. (2023) aplicaron modelos jerárquicos al análisis de sentimiento basado en aspectos. García-Méndez et al. (2023) propusieron un enfoque de análisis emocional dirigido a mensajes financieros de Twitter y Lee et al. (2023) introdujeron StockEmotions, un conjunto de 10.000 comentarios de StockTwits organizados en doce categorías emocionales de grano fino. Jiang y Zeng (2023) utilizaron FinBERT para extraer sentimiento de noticias financieras e incorporaron posteriormente esa información en un modelo LSTM orientado a la predicción de movimientos bursátiles.

En el ámbito metodológico, Peng et al. (2024) introdujeron un enfoque de aprendizaje contrastivo supervisado para mejorar representaciones de sentimiento financiero; Lin y Liao (2024) propusieron prompts enriquecidos con léxicos para análisis dimensional; y Mahdavi Ardekani et al. (2024) exploraron modelos de lenguaje de gran escala mediante FinSentGPT. Más recientemente, Sun et al. (2025) incorporaron conocimiento procedente de diccionarios y características específicas para mejorar la identificación de la clase neutral, mientras que Sheetal y Aithal (2025) integraron el diccionario Loughran-McDonald con BERT para analizar titulares financieros. De manera complementaria, Cui et al. (2025) evidenciaron diferencias entre el sentimiento de participantes retail e institucionales en empresas del CSI 300; aunque su medida de sentimiento se construye a partir de flujos de

negociación y no de textos, sus resultados respaldan la relevancia de distinguir empíricamente entre ambos tipos de participante.

A partir de esta evolución, se identifican tres espacios en los que se sitúa la contribución del presente estudio. En primer lugar, aunque existe una amplia literatura sobre sentimiento derivado de noticias y contenido generado por usuarios, resulta pertinente profundizar en comparaciones realizadas bajo un mismo esquema de clasificación entre fuentes retail e institucionales relacionadas con contextos bursátiles equivalentes. En segundo lugar, el desarrollo de recursos como StockEmotions y de enfoques emocionales dirigidos evidencia el creciente interés por representaciones más granulares que la polaridad tradicional, pero también muestra que una mayor granularidad incrementa la complejidad de la clasificación y requiere ser evaluada empíricamente. En tercer lugar, la literatura reciente sobre FinBERT y otros modelos especializados respalda el uso de modelos de lenguaje adaptados al dominio financiero para representar y clasificar información textual.

En este contexto, la presente investigación integra tres componentes: la caracterización de un corpus compuesto por contenidos retail e institucionales relacionados con el S&P 500 y Tesla Inc.; la clasificación de los textos en cinco categorías emocionales —enojo, entusiasmo, miedo, neutralidad y tristeza— mediante fine-tuning de FinBERT; y la comparación estadística de las distribuciones emocionales identificadas entre ambos tipos de fuente.

Hipótesis

H1: la distribución de las pseudoetiquetas emocionales —enojo, entusiasmo, miedo, neutralidad y tristeza— difiere significativamente entre fuentes retail e institucionales relacionadas con el S&P 500 y Tesla Inc.

Contraste estadístico: prueba Chi-cuadrado de independencia (χ^2), nivel de significancia $\alpha = 0,05$, complementada con la divergencia Kullback-Leibler (KL) para cuantificar las diferencias entre las distribuciones.

H2: Existen diferencias estadísticamente significativas en el desempeño de clasificación entre el modelo multiclase cuyas predicciones se agrupan en las categorías entusiasmo y no entusiasmo, y el clasificador binario entrenado independientemente para distinguir esas mismas categorías, evaluados sobre las mismas observaciones del conjunto de prueba.

Contraste estadístico: comparación de Macro-F1 entre modelo binario y multiclase. Prueba *de* McNemar para diferencias entre clasificadores. Intervalos de confianza *bootstrap*.

Variables

El estudio define dos variables principales alineadas con el modelo conceptual y un componente metodológico de clasificación. La primera corresponde al tipo de fuente textual financiera (F), utilizada para agrupar los textos según su origen retail o institucional. La segunda corresponde a la emoción identificada en cada texto (E), representada mediante cinco categorías discretas: enojo, entusiasmo, miedo, neutralidad y tristeza. Adicionalmente, el modelo FinBERT ajustado mediante fine-tuning ($M\theta$) funciona como componente metodológico para realizar la clasificación emocional, pero no se considera una variable moderadora.

Variable independiente o de agrupación: tipo de fuente textual financiera. **Definición conceptual:** origen del contenido textual financiero según el tipo de participante o canal de producción de la información. **Definición operacional:** F = 0 para fuentes retail, representadas principalmente por StockTwits y Twitter Financial; F = 1 para fuentes institucionales, representadas por Reuters, CNBC y otros medios financieros especializados.

Variable dependiente: emoción identificada en textos financieros. **Definición conceptual:** categoría emocional asignada al contenido textual a partir del proceso de clasificación multiclase. **Definición operacional:** $E \in$ {enojo, entusiasmo, miedo, neutralidad, tristeza}. Cada texto es asignado a una de las cinco categorías mediante el modelo FinBERT ajustado, y las

distribuciones resultantes se comparan entre fuentes retail e institucionales mediante estadísticas descriptivas, prueba Chi-cuadrado y divergencia de Kullback-Leibler.

Componente metodológico: modelo de clasificación Mθ.

Definición conceptual: modelo de lenguaje especializado en el dominio financiero utilizado para representar y clasificar textos.

Definición operacional: FinBERT ajustado mediante fine-tuning sobre el corpus de investigación y evaluado mediante métricas de clasificación multiclase, entre ellas Macro-F1, Micro-F1, Weighted-F1, precisión, recall, F1 por clase y MCC.

Tabla 3 *Operacionalización de las variables y del componente metodológico de clasificación*

Elemento	Definición conceptual	Definición operacional	Indicador	Tipo	Escala de medición
Tipo de fuente textual financiera (F)	Origen del contenido textual financiero según el tipo de fuente que genera o publica la información.	Variable binaria: F = 0, fuentes retail, representadas principalmente por StockTwits y Twitter Financial; F = 1, fuentes institucionales, representadas por Reuters, CNBC y otros medios financieros especializados.	Tipo de fuente: retail / institucional.	Cualitativa nominal dicotómica	Nominal
Emoción identificada (E)	Categoría emocional asignada al contenido de un texto financiero mediante un esquema de clasificación multiclase.	Cada texto es clasificado en una de cinco categorías: enojo, entusiasmo, miedo, neutralidad o tristeza. Las distribuciones de las categorías son posteriormente comparadas entre fuentes retail e institucionales.	Clase emocional; frecuencia y proporción de cada emoción por tipo de fuente.	Cualitativa nominal politómica	Nominal
Modelo de clasificación (Mθ)	Modelo de lenguaje especializado empleado como componente metodológico para representar y clasificar textos del dominio financiero.	FinBERT ajustado mediante fine-tuning sobre el corpus de investigación. Su desempeño se evalúa sobre datos no utilizados en el entrenamiento mediante métricas de clasificación multiclase.	Macro-F1, Micro-F1, Weighted-F1, MCC, precisión, recall y F1 por clase.	Componente metodológico / instrumental	No aplica como variable de medición

Nota. F representa el tipo de fuente textual financiera; E, la emoción identificada; y Mθ, el modelo de clasificación FinBERT ajustado mediante fine-tuning. El modelo Mθ no se considera una variable moderadora, sino un componente metodológico empleado para estimar la categoría emocional de cada texto. Las fuentes retail corresponden principalmente a StockTwits y Twitter Financial, mientras que las fuentes institucionales comprenden Reuters, CNBC y otros medios financieros especializados. La clasificación emocional considera cinco categorías: enojo, entusiasmo, miedo, neutralidad y tristeza. MCC = coeficiente de correlación de Matthews.

Metodología

Enfoque y alcance de la investigación

La investigación adopta un enfoque cuantitativo, sustentado en técnicas de procesamiento del lenguaje natural (PLN), clasificación de emociones y aprendizaje automático aplicadas a textos financieros. Este enfoque permite transformar información textual no estructurada en categorías emocionales, frecuencias, proporciones y métricas de desempeño susceptibles de análisis estadístico. La utilización de técnicas computacionales para extraer y analizar información de textos financieros se encuentra ampliamente documentada en la literatura sobre sentimiento textual y procesamiento del lenguaje financiero (Kearney y Liu, 2014; Araci, 2019).

El alcance de la investigación es descriptivo y comparativo. Es descriptivo porque caracteriza el corpus de textos financieros y examina la frecuencia y distribución de cinco categorías emocionales —enojo, entusiasmo, miedo, neutralidad y tristeza—. Es comparativo porque contrasta estadísticamente las distribuciones emocionales identificadas entre fuentes retail e institucionales mediante estadísticas descriptivas, prueba Chi-cuadrado de independencia y divergencia de Kullback-Leibler (KL).

Adicionalmente, el estudio incorpora un componente evaluativo orientado a determinar el desempeño del modelo FinBERT ajustado mediante fine-tuning en la tarea de clasificación emocional multiclase. Para ello se emplean métricas como Macro-F1, Micro-F1, Weighted-F1, precisión, recall, F1 por clase y coeficiente de correlación de Matthews

(MCC), junto con una comparación complementaria entre la formulación multiclase y una clasificación binaria de polaridad.

En consecuencia, la investigación no busca establecer relaciones causales ni correlacionales entre las emociones identificadas y variables de mercado como retornos o volatilidad. Su propósito metodológico se concentra en caracterizar el corpus, clasificar los textos y comparar las distribuciones emocionales entre tipos de fuente, en correspondencia directa con los objetivos específicos planteados**.**

Diseño y tipo de investigación

El diseño de la investigación es no experimental y retrospectivo. Es no experimental porque no se manipulan deliberadamente las variables ni se interviene sobre los textos analizados; el estudio trabaja con contenidos financieros generados previamente en sus fuentes originales. Es retrospectivo porque utiliza un corpus compuesto por textos producidos durante el periodo comprendido entre enero de 2022 y abril de 2026, los cuales son recopilados y analizados posteriormente mediante técnicas de procesamiento del lenguaje natural y clasificación automática.

El tipo de investigación es descriptivo-comparativo. Su componente descriptivo se orienta a caracterizar el corpus y las distribuciones de las cinco categorías emocionales identificadas —enojo, entusiasmo, miedo, neutralidad y tristeza—. Su componente comparativo examina las diferencias existentes entre las distribuciones emocionales de las fuentes retail e institucionales mediante estadísticas descriptivas y pruebas de comparación distribucional.

Aunque el corpus incluye textos relacionados con el S&P 500 y Tesla Inc. durante un periodo de más de cuatro años, los activos y el intervalo temporal funcionan como criterios de delimitación y contextualización del corpus. En los resultados finales no se realiza un contraste inferencial principal entre activos ni un análisis longitudinal de la evolución emocional por subperiodos. En consecuencia, el diseño no se plantea como correlacional ni busca establecer asociaciones entre las emociones identificadas y variables de mercado como retornos o volatilidad.

Modelo conceptual propuesto

En coherencia con el modelo conceptual presentado en la Figura 1 del marco teórico, la metodología operacionaliza el análisis mediante procedimientos orientados a clasificar los textos financieros y comparar las categorías emocionales identificadas. Para ello, el tipo de fuente textual se utiliza como variable de agrupación, diferenciando las fuentes retail de las institucionales, mientras que la emoción asignada a cada texto constituye la categoría de análisis.

La implementación metodológica comprende la conformación y depuración del corpus, la construcción de las cinco categorías emocionales, el ajuste fino de FinBERT y la evaluación de su desempeño. Posteriormente, las distribuciones emocionales se comparan entre los dos tipos de fuente para contrastar la hipótesis H1. De forma complementaria, se evalúa el desempeño de la clasificación multiclase frente a la formulación binaria en el marco de H2, conforme a las métricas y los procedimientos estadísticos definidos en las subsecciones correspondientes.

Esta operacionalización permite vincular el planteamiento conceptual con los procedimientos de análisis efectivamente desarrollados, sin atribuir relaciones causales al tipo de fuente ni asumir que las categorías asignadas automáticamente representan emociones humanas verificadas.

Especificación estadística formal

Con el propósito de fortalecer la coherencia entre el modelo conceptual, las hipótesis y el análisis empírico, el diseño metodológico incorpora una especificación estadística formal de las variables, las métricas de desempeño y los contrastes utilizados. Esta formalización permite representar de manera explícita la relación entre el tipo de fuente textual, la emoción clasificada y la evaluación del modelo FinBERT ajustado mediante fine-tuning.

En términos formales, $F \in \{0, 1\}$ representa el tipo de fuente textual financiera, donde $F = 0$ corresponde a fuentes retail y $F = 1$ a fuentes institucionales. La variable $E \in \{1, 2, 3, 4, 5\}$ representa la emoción clasificada, correspondiente a las categorías enojo, entusiasmo, miedo, neutralidad y tristeza. Por su parte, M_θ representa el modelo FinBERT ajustado mediante fine-tuning, utilizado como componente metodológico para estimar la categoría emocional de cada texto.

Contraste estadístico de H1

La primera hipótesis evalúa si la distribución de las emociones identificadas difiere significativamente entre fuentes retail e institucionales. En términos estadísticos se establecen las siguientes hipótesis:

$$H_{01}: P(E \mid F = 0) = P(E \mid F = 1)$$

$$H_{11}: P(E \mid F = 0) \neq P(E \mid F = 1)$$

La hipótesis nula H_{01} establece que la distribución de las emociones es independiente del tipo de fuente textual, mientras que la hipótesis alternativa H_{11} plantea que dicha distribución difiere significativamente entre fuentes retail e institucionales.

El contraste principal se realiza mediante la prueba Chi-cuadrado de independencia, apropiada para examinar la asociación entre dos variables categóricas a partir de las diferencias entre las frecuencias observadas y esperadas. El estadístico se expresa como:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

donde O_{ij} representa la frecuencia observada en la celda i, j de la tabla de contingencia y E_{ij} corresponde a la frecuencia esperada bajo la hipótesis de independencia. Para el contraste se adopta un nivel de significancia de $\alpha = 0{,}05$; en consecuencia, un valor de $p < 0.05$ conduce al rechazo de la hipótesis nula.

Divergencia de Kullback-Leibler

De manera complementaria a la prueba Chi-cuadrado, se utiliza la divergencia de Kullback-Leibler para cuantificar la separación entre las distribuciones de probabilidad emocional correspondientes a las fuentes retail e institucionales. La formulación propuesta por Kullback y Leibler (1951) se define como:

$$D_{KL}(P \parallel Q) = \sum_{i=1}^{K} P_i \ln\left(\frac{P_i}{Q_i}\right)$$

donde P_i representa la probabilidad de la emoción i en una de las distribuciones y Q_i la probabilidad correspondiente a esa misma emoción en la distribución de comparación. En esta investigación, $K = 5$, dado que se consideran cinco categorías emocionales.

La divergencia de Kullback-Leibler es asimétrica, por lo que, en general:

$$D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$$

Por esta razón, para la comparación entre fuentes se emplea una versión simétrica, calculada como el promedio de las divergencias obtenidas en ambas direcciones:

$$D_{SKL}(P, Q) = \frac{1}{2}[D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P)]$$

En el procedimiento computacional, las distribuciones se ajustan mediante una constante numérica pequeña cuando es necesario evitar probabilidades iguales a cero y posteriormente se normalizan antes del cálculo de la divergencia. Un valor de $D_{SKL} = 0$ representa distribuciones coincidentes, mientras que valores superiores a cero reflejan separación entre ellas. Debido a que la divergencia de Kullback-Leibler no dispone de umbrales universales para clasificar la magnitud de las diferencias, en esta investigación se utiliza como una medida descriptiva complementaria y no como una prueba de significancia estadística (Kullback y Leibler, 1951).

Evaluación del desempeño del modelo de clasificación

El desempeño del modelo se evalúa sobre un conjunto de prueba independiente mediante métricas derivadas de la matriz de confusión. Para cada clase i se consideran verdaderos positivos (TP_i), falsos positivos (FP_i), falsos negativos (FN_i) y verdaderos negativos (TN_i). Estas cantidades permiten calcular métricas globales y específicas por

clase para caracterizar el comportamiento del clasificador en un problema multiclase (Sokolova y Lapalme, 2009).

La exactitud global o *accuracy* corresponde a la proporción de observaciones correctamente clasificadas respecto al total de observaciones evaluadas:

$$Accuracy = \frac{\sum_{i=1}^{K} T\,P_i}{N}$$

donde N representa el número total de observaciones del conjunto de prueba.

La precisión de la clase i mide la proporción de observaciones predichas como pertenecientes a dicha categoría que fueron correctamente clasificadas:

$$Precision_i = \frac{TP_i}{TP_i + FP_i}$$

El *recall* o sensibilidad de la clase i representa la proporción de observaciones realmente pertenecientes a dicha categoría que fueron identificadas correctamente:

$$Recall_i = \frac{TP_i}{TP_i + FN_i}$$

El F1-score de la clase i corresponde a la media armónica entre precisión y *recall*:

$$F1_i = 2\frac{Precision_i \cdot Recall_i}{Precision_i + Recall_i}$$

Estas métricas permiten identificar diferencias de desempeño entre las distintas categorías emocionales, lo cual resulta especialmente relevante cuando la distribución de las clases no es homogénea.

Macro-F1

El Macro-F1 constituye la métrica principal de evaluación multiclase. Se calcula obteniendo el F1 de cada categoría y posteriormente realizando un promedio no ponderado entre las K clases:

$$MacroF1 = \frac{1}{K}\sum_{i=1}^{K} F\,1_i$$

Al asignar el mismo peso a todas las clases, el Macro-F1 permite que el desempeño de las categorías minoritarias tenga la misma influencia que el de las clases con mayor representación. Esta característica resulta particularmente útil en problemas multiclase con desbalance entre categorías (Sokolova y Lapalme, 2009).

Micro-F1

El Micro-F1 se obtiene agregando primero los verdaderos positivos, falsos positivos y falsos negativos de todas las categorías. La precisión y el *recall* micro se calculan como:

$$Precision_{micro} = \frac{\sum_{i=1}^{K} T\,P_i}{\sum_{i=1}^{K}(TP_i + FP_i)}$$

$$Recall_{micro} = \frac{\sum_{i=1}^{K} T\,P_i}{\sum_{i=1}^{K}(TP_i + FN_i)}$$

A partir de estas cantidades:

$$MicroF1 = 2\frac{Precision_{micro} \cdot Recall_{micro}}{Precision_{micro} + Recall_{micro}}$$

En problemas multiclase de etiqueta única, en los que cada observación pertenece a una sola categoría y recibe una única predicción, Micro-F1 coincide numéricamente con la *accuracy*. Ambas métricas se reportan para mantener una descripción completa y comparable del desempeño global.

Weighted-F1

El Weighted-F1 calcula un promedio de los valores F1 obtenidos por clase, ponderado según el número de observaciones reales de cada categoría:

$$WeightedF1 = \sum_{i=1}^{K} \frac{n_i}{N} F1_i$$

donde n_i representa el número de observaciones pertenecientes a la clase i y N corresponde al total de observaciones. A diferencia del Macro-F1, esta métrica otorga mayor peso a las categorías con mayor representación en el conjunto de evaluación.

Coeficiente de correlación de Matthews

Como medida complementaria global se utiliza el coeficiente de correlación de Matthews (MCC), que permite evaluar la concordancia entre las etiquetas reales y las predicciones considerando la estructura completa de la matriz de confusión. Para un

problema multiclase con matriz de confusión C, la generalización del coeficiente puede expresarse como:

$$MCC = \frac{c\,s - \sum_k p_k\,t_k}{\sqrt{\left(s^2 - \sum_k p_k^2\right)\left(s^2 - \sum_k t_k^2\right)}}$$

donde:

$$c = \sum_k C_{kk}$$

representa el número total de predicciones correctas;

$$s = \sum_{i,j} C_{ij}$$

corresponde al número total de observaciones;

$$p_k = \sum_i C_{ik}$$

representa el número de observaciones predichas como pertenecientes a la clase k; y

$$t_k = \sum_j C_{kj}$$

representa el número de observaciones reales correspondientes a dicha clase.

El MCC toma valores comprendidos generalmente entre -1 y 1. Valores próximos a 1 indican una elevada concordancia entre predicciones y clases reales; valores cercanos a 0, ausencia de asociación apreciable; y valores negativos, desacuerdo sistemático. Para el problema multiclase se utiliza la extensión del coeficiente propuesta para K categorías (Gorodkin, 2004).

Contraste estadístico de H2

La hipótesis H2 se aborda mediante la comparación del desempeño de la clasificación emocional multiclase y su evaluación en una formulación binaria. El modelo FinBERT ajustado mediante fine-tuning produce predicciones en cinco categorías emocionales: enojo, entusiasmo, miedo, neutralidad y tristeza. Para la evaluación binaria, las predicciones se agrupan en dos categorías: entusiasmo y resto de las emociones.

Esta transformación permite examinar cómo varía el desempeño al reducir el número de categorías de decisión. Es importante precisar que la formulación binaria utilizada en esta fase se obtiene mediante la agrupación de las predicciones del modelo multiclase; por tanto, no corresponde a un clasificador binario entrenado de manera independiente.

El desempeño de ambas formulaciones se describe mediante el Macro-F1, calculado sobre las mismas observaciones del conjunto de prueba. Esta comparación permite cuantificar la diferencia observada entre la clasificación de cinco categorías y la evaluación binaria derivada de sus predicciones. Sin embargo, dado que ambas métricas se calculan a partir de las salidas del mismo modelo, la diferencia debe interpretarse como una

comparación entre niveles de granularidad y no como evidencia de superioridad entre dos modelos independientes.

La interpretación de los resultados considera que la formulación binaria reduce la complejidad de la tarea al agrupar cuatro categorías emocionales en una sola clase de referencia. En consecuencia, un Macro-F1 más alto en la evaluación binaria no implica necesariamente una mayor capacidad para distinguir las emociones específicas que constituyen el objetivo principal de esta investigación.

Para establecer si las diferencias observadas entre los clasificadores son estadísticamente significativas se emplea la prueba de McNemar, apropiada para comparar resultados dicotómicos emparejados obtenidos sobre las mismas observaciones (McNemar, 1947).

La comparación puede representarse mediante una tabla de contingencia 2×2, en la que b corresponde al número de observaciones correctamente clasificadas por el primer enfoque e incorrectamente por el segundo, mientras que $c$representa el patrón inverso. El contraste se concentra en estas observaciones discordantes.

El estadístico asintótico clásico de McNemar se expresa como:

$$\chi^2_{McNemar} = \frac{(b-c)^2}{b+c}$$

La hipótesis nula establece que ambas probabilidades de discordancia son iguales:

$$H_{02}: b = c$$

mientras que la hipótesis alternativa establece:

$$H_{12}: b \neq c$$

Con un nivel de significancia de $\alpha = 0{,}05$, un valor de $p < 0{,}05$ permite rechazar la hipótesis nula y concluir que los clasificadores presentan diferencias estadísticamente significativas en sus patrones de acierto y error.

Es importante distinguir la función de ambos procedimientos: Macro-F1 permite describir la magnitud y dirección de las diferencias de desempeño, mientras que la prueba de McNemar determina si las discordancias observadas entre las predicciones apareadas son estadísticamente significativas.

Intervalos de confianza mediante bootstrap

La incertidumbre asociada a las métricas de desempeño se estima mediante un procedimiento *bootstrap*. Este método aproxima la distribución muestral de una estadística mediante múltiples muestras obtenidas con reemplazo a partir del conjunto original (Efron y Tibshirani, 1993).

Si $\hat{\theta}$ representa la métrica de interés y se generan B muestras *bootstrap*, cada réplica se expresa como:

$$\hat{\theta}^{*(b)} = T\left(X^{*(b)}\right),\ b = 1, \ldots, B$$

donde $X^{*(b)}$ representa una muestra obtenida mediante remuestreo con reemplazo y $T(\cdot)$ corresponde a la función empleada para calcular la estadística de interés.

En esta investigación se utilizan $B = 1.000$ iteraciones de remuestreo. Para un intervalo de confianza percentil del 95%, los límites se calculan como:

$$IC_{95\%} = \left[Q_{0.025}(\hat{\theta}^*), Q_{0.975}(\hat{\theta}^*)\right]$$

donde $Q_{0,025}$ y $Q_{0,975}$ representan los percentiles 2,5 y 97,5 de la distribución empírica obtenida mediante *bootstrap*. Los intervalos de confianza complementan las estimaciones puntuales al proporcionar información sobre la incertidumbre asociada al desempeño observado.

Integración de los procedimientos estadísticos

En conjunto, los procedimientos descritos permiten analizar dos dimensiones complementarias de la investigación. Para H1, la prueba Chi-cuadrado determina si existe asociación estadísticamente significativa entre el tipo de fuente y la distribución emocional, mientras que la divergencia simétrica de Kullback-Leibler cuantifica el grado de separación entre las distribuciones de las fuentes retail e institucionales.

En la evaluación del clasificador, *accuracy* y Micro-F1 describen el desempeño global; precisión, *recall* y F1 permiten examinar el comportamiento específico de cada emoción; Macro-F1 asigna igual importancia a todas las clases; Weighted-F1 incorpora la frecuencia relativa de las categorías; y MCC proporciona una medida global de concordancia basada en la matriz de confusión.

Para H2, la comparación de Macro-F1 permite identificar la dirección y magnitud de la diferencia de desempeño entre las formulaciones multiclase y binaria, mientras que la prueba de McNemar determina si las diferencias entre las predicciones apareadas son estadísticamente significativas. Finalmente, los intervalos de confianza mediante *bootstrap* permiten cuantificar la incertidumbre asociada a las métricas estimadas.

De esta manera, la evaluación estadística no depende de una única medida ni de umbrales arbitrarios de desempeño, sino de la interpretación conjunta de los indicadores, los contrastes estadísticos y la incertidumbre asociada a las estimaciones.

Aplicación analítica

La especificación estadística definida permite integrar de manera coherente los dos componentes analíticos principales de la investigación. En primer lugar, la hipótesis H1 se evalúa mediante la comparación de las distribuciones emocionales observadas en fuentes retail e institucionales. Para este propósito, la prueba Chi-cuadrado de independencia permite determinar si existe una asociación estadísticamente significativa entre el tipo de fuente textual y la categoría emocional identificada, mientras que la divergencia simétrica de Kullback-Leibler complementa este contraste al cuantificar el grado de separación entre ambas distribuciones.

En segundo lugar, el desempeño del modelo FinBERT ajustado mediante fine-tuning se evalúa sobre un conjunto de prueba independiente utilizando métricas globales y por clase. Macro-F1 constituye la métrica principal debido a que asigna igual importancia a cada categoría emocional, mientras que *accuracy*, Micro-F1, Weighted-F1, precisión, *recall*, F1 por clase y el coeficiente de correlación de Matthews (MCC) permiten caracterizar de

manera complementaria el comportamiento del clasificador. Los intervalos de confianza estimados mediante *bootstrap* aportan información adicional sobre la incertidumbre asociada a las métricas de desempeño.

De manera complementaria, la hipótesis H2 compara el desempeño de la formulación multiclase con una formulación binaria sobre las mismas observaciones del conjunto de prueba. La comparación descriptiva se realiza mediante Macro-F1 y la prueba de McNemar permite establecer si las diferencias entre los patrones de acierto y error de ambos clasificadores son estadísticamente significativas. Esta comparación se interpreta considerando que la tarea binaria presenta un espacio de decisión menos complejo que la clasificación de cinco emociones, por lo que un mayor desempeño del modelo binario no implica necesariamente una mayor pertinencia para el objetivo analítico del estudio.

Fases de la investigación

La investigación se desarrolló en seis fases, cada una alineada con los objetivos específicos del estudio.

Fase 1. Identificación y recolección del corpus

La primera fase consistió en la identificación y recopilación de textos financieros en inglés relacionados con el índice S&P 500 y Tesla Inc. (TSLA), correspondientes al periodo comprendido entre enero de 2022 y abril de 2026. El corpus se construyó a partir de dos tipos de fuente textual: contenidos generados por inversionistas retail y contenidos procedentes de fuentes institucionales.

Las fuentes retail estuvieron representadas principalmente por mensajes provenientes de StockTwits y Twitter Financial, mientras que las fuentes institucionales correspondieron a artículos y contenidos publicados por Reuters, CNBC y otros medios financieros especializados. La utilización de contenidos procedentes de plataformas de microblogging resulta pertinente en estudios financieros debido a que estas fuentes permiten analizar expresiones, expectativas y señales de sentimiento generadas por participantes del mercado (Oliveira et al., 2017; Wilksch y Abramova, 2023).

Origen y procedimiento de extracción de los datos

La recolección de noticias financieras institucionales se realizó mediante la API de Finnhub, utilizando el endpoint company-news. El procedimiento se implementó en Python con las bibliotecas requests y pandas, que permitieron realizar las solicitudes al servicio y organizar los registros obtenidos. La consulta se configuró mediante el símbolo del activo

financiero y los parámetros from y to, correspondientes a las fechas de inicio y finalización del periodo de consulta. La ventana temporal definida para la investigación comprendió desde enero de 2022 hasta abril de 2026.

En la configuración de la consulta se utilizaron los símbolos TSLA, correspondiente a Tesla Inc., y SPY, utilizado como instrumento de referencia asociado al mercado del S&P 500. Por tanto, la consulta de SPY debe entenderse como una aproximación basada en un instrumento negociado en bolsa, y no como una consulta directa a un índice bursátil mediante un símbolo propio. Los registros recuperados se incorporaron al conjunto de noticias financieras y se organizaron para su posterior depuración, clasificación por activo y análisis.

La investigación diferencia esta fuente institucional de las fuentes retail, representadas por contenidos de StockTwits y Twitter Financial. En consecuencia, Finnhub se documenta como el mecanismo de extracción de las noticias institucionales y no como el origen de todos los registros del corpus.

El corpus inicial quedó conformado por 10.000 textos. La distribución se estructuró de manera equilibrada según el tipo de fuente, con 5.000 registros provenientes de fuentes retail y 5.000 de fuentes institucionales. Asimismo, el corpus presentó un balance exacto por activo financiero, con 5.000 textos asociados con el S&P 500 y 5.000 relacionados con Tesla Inc. Esta estructura permitió evitar que una de las dos categorías de fuente o uno de los activos concentrara desproporcionadamente la muestra y facilitó la comparación posterior de las distribuciones emocionales.

Cada registro conservó información básica de trazabilidad, entre ella el contenido textual, la fuente de procedencia, el tipo de fuente y el activo financiero asociado. Estos metadatos permitieron identificar el origen de cada observación y mantener su correspondencia durante las etapas posteriores de depuración, clasificación y análisis.

La selección de fuentes retail se sustenta además en evidencia previa sobre la utilidad de contenidos financieros generados por usuarios para tareas de análisis de sentimiento. Oliveira et al. (2017) mostraron que variables derivadas de Twitter pueden aportar información relevante en diferentes tareas de análisis financiero, mientras que Wilksch y Abramova (2023) evidenciaron la importancia de considerar las particularidades del vocabulario utilizado en tweets financieros al construir modelos de sentimiento especializados.

Fase 2. Preprocesamiento y depuración

La segunda fase consistió en el preprocesamiento y depuración del corpus con el propósito de reducir ruido textual, eliminar registros poco informativos y conservar las características lingüísticas relevantes para el análisis financiero. Dado que el lenguaje financiero presenta vocabulario especializado, dependencias contextuales y expresiones cuyo significado puede diferir del lenguaje general, el procesamiento se diseñó procurando preservar términos relevantes del dominio (Daudert, 2021; Xiang et al., 2022; Consoli et al., 2022).

Inicialmente, se aplicó una función de limpieza textual para eliminar URLs, menciones y caracteres especiales, normalizar los espacios y convertir el contenido a minúsculas. En el tratamiento de los hashtags se eliminó el símbolo #, conservando la palabra asociada.

Asimismo, el procedimiento contempló la eliminación de números aislados, la tokenización, la lematización y la eliminación de palabras vacías (*stopwords*) en inglés mediante herramientas de NLTK. Estas operaciones se orientaron a reducir el ruido textual y preparar los registros para las etapas posteriores del análisis. No obstante, dado que las transformaciones lingüísticas pueden modificar información contextual relevante para los modelos basados en arquitecturas Transformer, se precisa en la fase siguiente el texto utilizado como entrada de FinBERT y el alcance de las operaciones aplicadas.

La eliminación de *stopwords* fue adaptada al dominio financiero mediante la conservación deliberada de términos y negaciones con posible valor semántico para la clasificación, entre ellos *bull*, *bear*, *fear*, *risk*, *crash*, *rally*, *buy*, *sell*, *up*, *down*, *not*, *no* y *never*. Esta adaptación buscó evitar que el procedimiento de normalización eliminara expresiones relevantes para interpretar el contenido financiero, en concordancia con la importancia de considerar las particularidades del vocabulario del dominio documentadas en la literatura sobre análisis textual financiero (Kearney y Liu, 2014; Araci, 2019).

Después de la limpieza se aplicó un filtro mínimo de longitud, mediante el cual se eliminaron los textos con menos de 10 tokens. Asimismo, se efectuó un procedimiento de deduplicación semántica por fuente utilizando representaciones TF-IDF y similitud coseno, considerando como altamente similares los registros con una similitud superior a 0.95. Este procedimiento complementó la eliminación previa de duplicados exactos realizada durante la conformación del corpus.

Posteriormente, se aplicaron filtros orientados a excluir contenido no financiero, spam y otros registros irrelevantes mediante patrones de expresiones regulares. El Trabajo de

Campo documenta además la aplicación de un filtro de idioma para conservar textos en inglés.

Se establecieron límites máximos de 100 tokens para los textos retail y 400 tokens para los textos institucionales como criterios operativos de preprocesamiento. Esta diferenciación buscó considerar las características de extensión de los contenidos provenientes de ambos tipos de fuente. No obstante, los límites no deben interpretarse como umbrales universales ni como valores cuya superioridad haya sido demostrada experimentalmente en esta investigación. Su aplicación pudo ocasionar la pérdida de información en textos que superaban los valores definidos, así como diferencias en la composición del corpus final. Por ello, se reconoce la necesidad de evaluar la sensibilidad de los resultados frente a otras configuraciones de longitud y de examinar las tasas de exclusión por fuente y activo financiero.

Como resultado del proceso de preprocesamiento, el corpus inicial de 10.000 textos se redujo a 7.752 registros, lo que representa la eliminación de 2.248 observaciones, equivalentes al 22,5% del corpus original. El mayor número de exclusiones correspondió al filtro de longitud máxima, seguido del filtro de contenido no financiero. El corpus depurado obtenido en esta fase constituyó la base para el procedimiento posterior de etiquetado y clasificación emocional.

Fase 3. Representación semántica de los textos financieros

La tercera fase consistió en transformar los textos preprocesados en representaciones compatibles con el modelo FinBERT y construir las etiquetas iniciales utilizadas en las etapas posteriores del análisis. Para ello se empleó el modelo ProsusAI/finbert, basado en BERT y adaptado al dominio financiero, cuya especialización permite representar contextualizadamente expresiones propias de noticias, reportes y otros textos financieros (Araci, 2019).

Inicialmente, se aplicó una función de limpieza textual para eliminar URLs, menciones y caracteres especiales, normalizar los espacios y convertir el contenido a minúsculas. En el tratamiento de los hashtags se eliminó el símbolo #, conservando la palabra asociada. Asimismo, el procedimiento contempló la eliminación de números aislados, la tokenización, la lematización y la eliminación de palabras vacías (stopwords) en inglés mediante herramientas de NLTK. Estas operaciones se orientaron a reducir el ruido textual y preparar los registros para las etapas posteriores del análisis. No obstante, dado que las transformaciones lingüísticas pueden modificar información contextual relevante para los modelos basados en arquitecturas Transformer, se precisa en la fase siguiente el texto utilizado como entrada de FinBERT y el alcance de las operaciones aplicadas.

La representación contextual generada por FinBERT fue utilizada posteriormente para obtener una clasificación inicial de sentimiento en las categorías positive, negative y neutral. Esta estrategia se sustenta en el uso de modelos de lenguaje especializados para capturar relaciones semánticas y contextuales del lenguaje financiero, aspecto relevante cuando el significado de una expresión depende del dominio y del contexto en el que aparece (Araci, 2019; Daudert, 2021). Estudios posteriores también han mostrado la

utilidad de modelos contextuales y jerárquicos para representar sentimientos y aspectos específicos en textos financieros (Lengkeek et al., 2023).

Debido a que la salida original de FinBERT corresponde a tres categorías generales de sentimiento y la investigación requiere cinco categorías emocionales, se implementó un procedimiento de mapeo para construir las pseudoetiquetas. Los textos clasificados como positivos se asociaron con entusiasmo y los clasificados como neutrales con neutralidad. Los textos clasificados inicialmente como negativos se subdividieron en miedo, enojo y tristeza mediante listas de términos léxicos definidas previamente en el código. La categoría se asignó según el mayor número de coincidencias con los términos de cada lista; cuando no se identificaron coincidencias, se asignó miedo como categoría predeterminada. Esta última regla constituye una decisión operativa para completar la asignación automática y no una evidencia de que el texto exprese necesariamente miedo. Por ello, las cinco categorías se interpretan como pseudoetiquetas generadas mediante reglas léxicas y predicciones de FinBERT, y no como emociones verificadas de manera independiente para la totalidad del corpus.

Este procedimiento produjo cinco clases emocionales: enojo, entusiasmo, miedo, neutralidad y tristeza. Las etiquetas obtenidas se consideran pseudoetiquetas de referencia, dado que fueron generadas mediante una combinación de la predicción de FinBERT y reglas léxicas, y no corresponden a una anotación humana exhaustiva del corpus. Esta distinción resulta importante para la interpretación posterior del desempeño del modelo.

Una vez construido el corpus etiquetado, se realizó una división estratificada en tres subconjuntos: 70% para entrenamiento, 15% para validación y 15% para prueba, utilizando

una semilla aleatoria fija para asegurar reproducibilidad. La estratificación permitió conservar aproximadamente la proporción de cada categoría emocional en los tres conjuntos. El corpus quedó distribuido en 5.426 observaciones de entrenamiento, 1.163 de validación y 1.163 de prueba.

El resultado de esta fase fue un corpus estructurado y etiquetado que sirvió como insumo para el fine-tuning posterior de FinBERT y para la evaluación independiente del clasificador.

Las pseudoetiquetas generadas en esta fase constituyeron la variable categórica utilizada posteriormente para caracterizar el corpus y contrastar H1 mediante la comparación de sus distribuciones entre fuentes retail e institucionales. En consecuencia, el contraste de H1 se realizó sobre estas pseudoetiquetas y no sobre las predicciones del modelo posteriormente ajustado mediante fine-tuning.

Fase 4. Identificación y clasificación de emociones

La cuarta fase consistió en el ajuste y entrenamiento del modelo FinBERT mediante fine-tuning supervisado sobre el corpus pseudoetiquetado generado en la fase anterior. Esta etapa tuvo como propósito entrenar y evaluar un clasificador multiclase capaz de reproducir las cinco categorías emocionales definidas en el corpus. El modelo fine-tuned no se utilizó para generar las distribuciones empleadas en la contrastación de H1; dichas distribuciones procedieron de las pseudoetiquetas construidas previamente mediante FinBERT base y reglas heurísticas. Se seleccionó FinBERT por tratarse de un modelo basado en BERT y adaptado al dominio financiero, diseñado para representar las

particularidades semánticas presentes en este tipo de textos y mejorar el desempeño en tareas de clasificación de sentimiento financiero (Araci, 2019).

El corpus etiquetado se dividió de manera estratificada en tres subconjuntos: 70% para entrenamiento, 15% para validación y 15% para prueba. Esta partición permitió conservar aproximadamente la distribución relativa de las cinco categorías emocionales en cada subconjunto y separar un conjunto de prueba independiente que no participó en el ajuste de los parámetros del modelo. La división final estuvo conformada por 5.426 textos de entrenamiento, 1.163 de validación y 1.163 de prueba.

El proceso de fine-tuning de FinBERT se desarrolló mediante dos configuraciones experimentales. La versión inicial (v1) utilizó una tasa de aprendizaje de 2e-5, un tamaño de lote de 16, una longitud máxima de entrada de 512 tokens y un máximo de cinco épocas. En la segunda configuración (v2), se estableció una tasa de aprendizaje de 1e-5, un tamaño de lote de 32, una longitud máxima de 128 tokens, un warmup ratio de 0,10 y un máximo de ocho épocas. Adicionalmente, se mantuvo una función de pérdida ponderada y se aplicó a la clase enojo, aumentando su representación de 39 a 200 observaciones dentro del conjunto de entrenamiento.

Para describir el desbalance de las categorías emocionales se utilizaron umbrales operativos definidos a partir de la proporción de cada clase en el corpus etiquetado: superior al 5% para las categorías representadas, entre el 1% y el 5% para las minoritarias e inferior al 1% para las críticas. La aplicación de estos criterios permitió identificar la categoría de enojo como la de menor representación, con 55 textos (0,7% del corpus). Por esta razón, se aplicó *oversampling* a dicha categoría en el conjunto de entrenamiento, con el propósito de aumentar su presencia durante el ajuste del modelo. Los umbrales se

utilizaron para describir la distribución y orientar el tratamiento de la clase menos frecuente; no constituyen límites universales para determinar el desbalance. La validación y el conjunto de prueba conservaron la distribución correspondiente a la división estratificada, de modo que la evaluación no se realizara sobre datos artificialmente sobremuestreados.

La versión final v2 utilizó una tasa de aprendizaje de 1×10^{-5}, un tamaño de lote de 32, una longitud máxima de entrada de 128 tokens, un warmup ratio de 0,10 y un máximo de ocho épocas. Asimismo, se mantuvo una función de pérdida ponderada para compensar el desbalance entre clases y se incorporó *early stopping* con una paciencia de tres épocas consecutivas sin mejora en Macro-F1 de validación. La reducción de la longitud máxima de 512 a 128 tokens disminuyó la cantidad de texto procesado por secuencia y pudo reducir el costo computacional del entrenamiento; sin embargo, también pudo provocar la pérdida de información en los textos que superaban dicho límite. Esta restricción resulta especialmente relevante para las fuentes institucionales, cuyo límite de longitud durante la depuración fue de 400 tokens, frente a 100 tokens para las fuentes retail. Por tanto, los textos institucionales de mayor extensión pudieron experimentar un truncamiento más pronunciado, mientras que una proporción mayor de los textos retail pudo conservarse dentro del límite del modelo. Este efecto constituye una limitación metodológica que debe considerarse al interpretar el desempeño del clasificador y las diferencias observadas entre fuentes.

Debido a la escasa representación de la categoría enojo, que contaba con 39 registros originales en el conjunto de entrenamiento, se aplicó *oversampling* exclusivamente sobre esta clase, aumentando temporalmente su representación a 200 observaciones durante el entrenamiento. Esta estrategia se utilizó únicamente en el conjunto de entrenamiento, sin

alterar los conjuntos de validación y prueba, con el propósito de reducir el efecto del desbalance sobre el aprendizaje del clasificador.

La selección del modelo v2 se realizó con base en el Macro-F1 del conjunto de validación. El entrenamiento registró ocho épocas y alcanzó su mayor Macro-F1 de validación en la época 8, con un valor de 0,5654. El modelo seleccionado se guardó para su posterior evaluación sobre el conjunto de prueba independiente.

El ajuste mediante fine-tuning de modelos de lenguaje especializados constituye una estrategia ampliamente utilizada para adaptar representaciones preentrenadas a tareas específicas del dominio financiero. Araci (2019) demostró la utilidad de FinBERT para clasificación de sentimiento financiero; posteriormente, otros trabajos han explorado estrategias de ajuste y especialización mediante BERT, conocimiento léxico y aprendizaje contrastivo supervisado para mejorar la representación y clasificación de textos financieros (Peng et al., 2024; Sheetal y Aithal, 2025).

El resultado de esta fase fue un modelo FinBERT ajustado para clasificar textos financieros en cinco categorías emocionales —enojo, entusiasmo, miedo, neutralidad y tristeza—, cuya versión final fue posteriormente evaluada sobre el conjunto de prueba independiente.

Fase 5. Comparación de la distribución emocional entre fuentes

La quinta fase consistió en comparar las distribuciones de las cinco categorías emocionales identificadas en los textos provenientes de fuentes retail e institucionales. Esta comparación respondió directamente a la hipótesis H1 y al tercer objetivo específico

de la investigación, orientado a determinar si ambos tipos de fuente presentan perfiles emocionales estadísticamente diferenciados.

Inicialmente se calcularon frecuencias absolutas y proporciones de enojo, entusiasmo, miedo, neutralidad y tristeza para cada tipo de fuente. Estas estadísticas descriptivas permitieron caracterizar la distribución emocional de los contenidos retail $(F = 0)$ y de los contenidos institucionales $(F = 1)$, utilizando el mismo esquema de clasificación emocional para ambos grupos.

Posteriormente se aplicó la prueba Chi-cuadrado de independencia para evaluar si la distribución de las categorías emocionales era independiente del tipo de fuente textual. El contraste se realizó con un nivel de significancia de $\alpha = 0.05$, de modo que un valor de $p < 0.05$ fue considerado evidencia suficiente para rechazar la hipótesis nula de independencia entre ambas variables.

De manera complementaria, se calculó la divergencia de Kullback-Leibler para cuantificar la separación entre las distribuciones emocionales de las dos fuentes. Debido al carácter asimétrico de la formulación original, se calcularon las divergencias $D_{KL}(P_{retail} \parallel P_{institucional})$ y $D_{KL}(P_{institucional} \parallel P_{retail})$, y posteriormente se obtuvo una versión simétrica mediante el promedio de ambas direcciones, siguiendo la formulación metodológica descrita previamente (Kullback y Leibler, 1951).

La combinación de ambos procedimientos permitió abordar dos dimensiones complementarias del contraste. La prueba Chi-cuadrado permitió determinar si existía una asociación estadísticamente significativa entre el tipo de fuente y la categoría emocional, mientras que la divergencia simétrica de Kullback-Leibler permitió cuantificar el grado de

separación entre las distribuciones. La KL se interpretó como una medida descriptiva de divergencia y no como una prueba adicional de significancia estadística.

En esta investigación, los activos S&P 500 y Tesla Inc. delimitaron el contexto financiero del corpus y permitieron reunir textos de ambas fuentes relacionados con los activos considerados. Sin embargo, la contrastación estadística principal se centró en las diferencias entre las distribuciones de pseudoetiquetas emocionales de las fuentes retail e institucionales. No se realizó un contraste inferencial independiente para cada activo ni se evaluó si la asociación entre el tipo de fuente y la categoría emocional variaba entre ellos. En consecuencia, los hallazgos permiten caracterizar las diferencias entre fuentes en el corpus conjunto, mientras que el análisis de las diferencias por activo constituye una ampliación posible de la investigación.

La pertinencia de distinguir entre tipos de fuente se encuentra respaldada por investigaciones que han mostrado que los contenidos generados por participantes retail y las fuentes institucionales pueden presentar características diferenciadas de sentimiento y comportamiento financiero. Asimismo, estudios sobre noticias financieras y redes sociales evidencian que ambos tipos de contenido constituyen fuentes relevantes para el análisis textual del mercado (Kelly y Ahmad, 2018; Costola et al., 2023).

La comparación inferencial de esta fase se centró en la relación entre el tipo de fuente textual —retail e institucional— y la categoría emocional, mediante la prueba chi-cuadrado de independencia. Aunque el corpus incluyó textos relacionados con el S&P 500 y Tesla Inc., el análisis principal no evaluó de manera independiente la asociación entre el activo financiero y la categoría emocional, ni si la diferencia entre fuentes variaba según el activo. Por consiguiente, los resultados de esta fase deben interpretarse como una comparación

entre tipos de fuente para el corpus analizado en conjunto, sin atribuir diferencias estadísticamente significativas a cada activo por separado.

Fase 6. Evaluación del desempeño de los modelos

La sexta fase consistió en evaluar el desempeño del modelo FinBERT ajustado mediante fine-tuning sobre el conjunto de prueba independiente, conformado por 1.163 textos que no participaron en las etapas de entrenamiento ni validación. Esta separación permitió estimar el comportamiento del clasificador sobre observaciones no utilizadas durante el ajuste de sus parámetros.

La evaluación incluyó *accuracy*, Macro-F1, Micro-F1, Weighted-F1, precisión, *recall* y F1 por clase, así como el coeficiente de correlación de Matthews (MCC) y la matriz de confusión. Estas métricas permiten examinar dimensiones complementarias del desempeño y resultan especialmente útiles en problemas multiclase con distribuciones desbalanceadas, donde una única medida global puede ocultar diferencias sustanciales entre categorías (Sokolova y Lapalme, 2009; Gorodkin, 2004).

Macro-F1 se utilizó como métrica principal de evaluación, debido a que calcula el promedio no ponderado del F1 de todas las clases y, por tanto, asigna el mismo peso a categorías frecuentes y minoritarias. Esta característica resulta particularmente pertinente en el presente estudio, dado que la distribución emocional del corpus presenta una elevada concentración de neutralidad frente a clases menos frecuentes como tristeza y enojo. El análisis se complementó con Micro-F1 y Weighted-F1 para caracterizar el rendimiento agregado del clasificador, y con precisión, *recall* y F1 por clase para identificar diferencias específicas en su capacidad discriminante.

El MCC se utilizó como medida adicional de concordancia entre las predicciones y las etiquetas de referencia, dado que permite resumir el desempeño de una clasificación multiclase considerando simultáneamente los aciertos y errores distribuidos entre todas las categorías (Gorodkin, 2004). La matriz de confusión complementó este análisis al permitir observar los patrones específicos de clasificación y confusión entre las cinco emociones.

Para cuantificar la incertidumbre asociada con el desempeño del modelo se estimó un intervalo de confianza del 95% para Macro-F1 mediante *bootstrap*. El procedimiento utilizó 1.000 iteraciones de remuestreo con reemplazo sobre el conjunto de prueba, manteniendo en cada iteración un tamaño de 1.163 observaciones. La construcción de intervalos mediante remuestreo *bootstrap* constituye una estrategia ampliamente utilizada para aproximar la variabilidad de una estadística sin depender estrictamente de supuestos paramétricos sobre su distribución (Efron y Tibshirani, 1993).

De manera complementaria, se evaluó la hipótesis H2 mediante la comparación entre la formulación multiclase de cinco emociones y una formulación binaria sobre las mismas observaciones del conjunto de prueba. Macro-F1 permitió describir la magnitud y dirección de la diferencia de desempeño entre ambas configuraciones. Para realizar el contraste inferencial, las predicciones multiclase se llevaron al mismo espacio binario utilizado como línea base y se compararon de forma apareada mediante la prueba de McNemar, que permite evaluar si existen diferencias significativas entre los patrones de acierto y error de dos clasificadores aplicados a las mismas observaciones (McNemar, 1947).

La interpretación de esta comparación consideró que la formulación binaria presenta una tarea de clasificación menos compleja que la formulación de cinco emociones. En consecuencia, un Macro-F1 superior en la tarea binaria no se interpretó como evidencia

de superioridad conceptual frente al enfoque multiclase, sino como una diferencia de desempeño que debe analizarse teniendo en cuenta la distinta granularidad y complejidad de ambos espacios de clasificación.

Finalmente, no se estableció un umbral fijo de aceptación para Macro-F1. En lugar de utilizar un valor arbitrario como criterio de suficiencia, el desempeño se interpretó de manera conjunta a partir de las métricas globales, las métricas por clase, el MCC, la matriz de confusión y el intervalo de confianza obtenido mediante *bootstrap*. Este enfoque evita reducir la evaluación del clasificador a una única frontera numérica y permite una interpretación más consistente con la estructura multiclase y desbalanceada del problema.

Manejo de desbalance de clases

La distribución de las categorías emocionales presentó un desbalance considerable, caracterizado por una elevada concentración de textos clasificados como neutralidad y una representación sustancialmente menor de categorías como tristeza y enojo. En el corpus etiquetado, neutralidad representó el 82,7% de los registros, mientras que enojo correspondió únicamente al 0,7%, con 55 observaciones en el corpus total. Esta estructura constituye un problema relevante para el aprendizaje supervisado, dado que una distribución altamente desbalanceada puede favorecer el desempeño sobre las clases mayoritarias y reducir la capacidad del modelo para reconocer categorías minoritarias (He y Garcia, 2009).

Para mitigar este efecto se utilizaron dos estrategias complementarias durante el entrenamiento. En primer lugar, se implementó una función de pérdida ponderada, de manera que los errores asociados con clases menos frecuentes recibieran mayor peso

durante la optimización del modelo. Esta estrategia se mantuvo tanto en la versión inicial v1 como en la versión final v2 del fine-tuning.

En segundo lugar, en la versión final del modelo se aplicó *oversampling* específicamente sobre la categoría enojo, que constituía la clase de menor representación. En el conjunto de entrenamiento, esta categoría estaba conformada originalmente por 39 observaciones y fue aumentada temporalmente a 200 ejemplos para el proceso de entrenamiento. El *oversampling* se restringió al conjunto de entrenamiento, preservando sin modificaciones los conjuntos de validación y prueba, con el propósito de evitar contaminación de la evaluación y mantener una estimación independiente del desempeño del modelo.

El uso de estrategias de remuestreo y ponderación constituye una práctica ampliamente reconocida para abordar problemas de aprendizaje con distribuciones desbalanceadas. He y Garcia (2009) señalan que este tipo de problemas puede tratarse mediante enfoques a nivel de datos, como el remuestreo, y mediante estrategias sensibles al costo durante el aprendizaje. De manera complementaria, Buda et al. (2018) muestran que técnicas como el *oversampling* pueden utilizarse para reducir los efectos adversos de la desproporción entre clases en modelos de aprendizaje profundo.

La evaluación del modelo se realizó utilizando Macro-F1 como métrica principal, dado que esta medida calcula el F1 de cada categoría por separado y posteriormente obtiene un promedio no ponderado, evitando que las clases mayoritarias dominen por completo la interpretación del desempeño global. Esta decisión se complementó con Micro-F1, Weighted-F1, MCC y métricas por clase, con el propósito de examinar el rendimiento desde distintas perspectivas (Sokolova y Lapalme, 2009).

El tratamiento del desbalance no tuvo como objetivo igualar artificialmente la distribución completa del corpus, sino reducir el efecto de la baja representación de determinadas emociones durante el aprendizaje. Por esta razón, las métricas finales fueron calculadas sobre el conjunto de prueba original, sin aplicar *oversampling* ni modificaciones en su distribución.

Riesgos metodológicos

Durante el desarrollo de la investigación se identificaron varios riesgos metodológicos asociados con la calidad de las etiquetas, el desbalance de las categorías emocionales, la complejidad semántica del lenguaje financiero y la representatividad del corpus. Estos riesgos fueron considerados durante el procesamiento, entrenamiento e interpretación de los resultados con el propósito de delimitar adecuadamente el alcance de las conclusiones.

Uno de los principales riesgos metodológicos estuvo relacionado con la calidad de las pseudoetiquetas utilizadas para el fine-tuning. Las cinco categorías emocionales fueron construidas a partir de las predicciones iniciales de FinBERT y de reglas heurísticas empleadas para subdividir la categoría negativa en miedo, enojo y tristeza. Debido a que estas etiquetas fueron generadas automáticamente y no proceden de una anotación humana exhaustiva, existe la posibilidad de que contengan errores o ambigüedades que posteriormente sean aprendidos por el modelo.

Para examinar este riesgo, se realizó una validación humana sobre una muestra de 301 textos, clasificados de manera independiente por los dos autores de la investigación. Los anotadores coincidieron en 142 casos, equivalentes al 47,18% de acuerdo observado, y obtuvieron un coeficiente kappa de Cohen de 0,2752, lo que evidencia una concordancia

limitada una vez descontado el acuerdo esperado por azar. En la comparación individual con las pseudoetiquetas, el anotador 1 coincidió en 106 de 301 textos (35,22%), con κ = 0,1375, mientras que el anotador 2 coincidió en 237 textos (78,74%), con κ = 0,7131. Esta diferencia indica que la correspondencia entre las anotaciones humanas y las etiquetas automáticas no fue uniforme. Por consiguiente, los porcentajes de coincidencia con las pseudoetiquetas no deben interpretarse como medidas directas de corrección de los anotadores, sino como indicadores de correspondencia con el procedimiento automático de etiquetado.

Adicionalmente, existe una posible circularidad metodológica, dado que las predicciones de FinBERT base contribuyeron a la construcción de las pseudoetiquetas que posteriormente se utilizaron para ajustar el modelo mediante fine-tuning. En consecuencia, el modelo ajustado puede reproducir parcialmente patrones y errores presentes en el etiquetador inicial. Aunque la validación humana permite contrastar las etiquetas automáticas con criterios definidos por los evaluadores, no constituye una referencia completamente independiente, dado que fue realizada por los propios autores del estudio. Esta limitación debe tenerse en cuenta al interpretar los resultados de clasificación y restringe la posibilidad de considerar las pseudoetiquetas como una verdad de terreno validada exhaustivamente.

Un segundo riesgo correspondió al desbalance estructural de las clases emocionales. La neutralidad concentró el 82,7% del corpus, mientras que el enojo representó únicamente el 0,7%. Esta distribución puede inducir al clasificador a favorecer las categorías mayoritarias y dificultar el aprendizaje de patrones asociados con clases poco frecuentes, problema ampliamente documentado en aprendizaje supervisado con datos

desbalanceados (He y Garcia, 2009). Para reducir este efecto se empleó una función de pérdida ponderada y se aplicó *oversampling* sobre la categoría enojo únicamente en el conjunto de entrenamiento. No obstante, estas estrategias no sustituyen la diversidad léxica que proporcionaría un mayor número de ejemplos originales de las clases minoritarias. Uno de los principales riesgos metodológicos estuvo relacionado con la calidad de las pseudoetiquetas utilizadas para el fine-tuning. Las cinco categorías emocionales fueron construidas a partir de las predicciones iniciales de FinBERT y de reglas heurísticas empleadas para subdividir la categoría negativa en miedo, enojo y tristeza. Debido a que estas etiquetas fueron generadas automáticamente y no proceden de una anotación humana exhaustiva, existe la posibilidad de que contengan errores o ambigüedades que posteriormente sean aprendidos por el modelo.

Para examinar este riesgo, se realizó una validación humana sobre una muestra de 301 textos, clasificados de manera independiente por los dos autores de la investigación. Los anotadores coincidieron en 142 casos, equivalentes al 47,18% de acuerdo observado, y obtuvieron un coeficiente kappa de Cohen de 0,2752, lo que evidencia una concordancia limitada una vez descontado el acuerdo esperado por azar. En la comparación individual con las pseudoetiquetas, el anotador 1 coincidió en 106 de 301 textos (35,22%), con κ = 0,1375, mientras que el anotador 2 coincidió en 237 textos (78,74%), con κ = 0,7131. Esta diferencia indica que la correspondencia entre las anotaciones humanas y las etiquetas automáticas no fue uniforme. Por consiguiente, los porcentajes de coincidencia con las pseudoetiquetas no deben interpretarse como medidas directas de corrección de los anotadores, sino como indicadores de correspondencia con el procedimiento automático de etiquetado.

Adicionalmente, existe una posible circularidad metodológica, dado que las predicciones de FinBERT base contribuyeron a la construcción de las pseudoetiquetas que posteriormente se utilizaron para ajustar el modelo mediante fine-tuning. En consecuencia, el modelo ajustado puede reproducir parcialmente patrones y errores presentes en el etiquetador inicial. Aunque la validación humana permite contrastar las etiquetas automáticas con criterios definidos por los evaluadores, no constituye una referencia completamente independiente, dado que fue realizada por los propios autores del estudio. Esta limitación debe tenerse en cuenta al interpretar los resultados de clasificación y restringe la posibilidad de considerar las pseudoetiquetas como una verdad de terreno validada exhaustivamente.

Un tercer riesgo estuvo asociado con la ambigüedad semántica y contextual del lenguaje financiero. Este dominio incorpora expresiones especializadas, negaciones, abreviaturas, metáforas, ironía y otras formas discursivas cuyo significado puede depender fuertemente del contexto. Estas características dificultan la clasificación automática de sentimiento y emociones incluso cuando se utilizan modelos especializados (Araci, 2019; Daudert, 2021; Wilksch y Abramova, 2023). Para reducir el ruido se implementaron procedimientos de limpieza, preservación de términos financieros relevantes, filtrado de contenido no financiero y deduplicación semántica. Sin embargo, no se realizó una detección automática específica de ironía o sarcasmo, por lo que estos fenómenos permanecen como una limitación interpretativa del modelo.

Finalmente, se identificó un riesgo de representatividad temporal y de fuente. Aunque el corpus cubre el periodo comprendido entre enero de 2022 y abril de 2026, la cantidad de registros disponibles no es uniforme a lo largo de todos los años, lo que limita la

posibilidad de asumir una representación homogénea del periodo completo. Asimismo, el componente retail se encuentra altamente concentrado en StockTwits, plataforma de la cual proviene la gran mayoría de los textos de inversionistas minoristas incluidos en el corpus. Por esta razón, los resultados deben interpretarse como representativos del corpus analizado y no necesariamente de la totalidad de inversionistas retail o de todas las fuentes institucionales del mercado.

En conjunto, estos riesgos no invalidan el análisis realizado, pero establecen límites claros sobre la generalización de los resultados. Su reconocimiento permite interpretar las métricas de clasificación y las diferencias distribucionales dentro del contexto específico del corpus, las fuentes y el procedimiento de etiquetado empleados en la investigación.

Contribución científica

La contribución científica de esta investigación consiste en aportar evidencia empírica sobre la aplicación de técnicas de procesamiento del lenguaje natural (PLN) y de un modelo de lenguaje especializado en el dominio financiero para la clasificación emocional multiclase de textos relacionados con el S&P 500 y Tesla Inc. A diferencia de los enfoques tradicionales centrados en identificar únicamente sentimiento positivo, negativo o neutral, el estudio aborda cinco categorías emocionales: entusiasmo, miedo, enojo, tristeza y neutralidad. Esta aproximación permite examinar las posibilidades y limitaciones de una clasificación emocional más granular en textos procedentes de fuentes retail e institucionales, dentro del corpus analizado.

Desde el punto de vista metodológico, la investigación articula la conformación y depuración del corpus, la generación de pseudoetiquetas mediante FinBERT, el ajuste fino

del modelo, el tratamiento del desbalance de clases, la evaluación multiclase y la comparación estadística de las distribuciones emocionales entre tipos de fuente. La documentación de estas etapas permite describir las decisiones adoptadas durante el estudio y delimitar el alcance de los procedimientos aplicados. La contribución metodológica se centra, por tanto, en la aplicación y evaluación de un enfoque de clasificación emocional multiclase sobre el corpus seleccionado.

Una segunda contribución consiste en aportar evidencia empírica sobre la existencia de diferencias en las distribuciones emocionales de contenidos retail e institucionales dentro de un corpus común relacionado con el S&P 500 y Tesla Inc. La utilización conjunta de estadísticas descriptivas, la prueba Chi-cuadrado de independencia y la divergencia de Kullback-Leibler permitió analizar no solo la existencia de diferencias estadísticamente significativas, sino también cuantificar descriptivamente la separación entre los perfiles emocionales de ambas fuentes.

Adicionalmente, el estudio contribuye al análisis de las limitaciones que surgen al extender una clasificación financiera tradicional de sentimiento hacia un esquema emocional de mayor granularidad. Los resultados mostraron que el desempeño multiclase está condicionado por la distribución desigual de las categorías, la limitada cantidad de ejemplos en clases minoritarias y la calidad de las pseudoetiquetas utilizadas para el entrenamiento. La validación humana realizada sobre una muestra de 301 textos evidenció una concordancia limitada entre los dos anotadores: el acuerdo observado fue del 47,18% y el coeficiente kappa de Cohen fue de 0,2752. Asimismo, la coincidencia individual con las pseudoetiquetas presentó diferencias importantes entre ambos evaluadores. Estos resultados muestran que la clasificación emocional requiere criterios operacionales más

precisos, procesos de calibración y mecanismos de consenso para fortalecer la consistencia de las etiquetas de referencia.

Este hallazgo constituye una contribución metodológica relevante, ya que muestra que el incremento de granularidad emocional no depende únicamente de utilizar modelos de lenguaje más avanzados, sino también de disponer de corpus con mayor equilibrio entre clases y etiquetas de referencia de mejor calidad. En este sentido, el estudio se alinea con la evolución reciente del análisis de sentimiento financiero hacia modelos especializados, arquitecturas profundas y estrategias de representación cada vez más sensibles a las particularidades del dominio (Nyakurukwa y Seetharam, 2023; Peng et al., 2024; Mahdavi Ardekani et al., 2024).

En conjunto, la investigación aporta una base empírica y metodológica para futuras aproximaciones de análisis emocional financiero multiclase que incorporen validación humana, corpus especializados y estrategias de aprendizaje orientadas a reducir el efecto del desbalance y mejorar la diferenciación entre emociones semánticamente próximas.

Trabajo de Campo

Recolección del corpus

La recolección del corpus constituyó la base empírica del análisis emocional multiclase desarrollado en esta investigación. Se recopilaron textos financieros en inglés relacionados con el índice S&P 500 y Tesla Inc. (TSLA), correspondientes al periodo comprendido entre enero de 2022 y abril de 2026. Los registros fueron organizados según dos tipos de fuente: retail ($F = 0$) e institucional ($F = 1$).

Tabla 4 *Distribución del corpus por fuente y activo financiero*

Fuente	Tipo	Activo	n textos	% total	F
stocktwits_retail	Retail	S&P 500	2.471	2,71%	0
stocktwits_tsla	Retail	TSLA	2.478	24,78%	0
twitter_financial	Retail	S&P 500	29	0,29%	0
twitter_financial_tsla	Retail	TSLA	6	0,06%	0
stocktwits_retail	Retail	TSLA	16	0,16%	0
financial_news	Institucional	S&P 500	2.500	25,00%	1
financial_news	Institucional	TSLA	28	0,28%	1
financial_news_tsla	Institucional	TSLA	2.472	24,72%	1
Total			**10.000**	**100%**	

Nota. F=0 corresponde a fuentes retail (inversores minoristas); F=1 corresponde a fuentes institucionales (Noticias financieras especializadas). n = número de textos por fuente.

Las fuentes retail estuvieron representadas principalmente por mensajes provenientes de StockTwits y Twitter Financial, mientras que las fuentes institucionales correspondieron a artículos y contenidos publicados por Reuters, CNBC y otros medios financieros especializados. Esta estructura permitió incorporar contenidos generados por inversionistas minoristas y textos de carácter informativo procedentes de medios

financieros, manteniendo la diferenciación de fuente requerida para la posterior comparación de sus distribuciones emocionales.

Tabla 5 *Balance del corpus por tipo de fuente y activo financiero*

Dimensión	Categoría	n textos	Porcentaje
Tipo de fuente	Retail (F=0)	5.000	50%
Tipo de fuente	Institucional (F=1)	5.000	50%
Activo financiero	S&P 500	5.000	50%
Activo financiero	TSLA	5.000	50%

Nota. El balance exacto 50/50 en ambas dimensiones fue diseñado para eliminar el sesgo de representación por activo y facilitar la comparación estadística entre fuentes.

El corpus inicial estuvo conformado por 10.000 textos. De acuerdo con el diseño metodológico del estudio, se estructuró una distribución equilibrada por tipo de fuente, con 5.000 registros retail y 5.000 institucionales. Asimismo, se estableció una distribución equilibrada por activo financiero, con 5.000 textos relacionados con el S&P 500 y 5.000 asociados con Tesla Inc. Esta estructura buscó evitar una representación global desproporcionada de alguno de los dos tipos de fuente o de los activos considerados y facilitar la comparación posterior de las distribuciones emocionales.

Cada registro conservó información de trazabilidad correspondiente al contenido textual, la fuente de procedencia, el tipo de fuente y el activo financiero asociado. Estos atributos permitieron mantener la correspondencia de cada observación durante las etapas posteriores de depuración, construcción de etiquetas, entrenamiento y análisis.

Durante la conformación inicial del corpus se efectuó la identificación de registros duplicados exactos mediante *hash* MD5. Los procedimientos adicionales de depuración —como el filtro mínimo de tokens, la deduplicación semántica y la exclusión de contenido no

financiero— se aplicaron posteriormente durante la etapa de preprocesamiento y se describen en la siguiente subsección.

Preprocesamiento y depuración del corpus

El preprocesamiento transformó el corpus inicial de 10.000 textos en un conjunto depurado de 7.752 registros, que posteriormente fue utilizado para la construcción de las categorías emocionales y el ajuste del modelo. El procedimiento combinó operaciones de limpieza y normalización textual con criterios de exclusión orientados a reducir registros poco informativos, duplicados o no pertinentes para el análisis financiero.

Tabla 6 *Resultados del proceso de preprocesamiento*

Filtro aplicado	Textos antes	Eliminados	Textos después
Corpus bruto inicial	10.000	—	10.000
Limpieza: URLs, menciones, caracteres especiales	10.000	0	10.000
Filtro mínimo 10 tokens	10.000	248	9.752
Deduplicación semántica (similitud coseno > 0.95)	9.752	0	9.752
Filtro contenido no financiero	9.752	373	9.379
Filtro idioma inglés	9.379	0	9.379
Filtro longitud máxima (100/400 tokens por tipo)	9.379	1.627	7.752
Corpus final preprocesado	10.000	2.248 (22,5%)	7.752

Nota. La etapa de limpieza textual modificó el contenido de los registros sin producir exclusiones de filas. El filtro de contenido no financiero utilizó patrones de expresiones regulares para identificar spam, publicidad, correos de reclutamiento y contenido personal. El criterio de longitud máxima se estableció en 100 tokens para fuentes retail y 400 tokens

para fuentes institucionales. Los umbrales empleados correspondieron a decisiones operativas del pipeline de preprocesamiento.

Inicialmente, se eliminaron URLs, menciones, caracteres especiales y números aislados, se normalizaron los espacios y el contenido se convirtió a minúsculas. En los hashtags se eliminó el símbolo #, conservando la palabra asociada. Posteriormente, se realizó tokenización, lematización y eliminación de *stopwords* en inglés. La lista de *stopwords* fue adaptada para conservar términos y negaciones con posible relevancia semántica en el dominio financiero, entre ellos *bull, bear, fear, risk, crash, rally, buy, sell, up, down, not, no* y *never*. La preservación de este vocabulario correspondió a una decisión del pipeline orientada a evitar la pérdida de información potencialmente útil para la clasificación, en consonancia con la literatura que destaca las particularidades semánticas del lenguaje financiero (Kearney y Liu, 2014; Araci, 2019).

Después de la limpieza textual, se aplicó un criterio mínimo de longitud que excluyó los textos con menos de 10 tokens, lo que permitió reducir el corpus de 10.000 a 9.752 registros. Posteriormente, se evaluó la presencia de textos semánticamente similares mediante el procedimiento de deduplicación definido para el estudio, utilizando un umbral de similitud coseno superior a 0,95. Según los resultados registrados, esta etapa no produjo exclusiones adicionales. De manera complementaria, se aplicó el filtro de idioma inglés, que tampoco generó eliminaciones en el conjunto procesado.

Posteriormente, se aplicaron patrones de expresiones regulares para identificar y excluir contenido no financiero, spam, publicidad, correos de reclutamiento y otros registros

irrelevantes. Este procedimiento eliminó 373 textos y redujo el corpus a 9.379 observaciones. El filtro de idioma no produjo exclusiones adicionales, dado que los registros conservados correspondían a textos en inglés.

Finalmente, se aplicó un criterio de longitud máxima diferenciado según el tipo de fuente, establecido en 100 tokens para contenidos retail y 400 tokens para contenidos institucionales. Este filtro produjo la mayor reducción del corpus, con 1.627 registros excluidos, obteniéndose un conjunto final de 7.752 textos.

Para consultar la tabla de distribución de eliminaciones por fuente del corpus final, ir al Anexo B.

Distribución Emocional del Corpus Etiquetado

Concluido el preprocesamiento, se aplicó el modelo FinBERT base (Araci, 2019) para asignar pseudoetiquetas emocionales a los 7.752 textos. Las etiquetas de FinBERT (*Positive*, *negative*, *neutral*) se mapearon a cinco emociones: entusiasmo, miedo, enojo, tristeza y neutralidad. Las clases negativas se subdividieron mediante análisis de palabras clave específicas para cada emoción.

Tabla 7 *Distribución de emociones en el corpus etiquetado*

Emoción	n total	% total	n retail	n institucional	Estado
Neutralidad	6.411	82,7%	3.110	3.301	Representada
Entusiasmo	547	7,1%	348	199	Representada
Miedo	432	5,6%	199	233	Representada
Tristeza	307	4,0%	87	220	Minoritaria
Enojo	55	0,7%	8	47	Crítica

Nota. Para describir el grado de representación de las categorías emocionales en el corpus, se establecieron umbrales operativos basados en su proporción respecto del total

de textos etiquetados. Se clasificaron como “representadas” las categorías con una proporción superior al 5%, como “minoritarias” las categorías con una proporción igual o superior al 1% y hasta el 5%, y como “críticas” las categorías con una proporción inferior al 1%. Estos umbrales se utilizaron como criterios descriptivos para identificar el nivel de desbalance de las clases, no como puntos de corte universales. El tratamiento mediante *oversampling* se dirigió a la categoría crítica de enojo durante el entrenamiento, debido a su baja frecuencia en el corpus. Esta decisión no implica que todas las categorías minoritarias requieran necesariamente sobremuestreo.

Tabla 8 *División estratificada del corpus: conjuntos de entrenamiento, validación y prueba*

Emoción	Entrenamiento (70%)	Validación (15%)	Prueba (15%)	Total
Neutralidad	4.487	962	962	6.411
Entusiasmo	383	82	82	547
Miedo	302	65	65	432
Tristeza	215	46	46	307
Enojo	39	8	8	55
Total	**5.426**	**1.163**	**1.163**	**7.752**

Nota. La división estratificada garantiza que la proporción de cada emoción se mantiene constante en los tres subconjuntos, condición necesaria para la comparabilidad de las métricas de evaluación.

La distribución emocional del corpus refleja una característica inherente al dominio financiero: la predominancia de la neutralidad informativa (82,7%). Este resultado es coherente con la naturaleza de las fuentes utilizadas, donde la mayoría de los mensajes de *StockTwits* y artículos de noticias reportan hechos del mercado sin carga emocional explícita, hallazgo consistente con Kearney y Liu (2014) y Araci (2019).

De acuerdo con los umbrales operativos definidos, la neutralidad (82,7%), el entusiasmo (7,1%) y el miedo (5,6%) se clasificaron como categorías representadas. La tristeza (4,0%) se identificó como categoría minoritaria, mientras que el enojo (0,7%) se clasificó como categoría crítica. Esta última presentó la menor frecuencia, con 55 textos en el corpus total, por lo que recibió un tratamiento específico mediante *oversampling* durante el entrenamiento. La clasificación por estados permite describir el desbalance de las pseudoetiquetas, pero no constituye por sí sola una evaluación de la dificultad de clasificación ni de la calidad de las etiquetas generadas automáticamente.

Las categorías obtenidas en esta etapa corresponden a pseudoetiquetas generadas mediante FinBERT base y reglas heurísticas de mapeo hacia cinco emociones, por lo que no deben interpretarse como etiquetas humanas de referencia. Con el propósito de evaluar su correspondencia con el criterio humano, posteriormente se realizó una validación independiente sobre una muestra de 301 textos del corpus, procedimiento que se describe en la subsección Validación humana de las pseudoetiquetas y cuya guía de anotación se presenta en el Anexo D.

La distribución presentada en esta subsección corresponde, por tanto, a las pseudoetiquetas generadas antes del fine-tuning. Estas mismas pseudoetiquetas constituyeron la base para la comparación estadística entre fuentes retail e institucionales y para la contrastación de H1. El modelo FinBERT fine-tuned fue evaluado posteriormente sobre el conjunto de prueba y no intervino en la construcción de estas distribuciones.

Análisis de resultados

Entrenamiento del modelo: fine-tuning de FinBERT

El fine-tuning del modelo FinBERT se desarrolló en dos iteraciones. La versión inicial (v1) utilizó la configuración base definida en la metodología. A partir del comportamiento observado durante el entrenamiento se implementó una segunda configuración (v2), que incorporó ajustes de hiperparámetros, *early stopping* y *oversampling* específico de la clase enojo. Estos cambios estuvieron orientados a mejorar la capacidad de generalización del modelo y reducir el efecto del desbalance existente entre las categorías emocionales.

Tabla 9 *Hiperparámetros configurados para el fine-tuning: comparación v1 vs. v2*

Hiperparámetro	Versión v1	Versión v2	Justificación del ajuste
Tasa de aprendizaje (lr)	2e-5	1e-5	Reducir sobreajuste observado en v1
Tamaño de lote (batch size)	16	32	Mayor estabilidad del gradiente
Longitud máxima (tokens)	512	128	Textos cortos; reduce ruido por padding
Warmup ratio	—	0.10	Arranque suave del aprendizaje
Número de épocas	5	8	*Early stopping* con paciencia = 3
Oversampling clase enojo	No	Sí (39 → 200)	Clase crítica con 0,7% de representación
Función de pérdida ponderada	Sí	Sí	Compensación del desbalance de clases
Métrica principal de evaluación	Macro-F1	Macro-F1	Equidad entre clases independiente del soporte

Nota. lr = *learning rate*. La versión v2 constituye el modelo final reportado en este estudio. El *early stopping* detuvo el entrenamiento cuando el Macro-F1 de validación no mejoró durante tres épocas consecutivas.

En la versión v2, la tasa de aprendizaje se redujo de $\mathbf{2 \times 10^{-5}}$a $\mathbf{1 \times 10^{-5}}$, el tamaño de lote aumentó de 16 a 32 y la longitud máxima de entrada se redujo de 512 a 128 tokens. Asimismo, se incorporó un warmup ratio de 0,10, se estableció un máximo de ocho épocas y se utilizó *early stopping* con una paciencia de tres épocas. La función de pérdida ponderada se mantuvo en ambas configuraciones. Para la clase enojo, que disponía de 39 ejemplos en el conjunto de entrenamiento, se aplicó *oversampling* hasta alcanzar 200 observaciones durante el entrenamiento.

Tabla 10 *Evolución del entrenamiento por época: versión v2 (Modelo Final)*

Época	Loss entrenamiento	Loss validación	Macro-F1	MCC	Mejor modelo
1	1,3080	1,1543	0,3663	0,3960	No
2	0,7936	0,8809	0,4900	0,5552	No
3	0,5803	0,8470	0,4845	0,5360	No
4	0,3936	0,7977	0,5419	0,6134	No
5	0,3220	0,8229	0,5545	0,6165	No
6	0,2410	0,8279	0,5485	0,6247	No
7	0,1842	0,8662	0,5441	0,6191	No
8	0,2047	0,8706	0,5654	0,6485	Sí

Nota. MCC = Coeficiente de Correlación de Matthews. El mejor modelo se seleccionó por el mayor Macro-F1 de validación.

Macro-F1 se utilizó como métrica principal de seguimiento debido a que calcula el promedio del F1 de las distintas categorías asignando igual ponderación a cada clase. El MCC se utilizó como medida complementaria para observar la correspondencia global entre predicciones y etiquetas de referencia.

El entrenamiento de la versión v2 se desarrolló durante ocho épocas. De acuerdo con los resultados de validación registrados en la Tabla 10, el mayor Macro-F1 se obtuvo en la época 8, con un valor de 0,5654, y estuvo acompañado de un MCC de 0,6485. Por tanto,

esta época se identificó como la mejor configuración entre las evaluadas, de acuerdo con el criterio de selección basado en el Macro-F1 de validación. Durante las últimas épocas se observó que la pérdida de entrenamiento disminuyó inicialmente, mientras que la pérdida de validación aumentó respecto de sus valores mínimos anteriores. Este comportamiento constituye una señal de posible sobreajuste y debe interpretarse conjuntamente con las métricas de validación. Aunque se configuró un mecanismo de *early stopping* con una paciencia de tres épocas y un máximo de ocho épocas, es necesario distinguir entre la configuración del mecanismo y su activación efectiva durante la ejecución.

La comparación entre v1 y v2 evidencia el impacto positivo del *oversampling*: los pesos de la clase enojo se redujeron de 27,83 (v1) a 5,59 (v2), lo que permitió una función de pérdida más equilibrada. El Macro-F1 de validación de 0,541 en v2 representó una mejora de 1,83 puntos porcentuales respecto a v1 (0,523), resultado consistente con las limitaciones estructurales del corpus.

Tabla 11 *Comparativa v1 vs. v2*

Métrica de validación	v1	v2
Macro-F1	0,5228	0,5654
MCC	0,6141	0,6485

Nota. Los resultados de v1 corresponden al registro previo del entrenamiento inicial.

Contrastación de la hipótesis H1: diferencias distribucionales entre fuentes

La hipótesis H1 planteó que la distribución de las pseudoetiquetas emocionales identificadas en los textos provenientes de fuentes retail difiere de la distribución observada en las fuentes institucionales. Para evaluar esta hipótesis se utilizaron las pseudoetiquetas generadas en la Fase 3 mediante FinBERT base y reglas heurísticas de mapeo emocional

sobre los 7.752 textos del corpus final. El modelo ajustado mediante fine-tuning no intervino en este contraste.

Para evaluar la existencia de asociación entre el tipo de fuente y la categoría emocional se utilizó la prueba Chi-cuadrado de independencia, complementada con la divergencia simétrica de Kullback-Leibler como medida descriptiva de separación entre distribuciones.

Como complemento a la significancia estadística, se estimó el V de Cramér para cuantificar la magnitud de la asociación entre el tipo de fuente y la categoría emocional.

Tabla 12 *Distribución emocional comparada: fuentes retail vs. Institucionales*

Emoción	Retail % (F=0)	Institucional % (F=1)	Diferencia (pp)	Dirección
Neutralidad	82,9%	82,5%	0,4	Similar
Entusiasmo	9,3%	5,0%	4,3	Retail > Institucional
Miedo	5,3%	5,8%	0,5	Institucional > Retail
Tristeza	2,3%	5,5%	3,2	Institucional > Retail
Enojo	0,2%	1,2%	1,0	Institucional > Retail

Nota. pp = puntos porcentuales. La diferencia se calcula como el valor retail menos el valor institucional; un valor positivo indica mayor proporción en fuentes retail.

La comparación descriptiva mostró que la neutralidad presentó proporciones similares en ambos tipos de fuente, con 82,9% en retail y 82,5% en institucional. Las mayores diferencias se observaron en entusiasmo, con 9,3 % en retail frente a 5,0% en institucional; tristeza, con 2,3% frente a 5,5%; y enojo, con 0,2% frente a 1,2%, respectivamente. El miedo mostró una diferencia menor, con 5,3% en retail y 5,8% en institucional.

Tabla 13 *Resultados de las pruebas estadísticas para la contrastación de H1*

Prueba estadística	Estadístico	Valor p / criterio	Decisión

Chi-cuadrado de independencia (χ^2)	126,422	$p < 0,0001$, $\alpha = 0,05$	Rechazar H_0
Grados de libertad	4	—	—
Divergencia KL simétrica	0,0356	Medida descriptiva	Cuantifica separación
KL retail ‖ institucional	0,0312	—	—
KL institucional ‖ retail	0,0400	—	—
V de Cramér	0,128	Tamaño del efecto	Magnitud de la asociación

Nota. La prueba Chi-cuadrado evalúa la existencia de asociación entre el tipo de fuente y la categoría emocional. El V de Cramér cuantifica la magnitud de dicha asociación. La divergencia KL se utiliza como una medida descriptiva de separación entre distribuciones y no como prueba de significancia estadística.

Hallazgo H1: distribuciones emocionales estadísticamente diferentes entre fuentes

La prueba Chi-cuadrado de independencia se aplicó a los conteos observados de las cinco categorías de pseudoetiquetas emocionales según el tipo de fuente —retail e institucional— en el corpus final de 7.752 textos. El contraste obtuvo un estadístico $\chi^2(4) = 126,422$, con $p < 0,0001$. Con un nivel de significancia de $\alpha = 0,05$, se rechazó la hipótesis nula de independencia entre el tipo de fuente y la categoría emocional. Este resultado aporta evidencia de una asociación estadística entre ambas variables en las pseudoetiquetas analizadas. El contraste no se realizó sobre las predicciones del modelo FinBERT ajustado mediante fine-tuning, sino sobre las etiquetas generadas previamente con FinBERT base y las reglas heurísticas de mapeo emocional.

Para complementar el contraste de significancia estadística, se estimó el V de Cramér, cuyo valor fue 0,128. Este resultado indica que la asociación entre el tipo de fuente y la categoría emocional tiene una magnitud limitada, por lo que la significancia estadística no

debe interpretarse como evidencia de una diferencia de gran tamaño. La comparación descriptiva mostró que el entusiasmo tuvo una mayor proporción en las fuentes retail (9,3%) que en las institucionales (5,0%), mientras que la tristeza presentó una mayor proporción en las fuentes institucionales (5,5%) que en las retail (2,3%). El enojo también fue más frecuente en las fuentes institucionales (1,2%) que en las retail (0,2%). Por su parte, el miedo mostró proporciones relativamente cercanas entre ambos tipos de fuente, al igual que la neutralidad (82,9% en retail y 82,5% en institucionales). De manera complementaria, la divergencia KL simétrica fue de 0,0356 y se interpretó como una medida descriptiva de separación entre las distribuciones, no como una prueba adicional de significancia estadística. Estos resultados describen diferencias entre las pseudoetiquetas generadas y no constituyen una verificación directa de las emociones reales de los autores de los textos.

De manera complementaria, la divergencia simétrica de Kullback-Leibler fue de 0,0356, calculada a partir de $D_{KL}(retail \parallel institucional) = 0{,}0328$ y $D_{KL}(institucional \parallel retail) = 0{,}0384$. Este valor cuantifica descriptivamente la separación entre ambas distribuciones, pero no constituye una prueba adicional de significancia estadística ni dispone de umbrales universales que permitan clasificar su magnitud como baja, moderada o alta

En conjunto, los resultados muestran que las fuentes retail e institucionales presentan perfiles emocionales distintos dentro del corpus analizado. La evidencia inferencial proviene de la prueba Chi-cuadrado, mientras que la divergencia KL complementa la interpretación al cuantificar la separación entre las distribuciones. Estas diferencias deben entenderse como asociaciones observadas en el corpus estudiado y no como evidencia

causal sobre el comportamiento de los inversionistas o sobre la naturaleza de las fuentes financieras.

Evaluación Final del Modelo y Contrastación de la Hipótesis H2

La evaluación final del modelo FinBERT ajustado mediante fine-tuning se realizó sobre un conjunto de prueba independiente de 1.163 textos, que no participó en el entrenamiento ni en la selección de hiperparámetros. Este conjunto permitió estimar el desempeño del clasificador sobre observaciones no utilizadas durante el proceso de ajuste.

Tabla 14 *Métricas globales de evaluación del modelo FinBERT fine-tuned v2*

Métrica global	Valor
Macro-F1	0,5564
Micro-F1	0,8813
F1 ponderado (Weighted-F1)	0,8901
Coef. Correlación Matthews (MCC)	0,6705
Exactitud (Accuracy)	0,8813
IC 95% bootstrap (Macro-F1)	[0,5198; 0,5919]

Nota. El intervalo de confianza del 95% para Macro-F1 se estimó mediante 1.000 iteraciones de remuestreo bootstrap con reemplazo, utilizando 1.163 observaciones en cada iteración. MCC = coeficiente de correlación de Matthews.

La evaluación del modelo FinBERT v2 sobre el conjunto de prueba independiente, conformado por 1.163 textos, arrojó un Macro-F1 de 0,5564, un Micro-F1 de 0,8813, un Weighted-F1 de 0,8901, un coeficiente de correlación de Matthews (MCC) de 0,6705 y una

exactitud de 0,8813. Mediante *bootstrap* se estimó un intervalo de confianza del 95% para el Macro-F1 entre 0,5198 y 0,5919.

Estos resultados muestran diferencias en el desempeño según la métrica utilizada. Aunque la exactitud y el F1 ponderado son elevados, el Macro-F1 es menor, lo que evidencia que el rendimiento no es homogéneo entre las cinco categorías emocionales. En particular, la clase enojo presentó un F1 de 0,00 en el conjunto de prueba, mientras que neutralidad alcanzó un F1 de 0,95. Por tanto, la interpretación del desempeño debe considerar el desbalance de clases y las limitaciones del modelo para identificar las emociones minoritarias.

Tabla 15 *F1-Score por clase emocional en el conjunto de prueba*

Emoción	Precisión	Recall	F1-Score	Soporte
Entusiasmo	0,66	0,87	0,75	82
Miedo	0,50	0,71	0,59	65
Enojo	0,00	0,00	0,00	8
Tristeza	0,44	0,59	0,50	46
Neutralidad	0,98	0,92	0,95	962
Promedio ponderado	**0,52**	**0,62**	**0,56**	**232,60**

Nota. El F1-Score corresponde a la media armónica entre precisión y *recall*. El promedio ponderado incorpora la frecuencia relativa de cada categoría.

El análisis por clase mostró diferencias importantes en el desempeño. Neutralidad alcanzó un F1 de 0,95, seguida de entusiasmo con 0, 75, miedo con 0,59, tristeza con 0,5 y enojo con 0. Las categorías con mayor número de ejemplos de entrenamiento tendieron a presentar un mejor desempeño, mientras que las clases con menor representación obtuvieron valores F1 inferiores. En particular, la clase enojo contó con 39 ejemplos originales en entrenamiento y ocho observaciones en el conjunto de prueba, por lo que el

resultado obtenido sugiere que el *oversampling* aplicado no compensó completamente la limitada disponibilidad de ejemplos originales de esta categoría.

Para cuantificar la incertidumbre asociada con Macro-F1 se estimó un intervalo de confianza del 95% mediante 1.000 iteraciones de remuestreo *bootstrap* con reemplazo sobre el conjunto de prueba. El intervalo obtenido fue [0,5198; 0,5919], lo que permite complementar la estimación puntual de Macro-F1 con una medida de su variabilidad muestral.

El análisis de la matriz de confusión del modelo FinBERT ajustado (véase el Anexo C) permite identificar patrones específicos de error entre las categorías emocionales. La confusión más evidente entre las categorías minoritarias se presenta entre miedo y tristeza: 10 textos etiquetados como miedo fueron clasificados como tristeza y otros 10 etiquetados como tristeza fueron clasificados como miedo. Asimismo, de los ocho textos cuya etiqueta real correspondía a enojo, cuatro fueron clasificados como miedo, uno como tristeza y tres correctamente como enojo. Estos resultados muestran que el modelo presenta dificultades para diferenciar algunas categorías emocionales semánticamente próximas, particularmente cuando su representación en el conjunto de prueba es reducida. La categoría enojo requiere especial cautela interpretativa, dado que su evaluación se basa en solo ocho observaciones, de modo que cada error tiene un efecto considerable sobre sus métricas. Adicionalmente, 37 textos etiquetados como neutralidad fueron clasificados como entusiasmo y 34 como miedo, lo que evidencia que los errores no se limitan a las categorías minoritarias. En conjunto, la matriz complementa las métricas agregadas al mostrar cómo se distribuyen los errores de clasificación, sin que estos resultados permitan concluir por sí solos que las categorías sean indistinguibles.

Para contrastar H2, se compararon dos enfoques evaluados sobre una formulación binaria común de entusiasmo frente a no entusiasmo: el clasificador binario entrenado independientemente y las predicciones del modelo multiclase original agrupadas posteriormente en esas dos categorías. Ambos enfoques se evaluaron sobre las mismas 1.163 observaciones del conjunto de prueba y utilizando la misma definición de las categorías. La diferencia entre sus valores de Macro-F1 se utilizó como medida descriptiva del desempeño, mientras que un procedimiento bootstrap pareado permitió cuantificar la incertidumbre de dicha diferencia. De manera complementaria, la prueba de McNemar evaluó las discordancias entre los patrones de acierto y error de ambos clasificadores.

Estos resultados muestran que el desempeño global del modelo varía de manera importante entre las cinco categorías emocionales, particularmente entre la clase mayoritaria, neutralidad, y las categorías minoritarias. Por esta razón, la interpretación del modelo no debe basarse únicamente en la exactitud o en las métricas ponderadas, sino considerar conjuntamente Macro-F1, MCC, las métricas por clase y el intervalo de confianza obtenido mediante *bootstrap*.

Hallazgo H2: comparación del clasificador binario independiente y el modelo multiclase agrupado

La evaluación del clasificador binario independiente sobre el conjunto de prueba de 1.163 observaciones obtuvo un Macro-F1 de 0,8311, una exactitud de 0,9587 y un coeficiente de correlación de Matthews (MCC) de 0,6647. El F1 de la categoría entusiasmo fue de 0,6842, lo que evidencia que la identificación de esta clase minoritaria continúa representando un desafío para el modelo.

Para comparar ambos enfoques en una tarea equivalente, las predicciones del modelo multiclase de cinco emociones se agruparon en dos categorías: entusiasmo y no entusiasmo. Esta evaluación produjo un Macro-F1 de 0,8624, superior al 0,8311 obtenido por el clasificador binario independiente. La diferencia descriptiva, calculada como el Macro-F1 del clasificador binario independiente menos el del modelo multiclase agrupado, fue de −0,0309, equivalente a −3,09 puntos porcentuales.

Estos resultados deben interpretarse considerando que la clasificación binaria independiente fue entrenada específicamente para distinguir entusiasmo del resto de las emociones, mientras que la evaluación binaria derivada del modelo multiclase corresponde a la agrupación de sus predicciones originales. Por tanto, la comparación permite examinar el comportamiento de ambos enfoques en un espacio de dos categorías, sin perder de vista que el modelo multiclase también fue diseñado para distinguir cinco emociones específicas.

Contraste H2: comparación del desempeño de los modelos

Tabla 16 *Comparación del desempeño de los modelos*

Indicador	Multiclase de cinco clases	Multiclase agrupado en dos clases	Binario independiente
Número de clases evaluadas	5	2	2
Macro-F1	0,5564	0,8624	0,8311
Diferencia frentre al multiclase agrupado	—	—	-0,0309
IC 95% de la diferencia Macro-F1	—	—	[-0,0788; 0,0128]
McNemar: p-valor	—	—	0,8955

Nota. El Macro-F1 del modelo multiclase de cinco clases corresponde a la evaluación original de las categorías enojo, entusiasmo, miedo, neutralidad y tristeza. La comparación inferencial utilizada para contrastar H2 se realizó exclusivamente entre el clasificador

binario independiente y las predicciones del modelo multiclase agrupadas posteriormente en entusiasmo y no entusiasmo, ambos evaluados sobre las mismas 1.163 observaciones del conjunto de prueba. La diferencia de Macro-F1 describe la dirección y magnitud del contraste, el intervalo de confianza bootstrap pareado cuantifica su incertidumbre y la prueba de McNemar evalúa las discordancias entre los patrones de acierto y error de ambos enfoques. McNemar no constituye una prueba directa de igualdad entre valores de Macro-F1.

Para contrastar H2 se comparó el desempeño del clasificador binario independiente con las predicciones del modelo multiclase agrupadas en dos categorías: entusiasmo y no entusiasmo, utilizando las mismas 1.163 observaciones del conjunto de prueba. El Macro-F1 del modelo multiclase agrupado fue de 0,8624, mientras que el clasificador binario independiente obtuvo un Macro-F1 de 0,8311. La diferencia, calculada como el Macro-F1 del clasificador binario independiente menos el del modelo multiclase agrupado, fue de −0,0309, equivalente a −3,09 puntos porcentuales.

La incertidumbre de esta diferencia se evaluó mediante un procedimiento bootstrap pareado de 2.000 réplicas, que produjo un intervalo de confianza del 95% entre −0,0788 y 0,0128. Debido a que este intervalo incluye el cero, no se encontró evidencia suficiente para afirmar que la diferencia entre los valores de Macro-F1 sea estadísticamente distinta de cero al nivel de significancia del 5%.

Adicionalmente, se aplicó la prueba de McNemar a los aciertos y errores pareados de ambos enfoques sobre las mismas observaciones. Se obtuvieron 28 casos en los que únicamente acertó el modelo multiclase agrupado y 30 casos en los que únicamente acertó el clasificador binario independiente. El contraste produjo un estadístico de 0,0172 y un

valor p de 0,8955. Dado que el valor p es superior a 0,05, no se rechazó la hipótesis nula de igualdad en las proporciones de aciertos pareados. Este resultado no demuestra equivalencia entre los modelos, sino que no se encontró evidencia estadísticamente significativa de una diferencia mediante esta prueba.

En conjunto, los resultados no respaldan H2 en su formulación de una diferencia estadísticamente significativa de desempeño entre los enfoques comparados. Aunque el Macro-F1 del modelo multiclase agrupado fue descriptivamente superior al del clasificador binario independiente, tanto el intervalo bootstrap como la prueba de McNemar indican que no existe evidencia estadística suficiente para sostener dicha diferencia al nivel de significancia del 5%. Esta interpretación debe considerar que ambos enfoques se evaluaron en una formulación binaria común, mientras que el modelo multiclase también tiene la capacidad de distinguir cinco categorías emocionales específicas.

Discusión de la hipótesis H2

Los resultados muestran que el clasificador binario independiente obtuvo un Macro-F1 de 0,8311 en el conjunto de prueba, mientras que las predicciones del modelo multiclase agrupadas en las categorías entusiasmo y no entusiasmo alcanzaron un Macro-F1 de 0,8624. La diferencia descriptiva fue de −0,0309, equivalente a −3,09 puntos porcentuales para el clasificador binario independiente respecto al modelo multiclase agrupado. Por tanto, los resultados no muestran una superioridad descriptiva del clasificador binario independiente.

El intervalo de confianza del 95% de la diferencia de Macro-F1, estimado mediante *bootstrap* pareado de 2.000 réplicas, se situó entre −0,0788 y 0,0128. Dado que este

intervalo incluye el cero, no se encontró evidencia suficiente para afirmar que la diferencia entre ambos valores de Macro-F1 sea estadísticamente distinta de cero. De forma consistente, la prueba de McNemar produjo un valor p de 0,8955, superior al nivel de significancia de 0,05. En consecuencia, no se rechazó la hipótesis nula de igualdad en las proporciones de aciertos pareados.

Estos resultados no respaldan H2 en su formulación de una diferencia estadísticamente significativa de desempeño entre los dos enfoques binarios comparados. Sin embargo, la ausencia de significancia estadística no demuestra que los clasificadores sean equivalentes. La interpretación se limita a indicar que, con las observaciones y los procedimientos utilizados, no se obtuvo evidencia suficiente para establecer una diferencia estadísticamente significativa.

Desde la perspectiva analítica, el modelo multiclase conserva la capacidad de distinguir cinco categorías emocionales —enojo, entusiasmo, miedo, neutralidad y tristeza—, mientras que el clasificador binario independiente se concentra en identificar entusiasmo frente al resto de las emociones. Esta diferencia de granularidad debe considerarse al interpretar los resultados, ya que el Macro-F1 de la tarea binaria no sustituye la evaluación de la capacidad del modelo multiclase para distinguir emociones específicas.

Asimismo, el desempeño de ambos enfoques debe analizarse considerando el desbalance de las categorías y la calidad de las pseudoetiquetas utilizadas como referencia. En el clasificador binario, la clase entusiasmo presentó menor representación que la clase no entusiasmo, y su F1 fue de 0,6842. Esto evidencia que la identificación de entusiasmo continúa siendo un reto de clasificación. Además, dado que las etiquetas de referencia proceden de un proceso de pseudoetiquetado, los resultados expresan el grado

de correspondencia de los modelos con dichas etiquetas y no constituyen por sí solos una validación definitiva de las emociones humanas presentes en los textos.

Validación humana de las pseudoetiquetas

Con el propósito de evaluar la correspondencia entre las pseudoetiquetas utilizadas para el ajuste del modelo y el criterio humano, se realizó una revisión independiente sobre una muestra de 301 textos del corpus. Los dos anotadores, Edison Fabián Moreno Hernández y Víctor Alfonso Rodríguez Guzmán, asignaron de manera independiente una de las cinco emociones definidas en el estudio: enojo, entusiasmo, miedo, neutralidad y tristeza. La decisión se realizó considerando la emoción predominante expresada en cada texto y sin utilizar la pseudoetiqueta automática como criterio de clasificación.

La guía de anotación estableció definiciones operacionales para cada categoría y tomó como unidad de análisis el texto completo dentro de su contexto financiero. Cada anotador asignó una única emoción predominante por observación, siguiendo los criterios definidos en el Anexo D.

La concordancia entre los dos anotadores se evaluó mediante el porcentaje de acuerdo observado y el coeficiente kappa de Cohen. Este coeficiente permite estimar el grado de acuerdo entre dos evaluadores en variables categóricas, ajustando el acuerdo observado por aquel que podría esperarse por azar (Cohen, 1960). En esta investigación, su aplicación permitió complementar el porcentaje de coincidencias y disponer de una medida de concordancia corregida por azar para las cinco categorías emocionales analizadas.

Sobre los 301 textos evaluados, los dos anotadores coincidieron exactamente en 142 casos, equivalentes al 47,18% de acuerdo observado. El coeficiente kappa de Cohen fue

0,2752, lo que evidencia una concordancia limitada entre ambos anotadores más allá del acuerdo esperado por azar.

La comparación individual con las pseudoetiquetas mostró resultados diferenciados. El Anotador 1 coincidió con las etiquetas automáticas en 106 de 301 textos, equivalentes al 35,22%, con $\kappa = 0,1375$. El Anotador 2 coincidió en 237 de 301 textos, equivalentes al 78,74%, con con $\kappa = 0,7131$.

Esta asimetría indica que uno de los criterios humanos presentó una mayor alineación con el procedimiento automático de pseudoetiquetado. No obstante, una mayor concordancia con las pseudoetiquetas no debe interpretarse como evidencia de una mayor corrección del anotador, sino únicamente como una mayor correspondencia con el criterio automático utilizado para generar las etiquetas.

Los resultados muestran que las pseudoetiquetas no reproducen de manera uniforme el criterio humano y, por tanto, no deben considerarse equivalentes a etiquetas humanas de referencia. La validación permite identificar una limitación metodológica relevante asociada tanto con la ambigüedad inherente a la clasificación de emociones financieras como con el procedimiento de generación automática de etiquetas.

Adicionalmente, el procedimiento presenta una posible circularidad metodológica, dado que FinBERT base se utilizó inicialmente para generar señales de polaridad que posteriormente fueron transformadas, mediante reglas de mapeo, en cinco pseudoetiquetas emocionales empleadas durante el fine-tuning. En consecuencia, el desempeño de FinBERT v2 frente a estas etiquetas debe interpretarse como una medida

de correspondencia con el etiquetador automático inicial y no como evidencia suficiente de que el modelo identifica correctamente las emociones expresadas en los textos. Esta precaución resulta especialmente relevante porque la concordancia entre los dos anotadores humanos fue de 47,18%, con un coeficiente kappa de Cohen de 0,2752, lo que evidencia una concordancia limitada entre sus criterios de clasificación. Por tanto, el Macro-F1 obtenido frente a las pseudoetiquetas no debe interpretarse como una medida directa de validez emocional humana. Para examinar esta limitación, resulta pertinente evaluar las predicciones de FinBERT v2 frente a una referencia humana definida mediante criterios explícitos de anotación y, cuando corresponda, analizar por separado los casos en los que ambos evaluadores coincidieron.

Tabla 17 *Evaluación de FinBERT v2 vs. frente a las anotaciones humanas*

Evaluación	Textos	Accuracy	Macro-F1
Modelo vs. Edison Fabián Moreno Hernández	301	0,35	0,26
Modelo vs. Edison Víctor Alfonso Rodríguez Guzmán	301	0,79	0,73
Modelo vs. Consenso humano	142	0,69	0,50

Con el propósito de complementar la evaluación del modelo frente a las pseudoetiquetas, se compararon sus predicciones con las anotaciones realizadas por los dos evaluadores humanos sobre los 301 textos revisados. La evaluación global mostró diferencias importantes según el anotador utilizado como referencia. Frente a las anotaciones de Edison Moreno, el modelo obtuvo un Macro-F1 de 0,2571 y una exactitud de 0,35; frente a las de Víctor Alfonso Rodríguez Guzmán, alcanzó un Macro-F1 de 0,7312 y una exactitud de 0,79. Esta divergencia evidencia una variación considerable en la correspondencia entre las predicciones del modelo y los criterios individuales de

clasificación, por lo que los resultados deben interpretarse considerando las diferencias de criterio observadas entre los anotadores.

Adicionalmente, se evaluó el modelo en los 142 textos en los que ambos anotadores coincidieron en la categoría emocional asignada, equivalentes al 47,18% de la muestra. En este subconjunto, FinBERT v2 obtuvo una exactitud de 0,69 y un Macro-F1 de 0,4954. La diferencia entre ambas métricas indica que el desempeño global no se distribuyó uniformemente entre las cinco categorías emocionales. En particular, el modelo presentó mejores resultados para neutralidad, con un F1 de 0,81, y entusiasmo, con un F1 de 0,71, seguidos de miedo, con 0,65, y enojo, con 0,31. Para tristeza, el F1 fue de 0,00 en los tres casos de referencia consensuada correspondientes a esta categoría.

Estos resultados muestran que la evaluación frente a una referencia humana consensuada aporta información complementaria a las métricas obtenidas frente a las pseudoetiquetas. Sin embargo, el desempeño observado debe interpretarse con cautela, debido al tamaño reducido de algunas categorías y a que el subconjunto consensuado representa menos de la mitad de los textos revisados. Asimismo, la concordancia entre anotadores fue limitada, con un acuerdo observado de 47,18% y un coeficiente kappa de Cohen de 0,2752. Por tanto, esta evaluación no constituye una validación definitiva de la clasificación emocional, pero permite documentar las diferencias entre el etiquetado automático y los criterios humanos disponibles.

Resumen consolidado de resultados

La Tabla 18 sintetiza los principales resultados obtenidos y su relación con las hipótesis del estudio. El corpus inicial fue sometido a un proceso de depuración que redujo el

conjunto de 10.000 a 7.752 textos. Posteriormente, mediante FinBERT base y reglas heurísticas de mapeo, se generaron pseudoetiquetas correspondientes a cinco categorías emocionales, con una marcada concentración en neutralidad.

Tabla 18 *Resumen consolidado de resultados*

Componente	Indicador	Resultado actualizado
FinBERT v2	Versión final	FinBERT v2
FinBERT v2	Mejor época de validación	8
FinBERT v2	Macro-F1 de validación	0,5654
FinBERT v2	MCC de validación	0,6485
FinBERT v2	Macro-F1 de prueba	0,5564
FinBERT v2	MCC de prueba	0,6705
FinBERT v2	Accuracy de prueba	0,8813
FinBERT v2	Weighted-F1 de prueba	0,8901
FinBERT v2	IC 95% del Macro-F1 de prueba	[0,5198; 0,5919]
H1	Chi-cuadrado de independencia	$\chi^2(4) = 126{,}422$
H1	Valor p	$p < 0{,}0001$
H1	V de Cramér	0,128
H1	Divergencia KL simétrica	0,0356
H1	Decisión	Evidencia a favor de H1
H2	Macro-F1 multiclase agrupado	0,8624
H2	Macro-F1 binario independiente	0,8311
H2	Diferencia de Macro-F1	−0,0309
H2	IC 95% de la diferencia	[−0,0788; 0,0128]
H2	Estadístico de McNemar	0,0172
H2	Valor p de McNemar	0,8955
H2	Decisión	H2 no respaldada

Nota. MCC = coeficiente de correlación de Matthews; KL = divergencia de Kullback-Leibler simétrica; IC = intervalo de confianza. Para H1, la prueba Chi-cuadrado de independencia sustenta la inferencia estadística sobre la asociación entre el tipo de fuente

y la categoría emocional, mientras que el V de Cramér cuantifica la magnitud de dicha asociación y la divergencia KL simétrica describe la separación entre las distribuciones. Para H2, la comparación se realizó entre las predicciones del modelo multiclase agrupadas en entusiasmo y no entusiasmo y el clasificador binario independiente, ambos evaluados sobre las mismas 1.163 observaciones del conjunto de prueba. La diferencia de Macro-F1 describe la dirección y magnitud del contraste, el intervalo de confianza bootstrap pareado cuantifica su incertidumbre y la prueba de McNemar evalúa las discordancias entre los patrones de acierto y error de ambos enfoques.

El modelo FinBERT ajustado mediante fine-tuning fue evaluado sobre un conjunto de prueba independiente de 1.163 textos, alcanzando un Macro-F1 de 0,5564, Accuracy de 0,8813, Weighted-F1 de 0,8901 y MCC de 0,6705. El análisis por clase evidenció diferencias importantes entre categorías, con mejor desempeño en neutralidad y menor desempeño en enojo.

Para H1, la prueba Chi-cuadrado de independencia mostró una asociación estadísticamente significativa entre el tipo de fuente y la categoría emocional, con $\chi^2(4) = 126{,}422$ y $p < 0{,}0001$. El V de Cramér fue de 0,128, lo que indica que, aunque existe evidencia estadística de asociación, su magnitud es limitada. Como complemento descriptivo, la divergencia KL simétrica fue de 0,0356 y permitió cuantificar la separación entre las distribuciones emocionales de las fuentes retail e institucionales. En consecuencia, los resultados aportan evidencia a favor de H1 dentro del corpus pseudoetiquetado analizado, sin implicar relaciones causales ni permitir generalizaciones más allá de las fuentes, activos, periodo e idioma considerados en la investigación.

Para H2, se comparó el desempeño de las predicciones del modelo multiclase agrupadas en las categorías entusiasmo y no entusiasmo con un clasificador binario entrenado independientemente para la misma tarea, utilizando las mismas 1.163 observaciones del conjunto de prueba. El modelo multiclase agrupado obtuvo un Macro-F1 de 0,8624, mientras que el clasificador binario independiente alcanzó un Macro-F1 de 0,8311. La diferencia descriptiva, calculada como el Macro-F1 del clasificador binario independiente menos el del modelo multiclase agrupado, fue de −0,0309, equivalente a −3,09 puntos porcentuales.

La incertidumbre de esta diferencia se evaluó mediante un bootstrap pareado de 2.000 réplicas, obteniéndose un intervalo de confianza del 95% de [−0,0788; 0,0128]. Dado que este intervalo incluye el cero, no se encontró evidencia suficiente para afirmar que la diferencia entre ambos valores de Macro-F1 es estadísticamente distinta de cero. De manera consistente, la prueba de McNemar produjo un estadístico de 0,0172 y un valor p de 0,8955, por lo que no se rechazó la hipótesis nula de igualdad en las proporciones de aciertos pareados.

En consecuencia, los resultados no respaldan H2 en su formulación de una diferencia estadísticamente significativa de desempeño entre los dos enfoques binarios comparados. Aunque el modelo multiclase agrupado presentó un Macro-F1 descriptivamente superior al clasificador binario independiente, la ausencia de significancia estadística no demuestra equivalencia entre ambos modelos. Además, el modelo multiclase conserva una mayor granularidad analítica al distinguir cinco categorías emocionales, mientras que la formulación binaria se limita a diferenciar entusiasmo frente a no entusiasmo.

Propuesta de solución a la problemática

Situación actual

La investigación desarrollada parte de la necesidad de analizar con mayor granularidad el contenido emocional presente en textos financieros provenientes de fuentes retail e institucionales. El estudio se concentró en información relacionada con el índice S&P 500 y Tesla Inc. (TSLA) durante el periodo comprendido entre enero de 2022 y abril de 2026, con el propósito de identificar y comparar cinco categorías emocionales —enojo, entusiasmo, miedo, neutralidad y tristeza— mediante un modelo FinBERT ajustado al dominio financiero. Esta estructura permitió abordar el problema desde una perspectiva que supera la clasificación tradicional de polaridad positiva, negativa o neutral y que busca conservar diferencias emocionales más específicas dentro del discurso financiero.

El diseño metodológico contempló inicialmente un corpus de 10.000 textos, posteriormente sometido a un proceso de limpieza, normalización y filtrado. Como resultado del preprocesamiento, el conjunto se redujo a 7.752 registros, lo que representó la exclusión de 2.248 observaciones, equivalentes al 22,5% del corpus inicial. Entre los criterios aplicados se incluyeron la eliminación de textos demasiado cortos, contenidos no financieros y documentos que superaban los límites de longitud establecidos para cada tipo de fuente. El filtro de longitud produjo la mayor reducción, con 1.627 textos excluidos.

Aunque el proceso de depuración permitió obtener un conjunto de textos más adecuado para el análisis, los resultados revelaron una segunda dificultad de mayor relevancia metodológica: la distribución de las pseudoetiquetas emocionales presentó un marcado desbalance. De los 7.752 textos procesados, 6.411 fueron clasificados como neutralidad, equivalentes al 82,7% del total, mientras que enojo estuvo representado únicamente por

55 textos, correspondientes al 0,7%. Esta concentración significa que la tarea de clasificación no estuvo compuesta por cinco categorías con disponibilidad semejante de información, sino por una clase mayoritaria ampliamente dominante y varias emociones con una representación considerablemente menor.

Esta situación afectó directamente la evaluación del modelo multiclase. Un clasificador puede alcanzar valores elevados de exactitud cuando reconoce correctamente la clase mayoritaria, sin que ello implique necesariamente un comportamiento equilibrado frente a las categorías minoritarias. Por esta razón, el estudio utilizó Macro-F1 como una métrica central de evaluación, dado que asigna el mismo peso a cada clase independientemente de su frecuencia. El uso de esta métrica permitió identificar diferencias que no son igualmente visibles cuando se observan exclusivamente la exactitud o las métricas ponderadas.

Para mitigar el desbalance se implementaron una función de pérdida ponderada y un procedimiento de *oversampling* sobre la clase enojo. Durante el entrenamiento, los 39 ejemplos originales de esta categoría fueron incrementados hasta 200 mediante remuestreo. Sin embargo, esta estrategia aumentó la frecuencia de observación de los mismos registros sin incorporar nueva diversidad lingüística. En el conjunto de prueba, enojo obtuvo un F1 de 0,00 y contó con únicamente ocho observaciones, resultado que muestra las limitaciones de utilizar remuestreo como respuesta principal frente a una clase con escasa representación original.

La evaluación final del modelo FinBERT ajustado sobre un conjunto independiente de 1.163 textos mostró una exactitud de 0,8813, un Weighted-F1 de 0,8901, un MCC de 0,6705 y un Macro-F1 de 0,5564. El intervalo de confianza bootstrap del 95% para Macro-

F1 fue [0,5198; 0,5919]. La diferencia entre el Macro-F1 y las métricas ponderadas confirma la importancia de analizar el comportamiento del modelo por categoría. Mientras las métricas globales reflejan un desempeño favorable en términos agregados, el análisis por clase muestra que las emociones minoritarias continúan representando el principal desafío de clasificación.

A esta dificultad se suma el origen de las etiquetas utilizadas durante el entrenamiento. FinBERT fue empleado inicialmente para producir las categorías *positive*, *negative* y *neutral*, que posteriormente fueron transformadas en las cinco emociones consideradas en la investigación. Las categorías negativas fueron subdivididas mediante palabras clave y reglas heurísticas. Este procedimiento hizo posible construir un corpus pseudoetiquetado a una escala suficiente para desarrollar el fine-tuning, pero también introdujo una limitación metodológica importante: el modelo utilizado como punto de partida participó en la generación de las señales que posteriormente sirvieron como referencia para su propio ajuste. En consecuencia, existe el riesgo de reproducir patrones, errores o sesgos del etiquetador inicial en lugar de aprender exclusivamente a partir de criterios humanos independientes.

La validación humana realizada sobre una muestra de 301 textos permitió examinar esta limitación desde una referencia externa. Los resultados muestran que los dos anotadores coincidieron en 142 casos, equivalentes a un acuerdo observado del 47,18%, con un coeficiente kappa de Cohen de 0,2752. Asimismo, la correspondencia entre cada anotador y las pseudoetiquetas fue desigual: uno de los anotadores presentó una coincidencia del 35,22%, mientras que el otro alcanzó el 78,74%. Estos resultados indican que las pseudoetiquetas no pueden considerarse equivalentes a una referencia humana

definitiva y que el esquema de clasificación requiere mecanismos adicionales de calibración, revisión y consenso.

No obstante, los resultados también muestran que el esquema multiclase de pseudoetiquetado permitió caracterizar diferencias entre las fuentes analizadas. Para H1, la prueba Chi-cuadrado de independencia mostró una asociación estadísticamente significativa entre el tipo de fuente y la categoría emocional, con $\chi^2(4) = 126{,}422$ y $p < 0{,}0001$. El V de Cramér fue de 0,128, lo que indica que la magnitud de la asociación fue limitada. De manera complementaria, la divergencia KL simétrica fue de 0,0356 y permitió cuantificar descriptivamente la separación entre ambas distribuciones. Las mayores diferencias descriptivas se observaron en entusiasmo, tristeza y enojo. En conjunto, estos resultados aportan evidencia a favor de H1 dentro del corpus pseudoetiquetado analizado y respaldan la utilidad del análisis multiclase para caracterizar diferencias en el contenido emocional de fuentes retail e institucionales, sin implicar relaciones causales.

La comparación entre los dos enfoques binarios aportó otro elemento relevante para la interpretación de los resultados. Las predicciones del modelo multiclase agrupadas en las categorías entusiasmo y no entusiasmo obtuvieron un Macro-F1 de 0,8624, mientras que el clasificador binario entrenado independientemente alcanzó un Macro-F1 de 0,8311. La diferencia descriptiva, calculada como el Macro-F1 del clasificador binario independiente menos el del modelo multiclase agrupado, fue de −0,0309, equivalente a −3,09 puntos porcentuales.

Sin embargo, esta diferencia descriptiva no fue estadísticamente significativa. El intervalo de confianza del 95% obtenido mediante bootstrap pareado de 2.000 réplicas fue [−0,0788; 0,0128], por lo que incluyó el cero. De manera consistente, la prueba de

McNemar produjo un estadístico de 0,0172 y un valor p de 0,8955. En consecuencia, no se encontró evidencia estadísticamente suficiente para establecer una diferencia de desempeño entre los dos enfoques binarios comparados, por lo que los resultados no respaldan H2.

Estos resultados deben interpretarse considerando que ambos enfoques fueron evaluados sobre una formulación binaria común, mientras que el modelo multiclase original conserva además la capacidad de distinguir cinco categorías emocionales específicas. La ausencia de significancia estadística entre los dos enfoques binarios no demuestra equivalencia entre ellos, sino que indica que, con las observaciones y los procedimientos utilizados, no se obtuvo evidencia suficiente para establecer una diferencia estadísticamente significativa.

Por tanto, la decisión entre una formulación binaria y una multiclase no debe basarse exclusivamente en una métrica global de desempeño. El problema de investigación busca caracterizar cómo se distribuyen distintas emociones financieras entre fuentes retail e institucionales. Reducir el análisis a entusiasmo frente a no entusiasmo implicaría perder parte de la granularidad emocional que constituye uno de los principales aportes analíticos del estudio.

En conjunto, la situación actual identificada por la investigación puede sintetizarse en tres necesidades fundamentales: mejorar la calidad y trazabilidad de las etiquetas utilizadas como referencia; fortalecer la representación y diversidad lingüística de las emociones minoritarias; y conservar un enfoque analítico que permita mantener la granularidad emocional cuando esta sea relevante para el objetivo de investigación, sin asumir que una formulación binaria constituye necesariamente una alternativa superior.

Estas necesidades constituyen el punto de partida para definir las oportunidades de mejora y construir una propuesta de solución que responda directamente a los hallazgos obtenidos.

Oportunidades

Los resultados obtenidos permiten identificar varias oportunidades de mejora orientadas a fortalecer la clasificación emocional multiclase y aumentar la confiabilidad del proceso analítico. Estas oportunidades se derivan directamente de las principales limitaciones observadas durante la investigación: la dependencia de pseudoetiquetas generadas automáticamente, la baja representación de algunas emociones, la incertidumbre asociada con las predicciones, la necesidad de mejorar la trazabilidad del proceso y la conveniencia de aprovechar las diferencias observadas entre fuentes retail e institucionales.

La primera oportunidad consiste en fortalecer la calidad de las etiquetas mediante validación humana asistida por inteligencia artificial. La revisión humana de 301 textos mostró que las pseudoetiquetas no reproducen de manera uniforme el criterio humano y que existe una concordancia limitada entre anotadores. Por tanto, una mejora metodológica no debería limitarse a producir un mayor volumen de pseudoetiquetas, sino incorporar un proceso progresivo de revisión humana que permita identificar errores, ambigüedades y desacuerdos. Este enfoque permitiría construir gradualmente un subconjunto de referencia humana de mayor calidad y reducir la dependencia de las decisiones automáticas iniciales.

En este sentido, la inteligencia artificial puede utilizarse como mecanismo de apoyo y priorización, no como sustituto del criterio humano. El sistema podría identificar

automáticamente los textos con menor confianza, aquellos pertenecientes a clases minoritarias o aquellos en los que existan inconsistencias entre diferentes reglas de clasificación, y enviarlos prioritariamente a revisión. Esta estrategia permitiría concentrar el esfuerzo de anotación humana en los casos con mayor valor informativo y disminuir la necesidad de revisar manualmente la totalidad del corpus.

La segunda oportunidad corresponde a mejorar la representación de las emociones minoritarias mediante la incorporación de ejemplos reales y lingüísticamente diversos. El *oversampling* aplicado a la clase enojo incrementó artificialmente el número de observaciones disponibles durante el entrenamiento, pero no generó nuevas formas de expresión emocional, debido a que el procedimiento se basó en la repetición de ejemplos existentes.

Por esta razón, una estrategia de mejora debería priorizar la incorporación de nuevos textos que representen distintas formas lingüísticas de expresar enojo, miedo y tristeza en contextos financieros. El objetivo no debería ser alcanzar necesariamente un equilibrio artificial entre las cinco categorías, sino aumentar la diversidad semántica de las clases menos representadas y reducir la dependencia de unos pocos ejemplos repetidos. Estos textos podrían seleccionarse de nuevas fuentes o periodos y posteriormente someterse a anotación humana.

La tercera oportunidad consiste en incorporar explícitamente la incertidumbre de clasificación. El desempeño obtenido mostró que una predicción categórica única no siempre refleja adecuadamente el nivel de certeza del modelo. En lugar de conservar únicamente la clase asignada, el sistema podría almacenar las probabilidades asociadas con cada una de las cinco emociones. Esto permitiría identificar textos en los que la

diferencia entre las dos emociones más probables sea reducida y, por tanto, exista mayor ambigüedad.

Esta información podría utilizarse para establecer mecanismos de revisión selectiva. Por ejemplo, una predicción con alta probabilidad de neutralidad y baja probabilidad para las demás categorías podría procesarse automáticamente, mientras que un texto con probabilidades similares entre miedo y tristeza podría enviarse a revisión humana. De esta manera, la incertidumbre se convertiría en un criterio operativo dentro del flujo de clasificación.

La cuarta oportunidad corresponde a implementar un esquema human-in-the-loop. La investigación evidenció una posible circularidad metodológica al utilizar FinBERT para generar las pseudoetiquetas que posteriormente se emplearon para ajustar el mismo modelo. Un esquema human-in-the-loop permitiría romper parcialmente esta dependencia mediante un ciclo iterativo de clasificación automática, selección de casos problemáticos, revisión humana, corrección de etiquetas y nuevo entrenamiento.

La ventaja de este enfoque consiste en que la intervención humana no se limitaría a una validación final, sino que pasaría a formar parte del proceso de aprendizaje. A medida que los anotadores corrijan casos ambiguos o erróneos, el conjunto de referencia podría mejorar progresivamente y proporcionar al modelo ejemplos más confiables para nuevas etapas de ajuste.

La quinta oportunidad consiste en aprovechar las diferencias observadas entre fuentes retail e institucionales como un componente analítico adicional. Los resultados de H1 mostraron una asociación estadísticamente significativa entre el tipo de fuente y la

categoría emocional dentro del corpus analizado, con $\chi^2(4) = 126{,}422$ y $p < 0{,}0001$. No obstante, el V de Cramér de 0,128 indicó que la magnitud de dicha asociación fue limitada. Estas diferencias pueden utilizarse para desarrollar indicadores comparativos que permitan observar cómo varía la composición emocional entre fuentes, sin interpretar tales diferencias como relaciones causales con precios o rendimientos bursátiles.

A partir de esta evidencia, podría construirse un indicador periódico de divergencia emocional entre fuentes retail e institucionales. Este indicador podría calcularse de forma agregada y, posteriormente, desagregarse por activo o periodo temporal. Su utilidad estaría en proporcionar una representación sistemática de las diferencias entre ambos tipos de discurso financiero, no en predecir movimientos futuros del mercado.

La sexta oportunidad consiste en fortalecer la representatividad temporal del corpus. Aunque el estudio comprende el periodo enero de 2022–abril de 2026, la disponibilidad de textos no fue uniforme a lo largo del tiempo. Por tanto, una implementación futura debería incorporar criterios de muestreo temporal que permitan evitar concentraciones excesivas en determinados años, meses o eventos de mercado.

La séptima oportunidad corresponde a mejorar la trazabilidad y reproducibilidad del pipeline. Cada texto debería conservar información sobre su fuente, activo, periodo, versión del preprocesamiento, pseudoetiqueta inicial, anotación humana —cuando exista—, probabilidades del modelo y versión del modelo utilizado. Esta trazabilidad permitiría reconstruir las decisiones realizadas en cada etapa y facilitaría auditorías posteriores sobre errores de clasificación o cambios entre versiones.

Finalmente, los resultados permiten identificar una oportunidad transversal relacionada con la selección de la granularidad de clasificación. La comparación entre el modelo multiclase agrupado y el clasificador binario independiente no mostró una diferencia estadísticamente significativa de desempeño, a pesar de que el primero presentó un Macro-F1 descriptivamente superior. Por tanto, la elección entre una formulación binaria y una multiclase no debería basarse exclusivamente en una diferencia puntual de desempeño, sino también en el nivel de detalle requerido por el objetivo analítico.

En este estudio, la clasificación multiclase conserva el valor de distinguir cinco categorías emocionales —enojo, entusiasmo, miedo, neutralidad y tristeza—, mientras que la formulación binaria simplifica el problema a entusiasmo frente a no entusiasmo. En consecuencia, el reto consiste en fortalecer la calidad y confiabilidad del sistema multiclase cuando la granularidad emocional sea relevante, sin asumir que una formulación binaria constituye necesariamente una alternativa superior.

En conjunto, estas oportunidades permiten avanzar desde un esquema centrado principalmente en pseudoetiquetas automáticas hacia una arquitectura metodológica más robusta, en la que la clasificación automática, la supervisión humana, la incertidumbre, la diversidad semántica, la trazabilidad y la selección adecuada de la granularidad funcionen de manera integrada. Este enfoque permitiría aumentar la confiabilidad del análisis sin perder de vista que la elección entre formulaciones multiclase y binarias debe responder al propósito analítico y a la calidad de la evidencia disponible.

Propuesta de solución al problema planteado

A partir de los resultados obtenidos, se propone el desarrollo de un sistema de análisis emocional financiero multiclase con validación humana progresiva, tratamiento explícito de la incertidumbre y comparación entre fuentes retail e institucionales. La propuesta busca responder a las principales limitaciones identificadas durante la investigación sin abandonar el enfoque de cinco emociones, dado que esta granularidad constituye un elemento central para responder la pregunta de investigación.

La propuesta parte de una decisión metodológica orientada por el propósito analítico: conservar la clasificación multiclase como componente central cuando sea necesario distinguir emociones financieras específicas, sin asumir que una formulación binaria constituye una alternativa superior. En la comparación equivalente realizada para H2, las predicciones del modelo multiclase agrupadas en entusiasmo y no entusiasmo obtuvieron un Macro-F1 de 0,8624, mientras que el clasificador binario independiente alcanzó 0,8311. Sin embargo, el intervalo de confianza bootstrap de la diferencia incluyó el cero y la prueba de McNemar produjo un valor p de 0,8955, por lo que no se encontró evidencia estadísticamente significativa de diferencia entre ambos enfoques binarios.

Esta evidencia indica que la selección de la granularidad no debería depender exclusivamente de una diferencia puntual de desempeño. La formulación multiclase conserva la capacidad de distinguir enojo, entusiasmo, miedo, neutralidad y tristeza, mientras que la formulación binaria responde a una tarea más específica de entusiasmo frente a no entusiasmo. Por tanto, la elección entre ambas debe responder al objetivo analítico, al nivel de detalle requerido y a la calidad de la evidencia disponible.

El objetivo de la solución no consiste únicamente en aumentar el desempeño predictivo global del modelo, sino en mejorar la confiabilidad, trazabilidad y capacidad de interpretación de la clasificación multiclase, especialmente en las categorías minoritarias.

El primer componente corresponde a la construcción progresiva de un corpus con referencia humana. En lugar de depender exclusivamente de pseudoetiquetas generadas de forma automática, se propone seleccionar una muestra estratificada de textos para anotación humana independiente. Esta muestra debería incluir representación de las cinco emociones, con especial atención a las categorías minoritarias, a los textos con baja confianza del modelo y a aquellos casos en los que las reglas heurísticas produzcan ambigüedad.

Las anotaciones humanas deberían utilizar definiciones operacionales consistentes con las establecidas en la investigación para enojo, entusiasmo, miedo, neutralidad y tristeza. Sin embargo, los resultados obtenidos en la validación de 301 textos muestran que estas definiciones requieren una etapa adicional de calibración entre anotadores antes de utilizar las etiquetas humanas como referencia definitiva. Por ello, se recomienda incorporar ejercicios de entrenamiento, revisión de desacuerdos y ajuste de la guía de anotación antes de ampliar el conjunto humano de referencia.

El segundo componente es la implementación de un esquema human-in-the-loop. En el procedimiento desarrollado durante la investigación, FinBERT produjo inicialmente las pseudoetiquetas que posteriormente sirvieron como referencia para el fine-tuning. La propuesta modifica este flujo mediante la incorporación sistemática de revisión humana en los casos de mayor incertidumbre. El proceso puede representarse de la siguiente manera:

Corpus financiero → Clasificación automática inicial → Identificación de casos inciertos

→ Validación humana → Corpus corregido → fine-tuning

→ Nueva evaluación

Este ciclo permitiría reducir progresivamente la dependencia de las pseudoetiquetas como referencia exclusiva y transformar la intervención humana en una parte activa del proceso de aprendizaje.

El tercer componente consiste en fortalecer las clases minoritarias mediante nuevos ejemplos reales y lingüísticamente diversos. El *oversampling* puede mantenerse como técnica complementaria, pero no debería constituir la principal estrategia para enfrentar el desbalance. La prioridad sería incorporar nuevos textos representativos de enojo, miedo y tristeza y someterlos a revisión humana. De esta forma, el modelo tendría acceso a una mayor diversidad léxica y semántica, en lugar de observar repetidamente los mismos ejemplos originales.

Este componente responde directamente al resultado obtenido para la clase enojo. Aunque durante el entrenamiento sus 39 ejemplos originales fueron ampliados artificialmente hasta 200 mediante remuestreo, en el conjunto de prueba la categoría enojo obtuvo un F1 de 0,00 con solo ocho observaciones. Este resultado muestra que el incremento numérico mediante repetición no equivale a una ampliación real de la diversidad emocional disponible para el aprendizaje y que, en esta evaluación, el modelo no logró identificar correctamente los casos de enojo.

El cuarto componente corresponde a la incorporación explícita de la incertidumbre de las predicciones. Cada resultado del modelo debería almacenar no únicamente la emoción seleccionada, sino también las probabilidades asignadas a las cinco categorías. El sistema podría representar cada observación mediante un vector:

$$E = \left(P_{\text{enojo}}, P_{\text{entusiasmo}}, P_{\text{miedo}}, P_{\text{neutralidad}}, P_{\text{tristeza}}\right)$$

La categoría final correspondería a la emoción con mayor probabilidad, pero el vector completo permitiría identificar casos ambiguos. Por ejemplo, si un texto presenta probabilidades próximas entre miedo y tristeza, la observación podría considerarse de baja confianza y enviarse a revisión humana. En cambio, las predicciones con una diferencia amplia entre la primera y segunda categoría podrían procesarse automáticamente.

Este mecanismo permitiría utilizar los recursos humanos de forma más eficiente, concentrando la revisión en los textos donde la intervención humana puede aportar mayor valor.

El quinto componente consiste en construir un indicador comparativo de la distribución emocional entre fuentes retail e institucionales. La investigación mostró una asociación estadísticamente significativa entre el tipo de fuente y la categoría emocional mediante la prueba Chi-cuadrado, con $\chi^2(4) = 126{,}422$ y $p < 0{,}0001$. No obstante, el V de Cramér de 0,128 indicó que la magnitud de dicha asociación fue limitada, mientras que la divergencia KL simétrica permitió cuantificar descriptivamente la separación entre las distribuciones.

A partir de estos resultados, la solución podría calcular periódicamente las proporciones correspondientes a enojo, entusiasmo, miedo, neutralidad y tristeza para cada tipo de

fuente y representar sus diferencias mediante indicadores descriptivos. El propósito no sería predecir movimientos de mercado, sino proporcionar una herramienta que permita caracterizar sistemáticamente cómo se distribuye el contenido emocional en ambos tipos de discurso financiero.

El sexto componente corresponde al análisis diferenciado por fuente, activo y periodo. La estructura del sistema debería conservar estas dimensiones por separado para evitar interpretaciones agregadas que oculten diferencias relevantes. La propuesta puede contemplar comparaciones retail frente a institucional para el S&P 500, retail frente a institucional para TSLA y análisis temporales de la evolución de las distribuciones emocionales.

Estas comparaciones deberán interpretarse de forma descriptiva y asociativa. El estudio no evaluó relaciones causales entre emociones y precios bursátiles, por lo que el sistema no debería presentar cambios emocionales como causas de movimientos de los activos.

El séptimo componente consiste en desarrollar un panel de resultados y trazabilidad. Este panel integraría en una misma interfaz la distribución porcentual de las cinco emociones, las comparaciones entre fuentes, el desempeño por categoría, los niveles de confianza, el número de textos revisados manualmente, la concordancia entre modelo y anotadores y las medidas descriptivas de divergencia entre distribuciones.

La trazabilidad constituye un elemento particularmente importante. Cada observación debería conservar información sobre su fuente, activo, periodo, versión de

preprocesamiento, pseudoetiqueta inicial, anotación humana cuando exista, probabilidades estimadas y versión del modelo responsable de la predicción. Esta estructura facilitaría la reproducibilidad del proceso y permitiría revisar posteriormente la procedencia de cualquier resultado.

En términos de evaluación, la propuesta no debería utilizar la exactitud global como único criterio de éxito. El desempeño tendría que evaluarse mediante Macro-F1, F1 por clase, MCC, intervalos de confianza y concordancia con un conjunto humano de referencia. La exactitud y el Weighted-F1 podrían mantenerse como métricas complementarias, pero no deberían ocultar posibles dificultades en las clases con menor representación.

La solución propuesta constituye, por tanto, una evolución del pipeline desarrollado en esta investigación. Se conservarían sus componentes centrales —corpus financiero, FinBERT, clasificación multiclase, comparación entre fuentes y análisis estadístico—, pero se modificaría el aspecto metodológico más sensible: la dependencia predominante de pseudoetiquetas automáticas como referencia para el entrenamiento.

El aporte esperado de esta propuesta consiste en establecer una ruta metodológica más robusta para la clasificación emocional financiera. Su valor no reside únicamente en producir métricas de desempeño más altas, sino en combinar automatización y validación humana para obtener etiquetas más confiables, fortalecer el aprendizaje de las categorías minoritarias, hacer explícita la incertidumbre de las predicciones y conservar la diferenciación emocional cuando esta resulte necesaria para comparar fuentes retail e institucionales.

En relación con el objetivo general, el “para qué” de esta propuesta radica en disponer de un procedimiento capaz de caracterizar y comparar de manera más confiable las distribuciones de emociones financieras presentes en diferentes tipos de fuentes. La granularidad de cinco categorías constituye el nivel de detalle principal del sistema cuando el objetivo sea distinguir emociones específicas, mientras que formulaciones binarias pueden utilizarse cuando respondan a preguntas analíticas más acotadas. De esta forma, la solución amplía la utilidad del modelo como herramienta de análisis del discurso financiero sin atribuirle capacidades predictivas o causales que no fueron evaluadas en la investigación.

Discusión

Limitaciones de la investigación

La interpretación de los resultados debe considerar varias limitaciones metodológicas relacionadas con la distribución de las emociones, la cobertura temporal del corpus y la concentración de las fuentes analizadas. Estas limitaciones no invalidan los resultados obtenidos, pero delimitan el alcance de las conclusiones y orientan las principales oportunidades de investigación futura.

Desbalance estructural del corpus

La primera limitación corresponde al marcado desbalance de las pseudoetiquetas emocionales. De los 7.752 textos utilizados después del preprocesamiento, 6.411 fueron clasificados como neutralidad (82,7%), mientras que enojo estuvo representado por 55 observaciones (0,7%). Durante el entrenamiento, esta última categoría contó originalmente con 39 ejemplos, lo que limitó considerablemente la cantidad y diversidad de información disponible para aprender sus patrones lingüísticos.

Para mitigar este problema se aplicó *oversampling* con reemplazo, aumentando los 39 ejemplos originales de enojo hasta 200 observaciones durante el entrenamiento, junto con una función de pérdida ponderada. Sin embargo, este procedimiento incrementó la frecuencia de presentación de los ejemplos existentes sin introducir nueva diversidad lingüística. En el conjunto de prueba, la clase enojo obtuvo un F1 de 0,00 sobre ocho observaciones, lo que evidencia la dificultad de evaluar y aprender de manera robusta una categoría con soporte reducido y escasa diversidad de ejemplos.

El bajo número de observaciones de enojo también implica una elevada incertidumbre en la estimación de su desempeño individual, dado que cambios en unas pocas predicciones pueden modificar de manera importante el F1 de esta categoría. Esta limitación debe distinguirse de la incertidumbre asociada con el desempeño multiclase global. Para este último, el modelo obtuvo un Macro-F1 de 0,5564 y un intervalo de confianza bootstrap del 95 % de [0,5198; 0,5919]. Este intervalo complementa la estimación puntual del desempeño global del clasificador, pero no constituye un intervalo de confianza específico para la clase enojo.

Representatividad temporal del corpus

La segunda limitación corresponde a la representatividad temporal del corpus. Aunque la investigación comprende el periodo entre enero de 2022 y abril de 2026, la disponibilidad de textos no fue uniforme a lo largo de todo el intervalo. La cantidad de observaciones varió entre años y, adicionalmente, 2026 corresponde únicamente al periodo enero–abril, por lo que su total no es directamente comparable con los años completos de 2022 a 2025. En consecuencia, la distribución temporal del corpus refleja la disponibilidad de las fuentes recopiladas y no un diseño de muestreo con igual número de observaciones por año, mes o evento de mercado.

Esta concentración desigual puede generar una mayor representación de determinados momentos del mercado y una menor cobertura de otros. Por tanto, los resultados deben interpretarse como correspondientes al corpus disponible durante el periodo estudiado y no como una representación temporal homogénea de todas las condiciones del mercado entre enero de 2022 y abril de 2026.

Esta limitación resulta relevante porque las formas de expresión, los temas dominantes y las condiciones del mercado pueden variar entre periodos. Una concentración elevada de observaciones en determinados meses o eventos puede influir en las distribuciones emocionales observadas y en los patrones lingüísticos disponibles para el entrenamiento y la evaluación del modelo. En futuras investigaciones, sería conveniente ampliar la cobertura temporal e implementar estrategias de muestreo estratificado por año y mes que permitan reducir concentraciones excesivas y evaluar con mayor precisión la estabilidad de las distribuciones emocionales y del desempeño del clasificador frente a cambios en el contexto financiero.

Tabla 19 *Distribución temporal del del corpus final por año*

Año	Total de textos	% del corpus
2022	1.231	15,88%
2023	1.747	22,54%
2024	1.583	20,42%
2025	1.718	22,16%
2026*	1.473	19,00%
Total	**7.752**	**100,00%**

Nota. La tabla presenta la cantidad de textos del corpus final agrupados por año. El total de 7.752 observaciones corresponde al periodo comprendido entre enero de 2022 y abril de 2026. Los valores de 2022 a 2025 corresponden a años calendario completos dentro del periodo de estudio, mientras que 2026 incluye únicamente los textos disponibles entre enero y abril; por tanto, los totales anuales no deben interpretarse como periodos de igual duración ni como evidencia de una cobertura temporal homogénea.

La distribución desagregada de textos por año, fuente, tipo y activo se presenta en el Anexo A.

Alcance geográfico y sesgo de fuente

La tercera limitación se relaciona con el alcance geográfico y lingüístico del corpus y con la concentración del componente retail en un número reducido de plataformas, principalmente StockTwits. Las fuentes retail estuvieron representadas mayoritariamente por mensajes provenientes de StockTwits y, en menor medida, de Twitter Financial, mientras que las fuentes institucionales correspondieron a contenidos de Reuters, CNBC y otros medios financieros especializados. Adicionalmente, el corpus estuvo compuesto por textos en inglés asociados principalmente con el mercado estadounidense. En consecuencia, la composición de las fuentes y el contexto lingüístico y geográfico delimitan el alcance de los resultados obtenidos.

Por tanto, los patrones observados deben interpretarse como características del corpus y de las fuentes analizadas, y no generalizarse automáticamente al conjunto de inversionistas minoristas o institucionales. En particular, la mayor proporción descriptiva de entusiasmo observada en las fuentes retail podría reflejar diferencias relacionadas con la composición del corpus, las características lingüísticas de las plataformas utilizadas o diferencias entre los tipos de fuente. Sin embargo, el diseño de esta investigación no permite identificar cuál de estos factores explica dicha diferencia ni establecer relaciones causales entre ellos. En consecuencia, este hallazgo debe entenderse como una asociación observada dentro del corpus pseudoetiquetado analizado.

Asimismo, el modelo no fue evaluado con textos en español ni validado específicamente en mercados latinoamericanos. Por esta razón, su aplicabilidad directa a contextos como Colombia u otros países de la región requiere validación adicional antes de asumir que las categorías emocionales, los patrones lingüísticos y los niveles de desempeño observados

en este estudio se transfieren de manera equivalente. Futuras investigaciones deberían incorporar corpus en español y fuentes financieras latinoamericanas para evaluar la estabilidad del esquema de clasificación en contextos lingüísticos y de mercado diferentes.

Reflexión final sobre las limitaciones

Las limitaciones descritas deben considerarse conjuntamente al interpretar los hallazgos. El desbalance emocional condiciona el aprendizaje de las clases minoritarias; la cobertura temporal desigual limita la representatividad homogénea de todo el periodo estudiado; y la concentración de las fuentes restringe la extrapolación de los patrones observados a otras poblaciones, plataformas, idiomas y mercados.

La prueba Chi-cuadrado de independencia proporcionó evidencia de una asociación estadísticamente significativa entre el tipo de fuente y la categoría emocional dentro del corpus pseudoetiquetado analizado, con $\chi^2(4) = 126{,}422$ y $p < 0{,}0001$. Sin embargo, el V de Cramér de 0,128 indicó que la magnitud de dicha asociación fue limitada. De manera complementaria, la divergencia KL simétrica de 0,0356 permitió cuantificar descriptivamente la separación entre las distribuciones, sin constituir una prueba adicional de significancia estadística. Estos resultados deben interpretarse dentro de las condiciones específicas del conjunto de datos utilizado y no como evidencia causal ni como una verificación directa de las emociones reales de los autores de los textos.

Reconocer estas restricciones permite delimitar con mayor precisión el aporte de la investigación. Los resultados ofrecen evidencia empírica sobre la clasificación y comparación de cinco categorías emocionales dentro del corpus estudiado, pero su interpretación está condicionada por el desbalance entre clases, la cobertura temporal

desigual, la concentración de las fuentes y la calidad de las pseudoetiquetas utilizadas como referencia. En consecuencia, los hallazgos no deben extrapolarse automáticamente a otros periodos, plataformas, idiomas, mercados o poblaciones de inversionistas.

Estas limitaciones también orientan las principales líneas de mejora: ampliar la diversidad de las categorías minoritarias, fortalecer la validación humana mediante protocolos independientes de anotación y consenso, mejorar la representatividad temporal, diversificar las fuentes y evaluar el enfoque en otros idiomas y contextos financieros. De esta forma, las limitaciones no invalidan los resultados obtenidos, sino que definen con mayor precisión su alcance y las condiciones bajo las cuales pueden interpretarse.

Conclusiones y trabajo futuro

Conclusiones

Conclusión 1. Caracterización del corpus y comportamiento emocional

La caracterización del corpus permitió establecer la composición de los textos financieros analizados y evidenció un marcado desbalance entre las cinco categorías de pseudoetiquetas emocionales. El corpus inicial estuvo compuesto por 10.000 textos, distribuidos en partes iguales entre fuentes retail e institucionales y entre S&P 500 y Tesla, con 5.000 textos por tipo de fuente y 5.000 por activo. Después del proceso de depuración y preprocesamiento, se obtuvieron 7.752 textos para las etapas de clasificación y análisis. La distribución de las pseudoetiquetas estuvo dominada por neutralidad, con 6.411 textos (82,7%), seguida de entusiasmo (7,1%), miedo (5,6%), tristeza (4,0%) y enojo (0,7%; 55 textos).

Estos resultados responden al primer objetivo específico, relacionado con la caracterización del corpus y la identificación de sus principales patrones de pseudoetiquetado emocional. El principal hallazgo fue que la elevada concentración en neutralidad y la reducida representación de algunas categorías, especialmente enojo, generan un escenario de clasificación desbalanceado. En consecuencia, el análisis multiclase del corpus requiere estrategias específicas para evitar que las categorías con mayor frecuencia dominen el proceso de aprendizaje y para fortalecer la representación de las emociones minoritarias.

Conclusión 2. Desempeño del modelo FinBERT en la clasificación emocional multiclase

El fine-tuning de FinBERT permitió construir un clasificador para las cinco categorías emocionales definidas en la investigación: enojo, entusiasmo, miedo, neutralidad y tristeza. Sobre un conjunto de prueba independiente de 1.163 textos, el modelo alcanzó una exactitud de 0,8813, un Micro-F1 de 0,8813, un Weighted-F1 de 0,8901, un MCC de 0,6705 y un Macro-F1 de 0,5564. El intervalo de confianza bootstrap del 95 % para Macro-F1 fue [0,5198; 0,5919].

El análisis por categoría evidenció un comportamiento desigual entre las cinco clases. Neutralidad presentó el mayor desempeño, con un F1 de 0,95, seguida de entusiasmo, con 0,75; miedo, con 0,59; tristeza, con 0,50; y enojo, con 0,00. Esta última categoría contó originalmente con 39 ejemplos en el conjunto de entrenamiento y únicamente ocho observaciones en el conjunto de prueba. A pesar del oversampling aplicado durante el entrenamiento, el modelo no identificó correctamente casos de enojo en el conjunto de prueba, lo que evidencia que el remuestreo no compensó la limitada disponibilidad y diversidad de ejemplos originales de esta categoría.

Estos resultados responden al segundo objetivo específico, relacionado con la clasificación de los textos y la evaluación del desempeño multiclase. La diferencia entre las métricas ponderadas y Macro-F1 muestra que la evaluación de este tipo de modelos no debe depender exclusivamente de la exactitud global, sino considerar conjuntamente Macro-F1, MCC y las métricas por clase, especialmente cuando existe un desbalance marcado entre las categorías.

Conclusión 3. Diferencias entre las distribuciones emocionales de las fuentes

La comparación estadística de las pseudoetiquetas emocionales mostró una asociación estadísticamente significativa entre el tipo de fuente y la categoría emocional. La prueba Chi-cuadrado de independencia obtuvo $\chi^2(4) = 126{,}422$ y $p < 0{,}0001$, por lo que se rechazó la hipótesis nula de independencia. No obstante, el V de Cramér fue de 0,128, lo que indica que la magnitud de la asociación fue limitada. De manera complementaria, la divergencia KL simétrica fue de 0,0356 y se utilizó como medida descriptiva para cuantificar la separación entre las distribuciones.

Las diferencias descriptivas más visibles se observaron en entusiasmo, con 9,3% en fuentes retail frente a 5,0% en fuentes institucionales; tristeza, con 2,3% en retail frente a 5,5% en institucionales; y enojo, con 0,2% frente a 1,2%, respectivamente. Neutralidad presentó proporciones similares en ambos grupos, con 82,9% en retail y 82,5% en fuentes institucionales.

Estos resultados responden al tercer objetivo específico y aportan evidencia a favor de H1 dentro del corpus pseudoetiquetado analizado. El hallazgo debe interpretarse como una asociación entre el tipo de fuente y la distribución de las pseudoetiquetas emocionales, y no como evidencia de una diferencia causal o de una medición directa de las emociones reales de los autores de los textos. Asimismo, el contraste de H1 se realizó sobre las pseudoetiquetas generadas previamente mediante FinBERT base y reglas heurísticas de mapeo, y no sobre las predicciones del modelo FinBERT ajustado mediante fine-tuning.

Conclusión 4. Diferencia de desempeño entre los enfoques multiclase y binario y relevancia de la granularidad emocional

Para contrastar H2, se compararon dos enfoques sobre una formulación binaria común de entusiasmo frente a no entusiasmo y utilizando las mismas 1.163 observaciones del conjunto de prueba. Las predicciones del modelo multiclase original, agrupadas posteriormente en estas dos categorías, obtuvieron un Macro-F1 de 0,8624, mientras que el clasificador binario entrenado independientemente alcanzó un Macro-F1 de 0,8311. La diferencia, calculada como el Macro-F1 del clasificador binario independiente menos el del modelo multiclase agrupado, fue de −0,0309, equivalente a −3,09 puntos porcentuales. En términos descriptivos, por tanto, el clasificador binario independiente no presentó un desempeño superior.

La incertidumbre de esta diferencia se evaluó mediante bootstrap pareado de 2.000 réplicas, obteniéndose un intervalo de confianza del 95 % de [−0,0788; 0,0128]. Dado que el intervalo incluye el cero, no se encontró evidencia suficiente para afirmar que la diferencia entre ambos valores de Macro-F1 fuera estadísticamente distinta de cero. De manera consistente, la prueba de McNemar produjo un estadístico de 0,0172 y un valor p de 0,8955, por lo que no se rechazó la hipótesis nula de igualdad en las proporciones de

En consecuencia, los resultados no respaldan H2 en su formulación de una diferencia estadísticamente significativa de desempeño entre los dos enfoques binarios comparados. Sin embargo, la ausencia de significancia estadística no demuestra equivalencia entre ellos. La interpretación debe considerar que la formulación binaria se limita a distinguir entusiasmo frente a no entusiasmo, mientras que el modelo multiclase original conserva la capacidad de diferenciar cinco categorías emocionales. Por tanto, la selección del enfoque

debe responder tanto al desempeño observado como al nivel de granularidad requerido por el objetivo analítico.

Finalmente, la validación humana realizada sobre 301 textos refuerza la necesidad de interpretar el desempeño del modelo considerando la calidad de las etiquetas de referencia. Los dos anotadores coincidieron en 142 casos, equivalentes al 47,18 %, con un kappa de Cohen de 0,2752. La concordancia individual con las pseudoetiquetas fue desigual, con 35,22% para uno de los anotadores y 78,74% para el otro. Estos resultados muestran que las pseudoetiquetas no deben considerarse equivalentes a una referencia humana definitiva y orientan futuras mejoras hacia corpus anotados mediante procedimientos de calibración, consenso y validación independiente.

Síntesis frente al objetivo general y al problema de investigación

En conjunto, los resultados permiten dar respuesta al objetivo general de la investigación al caracterizar el corpus financiero, clasificar los textos en cinco categorías emocionales mediante FinBERT ajustado mediante fine-tuning y comparar estadísticamente las distribuciones de las pseudoetiquetas emocionales entre fuentes retail e institucionales relacionadas con el S&P 500 y Tesla Inc.

En relación con H1, la comparación de las pseudoetiquetas emocionales mostró una asociación estadísticamente significativa entre el tipo de fuente y la categoría emocional, con $\chi^2(4) = 126{,}422$ y $p < 0{,}0001$. No obstante, el V de Cramér de 0,128 indicó que la magnitud de dicha asociación fue limitada. Las diferencias descriptivas más visibles correspondieron a una mayor proporción de entusiasmo en las fuentes retail y mayores

proporciones de tristeza y enojo en las fuentes institucionales. Estos resultados aportan evidencia a favor de H1 dentro del corpus pseudoetiquetado analizado, sin implicar relaciones causales ni una medición directa de las emociones reales de los autores de los textos.

Al mismo tiempo, la evaluación del clasificador mostró que el desempeño multiclase no fue uniforme entre las cinco categorías y que las clases minoritarias presentaron mayores dificultades de clasificación. Para H2, las predicciones del modelo multiclase agrupadas en entusiasmo y no entusiasmo obtuvieron un Macro-F1 de 0,8624, mientras que el clasificador binario independiente alcanzó 0,8311. La diferencia descriptiva fue de −0,0309, calculada como el Macro-F1 del clasificador binario independiente menos el del modelo multiclase agrupado.

Sin embargo, el intervalo de confianza bootstrap pareado del 95 % de la diferencia, [−0,0788; 0,0128], incluyó el cero, y la prueba de McNemar produjo un valor p de 0,8955. En consecuencia, no se encontró evidencia estadísticamente significativa de una diferencia entre los dos enfoques binarios comparados y los resultados no respaldan H2. La ausencia de significancia no demuestra equivalencia entre ambos enfoques, y la selección de la granularidad debe considerar tanto el desempeño observado como el nivel de detalle emocional requerido por el objetivo analítico.

La principal contribución del estudio no reside únicamente en los valores alcanzados por las métricas de clasificación, sino en establecer un procedimiento reproducible para caracterizar y comparar cinco categorías emocionales en textos financieros y, al mismo tiempo, identificar las principales dificultades metodológicas asociadas con el desbalance de clases, la calidad de las pseudoetiquetas, la incertidumbre de las predicciones y la

selección de la granularidad de clasificación. Los hallazgos proporcionan una base metodológica para desarrollar posteriormente corpus con mayor validación humana, mejorar la representación de las categorías minoritarias y evaluar el enfoque en otras fuentes, periodos, idiomas y contextos financieros.

Trabajo futuro

A partir de los resultados obtenidos en esta investigación se identifican varias líneas de trabajo futuro orientadas a fortalecer la clasificación emocional financiera, mejorar la calidad de las etiquetas de referencia y ampliar la aplicabilidad del enfoque a otros contextos, fuentes e idiomas. Estas oportunidades se derivan principalmente de las limitaciones observadas en la calidad de las pseudoetiquetas, el desbalance entre las categorías emocionales, la cobertura temporal del corpus, la concentración de las fuentes y la necesidad de evaluar diferentes estrategias de modelado y niveles de granularidad bajo condiciones comparables.

Una primera línea prioritaria corresponde a la construcción de un corpus financiero multiclase anotado manualmente y validado mediante un protocolo de consenso. La validación humana realizada en este estudio mostró que las pseudoetiquetas generadas automáticamente no reproducen de manera uniforme el criterio humano. Por ello, futuras investigaciones deberían desarrollar un conjunto de datos en el que varios anotadores clasifiquen directamente los textos utilizando una guía previamente calibrada, con mecanismos para resolver desacuerdos y documentar los criterios de decisión. Este corpus podría funcionar como referencia independiente tanto para el entrenamiento como para la evaluación de nuevos modelos, reduciendo la circularidad derivada de utilizar pseudoetiquetas generadas mediante FinBERT base y reglas heurísticas como referencia

para el posterior fine-tuning. Esta línea también permitiría estudiar con mayor precisión las fronteras semánticas entre emociones cercanas y mejorar la consistencia de las cinco categorías utilizadas.

Una segunda línea consiste en fortalecer la representación de las emociones minoritarias, especialmente enojo, miedo y tristeza. En el presente estudio, enojo presentó el menor soporte y el desempeño más bajo entre las categorías evaluadas. Aunque se aplicaron *oversampling* y ponderación de la función de pérdida, estas técnicas no incorporan nueva diversidad lingüística cuando se basan únicamente en la repetición de observaciones existentes. En futuras investigaciones sería conveniente ampliar de forma deliberada las clases minoritarias mediante nuevos textos y anotación humana. También podría evaluarse el uso de *active learning*, de manera que el modelo identifique las observaciones de mayor incertidumbre y las priorice para revisión, generando progresivamente un conjunto de entrenamiento más informativo.

Una tercera línea de investigación consiste en ampliar y mejorar la representatividad temporal del corpus. Aunque el estudio comprende el periodo entre enero de 2022 y abril de 2026, la distribución de observaciones no fue uniforme y 2026 corresponde únicamente a un periodo parcial. Una investigación posterior podría implementar una estrategia de muestreo temporal estratificado por año y mes orientada a reducir concentraciones excesivas y obtener una cobertura más comparable entre periodos. Esto permitiría evaluar si las distribuciones emocionales y el desempeño del clasificador se mantienen relativamente estables o presentan variaciones asociadas con diferentes condiciones del mercado. Asimismo, podrían incorporarse análisis longitudinales para estudiar la evolución temporal de las distribuciones emocionales.

También resulta pertinente extender el análisis a contextos latinoamericanos y a otros idiomas relevantes para la región, particularmente español y portugués. El corpus actual está compuesto principalmente por textos financieros en inglés asociados con el mercado estadounidense, por lo que una extensión regional permitiría evaluar hasta qué punto el esquema de clasificación mantiene su desempeño cuando cambian el idioma, las convenciones discursivas y el contexto económico. Futuras investigaciones podrían incorporar fuentes de mercados como Colombia, México, Brasil, Chile, Perú y Argentina, utilizando corpus específicos para cada contexto y procedimientos de validación humana adaptados lingüísticamente.

Una quinta línea de trabajo sería comparar FinBERT con modelos de lenguaje instruccionales especializados en análisis financiero, incluyendo la línea de investigación asociada con FinSentGPT. Mientras que el presente trabajo se fundamenta en el fine-tuning de un modelo preentrenado especializado en lenguaje financiero, los modelos instruccionales permiten explorar esquemas alternativos de clasificación mediante instrucciones, *few-shot learning* o *zero-shot learning*. Una comparación controlada permitiría determinar si estos modelos ofrecen mejoras en la clasificación multiclase, particularmente en las categorías minoritarias, y analizar el costo computacional y la necesidad de datos anotados de cada estrategia. Para garantizar la validez de esta comparación, los modelos deberían evaluarse sobre el mismo conjunto de prueba independiente y mediante métricas como Macro-F1, F1 por clase, MCC y matrices de confusión.

La longitud máxima de entrada constituye una posible fuente de diferencias en la representación de los textos según su procedencia. La versión final del modelo procesó

como máximo 128 tokens por texto, mientras que el procedimiento de depuración admitió contenidos institucionales de hasta 400 tokens y contenidos retail de hasta 100 tokens. En consecuencia, la pérdida de contenido por truncamiento pudo afectar de manera desigual a ambos grupos. No obstante, dado que no se realizó una comparación experimental específica entre diferentes longitudes máximas de entrada, no es posible cuantificar el efecto de esta decisión sobre las métricas del modelo ni atribuirle directamente las diferencias entre las fuentes. Como trabajo futuro, se recomienda comparar configuraciones de 128, 256 y 512 tokens, manteniendo constantes la partición de datos, los demás hiperparámetros y las condiciones de evaluación. Esta comparación permitiría determinar si una mayor longitud de entrada mejora la clasificación y si el efecto difiere entre los textos retail e institucionales.

Finalmente, una línea de mayor alcance podría estudiar la relación temporal entre las distribuciones emocionales y variables cuantitativas del mercado, como retornos y volatilidad. Esta extensión no forma parte del alcance del estudio actual y debería plantearse como una investigación posterior independiente. Podrían construirse indicadores temporales de intensidad emocional y analizar su asociación con variables bursátiles, manteniendo una distinción estricta entre correlación, asociación temporal y causalidad.

Otra línea de trabajo consiste en evaluar de forma sistemática diferentes niveles de granularidad emocional bajo una misma referencia humana y condiciones de evaluación equivalentes. Los resultados de H2 no mostraron evidencia estadísticamente significativa de diferencia entre el clasificador binario independiente y las predicciones del modelo multiclase agrupadas en entusiasmo y no entusiasmo. Futuras investigaciones podrían

comparar formulaciones binarias, multiclase y eventualmente jerárquicas utilizando el mismo corpus anotado, las mismas particiones de entrenamiento y prueba y métricas comparables. Esto permitiría determinar en qué condiciones una mayor granularidad aporta información analítica adicional sin comprometer de manera sustancial el

En conjunto, estas líneas de trabajo permitirían avanzar desde el análisis comparativo desarrollado en esta investigación hacia un sistema más robusto, validado mediante referencias humanas independientes, con mayor diversidad semántica, mejor representatividad temporal y aplicabilidad a distintos idiomas y mercados. Asimismo, permitirían evaluar de manera controlada diferentes modelos, longitudes de entrada y niveles de granularidad, y explorar posteriormente asociaciones entre las distribuciones emocionales y variables cuantitativas del mercado, manteniendo una distinción rigurosa entre asociación y causalidad. De esta manera, el estudio establece una base metodológica para futuras investigaciones orientadas tanto a mejorar la calidad de la clasificación emocional como a ampliar su utilidad analítica en diferentes contextos financieros.

Referencias

Anufriev, M., Chernulich, A., y Tuinstra, J. (2022). Asset price volatility and investment horizons: An experimental investigation. *Journal of Economic Behavior & Organization, 193,* 19-48. https://doi.org/10.1016/j.jebo.2021.11.019

Araci, D. (2019). *FinBERT: Financial sentiment analysis with pre-trained language models* [Preprint]. arXiv. https://arxiv.org/abs/1908.10063

Ashtiani, M., y Raahemi, B. (2023). News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. *Expert Systems with Applications*, *217*, 119509. https://doi.org/10.1016/j.eswa.2023.119509

Audrino, F., Sigrist, F., y Ballinari, D. (2020). The impact of sentiment and attention measures on stock market volatility. *International Journal of Forecasting*, *36(2)*, 334-357. https://doi.org/10.1016/j.ijforecast.2019.05.010

Bollen, J., Mao, H., y Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2(1),* 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Buda, M., Maki, A., y Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Consoli, S., Barbaglia, L., y Manzan, S. (2022). Fine-grained, aspect-based sentiment analysis on economic and financial lexicon. *Knowledge-Based Systems*, *247*, 108781. https://doi.org/10.1016/j.knosys.2022.108781

Costola, M., Hinz, O., Nofer, M., y Pelizzo, L. (2023). Machine learning sentiment analysis, COVID-19 news and stock market reactions. *Research in International Business and Finance*, *64*, 101881. https://doi.org/10.1016/j.ribaf.2023.101881

Cui, J., Wei, Q., y Gao, X. (2025). How retail vs. institutional investor sentiment differ in affecting Chinese stock returns. *Journal of Risk and Financial Management*, *18*(2), 95. https://doi.org/10.3390/jrfm18020095

Daudert, T. (2021). Exploiting textual and relationship information for fine-grained financial sentiment analysis. *Knowledge-Based Systems*, *230*, 107389. https://doi.org/10.1016/j.knosys.2021.107389

Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 4171–4186. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Efron, B., y Tibshirani, R. J. (1993)**.** *An introduction to the bootstrap*. Chapman & Hall. https://doi.org/10.1201/9780429246593

García-Méndez, S., de Arriba-Pérez, F., Barros-Vila, A., y González-Castaño, F. J. (2023). Targeted aspect-based emotion analysis to detect opportunities and precaution in

financial Twitter messages. *Expert Systems with Applications, 218*, 119611. https://doi.org/10.1016/j.eswa.2023.119611

Ge, Y., Qiu, J., Liu, Z., Gu, W., y Xu, L. (2020). Beyond negative and positive: Exploring the effects of emotions in social media during the stock market crash. *Information Processing & Management*, *57(4)*, 102218. https://doi.org/10.1016/j.ipm.2020.102218

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry, 28*(5–6), 367–374. https://doi.org/10.1016/j.compbiolchem.2004.09.006

He, H., y Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Jiang, T., y Zeng, Q. (2023). *Financial sentiment analysis using FinBERT with application in predicting stock movement* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2306.02136

Kearney, C., y Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis, 33,* 171–185. https://doi.org/10.1016/j.irfa.2014.02.006

Kelly, S., y Ahmad, K. (2018). Estimating the impact of domain-specific news sentiment on financial assets. *Knowledge-Based Systems*, *150*, 116-126. https://doi.org/10.1016/j.knosys.2018.03.004

Kullback, S., y Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Lee, J., Youn, H. L., Poon, J., y Han, S. C. (2023). *StockEmotions: Discover investor emotions for financial sentiment analysis and multivariate time series* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2301.09279

Lengkeek, M., van der Knaap, F., y Frasincar, F. (2023). Leveraging hierarchical language models for aspect-based sentiment analysis on financial data. *Information Processing & Management*, *60(5)*, 103435. https://doi.org/10.1016/j.ipm.2023.103435

Li, S., Shi, W., Wang, J., y Zhou, H. (2021). A deep learning-based approach to constructing a domain sentiment lexicon: A case study in financial distress prediction. *Information Processing & Management, 58(5)*, 102673. https://doi.org/10.1016/j.ipm.2021.102673

Li, X., Xie, H., Chen, L., Wang, J., y Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems, 69,* 14–23. https://doi.org/10.1016/j.knosys.2014.04.022

Lin, W., y Liao, L.-C. (2024). Lexicon-based prompt for financial dimensional sentiment analysis. *Expert Systems with Applications, 244*, 122936. https://doi.org/10.1016/j.eswa.2023.122936

Loughran, T., y McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Mahdavi Ardekani, A., Bertz, J., Bryce, C., Dowling, M., y Long, S. (2024). FinSentGPT: A universal financial sentiment engine? *International Review of Financial Analysis, 94*, 103291. https://doi.org/10.1016/j.irfa.2024.103291

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153–157. https://doi.org/10.1007/BF02295996

Nassirtoussi, A., Aghabozorgi, S., Wah, T., y Ngo, D. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, *41*(16), 7653-7670. https://doi.org/10.1016/j.eswa.2014.06.009

Nguyen, T. H., Shirai, K., y Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications, 42*(24), 9603–9611. https://doi.org/10.1016/j.eswa.2015.07.052

Nisar, T. M., y Yeung, M. (2018). Twitter as a tool for forecasting stock market movements: A short-window event study. *The Journal of Finance and Data Science*, *4*(2), 101-119. https://doi.org/10.1016/j.jfds.2017.11.002

Nyakurukwa, K., y Seetharam, Y. (2023). The evolution of studies on social media sentiment in the stock market: Insights from bibliometric analysis. *Scientific African*, *20*, e01596. https://doi.org/10.1016/j.sciaf.2023.e01596

Oliveira, N., Cortez, P., y Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, *73*, 125-144. https://doi.org/10.1016/j.eswa.2016.12.036

Peng, B., Chersoni, E., Hsu, Y., Qiu, L., y Huang, C. (2024). Supervised Cross-Momentum Contrast: Aligning representations with prototypical examples to enhance financial sentiment analysis. *Knowledge-Based Systems*, *295*, 111683. https://doi.org/10.1016/j.knosys.2024.111683

Rodríguez-Ibáñez, M., Casánez-Ventura, A., Castejón-Mateos, F., y Cuenca-Jiménez, P.-M. (2023). A review on sentiment analysis from social media platforms. *Expert Systems with Applications,* *223*, 119862. https://doi.org/10.1016/j.eswa.2023.119862

Ruan, Y., Durresi, A., y Alfantoukh, L. (2018). Using Twitter trust network for stock market analysis. *Knowledge-Based Systems, 145,* 207–218. https://doi.org/10.1016/j.knosys.2018.01.016

Sheetal, R., y Aithal, P. K. (2025). Enhancing financial sentiment analysis: Integrating the LoughranMcDonald dictionary with BERT for advanced market predictive insights. *Procedia Computer Science,* *258*, 2244–2257. https://doi.org/10.1016/j.procs.2025.04.477

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives*, *17*(1), 83–104. https://doi.org/10.1257/089533003321164967

Sokolova, M., y Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Sun, Y., Yuan, H., y Xu, F. (2025). Financial sentiment analysis for pre-trained language models incorporating dictionary knowledge and neutral features. *Natural Language Processing Journal*, *11*, 100148. https://doi.org/10.1016/j.nlp.2025.100148

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., y Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://arxiv.org/abs/1706.03762

Wang, W., y Duxbury, D. (2021). Institutional investor sentiment and the mean-variance relationship: Global evidence. *Journal of Economic Behavior & Organization, 191,* 415-441. https://doi.org/10.1016/j.jebo.2021.08.029

Wilksch, M., y Abramova, O. (2023). PyFin-sentiment: Towards a machine-learning-based model for deriving sentiment from financial tweets. *International Journal of Information Management Data Insights*, *3(1)*, 100171. https://doi.org/10.1016/j.jjimei.2023.100171

Xiang, C., Zhang, J., Li, F., Fei, H., y Ji, D. (2022). A semantic and syntactic enhanced neural model for financial sentiment analysis. *Information Processing & Management*, *59(4)*, 102943. https://doi.org/10.1016/j.ipm.2022.102943

Anexos

Anexo A. Distribución de textos por año, fuente, tipo y activo del corpus final

Tabla 20 *Distribución de textos por año, fuente, tipo y activo del corpus final*

año	fuente	tipo	activo	F	cantidad
2022	financial_news	Institucional	S&P 500	1	291
2022	financial_news	Institucional	TSLA	1	6
2022	financial_news_tsla	Institucional	TSLA	1	247
2022	stocktwits_retail	Retail	S&P 500	0	330
2022	stocktwits_retail	Retail	TSLA	0	2
2022	stocktwits_tsla	Retail	TSLA	0	351
2022	twitter_financial	Retail	S&P 500	0	3
2022	twitter_financial_tsla	Retail	TSLA	0	1
2023	financial_news	Institucional	S&P 500	1	499
2023	financial_news	Institucional	TSLA	1	7
2023	financial_news_tsla	Institucional	TSLA	1	416
2023	stocktwits_retail	Retail	S&P 500	0	400
2023	stocktwits_retail	Retail	TSLA	0	3
2023	stocktwits_tsla	Retail	TSLA	0	419
2023	twitter_financial	Retail	S&P 500	0	2
2023	twitter_financial_tsla	Retail	TSLA	0	1
2024	financial_news	Institucional	S&P 500	1	549
2024	financial_news	Institucional	TSLA	1	4
2024	financial_news_tsla	Institucional	TSLA	1	318
2024	stocktwits_retail	Retail	S&P 500	0	305
2024	stocktwits_retail	Retail	TSLA	0	4
2024	stocktwits_tsla	Retail	TSLA	0	401
2024	twitter_financial	Retail	S&P 500	0	2
2024	twitter_financial_tsla	Retail	TSLA	0	0
2025	financial_news	Institucional	S&P 500	1	407
2025	financial_news	Institucional	TSLA	1	7
2025	financial_news_tsla	Institucional	TSLA	1	514
2025	stocktwits_retail	Retail	S&P 500	0	286
2025	stocktwits_retail	Retail	TSLA	0	3
2025	stocktwits_tsla	Retail	TSLA	0	498
2025	twitter_financial	Retail	S&P 500	0	2
2025	twitter_financial_tsla	Retail	TSLA	0	1
2026	financial_news	Institucional	S&P 500	1	254
2026	financial_news	Institucional	TSLA	1	3
2026	financial_news_tsla	Institucional	TSLA	1	478

2026	stocktwits_retail	Retail	S&P 500	0	422
2026	stocktwits_retail	Retail	TSLA	0	3
2026	stocktwits_tsla	Retail	TSLA	0	312
2026	twitter_financial	Retail	S&P 500	0	1
2026	twitter_financial_tsla	Retail	TSLA	0	0
Total					**7.752**

Nota. La tabla presenta la cantidad de textos disponibles en el corpus, desagregada por año, fuente, tipo y activo. La suma de las categorías corresponde a 7.752 textos.

Anexo B. Distribución de eliminaciones por fuente

Tabla 21 *Distribución de eliminaciones por fuente*

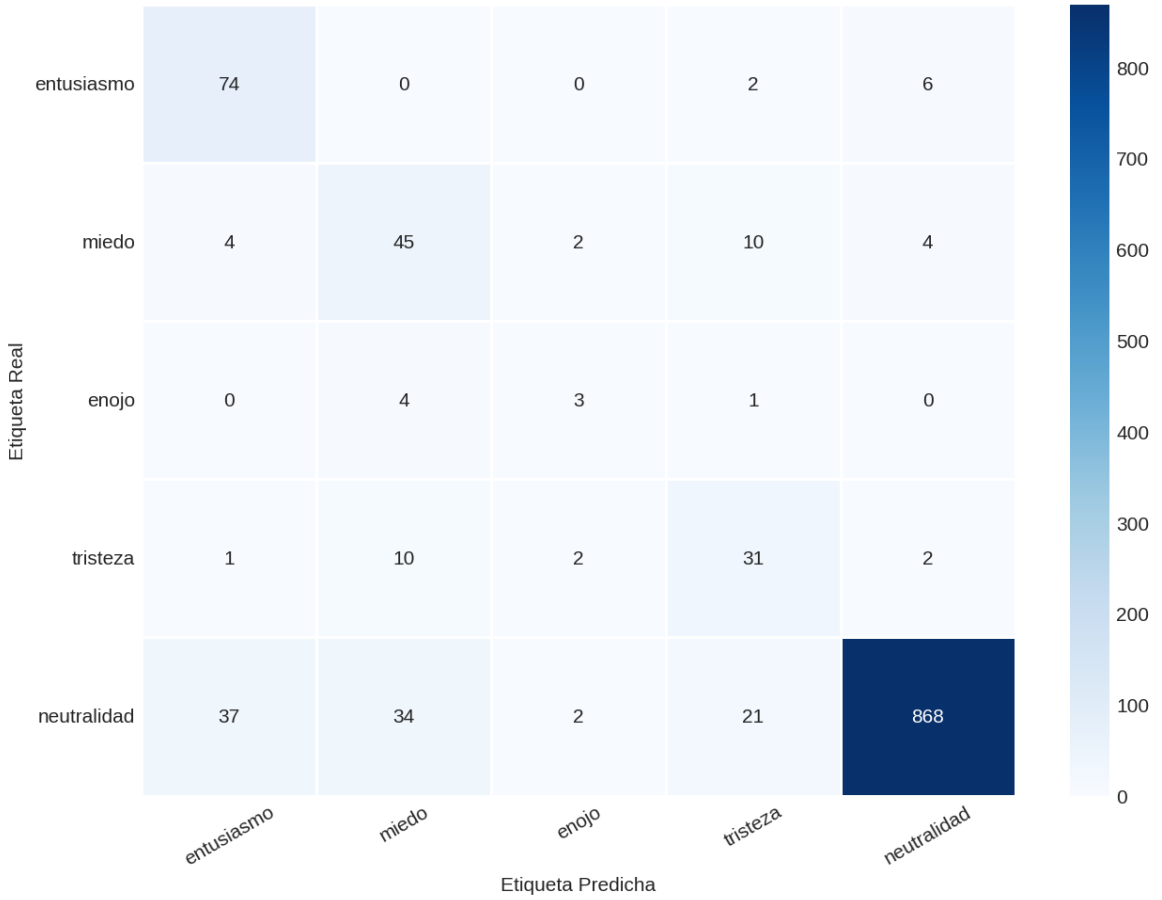
Fuente	n_inicial	n_final	eliminado	%eliminado
stocktwits_retail	2.487	1.758	729	29,3%
stocktwits_tsla	2.478	1.981	497	20,1%
twitter_financial	29	10	19	65,5%
twitter_financial_tsla	6	3	3	50,0%
financial_news	2.528	2.027	501	19,8%
financial_news_tsla	2.472	1.973	499	20,2%
Total	**10.000**	**7.752**	**2.248**	**22,5%**

Nota. La tabla presenta la cantidad de textos iniciales, preprocesados (eliminados) y los finales en el corpus desagregada por fuente.

En la tabla se puede observar la distribución de 10.000 textos iniciales, luego del preprocesamiento registra 7.752 textos, debido a que se eliminaron 2.248 (22,5%) registros de los textos iniciales del corpus.

Anexo C. Matriz de confusión del modelo FinBERT fine-tuned v2

Figura 3 *Matriz de confusión del modelo FinBERT fine-tuned v2*



Nota. La matriz de confusión corresponde a la evaluación final del modelo FinBERT fine-tuned v2 sobre el conjunto de prueba independiente de 1.163 textos. Las filas representan las pseudoetiquetas de referencia y las columnas las categorías predichas por el modelo. Los valores de la diagonal principal corresponden a clasificaciones correctas y los valores fuera de la diagonal representan errores de clasificación. Esta matriz debe interpretarse conjuntamente con las métricas globales de la evaluación final: Macro-F1 = 0,5564, Micro-F1 = 0,8813, Weighted-F1 = 0,8901, MCC = 0,6705 y Accuracy = 0,8813.

La matriz de confusión de la evaluación final complementa las métricas globales al permitir identificar la distribución de los aciertos y errores entre las cinco categorías emocionales. Su interpretación debe realizarse conjuntamente con las métricas por clase presentadas en la Tabla 15.

Los resultados por categoría muestran un desempeño desigual. Neutralidad obtuvo el mayor F1, con 0,95, seguida de entusiasmo, con 0,75; miedo, con 0,59; tristeza, con 0,50; y enojo, con 0,00. La categoría enojo presentó únicamente ocho observaciones en el conjunto de prueba y no registró identificaciones correctas en esta evaluación, por lo que sus resultados deben interpretarse considerando el reducido soporte disponible.

Estas diferencias refuerzan la necesidad de complementar las métricas globales con indicadores por categoría, particularmente en presencia de un marcado desbalance entre clases. Asimismo, dado que las etiquetas utilizadas como referencia proceden del proceso de pseudoetiquetado descrito en la metodología, los errores de clasificación representan desacuerdos respecto de dichas pseudoetiquetas y no necesariamente desacuerdos frente a emociones humanas verificadas.

Anexo D. Guía de anotación

Validación humana de las pseudoetiquetas

La validación humana se realizó con el propósito de evaluar la correspondencia entre las pseudoetiquetas utilizadas para el ajuste fino de FinBERT y el criterio de anotadores humanos independientes. Esta etapa permitió examinar la calidad de las etiquetas generadas automáticamente y, al mismo tiempo, reconocer explícitamente una posible limitación metodológica asociada con la circularidad del procedimiento de entrenamiento, dado que las señales iniciales utilizadas para construir las pseudoetiquetas procedieron del modelo FinBERT base.

En consecuencia, la validación humana no se planteó como un procedimiento para sustituir retrospectivamente las pseudoetiquetas utilizadas durante el entrenamiento, sino como una evaluación independiente de su correspondencia con criterios humanos aplicados a una muestra del corpus.

Diseño de la validación humana

La validación humana se realizó sobre una muestra de 301 textos del corpus final, equivalente aproximadamente al 3,88% de los 7.752 registros. Para cada observación se dispuso de la pseudoetiqueta generada mediante el procedimiento automático y de las anotaciones independientes de los dos autores del trabajo de grado. La clasificación humana se efectuó sin utilizar la pseudoetiqueta automática ni la decisión del otro anotador como criterio para asignar la emoción predominante. Esta condición buscó reducir la influencia directa de las etiquetas automáticas y preservar la independencia entre las decisiones humanas.

Las categorías consideradas fueron las mismas utilizadas durante la clasificación multiclase: entusiasmo, miedo, enojo, tristeza y neutralidad. No se realizaron imputaciones ni se identificaron valores faltantes en las tres variables categóricas consideradas para la comparación.

La validación humana debe interpretarse considerando dos limitaciones adicionales. En primer lugar, los dos anotadores fueron los propios autores de la investigación, por lo que sus decisiones, aunque independientes entre sí durante la clasificación, pueden compartir criterios interpretativos derivados de su formación y del diseño de la guía de anotación. En segundo lugar, el acuerdo observado del 47,18% y el coeficiente kappa de Cohen de 0,2752 evidencian que la muestra no proporciona una referencia humana de alta concordancia. Por tanto, esta validación permite identificar discrepancias y dimensionar las limitaciones de las pseudoetiquetas, pero no demuestra por sí sola la validez semántica definitiva de las cinco categorías.

Tabla 22 *Distribución de las pseudoetiquetas en la muestra de validación*

Categoría	Pseudoetiquetas en la muestra
Entusiasmo	82
Miedo	65
Enojo	8
Tristeza	46
Neutralidad	100
Total	301

Nota. La tabla presenta la distribución de las cinco pseudoetiquetas automáticas en la muestra seleccionada para la validación humana. Estas categorías corresponden a las

etiquetas utilizadas como referencia durante el ajuste del modelo y no deben interpretarse como anotaciones humanas de verdad de terreno.

Guía de anotación

Para favorecer una aplicación consistente de las categorías emocionales, se estableció una guía de anotación con definiciones operacionales y criterios orientativos para las cinco emociones. La unidad de análisis correspondió al texto completo dentro de su contexto financiero. Cada anotador debía identificar la emoción predominante y asignar una sola categoría por observación.

Los criterios se utilizaron como orientación para la toma de decisiones, sin asumir que las categorías fueran necesariamente excluyentes desde una perspectiva psicológica. En aquellos textos que contenían simultáneamente referencias a riesgos, pérdidas, expectativas, resultados o valoraciones, la clasificación debía corresponder a la emoción considerada predominante según el contenido global del texto.

Tabla 23 *Definiciones operacionales para la anotación de las cinco emociones*

Emoción	Definición operacional	Criterios orientativos
Entusiasmo	Expresión de optimismo, confianza, expectativa favorable o valoración positiva frente a una empresa, activo o mercado.	Resultados favorables, perspectivas de crecimiento, optimismo del mercado, confianza del inversionista.
Miedo	Expresión de preocupación, incertidumbre, riesgo o temor frente a pérdidas o acontecimientos adversos.	Advertencias, riesgos, posibles caídas, incertidumbre elevada, temor de los inversionistas.

Enojo	Expresión de indignación, molestia, rechazo o crítica intensa frente a una situación, empresa, mercado o decisión.	Quejas, indignación, acusaciones, rechazo fuerte o lenguaje marcadamente crítico.
Tristeza	Expresión de decepción, pesimismo, pérdida o deterioro de expectativas.	Resultados decepcionantes, pérdidas, deterioro, desilusión o pesimismo sin predominio de indignación.
Neutralidad	Contenido principalmente informativo o descriptivo sin una carga emocional predominante.	Datos, hechos, reportes, precios, resultados o acontecimientos presentados de forma predominantemente objetiva.

Nota. Las definiciones se utilizaron como criterios operacionales para favorecer la consistencia de la anotación independiente. Los anotadores no tuvieron acceso a la pseudoetiqueta como criterio para decidir la categoría asignada.

Resultados de la validación

La concordancia entre los dos anotadores humanos se evaluó mediante el porcentaje de acuerdo observado y el coeficiente kappa de Cohen. Este coeficiente permite estimar la concordancia entre dos evaluadores en variables categóricas, ajustando el acuerdo observado por aquel que podría esperarse por azar (Cohen, 1960).

Sobre los 301 textos evaluados, los dos anotadores coincidieron exactamente en 142 casos, equivalentes al 47,18% de acuerdo observado. El coeficiente kappa de Cohen fue $\kappa = 0{,}2752$, lo que indica que el acuerdo observado fue superior al esperado por azar, aunque la concordancia corregida por azar entre ambos evaluadores fue limitada.

La comparación individual evidenció una diferencia en la correspondencia de las anotaciones humanas con las pseudoetiquetas. El Anotador 1 coincidió con las etiquetas automáticas en el 35,22% de los casos (κ = 0,1375), mientras que el Anotador 2 coincidió en el 78,74% (κ = 0,7131). Esta diferencia indica que la relación entre el criterio humano y el etiquetador automático no fue uniforme. Sin embargo, estos indicadores no permiten concluir que uno de los anotadores sea más preciso que el otro, dado que las pseudoetiquetas no constituyen una referencia humana validada. Además, el acuerdo observado entre los dos anotadores fue del 47,18%, con κ = 0,2752, lo que evidencia una concordancia limitada entre sus decisiones. Por tanto, los resultados deben interpretarse como evidencia de discrepancias entre criterios de clasificación y no como una demostración definitiva de la validez semántica de las etiquetas.

Tabla 24 *Resultados de la validación humana y comparación con las pseudoetiquetas*

Métrica	Resultado
Textos evaluados	301
Acuerdos Anotador 1 vs. Anotador 2	142 (47,18%)
Kappa de Cohen	κ = 0,2752
Acuerdo Anotador 1 vs. pseudoetiquetas	106 (35,22%)
Acuerdo Anotador 2 vs. pseudoetiquetas	237 (78,74%)
Kappa Anotador 1 vs. pseudoetiquetas	κ = 0,1375
Kappa Anotador 2 vs. pseudoetiquetas	κ = 0,7131

Nota. κ = coeficiente kappa de Cohen. El porcentaje de acuerdo corresponde al número de coincidencias dividido entre el total de textos evaluados. Los valores de concordancia con las pseudoetiquetas expresan correspondencia con el etiquetador automático y no deben interpretarse como una medida directa de corrección del anotador humano.

La validación humana fue realizada por los dos autores de esta investigación, quienes clasificaron los textos de manera independiente siguiendo la guía de anotación definida para el estudio. Por tanto, aunque el procedimiento permitió contrastar las pseudoetiquetas con criterios humanos externos al etiquetador automático, no constituye una evaluación completamente independiente de los investigadores. Esta condición se reconoce como una limitación metodológica y debe considerarse al interpretar los resultados de concordancia.

Tabla 25 *Matriz de acuerdo entre los dos anotadores humanos*

Anotador 1 \ Anotador 2	**Entusiasmo**	**Miedo**	**Enojo**	**Tristeza**	**Neutralidad**
Entusiasmo	27	5	0	1	10
Miedo	6	28	0	19	16
Enojo	3	6	9	8	12
Tristeza	1	4	2	3	6
Neutralidad	30	23	2	5	75

Nota. Las filas corresponden al Anotador 1 y las columnas al Anotador 2. Los valores representan frecuencias absolutas. La suma de la diagonal principal es 142, correspondiente al número total de acuerdos entre ambos evaluadores.

La matriz muestra que las mayores cantidades absolutas de coincidencias se presentaron en neutralidad (75 casos), miedo (28) y entusiasmo (27). En tristeza y enojo se observaron menos coincidencias absolutas y una mayor dispersión de las decisiones hacia otras categorías. Este patrón sugiere la conveniencia de reforzar los criterios operacionales utilizados para distinguir especialmente miedo, tristeza y enojo en futuras etapas de anotación.

Comparación entre anotadores humanos y pseudoetiquetas

La comparación entre las anotaciones humanas y las pseudoetiquetas mostró una diferencia marcada entre los dos evaluadores. El Anotador 1 coincidió con las pseudoetiquetas en 106 de 301 textos (35,22%) y obtuvo $\kappa = 0,1375$, mientras que el Anotador 2 coincidió en 237 de 301 textos (78,74%) y obtuvo $\kappa = 0,7131$.

Esta asimetría indica que el criterio del segundo anotador estuvo considerablemente más alineado con el procedimiento automático de generación de pseudoetiquetas que el criterio del primero. Sin embargo, esta diferencia no debe interpretarse como evidencia de que el segundo anotador sea más correcto. Los valores únicamente expresan el grado de correspondencia entre cada criterio humano y las etiquetas producidas automáticamente.

La diferencia entre ambos anotadores también impide asumir que las pseudoetiquetas representan de manera uniforme un criterio humano independiente. En consecuencia, estas etiquetas deben entenderse como referencias automáticas utilizadas para el entrenamiento y no como equivalentes a una verdad de terreno humana.

Interpretación y discusión de los resultados

Los resultados muestran que la validación humana no produjo una concordancia elevada entre los dos anotadores. El acuerdo observado fue del 47,18% y el coeficiente kappa de Cohen fue 0,2752, evidenciando una correspondencia limitada entre ambos criterios una vez considerado el acuerdo esperado por azar.

Una posible explicación es que determinados textos financieros contengan simultáneamente información relacionada con resultados, riesgo, pérdidas, expectativas o

valoraciones, lo que puede dificultar la identificación de una única emoción predominante. En particular, las fronteras entre miedo, tristeza y enojo pueden depender del contexto discursivo y de la interpretación realizada por cada anotador.

La diferencia observada entre las concordancias individuales con las pseudoetiquetas constituye también un hallazgo metodológico relevante. El acuerdo de 78,74% del segundo anotador con las pseudoetiquetas, acompañado de $\kappa = 0{,}7131$, contrasta con el 35,22% y $\kappa = 0{,}1375$ obtenidos para el primero. Esta diferencia indica distintos grados de alineación con el procedimiento automático, pero no permite establecer cuál de los dos criterios humanos representa por sí solo una referencia definitiva.

En consecuencia, los resultados de esta validación muestran que las pseudoetiquetas no deben considerarse equivalentes a etiquetas humanas de referencia. La muestra evaluada funciona como un mecanismo independiente de control de calidad y evidencia la conveniencia de ampliar la anotación humana, fortalecer la guía de clasificación y realizar procesos de calibración y consenso antes de construir un conjunto humano destinado a servir como referencia definitiva.

Circularidad del procedimiento como limitación metodológica

El procedimiento de generación de etiquetas presenta una posible circularidad metodológica. En una primera etapa, FinBERT base produjo las señales de polaridad utilizadas para clasificar los textos como positivos, negativos o neutrales. Posteriormente, estas señales fueron transformadas en cinco categorías emocionales mediante el

esquema de mapeo y las reglas heurísticas establecidas en la investigación. Las pseudoetiquetas resultantes se utilizaron después para realizar el fine-tuning de FinBERT.

En consecuencia, el modelo ajustado aprende patrones derivados, al menos parcialmente, del mismo sistema automático que intervino en la generación de las etiquetas utilizadas durante su entrenamiento. Por esta razón, un desempeño elevado frente a las pseudoetiquetas no constituye por sí solo evidencia de una elevada validez semántica de las cinco emociones. El modelo puede reproducir de manera eficaz patrones del etiquetador inicial, incluyendo posibles errores o ambigüedades procedentes de FinBERT base o de las reglas utilizadas para transformar polaridades en emociones.

La validación humana incorporada en esta investigación mitiga parcialmente esta limitación al introducir una fuente de evaluación independiente del procedimiento automático. Sin embargo, la concordancia entre los dos anotadores fue limitada y la correspondencia individual con las pseudoetiquetas presentó diferencias importantes. En consecuencia, la circularidad debe mantenerse explícitamente como una limitación metodológica del estudio y el corpus pseudoetiquetado no debe presentarse como equivalente a un corpus completamente anotado por humanos.

Implicaciones para trabajos posteriores

Como extensión de esta investigación, se recomienda construir un subconjunto de referencia de mayor tamaño mediante anotación humana independiente, calibración previa y procedimiento de consenso. Una estrategia futura podría incluir una fase inicial de

entrenamiento de los anotadores, revisión conjunta de ejemplos ambiguos, ajustes de la guía operacional y una nueva ronda de anotación independiente.

Cuando existan desacuerdos, estos podrían resolverse mediante consenso o mediante la intervención de un tercer evaluador. El conjunto resultante podría utilizarse como referencia humana independiente para evaluar el modelo *fine-tuned* y estimar métricas de desempeño sin depender exclusivamente de las pseudoetiquetas utilizadas durante el entrenamiento.

Anexo E. Procedencia de los datos, evidencia de ejecución y trazabilidad metodológica

Procedencia de los datos

El código de la primera fase consulta conjuntos de datos disponibles en Hugging Face, entre ellos zeroshot/twitter-financial-news-sentiment, StephanAkkerman/stock-market-tweets-data, ashraq/financial-news-articles y BeIR/fiqa. La ejecución revisada confirma que estas cuatro fuentes se cargaron, mientras que otras fuentes inicialmente contempladas presentaron errores de carga. Por esta razón, se documentarán las fuentes efectivamente utilizadas, sus identificadores, los campos extraídos y los criterios de selección y depuración.

Software y configuración computacional

El código utiliza Python y bibliotecas como pandas, NumPy, scikit-learn, NLTK, Transformers, PyTorch y Datasets, además del entorno de ejecución de Google Colab y Google Drive para el almacenamiento de archivos. Se documentarán las versiones efectivamente utilizadas y los parámetros de procesamiento y entrenamiento. Para el ajuste fino se empleó como modelo base ProsusAI/finbert, con cinco categorías emocionales, una longitud máxima de 512 tokens, una tasa de aprendizaje de 2×10^{-5} y un tamaño de lote de 16 para la versión 1. Para la versión 2 se estableció una longitud máxima de 128 tokens, una tasa de aprendizaje 1e-5, un tamaño de lote de 32, un warmup ratio de 0,10 y un máximo de ocho épocas. El procedimiento contempla ponderación de clases y evaluación mediante Macro-F1 y MCC.

Construcción de las categorías emocionales

Se utilizó ProsusAI/finbert para generar pseudoetiquetas a partir de las categorías positivas, neutrales y negativas. Las predicciones positivas se asociaron con entusiasmo y las neutrales con neutralidad. Las predicciones negativas se subdividieron en miedo, enojo y tristeza mediante listas de términos léxicos definidas en el código. Para miedo se utilizaron: fear, risk, crash, volatile, panic, warning, concern, uncertain, worry, danger, threat, anxiety. Para enojo: outrage, ridiculous, fraud, manipulation, corrupt, scam, cheat, betray, fake, lie, disgrace, angry. Para tristeza: loss, disappoint, fail, regret, sad, unfortunate, decline, disaster, trouble, suffer, poor, down. La categoría se asignó según el mayor número de coincidencias y, cuando no se encontraron coincidencias, se asignó miedo por defecto. Por tanto, las cinco categorías se obtuvieron mediante seudoe-tiquetado automático y no representan, por sí solas, emociones verificadas para la totalidad de los textos mediante anotación humana independiente.

Listas léxicas y reglas de asignación de pseudoetiquetas

La construcción de las cinco categorías emocionales se realizó a partir de las predicciones iniciales de ProsusAI/finbert y de reglas léxicas aplicadas a los textos clasificados inicialmente como negativos. Las listas utilizadas fueron las siguientes:

- Miedo: fear, risk, crash, volatile, panic, warning, concern, uncertain, worry, danger, threat, anxiety.

- Enojo: outrage, ridiculous, fraud, manipulation, corrupt, scam, cheat, betray, fake, lie, disgrace, angry.

- Tristeza: loss, disappoint, fail, regret, sad, unfortunate, decline, disaster, trouble, suffer, poor, down.

La categoría se asignó según el mayor número de coincidencias léxicas. Cuando no se encontraron coincidencias en ninguna de las listas, se asignó miedo como categoría predeterminada. Esta regla permite completar la clasificación automática, pero puede incrementar artificialmente la frecuencia de miedo y afectar la distribución de las pseudoetiquetas. En consecuencia, los resultados de las categorías miedo, enojo y tristeza deben interpretarse considerando que su asignación depende parcialmente de estas reglas y no exclusivamente del contenido contextual de cada texto.

Para facilitar la reproducibilidad del procedimiento, se conserva la correspondencia entre las listas léxicas, las reglas implementadas en el código y las pseudoetiquetas generadas. Las listas anteriores deben coincidir exactamente con la versión del código que produjo el corpus analizado.

Evidencias de ejecución y trazabilidad

Con el propósito de documentar el procedimiento realizado, este anexo reúne las evidencias disponibles de las fuentes de datos, el entorno de trabajo, las salidas de procesamiento y los resultados del análisis. Las evidencias se identifican mediante numeración consecutiva y una descripción de su contenido.

Las tablas de trazabilidad relacionan cada fase con su objetivo, los archivos de entrada y salida conocidos y el resultado documentado. Las cantidades de registros se reportan únicamente cuando existe soporte en los registros de ejecución o en los archivos de resultados conservados. Cuando no se dispone de una salida original que permita verificar una cifra, esta circunstancia se señala expresamente.

Las capturas de pantalla se presentan como documentación complementaria y no como sustituto del código fuente ni como demostración suficiente de la reproducibilidad integral del procedimiento.

Figura 4 *Fase 1 - Recolección del corpus (Mínimo 5.000 textos por fuente, ene 2022 – abr 2026, S&P 500 y TSLA)*

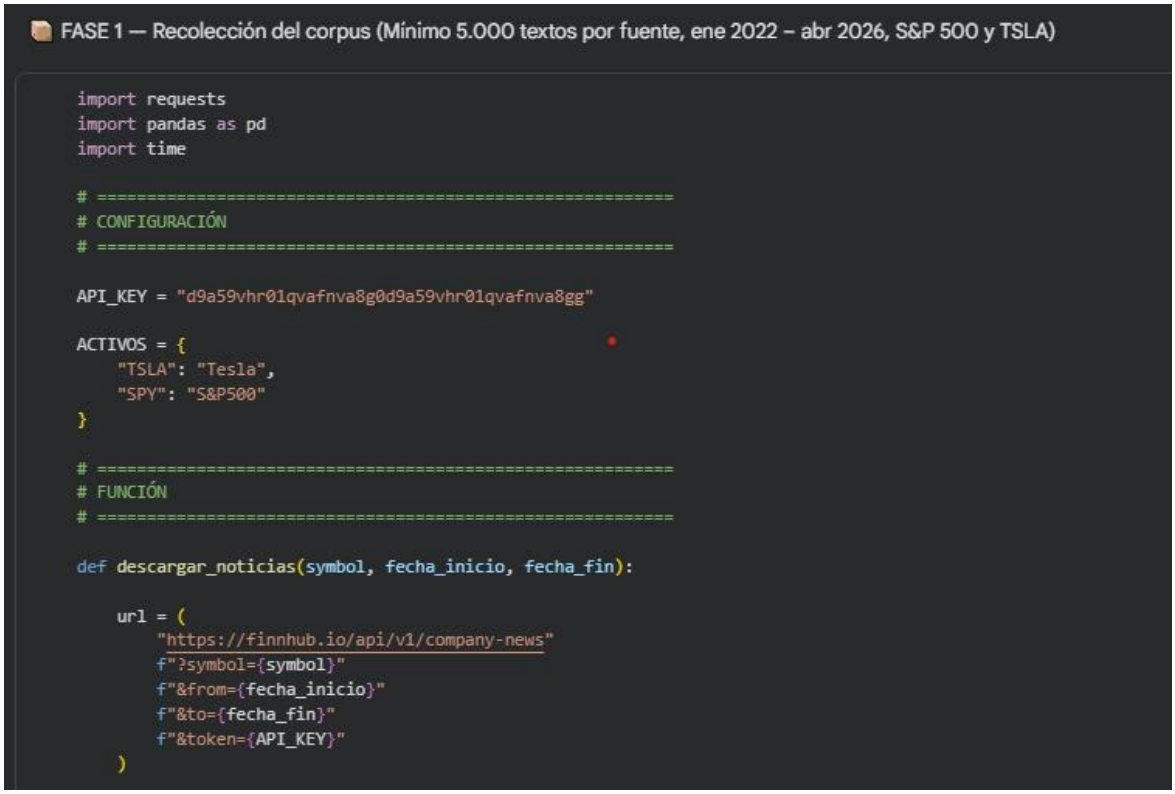


Figura 5 *Fase 1 - Consolidando corpus final*

```
================================================
CONSOLIDANDO CORPUS FINAL
================================================
Datasets que cargaron exitosamente: 4
Total bruto: 1299484
Tras limpieza de nulos: 1299427
Con 10+ tokens: 1222578
Duplicados eliminados: 240743
Tras deduplicacion: 981835

Retail (F=0) disponible: 677852
Institucional (F=1) disponible: 303983
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

================================================
FASE 1 COMPLETADA
================================================
Total textos finales: 10000
Guardado en: /content/drive/MyDrive/corpus_raw.csv

Distribucion F=0 (retail) vs F=1 (institucional):
F
0    5000
1    5000
```

Figura 6 *Fase 2 – Preprocesamiento y depuración*

FASE 2 — Preprocesamiento y depuración

```
# ============================================================
# FASE 2 — Preprocesamiento y depuracion
# ============================================================
!pip install nltk pandas scikit-learn -q

import pandas as pd
import re
import hashlib
import nltk
nltk.download('stopwords', quiet=True)
nltk.download('wordnet', quiet=True)
nltk.download('omw-1.4', quiet=True)
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import numpy as np

# Cargar corpus
from google.colab import drive
drive.mount('/content/drive', force_remount=False)
df = pd.read_csv('/content/drive/MyDrive/corpus_raw.csv')
print("Corpus cargado: " + str(len(df)) + " textos")
```

Figura 7 *Fase 2 – Validación por fuente y activo*

```
VALIDACIÓN POR FUENTE Y ACTIVO
                fuente  activo  cantidad  total_fuente  coincide
     financial_news S&P 500      2000          2000      True
     financial_news    TSLA        27            27      True
  financial_news_tsla    TSLA      1973          1973      True
    stocktwits_retail S&P 500      1743          1743      True
    stocktwits_retail    TSLA        15            15      True
      stocktwits_tsla    TSLA      1981          1981      True
    twitter_financial S&P 500        10            10      True
twitter_financial_tsla    TSLA         3             3      True

Total de registros distribuidos: 7752
```

Figura 8 *Fase 3 – Representación semántica y etiquetado con FinBERT*

```
# ============================================================
# FASE 3 — Representacion semantica y etiquetado con FinBERT
# Lee: corpus_preprocessed.csv (generado en Fase 2)
# Genera: corpus_labeled.csv + split_train/val/test.csv
# ============================================================
!pip install transformers torch scikit-learn -q

import os
import pandas as pd
import numpy as np
import torch
from transformers import AutoTokenizer, pipeline
from sklearn.model_selection import train_test_split
from google.colab import drive

drive.mount('/content/drive', force_remount=False)
BASE = '/content/drive/MyDrive/'

df = pd.read_csv(BASE + 'corpus_preprocessed.csv')
print("Corpus cargado: " + str(len(df)) + " textos")

if 'text_clean' in df.columns:
    df['input_text'] = df['text_clean'].fillna(df['text']).astype(str)
else:
    df['input_text'] = df['text'].astype(str)
```

Figura 9 *Fase 3 – Distribución de emociones*

```
Distribucion de emociones:
  neutralidad   | ######################################## 82.7% (6412)
  entusiasmo    | ### 7.2% (558)
  miedo         | ## 5.6% (431)
  tristeza      | # 3.8% (293)
  enojo         |  0.8% (59)

Por asset:
asset     emotion    n
SP500       enojo   32
SP500  entusiasmo  366
SP500       miedo  244
SP500 neutralidad 2952
SP500    tristeza  159
 TSLA       enojo   27
 TSLA  entusiasmo  192
 TSLA       miedo  187
 TSLA neutralidad 3460
 TSLA    tristeza  134
```

Figura 10 *Fase 4 – Fine-tuning de FinBERT*

```
# ============================================================
# FASE 4 — Fine-tuning de FinBERT
# AUTOCONTENIDA: conecta Drive, carga splits Fase 3, guarda modelo
# Lee: split_train.csv + split_val.csv
# Guarda: modelo fine-tuned en Drive
# ============================================================
!pip install transformers torch accelerate scikit-learn pandas -q

import os
import gc
import numpy as np
import pandas as pd
import torch
from google.colab import drive
from torch.nn import CrossEntropyLoss
from transformers import (
    AutoTokenizer,
    AutoModelForSequenceClassification,
    TrainingArguments,
    Trainer,
    EarlyStoppingCallback
)
from datasets import Dataset
from sklearn.utils.class_weight import compute_class_weight
from sklearn.metrics import f1_score, matthews_corrcoef

# ============================================================
# CONEXION A DRIVE
# ============================================================
drive.mount('/content/drive')
BASE = '/content/drive/MyDrive/Proyecto de Grado/fase3/'
MODEL_OUT  = BASE + 'finbert_finetuned/'
MODEL_NAME = "ProsusAI/finbert"

# Verificar archivos de Fase 3
for f in ['split_train.csv', 'split_val.csv']:
    ruta = BASE + f
    if not os.path.exists(ruta):
        raise FileNotFoundError("No se encuentra " + f + ". Ejecuta Fase 3 primero.")
```

Figura 11 *Fase 4 – Evaluación sobre set de validación*

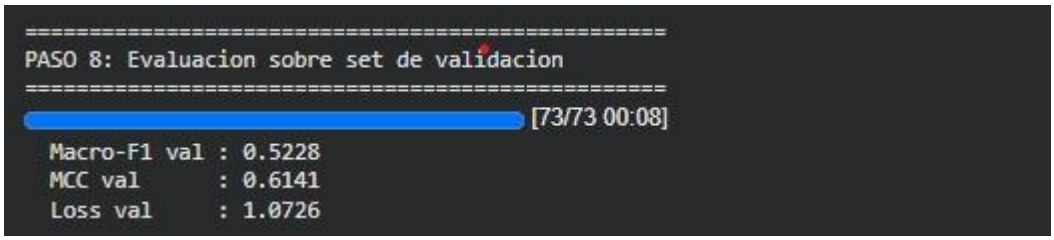


Figura 12 *Fase 4 – v2 – Reentrenamiento con oversampling*

```
# ============================================================
# FASE 4B — Reentrenamiento con oversampling
# AUTOCONTENIDA — lee desde ruta correcta del Drive
# ============================================================
!pip install transformers torch accelerate scikit-learn imbalanced-learn pandas datasets -q

import os
import gc
import numpy as np
import pandas as pd
import torch
from google.colab import drive
from torch.nn import CrossEntropyLoss
from transformers import (
    AutoTokenizer,
    AutoModelForSequenceClassification,
    TrainingArguments,
    Trainer,
    EarlyStoppingCallback
)
from datasets import Dataset
from sklearn.utils.class_weight import compute_class_weight
from sklearn.metrics import f1_score, matthews_corrcoef

drive.mount('/content/drive')

# ============================================================
# RUTA CORRECTA
# ============================================================
BASE      = '/content/drive/MyDrive/Proyecto de Grado/fase3/'
MODEL_OUT = '/content/drive/MyDrive/Proyecto de Grado/fase3/finbert_finetuned_v2/'
MODEL_NAME = "ProsusAI/finbert"
```

Figura 13 *Fase 4 – v2 – finbert fintuned*

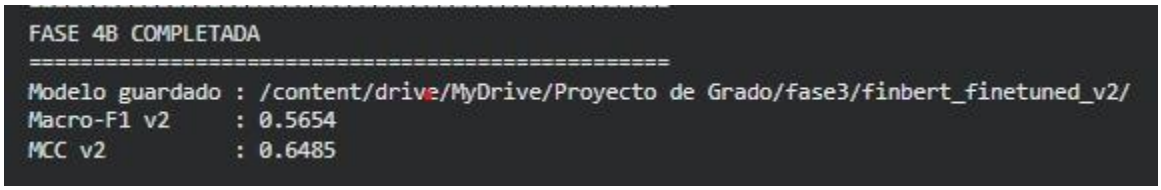


Figura 14 *Fase 5 – Comparación de distribuciones emocionales*

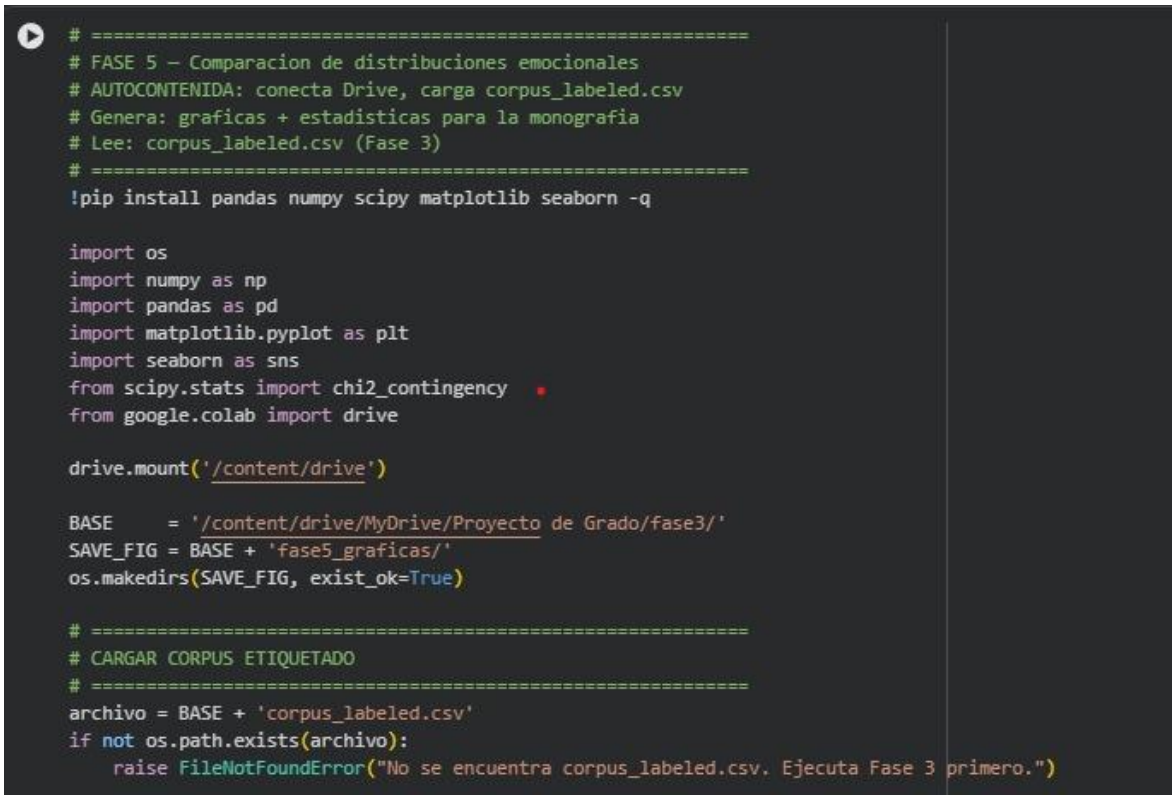


Figura 15 *Fase 5 – Resultado Chi – cuadrado*

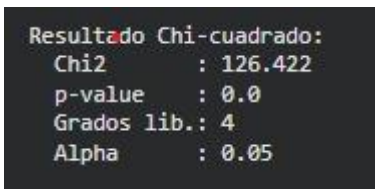


Figura 16 *Fase 6 – Evaluación final del modelo*

```
# ============================================================
# FASE 6 — Evaluacion final del modelo
# AUTOCONTENIDA: conecta Drive, carga modelo v2 + split_test
# Lee: split_test.csv + finbert_finetuned_v2/
# Guarda: reporte final, matriz de confusion, bootstrap IC
# ============================================================
!pip install transformers torch scikit-learn pandas matplotlib seaborn -q

import os
import gc
import numpy as np
import pandas as pd
import torch
import matplotlib.pyplot as plt
import seaborn as sns
from google.colab import drive
from transformers import AutoTokenizer, AutoModelForSequenceClassification
from sklearn.metrics import (
    classification_report,
    confusion_matrix,
    f1_score,
    matthews_corrcoef,
    accuracy_score
)
from datasets import Dataset

drive.mount('/content/drive')

BASE      = '/content/drive/MyDrive/Proyecto de Grado/fase3/'
MODEL_DIR = BASE + 'finbert_finetuned_v2/'
SAVE_FIG  = BASE + 'fase6_graficas/'
os.makedirs(SAVE_FIG, exist_ok=True)

# Verificar archivos
print("Verificando archivos...")
for f in ['split_test.csv']:
    ruta = BASE + f
    if not os.path.exists(ruta):
        raise FileNotFoundError("No se encuentra: " + ruta)
    kb = round(os.path.getsize(ruta) / 1024, 1)
    print("  OK: " + f + " (" + str(kb) + " KB)")

if not os.path.exists(MODEL_DIR):
    raise FileNotFoundError("No se encuentra el modelo: " + MODEL_DIR)
print("  OK: finbert_finetuned_v2/")
```

Figura 17 *Fase 6 – Métricas globales*

```
Metricas globales:
  Macro-F1    : 0.5564
  Micro-F1    : 0.8813
  Weighted-F1 : 0.8901
  MCC         : 0.6705
  Accuracy    : 0.8813
```

Figura 18 *Fase 7 – Crear etiquetas binarias*

```python
# FASE 7 — PASO 2: Crear etiquetas binarias
# 1 = entusiasmo | 0 = no entusiasmo

def crear_etiqueta_binaria(emocion):
    emocion = str(emocion).strip().lower()
    return int(emocion == "entusiasmo")

df_train_bin = df_train.copy()
df_val_bin = df_val.copy()
df_test_bin = df_test_bin_base.copy()

for df in [df_train_bin, df_val_bin, df_test_bin]:
    df["label_bin"] = df["emotion"].apply(
        crear_etiqueta_binaria
    )

print("Distribución binaria — entrenamiento:")
print(
    df_train_bin["label_bin"]
    .map({0: "No entusiasmo", 1: "Entusiasmo"})
    .value_counts()
)

print("\nDistribución binaria — validación:")
print(
    df_val_bin["label_bin"]
    .map({0: "No entusiasmo", 1: "Entusiasmo"})
    .value_counts()
)
```

Figura 19 *Fase 7 – Distribución binaria - entrenamiento*

```
Distribución binaria — entrenamiento:
label_bin
No entusiasmo    5043
Entusiasmo        383
Name: count, dtype: int64

Distribución binaria — validación:
label_bin
No entusiasmo    1081
Entusiasmo         82
Name: count, dtype: int64

Distribución binaria — prueba:
label_bin
No entusiasmo    1081
Entusiasmo         82
Name: count, dtype: int64

Etiquetas binarias creadas correctamente.
```

Custodia del código fuente y alcance de la reproducibilidad

El código fuente empleado en el desarrollo de la investigación se conserva bajo custodia de los autores y no se publica en un repositorio abierto de GitHub ni se distribuye como cuaderno ejecutable de Google Colab. En consecuencia, el presente anexo documenta la procedencia de los datos, las principales etapas metodológicas y las evidencias de ejecución disponibles, sin ofrecer acceso público al código fuente completo.

Esta decisión delimita el alcance de la verificación que puede realizarse a partir de los materiales incluidos. La documentación permite examinar la procedencia declarada de las fuentes y contrastar las evidencias disponibles con las cifras reportadas, pero no garantiza que un tercero pueda reconstruir íntegramente el procesamiento y obtener exactamente los mismos resultados sin acceso al código, los archivos de entrada y las configuraciones efectivamente utilizadas.