Aplicación de redes neuronales convolucionales y técnicas de procesamiento de lenguaje natural para el análisis de sentimientos en datos financieros

Estudiante: Juan Manuel Fernandez Ceballos

Director: Paula Maria Almonacid Hurtado

palmona1@eafit.edu.co

Área de Macroeconomía y Sistemas Financieros

Palabras clave: análisis financiero, minería de texto, procesamiento de lenguaje natural (NLP), predicción bursátil.

Maestría en Ciencia de los Datos y Analítica
Universidad EAFIT.

Resumen

Este estudio entrelaza los métodos tradicionales de análisis financiero con la fuerza del procesamiento del lenguaje natrual (PLN) y el aprendizaje profundo, aplicando una adaptación del marco CRISP-DM. Esta investigación se concentra en el sector de electrónica de consumo del mercado bursátil estadounidense, para lo cual se reunieron noticias financieras y datos históricos de cotizaciones. Con el fin de comprender con mayor detalle la naturaleza del lenguaje en el ámbito financiero, se llevaron a cabo procesos de limpieza y preprocesamiento textual, así como diversos experimentos con arquitecturas de redes neuronales convolucionales (CNN). Estos experimentos compararon tres enfoques de embeddings (aleatorios, GloVe y fastText) para capturar los matices de sentimiento.

Para determinar la influencia del sentimiento del mercado en las cotizaciones, se empleó un algoritmo de clasificación (Random Forest), con el propósito de pronosticar tendencias de precios, revelando el valor adicional que las emociones pueden aportar al mercado. Los resultados nos muestran la contribución positiva de los embeddings preentrenados a la clasificación de sentimientos y, a su vez, recalcan la importancia de incorporar otros indicadores para refinar la predicción de tendencias. De esta manera, el estudio sienta las bases para diseñar sistemas de análisis financiero más avanzados.

Contents

Índice de figuras

Índice de tablas

1 Descripción del proyecto

1.1 Planteamiento del problema

En el entorno financiero actual, los mercados se ven influenciados por una gran cantidad de información no estructurada, como noticias, publicaciones en redes sociales, informes de analistas y otros contenidos textuales. Esta información contiene opiniones, emociones y percepciones que pueden tener un impacto significativo en el comportamiento del mercado. Estudios han demostrado que el sentimiento expresado en estos medios puede prever movimientos de mercado significativos (Kanungsukkasem & Leelanupab, 2019).

Sin embargo, los métodos tradicionales de análisis financiero a menudo no logran capturar y procesar eficientemente estos datos subjetivos y complejos. Estos métodos se basan principalmente en datos estructurados y numéricos, como precios históricos y volúmenes de transacción, dejando de lado la riqueza de la información contenida en el texto (Kanungsukkasem & Leelanupab, 2019). La incapacidad de estos métodos para incorporar datos textuales no estructurados resulta en una pérdida significativa de información valiosa que podría mejorar las predicciones del mercado.

Existe una necesidad creciente de herramientas avanzadas para analizar automáticamente los sentimientos y las opiniones expresadas en estos datos no estructurados. La capacidad de procesar y comprender grandes cantidades de texto no estructurado, como noticias financieras y publicaciones en redes sociales, permitirá a los analistas financieros e inversores obtener una comprensión más profunda de los factores emocionales y subjetivos que impulsan los movimientos del mercado. Esto no solo mejoraría la precisión de las predicciones de mercado, sino que también permitiría una toma de decisiones de inversión más informada y oportuna (Kanungsukkasem & Leelanupab, 2019).

El análisis de sentimientos ha surgido como una técnica poderosa para resolver este problema. Mediante la aplicación de procesamiento del lenguaje natural (PNL) y técnicas de aprendizaje profundo, se pueden extraer y cuantificar las emociones expresadas en el texto. Las redes neuronales convolucionales (CNN) han demostrado ser particularmente efectivas en la captura de patrones y características locales en los datos textuales, mejorando así la capacidad de los modelos para predecir tendencias basadas en el sentimiento del mercado (Souma et al., 2019).

Además, integrar el análisis de sentimientos con otros indicadores financieros podría proporcionar una visión más completa y precisa de las dinámicas del mercado. La combinación de datos de sentimiento con datos estructurados, como precios de acciones y volúmenes de transacción, permite a los modelos predecir mejor las tendencias del mercado y responder a las fluctuaciones con mayor precisión (Lien Minh et al., 2018).

La necesidad de herramientas avanzadas para el análisis de sentimientos en datos financieros no estructurados es crítica. Estas herramientas deben ser capaces de procesar grandes volúmenes de datos textuales, extraer y cuantificar sentimientos, y combinar esta información con datos financieros tradicionales para mejorar las predicciones del mercado. El desarrollo de un sistema basado en CNN y técnicas avanzadas de PLN es una solución prometedora para satisfacer esta necesidad creciente.

1.2 Justificación

En los últimos años, el análisis del sentimiento de los datos de texto relacionados con las finanzas se ha vuelto cada vez más importante, ya que los sentimientos y opiniones expresados por los inversores, analistas en los medios y las redes sociales pueden afectar significativamente las tendencias del mercado de valores. Por ejemplo, Lien Minh et al. (2018) propusieron un modelo de dos corrientes basado en GRU y un embedding de palabras especializado en noticias financieras

(Stock2Vec), demostrando mejoras significativas en la predicción de tendencias del mercado a corto plazo.

Tradicionalmente, los métodos de análisis financiero se han basado en indicadores cuantitativos y fundamentales, sin tener en cuenta factores emocionales y percepciones subjetivas que pueden impulsar los movimientos del mercado. Sin embargo, a medida que hay más información disponible en línea y en las plataformas de redes sociales, las opiniones y sentimientos de los inversores se han vuelto más accesibles y relevantes.

Las técnicas de procesamiento de lenguaje natural (PLN) y aprendizaje automático (ML), especialmente las redes neuronales convolucionales (CNN), ofrecen un enfoque prometedor para abordar este desafío. Las CNN han demostrado un alto rendimiento en tareas de PLN, como la clasificación de texto y el análisis de sentimientos, debido a su capacidad para capturar características locales y patrones en los datos de entrada (Kraus & Feuerriegel, 2017; Ouyang et al., 2015).

Al aplicar CNN al análisis de sentimientos en datos financieros, como noticias y publicaciones en redes sociales, se puede extraer y cuantificar la información emocional y subjetiva contenida en estos textos. Esto puede proporcionar una visión más completa de los factores que impulsan los movimientos del mercado, complementando los enfoques tradicionales basados en indicadores cuantitativos.

Además, las CNN tienen la capacidad de manejar grandes volúmenes de datos y procesar información en tiempo real, lo cual es crucial en el entorno financiero rápidamente cambiante. Al combinar el análisis de sentimientos basado en CNN con otros indicadores financieros, se puede desarrollar un sistema de predicción de tendencias del mercado más preciso y oportuno.

El análisis de sentimientos basado en CNN en datos financieros tiene el potencial de proporcionar una ventaja competitiva significativa al permitir a los inversores y analistas anticipar mejor las tendencias del mercado, aprovechando la información emocional y subjetiva expresada en medios de comunicación y redes sociales. Este enfoque puede complementar los métodos tradicionales de análisis financiero y mejorar la toma de decisiones de inversión.

2 Objetivos

2.1 Objetivo general

Desarrollar un sistema basado en redes neuronales convolucionales y procesamiento de lenguaje natural para el análisis de sentimientos de noticias financieras relacionadas con las acciones de la industria de electrónica de consumo del sector tecnológico del mercado de valores de Estados Unidos de los últimos dos años, con el propósito de predecir tendencias del mercado en función del sentimiento y respaldar la toma de decisiones de inversión.

2.2 Objetivos específicos

- Recopilar y preparar un conjunto de datos de noticias financieras, redes sociales, informes de analistas y otros datos relacionados con las acciones de la industria de electrónica de consumo del sector tecnológico del mercado de valores de Estados Unidos de los últimos dos años.
- Desarrollar e implementar una arquitectura de CNN adaptada al análisis de sentimientos en datos financieros, utilizando técnicas como transferencia de aprendizaje y modelos de atención.
- Integrar técnicas avanzadas de PLN, como Word Embeddings y modelos de lenguaje pre-entrenados, para mejorar la comprensión del contexto y significado en los datos textuales.

3 Marco teórico y estado del arte

3.1 Mercado de valores

Los mercados de valores son una parte fundamental de la economía global. En este mercado se negocian diversos instrumentos financieros, como acciones, bonos e índices. El mercado permite a las empresas obtener financiación mediante la emisión de acciones, bonos y ofrece a los inversores la oportunidad de comprar acciones de estas empresas. Los mercados de valores son cruciales para la asignación eficiente de recursos en una economía, ayudando a mover fondos de los ahorradores a las inversiones más prometedoras.

Las acciones representan la propiedad parcial en una empresa. Los accionistas pueden obtener beneficios a través de dividendos y la apreciación del valor de las acciones. Las acciones son uno de los instrumentos más comunes y líquidos en los mercados de valores (Lee et al., 2024). Por otro lado, los bonos son instrumentos de deuda emitidos por empresas o gobiernos para recaudar capital. Los bonos pagan intereses a los tenedores durante un período específico y devuelven el capital al vencimiento (Roostaee & Abin, 2023). Además, los índices, como el S&P 500 o el Dow Jones Industrial Average, reflejan el rendimiento general de un grupo de acciones representativas del mercado. Los índices son utilizados por los inversores para medir el desempeño del mercado en su conjunto y comparar la rentabilidad de sus inversiones (Dessain, 2022)

El análisis del mercado de valores se realiza mediante varias metodologías. El análisis fundamental evalúa el valor intrínseco de una empresa examinando factores económicos, financieros y cualitativos. Este análisis incluye el estudio de estados financieros, la gestión de la empresa, la competencia, las condiciones de la industria y otros factores relevantes (Kanungsukkasem & Leelanupab, 2019). En contraste, el análisis técnico se basa en el estudio de gráficos de precios y

patrones de trading para predecir movimientos futuros de precios. Utiliza herramientas como medias móviles, líneas de tendencia y otros indicadores técnicos para analizar la acción del precio.

3.2 Análisis de sentimientos

El análisis de sentimientos, también conocido como minería de opiniones, es el proceso de determinar la actitud o sentimiento expresado por el autor de un texto sobre un tema específico. En finanzas, se utiliza para evaluar las opiniones expresadas por inversores y analistas en noticias, informes financieros y publicaciones en redes sociales. Este análisis nos permite comprender mejor los factores emocionales que influyen en los movimientos del mercado y puede ser una herramienta valiosa para predecir tendencias.

El análisis de sentimientos se divide en varios componentes. La polaridad es la clasificación del sentimiento como positivo, negativo o neutral, lo cual ayuda a identificar la dirección del sentimiento predominante en el texto (Ahmad & Umar, 2023). La intensidad mide la fuerza del sentimiento expresado, proporcionando una idea de cuán fuerte es la emoción o la opinión (Nguyen & Huynh, 2022). Además, los aspectos específicos identifican aspectos particulares de un tema y el sentimiento asociado a ellos, lo que permite un análisis más detallado y contextualizado del sentimiento (Du et al., 2023).

Las aplicaciones del análisis de sentimientos en finanzas son diversas. Una de las principales aplicaciones es la predicción de movimientos del mercado. Identificar tendencias emergentes y predecir movimientos de precios basados en el sentimiento agregado puede anticipar un aumento en el precio de las acciones de una empresa cuando hay un aumento en el sentimiento positivo sobre ella (Souma et al., 2019). Otra aplicación importante es la gestión de riesgos, donde se evalúa el sentimiento del mercado para anticipar posibles caídas y gestionar mejor el riesgo. Los

inversores pueden ajustar sus carteras en función del sentimiento del mercado para minimizar pérdidas potenciales (Kraus & Feuerriegel, 2017).

3.3 Procesamiento de lenguaje natural (PLN)

El procesamiento de lenguaje natural (PLN) es un campo de la inteligencia artificial que se enfoca en la interacción entre las computadoras y el lenguaje humano. Se utiliza para analizar, comprender y generar texto de una manera que sea valiosa para diversas aplicaciones. El PLN permite que las máquinas procesen y entiendan el lenguaje humano de manera efectiva, lo cual es crucial para tareas como el análisis de sentimientos en datos financieros.

El PLN emplea varias técnicas fundamentales. La tokenización es el proceso de dividir el texto en unidades más pequeñas como palabras o frases, lo cual es un paso fundamental en el análisis de texto, ya que permite el procesamiento de las palabras de manera individual (Ouyang et al., 2015). La lematización y el stemming son técnicas que reducen las palabras a sus formas base o raíces, considerando el contexto de la palabra o simplemente cortando los sufijos para encontrar la raíz (Ahmad & Umar, 2023). Los embeddings de palabras representan palabras como vectores en un espacio de alta dimensión utilizando técnicas como Word2Vec, GloVe y FastText, capturando relaciones semánticas entre las palabras y mejorando la precisión en tareas de PLN (Ouyang et al., 2015).

Además, los modelos de lenguaje como BERT y GPT-3 se utilizan para entender el contexto y significado de las palabras en oraciones, captando matices y contextos complejos en el lenguaje humano (Du et al., 2023). El análisis sintáctico y semántico analiza la estructura gramatical y el significado de las frases, ayudando a comprender mejor el texto y extraer información relevante (Nguyen & Huynh, 2022). El PLN es esencial para el análisis de grandes volúmenes de datos

textuales, permitiendo extraer información valiosa y comprender mejor las tendencias y percepciones en los mercados financieros (Ahmad & Umar, 2023).

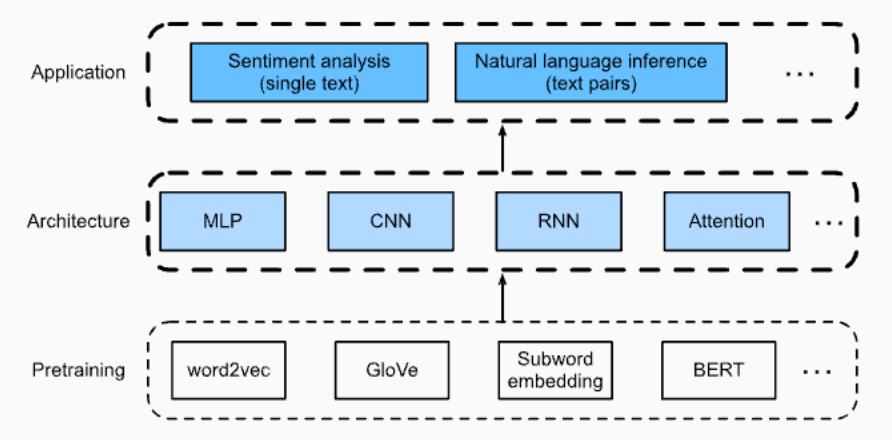
3.4 Redes neuronales convolucionales (CNN)

Las redes neuronales convolucionales (CNN) son un tipo de modelo computacional inspirado en la estructura y función del cerebro humano, especialmente diseñadas para reconocer patrones complejos y realizar predicciones basadas en datos de entrenamiento. Las CNN son particularmente efectivas en el procesamiento de datos con una estructura de cuadrícula, como imágenes y datos textuales organizados.

Las CNN están compuestas por varios componentes clave. Las neuronas son las unidades básicas que reciben entradas, las procesan y generan salidas. Cada neurona en una capa convolucional procesa una pequeña parte del dato de entrada y extrae características locales (Ouyang et al., 2015). Las capas, por su parte, están compuestas por neuronas y pueden incluir capas de entrada, ocultas y de salida. Las capas convolucionales y de pooling son comunes en las CNN (Abdelhady et al., 2024). Las capas de convolución aplican filtros para detectar características locales, mientras que las capas de pooling reducen la resolución espacial de las características, extrayendo características jerárquicas y reduciendo la dimensionalidad de los datos (Souma et al., 2019) .

Las CNN han demostrado ser altamente efectivas en tareas de PLN como la clasificación de texto y el análisis de sentimientos debido a su capacidad para capturar características locales y patrones en los datos de entrada. Estas redes son esenciales para desarrollar modelos precisos y eficientes que puedan analizar grandes volúmenes de datos textuales y predecir tendencias en los mercados financieros (Kraus & Feuerriegel, 2017).

Figura 1: Estructura de modelos de procesamiento de lenguaje natural desde el preentrenamiento hasta la aplicación



Fuente: Tomado de "Dive into Deep Learning" (Zhang et al., 2023)

Otro enfoque destacado para la clasificación de texto con CNN lo presenta (Kim, 2014), quien demuestra que un modelo convolucional, entrenado sobre vectores de palabras preentrenados, puede lograr resultados competitivos en diversas tareas de clasificación a nivel de oraciones. En su propuesta, cada oración se transforma en una matriz, donde cada fila corresponde al vector de embeddings de una palabra. Luego, se aplican filtros de distinto tamaño, lo cual permite detectar diferentes patrones sintácticos y semánticos.

4 Metodología

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) es un marco de trabajo estructurado que se utiliza ampliamente para guiar proyectos de minería de datos y aprendizaje automático, para este proyecto, se plantea una versión de esta metodología excluyendo el despliegue

4.1 Comprensión del negocio

En el entorno financiero actual, los mercados se ven influenciados por una gran cantidad de información no estructurada, como noticias, publicaciones en redes sociales, informes de analistas y otros contenidos textuales. Este tipo de información contiene opiniones, emociones y

percepciones que pueden tener un impacto significativo en el comportamiento del mercado. Sin embargo, capturar y procesar estos datos para integrarlos con métodos de análisis financiero tradicionales sigue siendo un desafío considerable. Este proyecto busca abordar esta brecha mediante la combinación de datos financieros estructurados y datos textuales no estructurados, proporcionando una visión más amplia y contextual del mercado.

4.2 Comprensión de los datos

Para llevar a cabo este proyecto, se inició con la identificación de las acciones pertenecientes a la industria de Electrónica de Consumo, dentro del sector tecnológico del mercado de valores de los Estados Unidos. Este paso fue fundamental para delimitar el conjunto de datos y garantizar la relevancia de los análisis posteriores.

La identificación de las acciones se realizó utilizando el portal Yahoo Finance, específicamente en la sección "Markets", dentro de la subsección "Sectors", seleccionando la categoría "Technology/Consumer Electronics" (https://finance.yahoo.com/sectors/technology/consumer-electronics/).

La información recopilada fue almacenada en una base de datos SQLite, en distintas tablas, donde la tabla "stock_market_company" se convirtió en el insumo principal para las siguientes fases del proyecto, incluyendo la extracción de cotizaciones y noticias de cada una de las acciones con rangos de 1 hora. Posteriormente, se utilizó la API de Polygon (https://polygon.io/) para extraer las cotizaciones históricas de las acciones identificadas, cubriendo el período desde el 1 de enero de 2022 hasta agosto de 2024, dichos datos fueron alojados en la tabla "stock_market_stockprice". Las cotizaciones incluyeron datos detallados, como precios de apertura, cierre, máximos, mínimos y volúmenes transaccionados. Este intervalo de tiempo fue seleccionado para garantizar que el

análisis incluyera datos recientes y permitiera identificar tendencias relevantes en el comportamiento del mercado.

Adicionalmente, la API de Polygon también permitió extraer noticias relacionadas con cada una de las acciones, asegurando que estas correspondieran al mismo rango temporal que las cotizaciones, las cuales fueron alojadas en la tabla "stock_market_news". Estos datos son el eje principal del proyecto, ya que permite analizar las emociones y percepciones expresadas en los medios de comunicación para determinar el impacto que las emociones y percepciones tienen en los movimientos del mercado.

Figura 2: Modelo de datos

Para el desarrollo del proyecto se cuenta con información proveniente de 25 acciones pertenecientes a la industria de electrónica de consumo del sector tecnológico del mercado de valores de Estados Unidos. En total, se han recolectado 78.455 registros en el conjunto de datos "stock_market_stockprice" y 15.205 registros en el conjunto de datos "stock_market_news". Estos datos, proporcionan la base necesaria para combinar análisis financieros tradicionales con

el estudio del sentimiento del mercado, con el fin de mejorar la capacidad predictiva y la toma de decisiones de inversión.

Cotizaciones

Accion	Cantidad de registros
AAPL	10.548
GPRO	7.816
VUZI	7.710
SONO	6.493
VZIO	5.889
WLDS	5.773
KOSS	5.649
HEAR	5.530
VOXX	5.092
UEIC	5.048
MICS	3.984
MSN	3.099
WTO	2.492
FEBO	1.099
ANDR	828
MOBO	688
AXIL	565
EXEO	81
NYXO	20
ZVTK	16
GAXYQ	12
LTEC	12
BZIC	7
CTXV	2
COLCF	2

Tabla 1 Cantidad de cotizaciones

Noticias

Accion	Cantidad de registros
AAPL	14.135
SONO	224
GPRO	211
HEAR	150
WLDS	127
VZIO	121
VUZI	64
VOXX	46
MICS	40
UEIC	36
KOSS	31
WTO	10
AXIL	4
MSN	4
FEBO	2

Tabla 2 Cantidad de noticias

4.3 Preparación de los datos

La preparación de los datos constituye un componente esencial del desarrollo de este proyecto, pues asegura que las fuentes de información sean consistentes, completas y adecuadas para el modelado. Este proceso comenzó con la integración de datos textuales y numéricos provenientes de noticias financieras y cotizaciones de acciones, respectivamente, almacenados en una base de datos relacional para facilitar su análisis.

Como se observa en las tablas 1 y 2, la disponibilidad de datos textuales (noticias) varía significativamente entre las 25 acciones consideradas. Esta disparidad en la cantidad de información textual sugiere la necesidad de aplicar un filtrado adicional, de modo que se puedan seleccionar aquellas acciones con un volumen suficiente de noticias.

Con el fin de asegurar la integridad de la información en las noticias, se imputaron los valores nulos de la columna *content* utilizando los datos disponibles en la columna *description*. Este paso permite preservar la mayor cobertura posible de texto para los análisis posteriores.

Una vez completos los registros en la columna *content*, se procedió a limpiar el texto eliminando cualquier carácter que no correspondiera a letras del alfabeto. Para ello, se utilizó la función re.sub('[^a-zA-Z]', ' ', text), que reemplaza caracteres no alfabéticos (números, signos de puntuación, símbolos especiales, etc.) por un espacio. De este modo, el texto resultante contiene únicamente letras y espacios, lo cual facilita el procesamiento en etapas posteriores.

Seguidamente, se llevaron a cabo dos pasos adicionales para preparar el texto. En primer lugar, se aplicó la tokenización mediante la función nltk.word_tokenize(text), dividiendo el contenido en una lista de palabras y permitiendo tratar cada una de manera independiente.

Posteriormente, se realizó la remoción de *stop words* y la lematización. Las *stop words* — palabras muy frecuentes que no suelen aportar valor analítico— se eliminaron comparándolas con una lista llamada stop_words. En simultáneo, la lematización transformó cada palabra a su forma base mediante el objeto lemmatizer, de manera que términos como “running” o “ran” se redujeron a “run”. Esta estrategia garantiza una mayor uniformidad y consistencia semántica en el conjunto de datos.

Para realizar un etiquetado inicial de las noticias financieras desde un enfoque de análisis de sentimientos, se emplea el diccionario de Loughran y McDonald a través de la biblioteca *pysentiment2*. Este recurso, usado en el ámbito del análisis de textos financieros, recopila y clasifica términos que aparecen con frecuencia en informes, comunicados y noticias corporativas, asignándoles categorías específicas (positivas, negativas, neutras).

4.4 Modelado

En esta fase de modelación, se establecieron tres escenarios de experimentación con distintos tipos de embeddings, con el fin de identificar cuál brinda el mejor desempeño en la clasificación de sentimientos. El primero utiliza embeddings aleatorios que se entrenan desde cero, permitiendo a la red aprender representaciones específicas de los datos. El segundo integra los embeddings preentrenados de GloVe, lo cual suele acelerar la convergencia y captar relaciones semánticas en corpus extensos. Finalmente, el tercero adopta fastText, que maneja subpalabras para mejorar la representación de términos poco frecuentes, especialmente relevantes en lenguaje financiero. Con esta comparativa, se busca determinar el enfoque más efectivo para la tarea de análisis de sentimientos en el dominio bursátil.

4.4.1 CNN con Embeddings aleatorios

Embeddings aleatorios

Se emplea una capa Embedding que inicializa los vectores de forma aleatoria (embedding_dim=64) y los entrena junto con la red. Esto permite que el modelo aprenda representaciones específicas del conjunto de datos, adaptadas a la tarea de análisis de sentimientos.

Capas convolucionales (Conv1D)

Tras la capa de embedding, se aplican dos convoluciones de 128 filtros cada una y un kernel_size de 3, con padding=”same” para preservar la dimensionalidad a lo largo de la secuencia. Las activaciones ReLU capturan patrones locales en la secuencia de tokens (por ejemplo, expresiones que podrían indicar un tono positivo o negativo). Cada capa convolucional va seguida de un Dropout(0.3), que ayuda a controlar el sobreajuste, y posteriormente se añade BatchNormalization para estabilizar y acelerar el entrenamiento.

Capa recurrente bidireccional

En lugar de aplicar un mecanismo de atención en esta arquitectura, se incorporó una capa Bidirectional(LSTM(32)) que puede aprovechar dependencias tanto hacia adelante como hacia atrás dentro de la secuencia. Esta aproximación combina la capacidad de las redes convolucionales para detectar características locales con la fortaleza de las LSTM en el modelado de dependencias de largo plazo.

Pooling global

Después de la capa LSTM, se emplea un GlobalMaxPooling1D para reducir la secuencia a un único vector por muestra, tomando el valor máximo de cada filtro. Esta operación conserva las características de mayor activación y suprime los detalles irrelevantes, simplificando la representación de la secuencia a un vector compacto.

Capas densas con regularización

Se añaden dos capas densas de 128 y 64 neuronas respectivamente, activadas con ReLU. Para ambas se define una regularización L2 (kernel_regularizer=l2(1e-4)) y se incluyen capas Dropout(0.2) y Dropout(0.1) para mitigar el riesgo de sobreajuste.

Capa de Salida

Finalmente, se utiliza una capa densa con 3 neuronas y activación softmax, que genera la distribución de probabilidad sobre las tres clases de sentimiento.

Compilación y entrenamiento

El modelo se compila con la función de pérdida sparse_categorical_crossentropy, el optimizador Adam y la métrica de exactitud (accuracy). Para el entrenamiento, se estableció un EarlyStopping que detiene el proceso si no se observan mejoras en la val_loss durante 10 épocas, restaurando los mejores pesos encontrados.

4.4.2 CNN con Glove

Embeddings preentrenados (Glove)

Se cargan los vectores GloVe glove.6B.100d (100 dimensiones) en memoria, guardándolos en un diccionario (embeddings_index). Para cada palabra en el vocabulario aprendido, se recupera su vector correspondiente y se asigna en una matriz de embeddings (embedding_matrix).

Finalmente, la capa Embedding se inicializa con esta matriz (weights=[embedding_matrix]), y con el parámetro trainable=True se permite un ajuste fino (fine-tuning) durante el entrenamiento.

Entrada y Capa Embedding

La capa de Embedding emplea la embedding_matrix de GloVe para transformar cada índice en un vector de 100 dimensiones, ofreciendo así representaciones semánticas preentrenadas de cada palabra.

Capas Convolucionales

A partir de la salida de la capa Embedding, se construyen tres caminos paralelos de Conv1D, cada uno con 128 filtros y padding='same', pero con distintos tamaños de kernel (3, 5 y 7). Las salidas se combinan mediante concatenate([x1, x2, x3]).

Capa Recurrente Bidireccional

A la concatenación resultante se le aplica una capa Bidirectional(LSTM(64, return_sequences=True)). Esto permite procesar la secuencia en dos direcciones —hacia adelante y hacia atrás—, capturando dependencias contextuales de largo alcance en ambos sentidos.

Capa de Atención (Self-Attention)

La atención se implementa con Attention()([x, x]), recibiendo a x como query, key y value. Este mecanismo asigna más peso a aquellas partes de la secuencia consideradas relevantes para la clasificación de sentimientos.

Pooling Global

Tras la atención, se emplea un GlobalMaxPooling1D() que reduce la secuencia a un único vector por muestra, tomando el valor máximo a través del tiempo.

Capas Densas con Regularización

Se añaden dos capas densas de 128 y 64 neuronas, respectivamente, cada una con activación ReLU y regularización L2 (kernel_regularizer=l2(1e-4)). Entre estas capas se incluyen capas Dropout(0.3) y Dropout(0.2) para mitigar el riesgo de sobreajuste.

Capa de Salida

Finalmente, se utiliza una capa densa con 3 neuronas y activación softmax, que genera la distribución de probabilidad sobre las tres clases de sentimiento.

Compilación y entrenamiento

El modelo se compila con la función de pérdida sparse_categorical_crossentropy, el optimizador Adam y la métrica de exactitud (accuracy). Para el entrenamiento, se estableció un EarlyStopping que detiene el proceso si no se observan mejoras en la val_loss durante 10 épocas, restaurando los mejores pesos encontrados.

4.4.3 CNN con FastText

Carga de Embeddings Preentrenados (FastText)

Se cargan los vectores FastText (por ejemplo, cc.en.300.vec) con 300 dimensiones, almacenándolos en un diccionario (embeddings_index). Se construye una embedding_matrix de tamaño vocab_size, embedding_dim, donde cada fila corresponde al vector de FastText para la palabra con dicho índice. Para las palabras sin vector disponible, se deja la fila en ceros, lo que indica que no se ha encontrado su representación en el diccionario FastText.

Entrada y Capa Embedding

Se define la capa de entrada Input(shape=(max_length,)) y, acto seguido, la capa Embedding con weights=[embedding_matrix] para inicializar los vectores con FastText. Se fija trainable=True para que estos vectores puedan ajustarse durante el entrenamiento (fine-tuning).

Capas Convolucionales

Se utilizan tres capas Conv1D en paralelo (cada una con 128 filtros y padding='same'), pero con kernel_size de 3, 5 y 7. Las salidas se combinan mediante concatenate([x1, x2, x3]).

Capa Recurrente Bidireccional (GRU)

Se aplica Bidirectional(GRU(64, return_sequences=True)) sobre la concatenación previa, capturando dependencias contextuales en ambas direcciones (hacia adelante y hacia atrás). Se añade una capa BatchNormalization para estabilizar y acelerar el entrenamiento, normalizando las activaciones intermedias.

Capa de Atención (Self-Attention)

La atención se implementa con Attention()([x, x]), recibiendo a x como query, key y value. Este mecanismo asigna más peso a aquellas partes de la secuencia consideradas relevantes para la clasificación de sentimientos.

Pooling Global

Tras la atención, se emplea un GlobalMaxPooling1D() que reduce la secuencia a un único vector por muestra, tomando el valor máximo a través del tiempo.

Capas Densas con Regularización

Se añaden dos capas densas de 128 y 64 neuronas, respectivamente, cada una con activación ReLU y regularización L2 (kernel_regularizer=l2(1e-4)). Entre estas capas se incluyen capas Dropout(0.3) y Dropout(0.2) para mitigar el riesgo de sobreajuste.

Capa de Salida

Finalmente, se utiliza una capa densa con 3 neuronas y activación softmax, que genera la distribución de probabilidad sobre las tres clases de sentimiento.

Compilación y entrenamiento

El modelo se compila con la función de pérdida sparse_categorical_crossentropy, el optimizador Adam y la métrica de exactitud (accuracy). Para el entrenamiento, se estableció un EarlyStopping que detiene el proceso si no se observan mejoras en la val_loss durante 10 épocas, restaurando los mejores pesos encontrados.

4.4.4 Random Forest

Con el fin de predecir la tendencia de precios (sube, baja o neutra) de acciones en el corto plazo, se implemento un modelo Random Forest el cual se construyó utilizando el sentimiento promedio diario como variable predictora para anticipar la tendencia de precio (sube, baja o neutra) en las acciones de AAPL y GPRO. En primer lugar, se partió de un conjunto de entrenamiento conformado por el 80 % más antiguo de la serie temporal y un 20 % restante como validación. Durante el proceso de búsqueda de hiperparámetros, se evaluaron diferentes combinaciones de profundidad máxima, número de árboles (n_estimators) y criterio de división. La mejor configuración hallada fue: n_estimators=100, max_depth=20 y min_samples_split=2, resultados que equilibraron la complejidad del modelo con su capacidad para capturar patrones a partir del sentimiento.

4.5 Evaluación

4.5.1 CNN con Embeddings Aleatorios

Modelo	Accuracy	Precision	Recall
CNN + Embed aleatorios	71,0%	82,3%	85,7%

Tabla 3 Resultado escenario 1

En este primer escenario, los vectores de palabras se inicializan de forma aleatoria y se entrenan junto con red neuronal. El modelo logra una exactitud global del 71%, lo que significa que, de cada 100 noticias, clasifica correctamente cerca de 71. Este comportamiento sugiere que el conjunto de datos podría tener particularidades que permiten al modelo aprender representaciones muy enfocadas en la tarea de análisis de sentimientos, sin necesidad de partir de conocimiento lingüístico previo.

En términos de precisión con un 82.3 % y recall con un 85.7 %, la red muestra un equilibrio entre ambas métricas, reflejando que el modelo no solo identifica correctamente las clases, sino que también logra capturar la mayoría de los ejemplos relevantes.

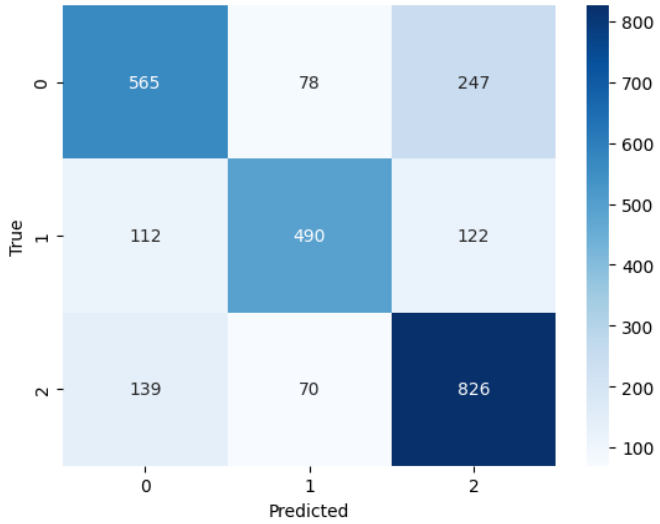


Figura 3 Matriz de confusión escenario 1

Este modelo presenta una diagonal relativamente buena, lo que significa que la red acierta un número considerable de casos en cada clase, aunque tambien es posible encontrar confusiones entre la clase 0 (negativo) y la clase 2 (positivo) en aquellas noticias donde el tono está menos claro o tiene matices ambiguos. Este comportamiento es típico cuando el vocabulario no está estandarizado y la red se apoya en *embeddings* aprendidos desde cero.

- **Clase 0 (Negativo):** En general, la mayoría de los ejemplos negativos se identifican correctamente, pero puede haber confusiones puntuales con la clase 1 (Neutro) y la 2 (Positivo) especialmente cuando el lenguaje es más suave o cuando hay opiniones contradictorias.
- **Clase 1 (Neutro):** Esta categoría es la más difícil para el modelo. A veces, noticias con un ligero toque negativo o positivo pueden ser clasificadas como neutras, y viceversa.
- **Clase 2 (Positivo):** El modelo logra acertar, pero puede confundirse cuando hay palabras muy fuertes que podrían indicar un sentimiento negativo.

4.5.2 CNN con GloVe

Modelo	Accuracy	Precision	Recall
CNN + Embed GloVe	68,6%	83,0%	80,0%

Tabla 4 Resultado escenario 2

La integración de embeddings preentrenados (GloVe) en el segundo escenario proporcionó al modelo una base semántica sólida, lo que se tradujo en un aumento significativo en la precisión. Al incorporar el conocimiento lingüístico previo capturado en los embeddings de GloVe, el modelo demostró una mayor capacidad para discriminar entre clases, especialmente en la identificación de noticias positivas.

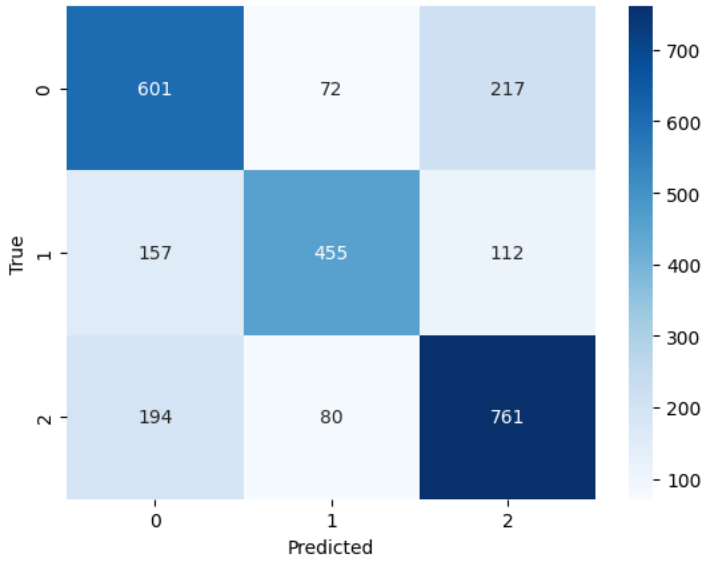


Figura 4 Matriz de confusión escenario 2

El modelo alcanza un número considerable de aciertos en la diagonal principal. Sin embargo, el mayor diferenciador es que al marcar una noticia como, por ejemplo, “positiva”, reduce la probabilidad de equivocarse comparado con otros métodos.

- **Clase 0 (Negativo):** Se observan menos clasificaciones erróneas de textos que eran neutrales o positivos etiquetados como negativos, lo que habla de la capacidad de GloVe para captar palabras muy orientadas a un sentimiento adverso.
- **Clase 1 (Neutro):** Sigue habiendo cierta confusión con las clases 0 y 2, principalmente por la ambigüedad de algunos textos; De igual forma, GloVe mitiga parte de esta confusión al haber aprendido contexto general en grandes corpus.
- **Clase 2 (Positivo):** La elevada precisión del modelo al clasificar las instancias como “positivas” demuestra una alta confiabilidad. La matriz de confusión revela una baja tasa de falsos positivos, lo que indica que cuando el modelo asigna una etiqueta positiva, existe una alta probabilidad de que la noticia sea efectivamente positiva.

4.5.3 CNN con FastText

Modelo	Accuracy	Precision	Recall
CNN + Embed FastText	67,0%	77,3%	92,0%

Tabla 5 Resultado escenario 3

En el tercer escenario, al implementar embeddings de FastText de 300 dimensiones permitió al modelo aprovechar las ricas representaciones semánticas de las palabras, capturando así los matices y contexto de las expresiones. Con este enfoque se obtuvo un recall del 92%, lo que indica una capacidad sobresaliente para identificar la mayoría de los ejemplos positivos reales. Sin embargo, esta alta sensibilidad conlleva una disminución marginal en la precisión, sugiriendo que el modelo, en su búsqueda por no omitir ningún caso, puede incluir algunos falsos positivos.

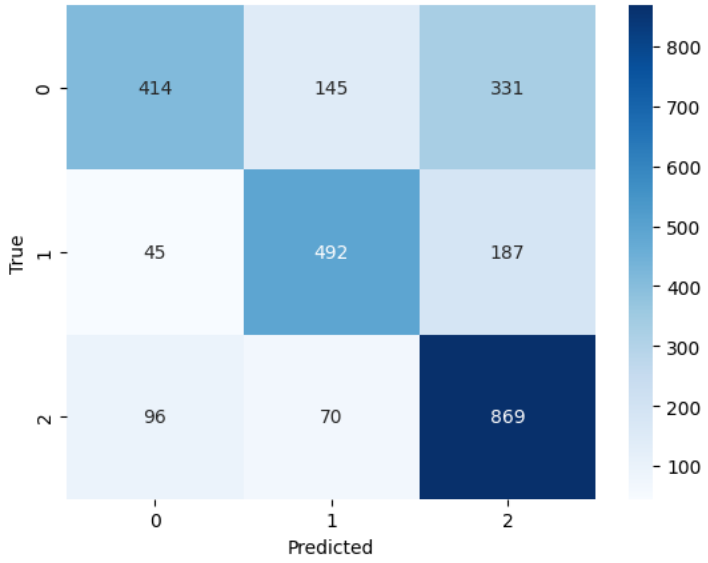


Figura 5 Matriz de confusión escenario 3

Para este modelo se refleja un gran número de muestras positivas reales de cada categoría clasificadas correctamente (debido al elevado *recall*). Sin embargo, también se ven algunos errores que llevan a clasificar noticias con matices neutros o negativos como positivas, o viceversa, elevando los *falsos positivos*.

- **Clase 0 (Negativo):** A la red no se le escapan muchos casos de negativo (pocos falsos negativos), pero en ocasiones, textos con alguna connotación fuerte podrían ser "exageradamente" asignados a esta clase.
- **Clase 1 (Neutro):** Es, probablemente, la categoría donde se observa más confusión. El texto con matices leves puede clasificarse como negativo o positivo, por el afán de "no dejar escapar" ejemplos.
- **Clase 2 (Positivo):** Mantiene un elevado acierto, pero la granularidad de sentimientos más sutiles puede confundirlo con lo neutro, generando *falsos positivos* si el texto no es claramente positivo.

4.5.4 Random Forest

En el conjunto de prueba, la exactitud del modelo alcanzó alrededor del 47,62 %. la matriz de confusión sugiere que el modelo tiene dificultades para distinguir entre escenarios de "Baja" y "Sube" cuando el sentimiento no es lo suficientemente claro o cuando el movimiento real del precio no presenta un comportamiento marcadamente direccional. Además, se evidencia una marcada debilidad en la detección del estado "Neutro", ya que los casos correspondientes a esta categoría tienden a ser confundidos con alzas o bajas.

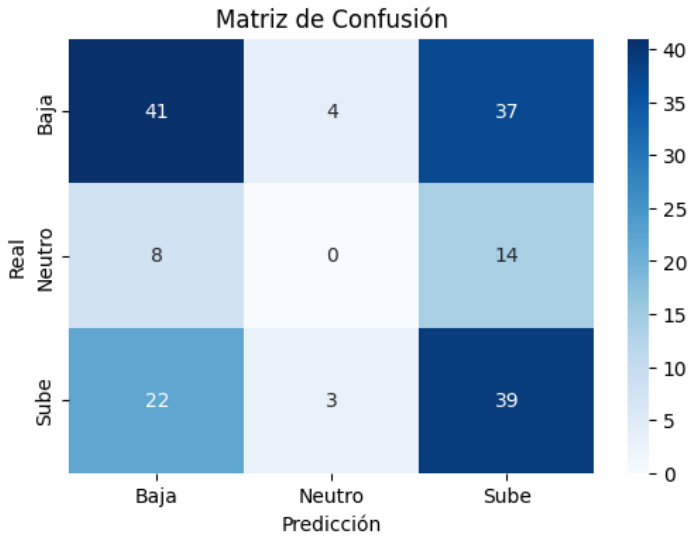


Figura 6 Matriz de confusión Random Forest

5 Conclusiones y trabajos futuros

Este estudio demuestra el potencial de las redes neuronales convolucionales (CNN) para analizar con precisión los sentimientos expresados en noticias financieras. Los resultados obtenidos revelan que la elección de los embeddings y la complejidad de los datos utilizados tienen un impacto significativo en la capacidad de los modelos para asignar el sentimiento de los textos.

Cada tipo de embedding evaluado —aleatorios, GloVe y FastText— aportó una perspectiva única. Los embeddings aleatorios demostraron que las redes pueden aprender a reconocer patrones directamente de los datos, mientras que los preentrenados como GloVe y FastText permitieron aprovechar el conocimiento lingüístico existente, mejorando la detección de matices en el lenguaje financiero.

Sin embargo, el en el proyecto también se destaca la importancia de la calidad y diversidad de los datos. Al igual que en otros campos de la inteligencia artificial, cuanto más variados y representativos sean los datos de entrenamiento, más precisas y confiables serán las predicciones del modelo.

Al incluir solo la variable de sentimiento promedio en el modelo de tendencia, la exactitud final fue modesta (cercana al 47 %). Esto indica que el sentimiento, si bien aporta información, no basta para explicar la dinámica de los precios. Se requieren variables adicionales para mejorar la capacidad predictiva.

Mirando al futuro, la incorporación de arquitecturas más avanzadas, como Transformers o modelos de lenguaje específicos para el ámbito financiero (por ejemplo, FinBERT), se convierten en una vía prometedora para capturar de forma más precisa la sutileza del lenguaje económico. Asimismo, la combinación de datos textuales con indicadores cuantitativos, tales como métricas bursátiles, índices macroeconómicos o información proveniente de redes sociales, podría dotar al sistema de una visión más holística de la dinámica del mercado.

En resumen, este estudio sienta las bases para el desarrollo de sistemas de análisis de sentimiento más sofisticados y confiables en el ámbito financiero. Los resultados obtenidos abren nuevas vías de investigación y ofrecen un panorama prometedor para el futuro de la inteligencia artificial aplicada a las finanzas.

6 Referencias bibliográficas

Abdelhady, N., Hassan A. Soliman, T., & F. Farghally, M. (2024). Stacked-CNN-BiLSTM-COVID: An effective stacked ensemble deep learning framework for sentiment analysis of Arabic COVID-19 tweets. *Journal of Cloud Computing*, *13*(1), 85. https://doi.org/10.1186/s13677-024-00644-6

Ahmad, H. O., & Umar, S. U. (2023). Sentiment Analysis of Financial Textual data Using Machine Learning and Deep Learning Models. *Informatica*, *47*(5), Article 5. https://doi.org/10.31449/inf.v47i5.4673

Dessain, J. (2022). Machine learning models predicting returns: Why most popular performance metrics are misleading and proposal for an efficient metric. *Expert Systems with Applications*, *199*, 116970. https://doi.org/10.1016/j.eswa.2022.116970

Du, K., Xing, F., & Cambria, E. (2023). Incorporating Multiple Knowledge Sources for Targeted Aspect-based Financial Sentiment Analysis. *ACM Transactions on Management Information Systems*, *14*(3), 23:1-23:24. https://doi.org/10.1145/3580480

Kanungsukkasem, N., & Leelanupab, T. (2019). Financial Latent Dirichlet Allocation (FinLDA): Feature Extraction in Text and Data Mining for Financial Time Series Prediction. *IEEE Access*, *7*, 71645–71664. Scopus. https://doi.org/10.1109/ACCESS.2019.2919993

Kim, Y. (2014). *Convolutional Neural Networks for Sentence Classification* (arXiv:1408.5882). arXiv. https://doi.org/10.48550/arXiv.1408.5882

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

Lee, H., Kim, J. H., & Jung, H. S. (2024). Deep-learning-based stock market prediction incorporating ESG sentiment and technical indicators. *Scientific Reports*, *14*(1). Scopus. https://doi.org/10.1038/s41598-024-61106-2

Lien Minh, D., Sadeghi-Niaraki, A., Huy, H. D., Min, K., & Moon, H. (2018). Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. *IEEE Access*, *6*, 55392–55404. Scopus. https://doi.org/10.1109/ACCESS.2018.2868970

Nguyen, B.-H., & Huynh, V.-N. (2022). Textual analysis and corporate bankruptcy: A financial dictionary-based sentiment approach. *Journal of the Operational Research Society*, *73*(1), 102–121. Scopus. https://doi.org/10.1080/01605682.2020.1784049

Ouyang, X., Zhou, P., Li, C. H., & Liu, L. (2015). Sentiment Analysis Using Convolutional Neural Network. *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, 2359–2364. https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.349

Roostaee, M. R., & Abin, A. A. (2023). Forecasting financial signal for automated trading: An interpretable approach. *Expert Systems with Applications*, *211*, 118570. https://doi.org/10.1016/j.eswa.2022.118570

Souma, W., Vodenska, I., & Aoyama, H. (2019). Enhanced news sentiment analysis using deep learning methods. *Journal of Computational Social Science*, *2*(1), 33–46. Scopus. https://doi.org/10.1007/s42001-019-00035-x

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). *Dive into Deep Learning*. Cambridge University Press.