FINTRACKER

Propuesta de Proyecto

Mateo Pérez Suarez (79439806Z)
Iñigo Peña de las Heras (20983346D)

1 de octubre de 2025

Índice

1. Abstract **2**

2. Resumen **3**

3. Análisis de los Datos **4**

4. Related Work **5**

4.0.1. Financial Event Extraction Using Wikipedia-Based Weak Supervision (Ein-Dor et al., 2019) 5

4.0.2. Extracting Fine-Grained Economic Events from Business News (Jacobs et al., 2020) 5

4.0.3. A Review of Sentiment, Semantic and Event-Extraction-Based Approaches in Stock Forecasting 6

4.0.4. SEntFiN 1.0: Entity-Aware Sentiment Analysis for Financial News 6

4.0.5. LLM-based Extraction and Summarization for Financial News . . 6

1 Abstract

Este proyecto tiene como objetivo desarrollar un sistema de detección automática de eventos financieros relevantes a partir de noticias y comunicados de distintas compañías. La motivación surge de la necesidad de analistas e inversores de procesar grandes volúmenes de información en tiempo real, identificando rápidamente hechos con impacto potencial en los mercados. Para fijar el alcance y garantizar viabilidad, el sistema se centrará en un conjunto reducido de eventos los cuales consideramos de mayor impacto: fusiones y adquisiciones, cambios de directiva y anuncios de resultados financieros.

El pipeline planteado combina varias tareas del Procesamiento de Lenguaje Natural. En primer lugar, se aplicará un clasificador de documentos que determine si un artículo contiene alguno de los eventos de interés. Posteriormente, se tratará extraer de manera automática entidades clave como nombres de empresas, directivos, fechas o cifras económicas. Con esta información, el sistema construirá una representación estructurada del evento (tipo, entidades implicadas y sus roles). Finalmente, se explorará la generación de un resumen breve o titular automático que sintetice la noticia en una sola frase clara y comprensible, facilitando su consulta rápida.

De este modo, el sistema no solo clasificará eventos, sino que también proporcionará contexto estructurado y una visión condensada de cada noticia, acercándose a una herramienta práctica para profesionales financieros.

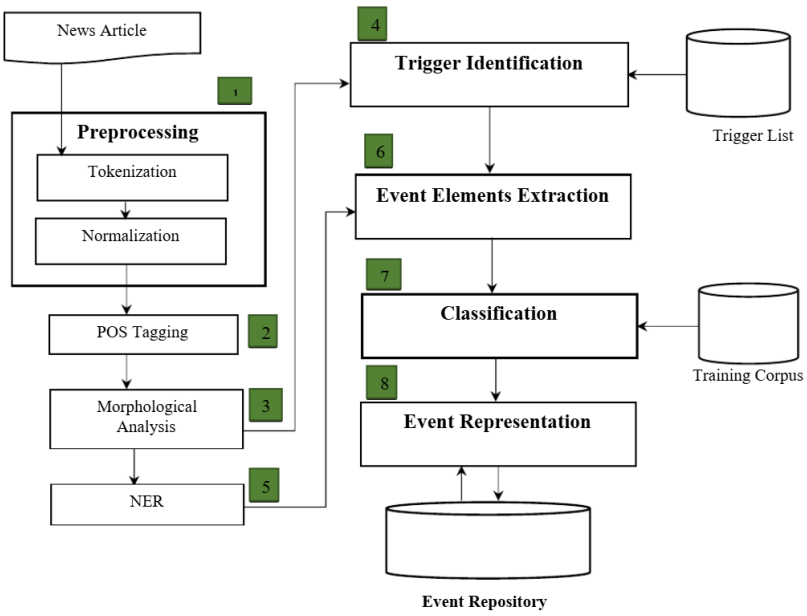

Figura 1: Arquitectura del sistema

2 Resumen

El sistema combinará varias técnicas de procesamiento de lenguaje natural. En primer lugar, mediante un clasificador de documentos identificamos si un texto contiene alguno de los eventos de interés. Para abordar esta tarea, evaluaremos el uso de modelos preentrenados en dominio financiero disponibles en Hugging Face, que pueden ser adaptados mediante fine-tuning a nuestro conjunto de clases. En concreto, hemos identificado tres candidatos relevantes:

- `leonas5555/finnews-topic-single-classify` [1]
- `fuchenru/Trading-Hero-LLM` [2]
- `farhanage/MTL-FinancialNews-Topic-Sentiment` [3]

Estos modelos pueden servir como base para nuestra tarea, aunque cada uno presenta limitaciones y diferencias en las etiquetas originales. Según el modelo elegido, podríamos tener que ajustar los tópicos de clasificación si no se pudiera entrenar con nuestras propias categorías. Esta flexibilidad nos permitirá valorar cuál se adapta mejor a los objetivos del proyecto.

La segunda funcionalidad de nuestro sistema consiste en la extracción automática de entidades relevantes dentro de las noticias financieras, como nombres de empresas, directivos, fechas o cifras económicas. Para esta tarea hemos identificado dos modelos preentrenados en Hugging Face que pueden servir como base:

- `engibeer/financial-ner-entities-bist30` [4] especializado en noticias de empresas del índice BIST30, lo que le da ventaja en dominio financiero, aunque con cobertura limitada a un mercado específico.
- `boltuix/NeuroBERT-NER` [5] modelo más general de NER basado en BERT, con mayor versatilidad, pero que requiere adaptación para reconocer entidades financieras específicas.

Una vez más, será necesario evaluar cuál se ajusta mejor a las necesidades de nuestro proyecto, ya que existen modelos capaces de extraer entidades de forma no supervisada, pero dependiendo de sus resultados, probablemente limitemos las entidades a la extraccion de la empresa mencionada y la fecha de publicacion, puesto que son datos que si venen etiquetados en nuestro dataset.

Finalmente, para la generación de un titular o resumen breve de cada noticia, podemos seguir dos caminos; podemos usar directamente alguno de los muchos modelos de summarization que ya existen (como BART, PEGASUS o T5), o probar a montar nuestro propio modelo basado en arquitecturas encoder–decoder. La idea sería tomar el texto de la noticia, representarlo con embeddings contextuales y, a partir de ahí, condensar la información en la fase de codificación y producir una versión más breve en la de decodificación. Esta tarea la podemos enfocar de dos formas: extractiva, quedándonos con las frases más relevantes, o abstractive, generando frases nuevas a partir del contenido.

Para medir la calidad de los resúmenes, utilizaremos métricas habituales como ROUGE o BLEU. La clasificación de noticias con eventos se evaluará mediante precision, recall y F1-score sobre un conjunto anotado, mientras que la extracción de entidades se validará con métricas estándar de Named Entity Recognition (como: precision, recall y F1).

3 Análisis de los Datos

Para este proyecto hemos recopilado un conjunto de noticias financieras usando la API oficial de Finnhub. Nos hemos centrado en siete compañías tecnológicas de referencia: Apple, Microsoft, Tesla, Meta, Google, Nvidia y Amazon.

El dataset resultante cuenta con unas 1.300 noticias (filas) y 9 atributos (columnas), todas publicadas en los últimos tres meses, que es el límite que permite la versión gratuita de la API.

Cada noticia del dataset incluye la siguiente información:

- **Category**: categoría general de la noticia.
- **datetime**: fecha y hora de publicación en formato UNIX.
- **headline**: titular.
- **id**: identificador único asignado por Finnhub.
- **image**: enlace a la imagen asociada (si la hay).
- **related**: tickers relacionados (a veces vacío).
- **source**: medio que publica la noticia.
- **summary**: resumen breve.
- **url**: enlace a la noticia original.

Este dataset lo estaremos generando directamente con la API de Finnhub, recopilando noticias de los últimos tres meses de las 7 empresas mencionadas.

El enlace URL de cada noticia que devuelve la API no apunta de forma directa al

artículo, sino que es un redireccionamiento a una request HTTP hacia la fuente original. Por eso, también tendremos que procesar este enlace para obtener la URL final.

El contenido de cada noticia viene resumido en la variable summary, asique con el URL directo que hemos conseguido, vamos a scrapear el contenido completo de cada noticia para trabajar textos mas extensos.

Otra limitación con la que nos hemos encontrado es que los artículos de Finnhub no incluyen etiquetas sobre el tipo de evento (fusiones, cambios de directiva, resultados, etc.), por lo que tendremos que generarlas nosotros, ya sea de forma manual o semiautomática.

4 Related Work

4.0.1. Financial Event Extraction Using Wikipedia-Based Weak Supervision (Ein-Dor et al., 2019)

Este trabajo introduce un enfoque de **supervisión débil** para la detección de eventos financieros, evitando la necesidad de grandes corpus anotados manualmente. Utilizan secciones de Wikipedia de miles de compañías para generar ejemplos de entrenamiento y entrenan clasificadores basados en **BERT**. Sus modelos alcanzan un **F1 de 0,94** en test internos y mantienen un *recall* superior al 70 % en noticias reales, superando a enfoques tradicionales. El estudio demuestra que la supervisión débil puede producir clasificadores efectivos y transferibles, aunque con menor rendimiento al pasar de Wikipedia a noticias por diferencias de estilo y cobertura limitada de eventos.

4.0.2. Extracting Fine-Grained Economic Events from Business News (Jacobs et al., 2020)

Los autores presentan el corpus **SentiVent**, anotado manualmente con diferentes tipos de eventos económicos y sus argumentos, para avanzar en la extracción de información granular en noticias financieras. Aplican modelos de clasificación y etiquetado de secuencia, incluyendo **BiLSTM-CRF** y variantes con embeddings contextuales como **BERT**, en tareas de detección de *triggers* y extracción de argumentos. Los resultados muestran **F1 de 60–70 %** en triggers, pero un rendimiento más bajo en argumentos, reflejando la dificultad de capturar relaciones semánticas finas en textos financieros. Este trabajo aporta un corpus valioso y subraya la necesidad de métodos capaces de manejar dependencias largas y contextos documentales completos.

4.0.3. A Review of Sentiment, Semantic and Event-Extraction-Based Approaches in Stock Forecasting

Este estudio revisa los enfoques de predicción financiera basados en sentimiento, análisis semántico y extracción de eventos, ofreciendo una **taxonomía comparativa** de técnicas y fuentes textuales. Se destaca que los métodos híbridos, que combinan las tres dimensiones, suelen producir mejores resultados que los enfoques aislados. La revisión señala, además, retos persistentes como la heterogeneidad de los datos, la falta de métricas de evaluación estandarizadas y la dificultad de aplicar estas técnicas en tiempo real. En conjunto, proporciona un mapa actualizado del estado del arte y marca direcciones futuras, como la creación de datasets más amplios y modelos capaces de integrar múltiples fuentes de información.

4.0.4. SEntFiN 1.0: Entity-Aware Sentiment Analysis for Financial News

El trabajo propone un análisis de sentimiento financiero a nivel de entidad, superando la limitación de enfoques globales que asignan una única etiqueta por documento. Para ello, presentan el corpus **SEntFiN 1.0**, con más de 10.700 titulares y 14.400 anotaciones entidad-sentimiento, el primero a gran escala en este dominio. Los resultados muestran que modelos contextuales como **RoBERTa** y **FinBERT** alcanzan hasta un **94 % de exactitud**, superando ampliamente a enfoques léxicos y tradicionales. La principal contribución reside en evidenciar la necesidad de análisis conscientes de entidades para captar impactos diferenciados, aunque el corpus aún se limita a titulares en inglés.

4.0.5. LLM-based Extraction and Summarization for Financial News

Este trabajo propone un pipeline de extracción estructurada que combina detección de entidades, análisis de sentimiento y **resumen automático** mediante **Large Language Models (LLMs)**. A partir de un dataset de 5.500 artículos financieros, los autores muestran que en el **90 % de los casos** los tickers detectados coinciden con los de feeds comerciales, y en un **22 % descubren entidades adicionales**. Los resúmenes generados resultan más informativos y específicos que métodos extractivos simples. El enfoque demuestra la viabilidad práctica de integrar LLMs en extracción financiera, aunque plantea desafíos de coste computacional y diseño de *prompts*. La contribución principal es mostrar cómo los LLMs mejoran cobertura y utilidad en el análisis financiero.

Paper	**Principales aportaciones / Resultados**
Financial Event Extraction Using Wikipedia-Based Weak Supervision (Ein-Dor et al., 2019)	Uso de supervisión débil con Wikipedia para generar datos de entrenamiento. Modelos basados en BERT alcanzan **F1 = 0.94** en Extended-wiki y **recall = 0.77** en noticias de 2019, superando a modelos supervisados tradicionales.
Extracting Fine-Grained Economic Events from Business News (Jacobs et al., 2020)	Creación del copus **SENTiVENT**. Evaluación con BiLSTM-CRF y BERT: detección de triggers con F1 de **60–70 %**, pero extracción de argumentos con rendimiento mucho más bajo, mostrando la dificultad del dominio financiero.
Extracting Structured Insights from Financial News: An Augmented LLM Driven Approach (Dolphin et al., 2024)	Pipeline con LLMs para extracción de entidades, sentimiento y resúmenes. Cobertura de **90 % de tickers** respecto a feeds comerciales y descubrimiento de **22 % de tickers adicionales**. Generación de resúmenes más informativos que métodos de extraccón.
SEntFiN 1.0: Entity-Aware Sentiment Analysis for Financial News (Sinha et al., 2023)	Dataset con 10.700 titulares y 14.400 anotaciones entidad-sentimiento. Modelos FinBERT y RoBERTa logran **94 % de exactitud y 93 % de F1**, muestran la eficacia de análisis de sentimiento.
A Review of Sentiment, Semantic and Event-Extraction-Based Financial Forecasting Approaches (Cheng et al., 2022)	Revisión de enfoques de predicción financiera basados en sentimiento, semántica y extracción de eventos. Se concluye que los **métodos híbridos** ofrecen mejores resultados, aunque persisten problemas de generalización y comparabilidad.

Cuadro 1: Principales aportaciones y resultados del related work

Bibliografía

[1] Hugging Face. *finnews-topic-single-classify*. Disponible en: `https://huggingface.co/leonas5555/finnews-topic-single-classify`

[2] Hugging Face. *Trading-Hero-LLM*. Disponible en: `https://huggingface.co/fuchenru/Trading-Hero-LLM`

[3] Hugging Face. *MTL-FinancialNews-Topic-Sentiment*. Disponible en: `https://huggingface.co/farhanage/MTL-FinancialNews-Topic-Sentiment`

[4] Hugging Face. *financial-ner-entities-bist30*. Disponible en: `https://huggingface.co/engibeer/financial-ner-entities-bist30`

[5] Hugging Face. *NeuroBERT-NER*. Disponible en: `https://huggingface.co/boltuix/NeuroBERT-NER`

[6] Ein-Dor, L., Gera, A., Halfon, A., ... (2019). *Financial Event Extraction Using Wikipedia-Based Weak Supervision* Disponible en: `https://arxiv.org/pdf/1911.10783`

[7] Jacobs, G., Lefever, E., Hoste, V. (2020). *Extracting Fine-Grained Economic Events from Business News* Disponible en: `https://aclanthology.org/2020.fnp-1.36.pdf`

[8] Dolphin, T., Jiang, D., et al. (2024). *Extracting Structured Insights from Financial News: An Augmented LLM Driven Approach* Disponible en: `https://arxiv.org/pdf/2407.15788`

[9] Sinha, A., Mittal, A., et al. (2023). *SEntFiN 1.0: Entity-Aware Sentiment Analysis for Financial News* Disponible en: `https://arxiv.org/pdf/2305.12257`

[10] Cheng, M., Xu, Y., et al. (2022). *A Review of Sentiment, Semantic and Event-Extraction-Based Financial Forecasting Approaches* Disponible en: `https://www.mdpi.com/2227-7390/10/14/2437`