UNIVERSIDAD EAFIT®

Procesamiento de lenguaje natural aplicado a reportes financieros.

David Trefftz
Lucas Martines
Johann Ruiz

Línea de énfasis de ciencia de datos.

31/05/2024

UNIVERSIDAD EAFIT®

Introducción.

En el entorno financiero actual, los analistas enfrentan un desafío significativo al tratar de procesar y evaluar la enorme cantidad de información contenida en los informes anuales 10-K de las empresas del S&P 500. Este proceso no solo es laborioso y requiere mucho tiempo, sino que también puede retrasar la toma de decisiones críticas para las inversiones. Ante esta problemática, nuestro equipo ha desarrollado un proyecto que implementa un algoritmo avanzado de procesamiento de lenguaje natural (PLN) para analizar automáticamente estos informes, optimizando así la eficiencia y precisión del análisis financiero.

El objetivo principal de este proyecto es transformar la forma en que los analistas financieros acceden y procesan la información crítica contenida en los informes 10-K, proporcionando una herramienta que no solo reduce drásticamente el tiempo y esfuerzo necesarios, sino que también mejora la calidad y profundidad del análisis. Esto permitirá a los analistas enfocarse en el aprovechamiento estratégico de la información y tomar decisiones informadas de manera más rápida y precisa.

Nuestro enfoque utiliza técnicas avanzadas de PLN para clasificar los informes en términos de percepciones positivas, negativas o neutras, y correlacionar estos sentimientos con indicadores financieros clave como el Valor Económico Agregado (EVA). Esta metodología no solo verifica la veracidad y relevancia de las conclusiones del análisis, sino que también fortalece la confianza en las decisiones estratégicas basadas en estos datos.

En esta propuesta se detallan los problemas a resolver, la solución propuesta, el impacto esperado, el marco teórico, la descripción de los datos a utilizar y la metodología a emplear. Asimismo, se presentan los modelos y técnicas a utilizar en cada una de las materias del proyecto integrador, así como un cronograma de entregables y las referencias bibliográficas utilizadas para sustentar nuestra investigación.

Con esta solución innovadora, pretendemos catalizar una revolución en el análisis financiero, marcando el comienzo de una era en la que los analistas pueden anticipar y

responder con mayor precisión y rapidez a las tendencias del mercado, potenciando así el éxito de las inversiones.

Marco teórico

Análisis de sentimiento en finanzas:

El análisis de sentimientos en finanzas ha sido un área activa de investigación, utilizando métodos que van desde el uso de diccionarios especializados hasta modelos de aprendizaje profundo.

Se ha demostrado que los métodos basados en diccionarios, como el diccionario Loughran-McDonald (LM), son eficaces para extraer opiniones de textos financieros, pero requieren extensas anotaciones manuales.

Los modelos de aprendizaje profundo como Transformer son populares por su rendimiento superior, pero requieren grandes cantidades de datos y recursos informáticos. - El método XLex propuesto (léxicos explicados) combina las ventajas de los métodos basados en diccionarios y los modelos transformadores que utilizan explicaciones complementarias SHApley (SHAP) para mejorar la interpretación y la eficacia del análisis de sentimiento en textos financieros.

El PNL en finanzas:

Las técnicas de procesamiento del lenguaje natural (PNL) se han generalizado en el sector financiero, pero enfrentan desafíos particulares en términos de solidez e interpretación.

Los avances recientes en grandes modelos de lenguaje como ChatGPT brindan nuevas oportunidades para mejorar las aplicaciones financieras basadas en PNL.

El taller propuesto sobre PNL Robusta en Finanzas (RobustFin) se centra en abordar los desafíos de solidez e interpretabilidad de la PNL para aplicaciones financieras, reconociendo la necesidad de cumplir con estándares regulatorios y operativos específicos de la industria financiera.

Modelos de clasificación en finanzas:

Los modelos de clasificación desempeñan un papel crucial a la hora de recomendar información financiera, pronosticar el mercado de valores y detectar fraude financiero.

Los modelos de clasificación de clase única han demostrado ser eficaces para recomendar información dentro de una industria del sector financiero, superando los problemas de aislamiento de la industria y sobrecarga de información.

La combinación de firmas financieras y análisis de sentimiento utilizando modelos híbridos como redes neuronales recurrentes (HyRNN) ha demostrado un gran potencial para mejorar la precisión de las previsiones de precios de las acciones y la detección de fraudes.

Desarrollo metodológico.

1. **Recopilación y preparación de datos:**

Recopilación de datos:
Los informes 10-K se obtuvieron a través de la base de datos financiera y económica de BLOOMBERG. Cada informe se descargó en formato PDF y se le asignó un valor numérico antes del título para etiquetar el documento. Posteriormente, se reunieron 110 documentos en un archivo CSV para ser procesados. Este archivo contiene todos los informes 10-K, y un segundo archivo CSV contenía el rendimiento respectivo de cada empresa. Esta estructura facilita la concatenación de cada etiqueta y su correspondiente rendimiento.

Limpieza de Datos:
Se comenzó tokenizando cada informe y eliminando palabras inusuales, URL y números. Se añadió un filtro que solo incluía palabras del diccionario en inglés utilizando la biblioteca Natural Language Toolkit (NLTK). Después, se eliminaron las palabras vacías en inglés, seguidas de la lematización y derivación de los tokens restantes. Estas técnicas permitieron estandarizar el texto y reducirlo a sus componentes esenciales, facilitando su análisis posterior.

2. **Vectorización y Reducción de Dimensionalidad:**

   Vectorización:
   Se realizó el proceso de vectorización de los documentos utilizando TF-IDF (Term Frequency-Inverse Document Frequency). La matriz TF-IDF resultante tenía una dimensionalidad de 110 filas por 11,922 columnas, donde las filas representan cada documento y las columnas representaban cada token convertido en un valor numérico.

   Embeddings:
   Además de la vectorización, se implementaron Representaciones de Codificador Bidireccional de Transformadores (BERT) para crear embeddings para los informes 10-K. Estos embeddings se guardaron y se importaron posteriormente para los modelos de clasificación.

   Reducción de Dimensionalidad:
   Para facilitar la visualización de datos y verificar la presencia de valores atípicos, se implementaron dos algoritmos de reducción de dimensionalidad: PCA (Análisis de Componentes Principales) y UMAP (Uniform Manifold Approximation and Projection). Se creó un diagrama de dispersión con PCA utilizando dos dimensiones y las etiquetas de los datos, y luego se utilizó UMAP para otra forma de visualizar los datos. Las distancias calculadas (Euclidiana, Manhattan y Chebyshev) mostraron que PCA y la distancia Euclidiana arrojaron los mejores resultados.

3. **Detección y Eliminación de Valores Atípicos:**
   Se utilizó un mapa de calor para visualizar las distancias entre documentos y establecer un umbral calculado con la media y la desviación estándar de las distancias. Si un documento estaba lo suficientemente lejos de la media, se elimina del conjunto de datos. Con este método, se eliminaron 7 valores atípicos, mejorando así la consistencia y calidad del conjunto de datos.

4. **Modelado y Evaluación:**

   Modelos de Aprendizaje Automático:

Se implementaron varios modelos de clasificación para predecir la métrica de rendimiento basada en la matriz TF-IDF o el embedding, incluyendo Regresión Logística, clasificación de vector de soporte lineal, Support Vector Machines (SVM), Random Forest y Gradient Boosting. Cada modelo se evaluó utilizando validación cruzada por k-folds (10 grupos), donde el conjunto de datos se dividió en 10 grupos con TF-IDF y 5 para embeddings, se entrenó cada modelo con nueve de los grupos y se utilizó el último para probar. Este proceso se repitió diez veces, devolviendo el promedio de precisión, recall y puntuaciones F1 para cada modelo, y generando la matriz de confusión para los resultados.

Los mejores resultados utilizando TF-IDF se encontraron en el modelo de regresión logística, con una precisión del 68.2%, recall del 63.9% y puntuación F1 del 61.6%. Con embeddings, el mejor modelo fue Support Vector Machines, con una precisión en promedio del 61%, recall de 72% y F1 score de 63%.

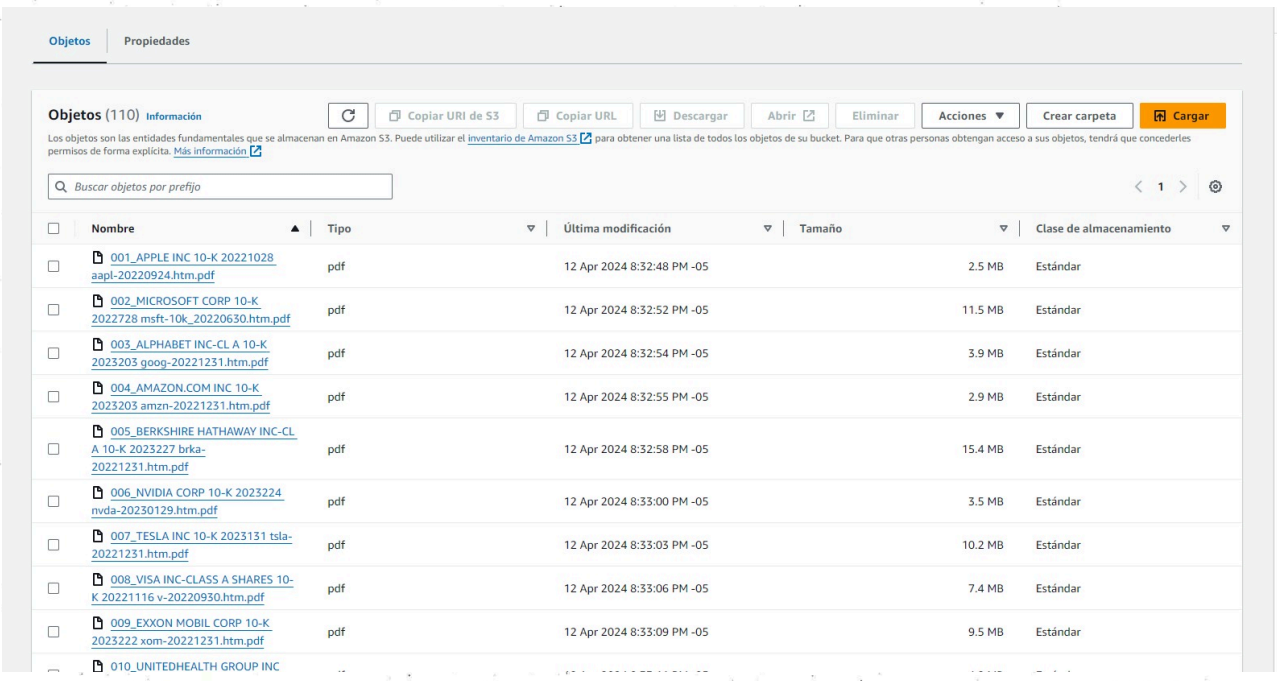
5. **Implementación y Monitoreo:**

Despliegue del Modelo:
El modelo de análisis de sentimientos se implementó en un entorno de producción, asegurando su integración con sistemas existentes. Se estableció un proceso de monitoreo continuó para supervisar el rendimiento del modelo y realizar actualizaciones o ajustes según sea necesario, garantizando su eficacia y relevancia en el tiempo.

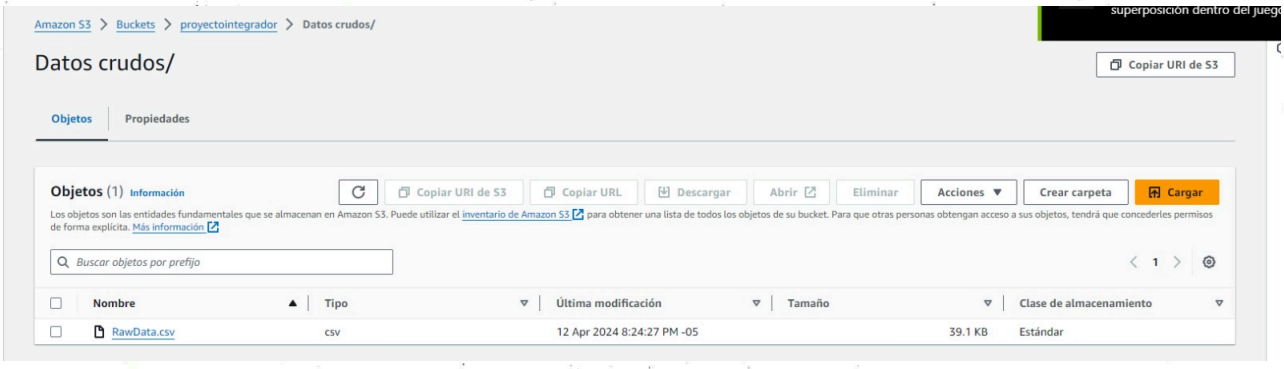
6. **Análisis y conclusiones:**
En conclusión, el equipo logró obtener buenos resultados, que podrían mejorarse con el crecimiento de la base de datos. Además, se observa y recalca el potencial del procesamiento de lenguaje natural aplicado a reportes financieros, ya que, si se mejora, reduciría considerablemente el tiempo para la toma de decisiones, lo cual en el mundo financiero se traduce en una mayor rentabilidad.

7. Tecnología:

La ingesta de datos se realizó de forma manual, donde el equipo descargo de manera individual cada reporte 10k de la plataforma BLOOMBERG. A cada reporte se le asignó un valor numérico para la organización de los documentos. Al tener un tamaño de muestra considerable, se almacenó de forma masiva los pdf en un data lake en AWS S3.



Para los datos crudos, se pudo realizar una búsqueda en conjunto de el ticker de la empresa, su nombre, descripción, WACC(Weighted Average Cost of Capital), EVA(Economic Value Added) y Market Cap utilizando igualmente la plataforma bloomberg. Este archivo inicialmente se descargó cómo xlsx, pero se transformó en CSV y se almacenó dentro del mismo bucket en S3.



Al ser importados en conjunto, la naturaleza de la importación de datos se puede considerar de forma **batch**. La base de datos es **no estructurada,** ya que contiene documentos pdf, los cuales contienen imágenes, y textos cuya longitud varía en gran cantidad.

Para procesar la información se utilizó Google Colab. Donde, al importar los documentos del bucket de S3, se pudo condensar todos los reportes dentro de un csv a través de un proceso de limpieza donde se removieron las imágenes y se conservó únicamente el texto de cada reporte y tu título, los cuales luego serían catalogados. De igual manera, los datos crudos fueron importados y catalogados.

Al ofrecer una amplia selección de librerías, Python permitió que el equipo usara herramientas de visualización como Matplotlib, Seaborn, ggplot y Plotly. Estas librerías facilitaron la exploración de datos de forma bidimensional, ver cómo cambió el dataset dependiendo de las herramientas de reducción de dimensionalidad y analizar los resultados de los modelos de clasificación.

8. Conclusiones

Se realizó un proceso riguroso de ingesta, procesamiento y visualización de datos. El principal desafío encontrado fue la ingesta de los reportes financieros, que no pudieron ser importados de forma masiva y debieron ser categorizados manualmente para un procesamiento adecuado. Este inconveniente obligó al equipo a trabajar con una base de datos más pequeña de lo deseado, lo cual dificultó la obtención de resultados que escalaran adecuadamente con el incremento del tamaño del dataset.

A pesar de estas limitaciones, los resultados obtenidos en los modelos de clasificación son prometedores. Los modelos demostraron una capacidad considerable para predecir métricas de rendimiento financiero utilizando tanto TF-IDF como embeddings de BERT. La Regresión Logística se destacó con TF-IDF, mientras que los Support Vector Machines mostraron mejores resultados con embeddings.

Este estudio resalta el potencial del procesamiento de lenguaje natural aplicado a reportes financieros. Si se mejora y amplía la base de datos, los modelos podrían proporcionar una herramienta valiosa para la toma de decisiones en el sector financiero, reduciendo significativamente el tiempo necesario para analizar grandes volúmenes de información y, en última instancia, aumentando la rentabilidad.

El equipo sugiere que futuras investigaciones se enfoquen en la automatización del proceso de ingesta de datos y en la expansión de la base de datos. Además, se podrían explorar técnicas más avanzadas de procesamiento de lenguaje natural y modelos híbridos para mejorar aún más la precisión y robustez de las predicciones. Con estas mejoras, se espera que los modelos de clasificación en finanzas puedan ofrecer recomendaciones y pronósticos aún más precisos y útiles para los profesionales del sector.

UNIVERSIDAD EAFIT®

Referencias bibliográficas

M. Rizinski, H. Peshov, K. Mishev, M. Jovanovik and D. Trajanov, "Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (XLex)," in IEEE Access, vol. 12, pp. 7170-7198, 2024, doi: 10.1109/ACCESS.2024.3349970.

JOHN, A.; LATHA, T. Stock market prediction based on deep hybrid RNN model and sentiment analysis. Automatika: Journal for Control, Measurement, Electronics, Computing & Communications, [s. l.], v. 64, n. 4, p. 981–995, 2023. DOI 10.1080/00051144.2023.2217602.

DAY, M.-Y.; LEE, C.-C. Deep learning for financial sentiment analysis on finance news providers. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on, [s. l.], p. 1127–1134, 2016. DOI 10.1109/ASONAM.2016.7752381.

SHAH, S. et al. Robust NLP for Finance (RobustFin). Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, [s. l.], p. 5884–5885, 2023. DOI 10.1145/3580305.3599211.

ADAM ZAREMBA; ENDER DEMIR. ChatGPT: Unlocking the future of NLP in finance. Modern Finance, [s. l.], v. 1, n. 1, 2023. DOI 10.61351/mf.v1i1.43.

CHANG, J.-W.; HUNG, J. C.; YEN, N. Design of a NLP-empowered finance fraud awareness model: the anti-fraud chatbot for fraud detection and fraud classification as an instance. Journal of Ambient Intelligence and Humanized Computing, [s. l.], v. 13, n. 10, p. 4663–4679, 2022. DOI 10.1007/s12652-021-03512-2

JIANG, Y.; WANG, H.; XIE, Q. Classification model of companies' financial performance based on integrated support vector machine. 2009 International Conference on Management Science and Engineering, Management Science and Engineering, 2009. ICMSE 2009. International Conference on, [s. l.], p. 1322–1328, 2009. DOI 10.1109/ICMSE.2009.5318030