6. Generación de resúmenes extractivos por medio de grafos de oraciones ponderadas en el sector financiero

Andrés Felipe Hernández Giraldo*
Cristian David Rocha Fernández**
Juan Carlos Montes Estrada***

DOI: https://doi.org/10.52501/cc.349.06

Resumen

El proyecto "Generación de resúmenes extractivos por medio de grafos de oraciones ponderadas en el sector financiero" tiene como objetivo principal desarrollar y evaluar un modelo que permita la generación automática de resúmenes extractivos para documentos financieros, utilizando grafos de oraciones ponderadas para mejorar la eficiencia y precisión en la síntesis de información relevante.

En primer lugar, se diseñó un algoritmo basado en grafos de oraciones ponderadas que selecciona las oraciones más significativas de los documentos, optimizando la construcción de resúmenes concisos. El preprocesamiento del texto fue un paso crucial en este proceso, asegurando que los corpus fueran limpiados y estructurados adecuadamente para facilitar la identificación y priorización de la información relevante. Este paso inicial fue fundamental para reducir la variabilidad y simplificar el procesamiento posterior, asegurando que el modelo pudiera reflejar con precisión los resultados clave.

* Magíster en Ciencia de Datos. Profesor en la Universidad Pontificia Bolivariana, Colombia. ORCID: https://orcid.org/0000-0002-7976-5326
** Magíster en Ciencia de Datos por la Pontificia Universidad Javeriana de Cali, Colombia. ORCID: https://orcid.org/0000-0001-5508-6452
*** Magíster en Ciencia de Datos de la Pontificia Universidad Javeriana de Cali, Colombia. ORCID: https://orcid.org/0009-0005-6458-170X

En segundo lugar, el modelo fue implementado y entrenado en un conjunto de datos específicos del sector financiero. Utilizando estructuras de grafos, se logró una visualización eficaz y estructurada de las relaciones entre oraciones, lo que permitió detectar comunidades de contenido relacionado y mejorar significativamente la coherencia y precisión de los resúmenes, generando así una claridad y comprensión de dichos resúmenes.

Palabras clave: *grafos, resúmenes, documentos financieros, procesamiento de lenguaje natural, extractivo.*

Introducción

En la última década, diversos estudios han explorado enfoques basados en procesamiento del lenguaje natural (NLP) y técnicas de aprendizaje automático *(Machine Learning [ML])* para mejorar la calidad de los resúmenes generados automáticamente (Nallapati et al., 2016). A pesar de los avances logrados, persisten desafíos significativos debido a la complejidad inherente a la comprensión del significado contextual y la relevancia de la información en diversos contextos. Las técnicas mencionadas buscan enfrentar el desafío de crear resúmenes precisos y coherentes, especialmente en documentos complejos y ricos en información, como los textos financieros; sin embargo, aún persisten dificultades considerables relacionadas con la comprensión contextual y la relevancia de la información, lo que crea una brecha en la capacidad para generar resúmenes automáticos que capturen de manera efectiva los puntos clave.

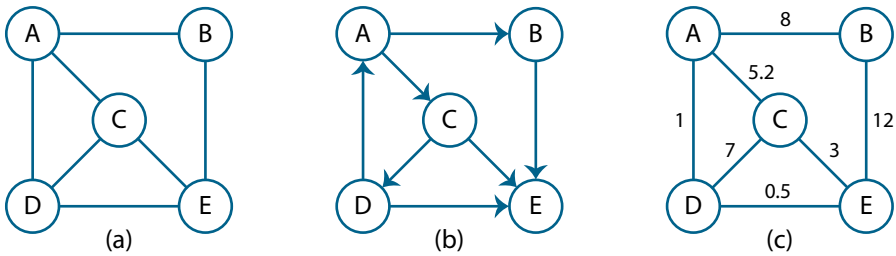
El concepto de resumen implica un proceso metodológico conocido como condensación. Se trata de una acción sobre el texto del documento para reducir la cantidad de información contenida en él y resaltar las partes más importantes del mensaje que sean convenientes para el usuario. Los principales objetivos del resumen son informar de forma completa y precisa a los usuarios sobre el contenido de los documentos proporcionando elementos cruciales que permiten al usuario decidir si desea ver o descartar el documento completo. Además, el resumen es útil para la recuperación

de información mediante palabras clave (Moreiro, 1988). La intención es que estos resúmenes no solo reflejen con precisión el contenido relevante, sino que también destaquen datos cruciales en cifras, optimizando así la interpretación y toma de decisiones en el ámbito financiero.

En respuesta a estas limitaciones, este estudio explora la técnica de grafos como una solución para la generación de resúmenes extractivos de documentos financieros. La metodología basada en grafos no solo permite representar de manera estructurada las relaciones entre diferentes oraciones, sino que también facilita la identificación de patrones y la extracción de información clave (Gómez Adorno, 2018).

La teoría de grafos, como base de este proyecto, permite modelar relaciones entre elementos a través de nodos y enlaces. Un grafo se compone de un conjunto de nodos (o vértices) conectados por aristas (enlaces), y puede ser dirigido o no dirigido, según si las aristas tienen una dirección específica o no (Mihalcea y Radev, 2011) como se muestra en la figura 6.1.

Figura 6.1. *Esquema de un grafo: (a) Dirigido, (b) No dirigido, (c) No dirigido ponderado*



Fuente: Mihalcea y Radev (2011).

Los grafos dirigidos son clave en este proyecto, ya que permiten representar relaciones asimétricas entre oraciones en un documento, lo que es fundamental para identificar y priorizar información relevante en los resúmenes. La unidireccionalidad de los grafos dirigidos facilita el análisis de la secuencia y dependencia contextual entre las oraciones, permitiendo una extracción precisa de la información clave en documentos financieros.

Existen varios desafíos técnicos en la representación de grafos y la optimización de los algoritmos para la generación de resúmenes coherentes y relevantes. A pesar de los avances en NLP y ML, la selección de la mejor re-

presentación de grafos y la adecuada ponderación de las oraciones siguen siendo problemas abiertos. En este sentido, este trabajo busca llenar esa brecha mediante el desarrollo de un modelo de resúmenes extractivos que utilice grafos de oraciones ponderadas para mejorar la eficiencia y precisión de la síntesis de información financiera.

El objetivo de este proyecto es desarrollar un modelo automático de generación de resúmenes extractivos, utilizando grafos de oraciones ponderadas, para mejorar la eficiencia y precisión en la síntesis de información clave en documentos financieros. Este avance facilitará la toma de decisiones en el sector financiero y optimizará el análisis de datos, lo que contribuirá a un mejor uso de los recursos debido a que se pueden detectar tendencias, patrones y señales de alerta temprana, lo cual permite tomar acciones preventivas o correctivas (Fornero, 2017).

El proyecto está alineado con el Objetivo de Desarrollo Sostenible (ODS) número 9: Industria, Innovación e Infraestructura, promovido por la ONU (2015), ya que impulsa la innovación tecnológica y mejora los procesos de análisis de datos en el sector financiero. Al proporcionar resúmenes automatizados más eficientes, se optimiza el tiempo y los recursos empleados en la toma de decisiones, lo que no solo mejora la productividad, sino que también contribuye a la conservación de recursos y la reducción del impacto ambiental. Esto es particularmente relevante en un entorno empresarial donde la precisión en el análisis financiero puede evitar inversiones insostenibles.

Este trabajo tiene el potencial de transformar la forma en que se manejan grandes volúmenes de información financiera, contribuyendo a un entorno más equitativo y sostenible. La optimización de recursos a través de la generación de resúmenes automáticos no solo mejora la eficiencia operativa, sino también puede generar un impacto positivo en la sostenibilidad ambiental y social al promover decisiones más informadas y responsables.

Metodología

Introducción a la metodología

En este estudio se empleó un enfoque mixto combinando técnicas de análisis cualitativas y cuantitativas, para desarrollar un modelo automatizado de generación de resúmenes extractivos, mediante el uso de algoritmos de grafos. El objetivo principal es preservar los conceptos más significativos y coherentes de los documentos financieros, ofreciendo una herramienta eficaz para la síntesis de información clave en el sector financiero. Este enfoque fue elegido debido a la necesidad de mejorar la comprensión y análisis de grandes volúmenes de información financiera, optimizando la toma de decisiones.

Diseño del estudio

El diseño de este estudio es de naturaleza mixta, combinando tanto la recolección de datos cuantitativos como el análisis cualitativo de las oraciones en los documentos. Para la parte cuantitativa, se trabajó con datos financieros estructurados en forma de textos que fueron procesados mediante algoritmos de grafos. El estudio no incluyó grupos de control, sino que se centró en la implementación y evaluación del modelo propuesto, analizando su eficacia en la extracción de información clave. La duración del estudio fue de 12 meses, tiempo en el cual se ajustaron los parámetros del algoritmo y se realizaron evaluaciones periódicas de los resultados generados por el sistema.

Métodos utilizados

Técnicas cualitativas: se realizó un análisis cualitativo a los textos origen consistente en un preprocesamiento de ellos para asegurar que los resúmenes preservaran el sentido y los conceptos más importantes de los documentos financieros.

Técnicas cuantitativas: se procesaron documentos financieros utilizando un algoritmo basado en grafos dirigidos. Este algoritmo mapea las oraciones como nodos y las relaciones entre ellas como aristas, ponderando cada nodo según su relevancia en el documento. Para determinar la relevancia de las oraciones, se aplicaron métricas como la frecuencia de términos y la similitud semántica entre los nodos. Los datos resultantes fueron analizados estadísticamente para evaluar la precisión y coherencia de los resúmenes generados.

Algoritmos de grafos: el modelo se basa en la teoría de grafos, donde cada oración del texto es representada como un nodo y las relaciones entre oraciones como aristas. Se utilizaron grafos dirigidos para representar la secuencia y dependencias entre las oraciones. El algoritmo pondera las oraciones en función de su importancia, identificando aquellas que deben ser incluidas en el resumen final (Cormen et al., 2009).

Proceso de validación y análisis de datos: para validar el modelo, se seleccionó una muestra de 50 documentos financieros de un corpus de 18 755. Los resúmenes generados fueron comparados con resúmenes elaborados manualmente, midiendo la coherencia y precisión de los resultados.

Esta combinación de enfoques cuantitativos y cualitativos asegura que el modelo no solo sea preciso en términos de selección de información, sino que también sea útil y aplicable en contextos financieros reales.

Limitaciones del estudio

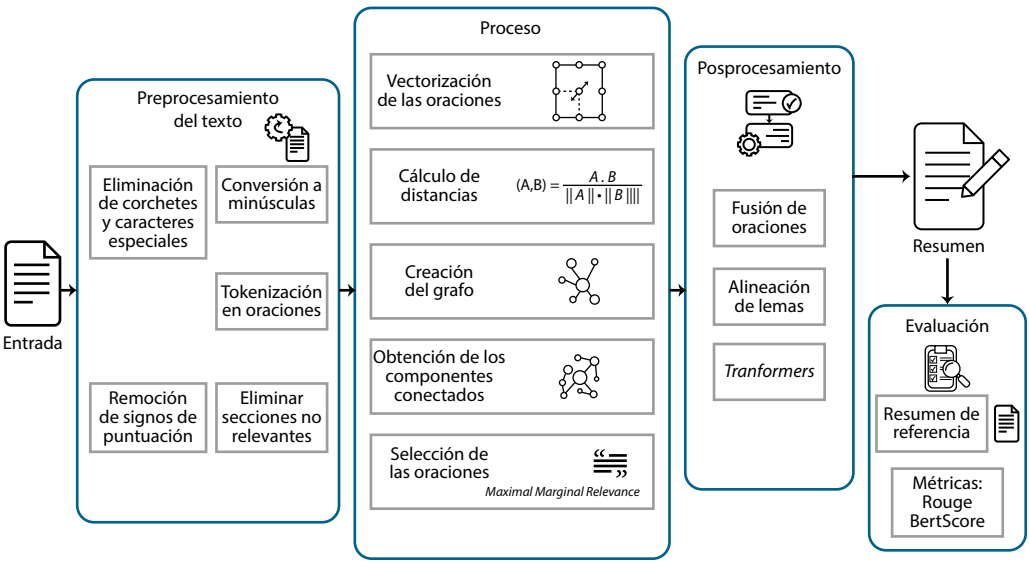
A pesar de los resultados positivos, el estudio presenta algunas limitaciones metodológicas. En primer lugar, el corpus utilizado estaba restringido a transcripciones en inglés, lo que limita la generalización del sistema a otros idiomas y contextos financieros. Además, aunque las métricas automáticas como ROUGE y BertScore son útiles para evaluar la calidad de los resúmenes, una validación más profunda por parte de expertos financieros podría proporcionar una mejor evaluación de la utilidad práctica del sistema.

Para futuras investigaciones, se sugiere ampliar el corpus a diferentes tipos de documentos financieros, como informes de mercado y estados financieros anuales, así como extender su aplicación a otros idiomas. También

se recomienda explorar el uso de modelos de aprendizaje profundo más avanzados, como *transformers* (BERT o T5), para mejorar la generación abstractiva de resúmenes. Adicionalmente, sería relevante evaluar la utilidad de los resúmenes en la toma de decisiones empresariales a través de la retroalimentación de expertos financieros. Por último, se propone la implementación de *embeddings* contextuales y técnicas híbridas para optimizar la selección de oraciones y mejorar la coherencia de los resúmenes generados.

Resultados y discusión

Figura 6.2. *Modelo desarrollado*



Nota: pasos desarrollados durante el proyecto.
Fuente: Elaboración propia.

El sistema desarrollado para la generación automática de resúmenes se basa en la técnica de grafos de oraciones ponderadas, seleccionada por su capacidad de capturar de manera precisa las relaciones semánticas entre las oraciones de un texto financiero. Este sistema se ilustra en la figura 6.2, que

detalla la arquitectura utilizada para identificar y combinar las oraciones más relevantes en un resumen coherente y fluido. El corpus utilizado proviene de transcripciones financieras de llamadas de resultados *earnings calls,* con 18 755 documentos que incluyen información clave sobre el desempeño financiero de las empresas (Chen, s/f), estos documentos están escritas en el idioma inglés, distribuidas en un rango temporal que abarca varios años.

Preprocesamiento de datos

El preprocesamiento del texto fue esencial para garantizar la calidad de los resúmenes generados. Se eliminó información irrelevante (corchetes, signos de puntuación) y se descartaron secciones como preguntas y respuestas que no aportaban valor al resumen Thanaki (2017). Este paso redujo la complejidad del corpus, enfocando el análisis en las partes más significativas de las transcripciones. En la figura 6.3, se muestra un fragmento de las oraciones resultantes tras este proceso, que sirvieron como base para la construcción del grafo.

Figura 6.3. *Fragmento de oraciones obtenidas después del preprocesamiento de un texto*

```
Texto 1 — 201 Oraciones:
Oración 1:  operator hello everyone and thank you for standing by and welcome to the brooks aut
Oración 2: now during todays conference all telephone participants will be in a listenonly mode
Oración 3: now a quick reminder todays call is being recorded its august 1st 2019
Oración 4: it is now my pleasure to turn the call over to mark namaroff director of investor re
Oración 5: please go ahead sir
Oración 6: mark namaroff  director of investor relations thank you david and good afternoon eve
Oración 7: we would like to welcome you to our third quarter earnings conference call for brook
Oración 8: our earnings press release was issued after the close of the market today and is ava
Oración 9: i would like to remind everyone that during the course of the call we will be making
Oración 10: there are many factors that may cause actual financial results or other events to d
Oración 11: i would refer you to the section of our earnings release titled safe harbor stateme
Oración 12: we make no obligation to update these statements should future financial data or ev
Oración 13: in addition we may refer to a number of nongaap financial measures which are used i
Oración 14: we believe that nongaap financial measures provide an additional way of viewing asp
Oración 15: but when considered with gaap financial results and the reconciliation of gaap meas
```

Nota: Ejemplo capturado del corpus.
Fuente: Elaboración propia.

Proceso de selección de oraciones

El siguiente paso fue la construcción del grafo, donde las oraciones se representan como nodos, y las relaciones de similitud entre ellas, como aristas. El algoritmo Maximal Marginal Relevance (MMR) fue elegido para eliminar la redundancia y asegurar la relevancia en la selección de las oraciones más representativas dentro de cada componente del grafo *(vid.* figura 6.4). Este proceso permitió seleccionar una oración representativa por cada grupo de oraciones relacionadas, manteniendo un equilibrio entre diversidad y relevancia en los resúmenes generados.

Para mejorar la identificación de las oraciones clave, se empleó la técnica de TF-IDF (Term Frequency-Inverse Document Frequency). Este es un método estadístico que mide la relevancia de una palabra en un documento dentro de un corpus. Se calcula en función de dos componentes:

a) Frecuencia de términos (TF): indica cuántas veces aparece una palabra en un documento específico. Las palabras que aparecen con mayor frecuencia tienen un mayor peso en ese documento.
b) Frecuencia inversa de documentos (IDF): reduce el peso de palabras comunes (como artículos y preposiciones) que aparecen en múltiples documentos del corpus. Las palabras frecuentes, en muchos documentos tienen un peso menor, mientras que las más específicas a un documento reciben mayor importancia.

En el contexto de este proyecto, TF-IDF permitió identificar las palabras más relevantes en las oraciones de cada documento financiero, otorgando un mayor peso a las que son significativas en el texto, pero no comunes en el resto del corpus. Esta técnica ayudó a optimizar el proceso de selección de oraciones, ya que aquellas con términos de mayor peso fueron priorizadas en el resumen final. En la tabla 6.1, se muestran ejemplos de oraciones seleccionadas, demostrando cómo TF-IDF contribuyó a mejorar la precisión y relevancia de los resúmenes generados.

Tabla 6.1. *Oraciones correspondientes a cada nodo del grafo*

Índice del nodo	*Oración asociada*
1	it is now my pleasure to turn the call over to mark namaroff director of investor relations
2	mark namaroff director of investor relations thank you david and good afternoon everyone on the line today
3	stephen s schwartz president and chief executive officer thank you mark and good afternoon to everyone on the call with us today
…	…
55	lets turn to slide 9 and look at the guidance for our fourth quarter fiscal fourth fiscal quarter of 2019
56	we paid out 7 million in dividends to shareholders in the quarter and finished the quarter by adding 20 million of net cash to the balance sheet to arrive at a total of 160 million of cash and marketable securities
57	i can net this out to say that as we entered the fourth quarter we have approximately 215 million of cash and cash equivalents available for use in operations and investments while only carrying 52 million of gross debt

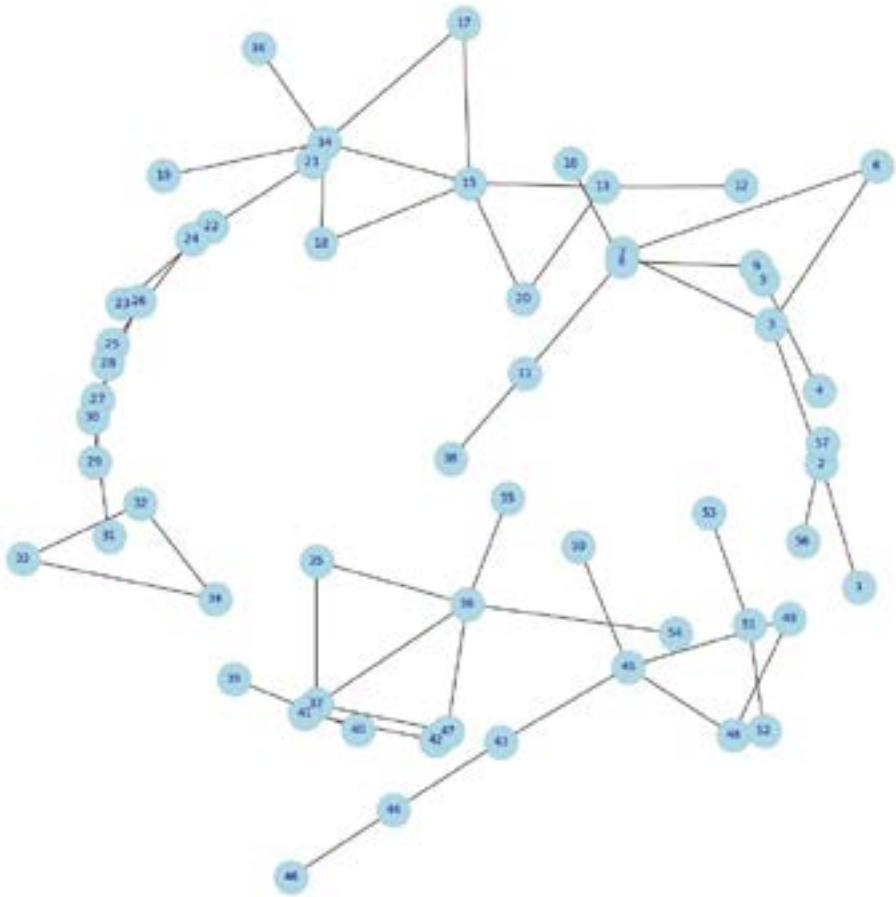
Nota. Fragmento de oraciones seleccionadas por técnica TF-IDF.
Fuente: Elaboración propia.

Esquema del grafo de oraciones

En la figura 6.4, se muestra un grafo de ejemplo que contiene 57 nodos y 49 aristas, ilustrando las relaciones entre las oraciones del texto de entrada. La técnica de grafos dirigidos permitió capturar las dependencias entre las oraciones, lo cual es fundamental para representar la estructura secuencial y lógica del texto.

En la tabla 6.1, se presentan las oraciones asociadas a cada nodo, lo que demuestra la capacidad del sistema para identificar las partes más importantes de un documento financiero.

Figura 6.4. *Esquema general del grafo de oraciones*



Nota: Grafo obtenido que contiene 57 nodos y 49 aristas.
Fuente: Elaboración propia.

Posprocesamiento y evaluación

Tras la selección de las oraciones más relevantes mediante el algoritmo MMR, se procedió a unirlas para formar un resumen coherente y fluido, buscando no solo una simple colección de oraciones, sino que presentara una narrativa comprensible. El posprocesamiento permitió combinar las oraciones seleccionadas de manera fluida, asegurando que el resumen final fuera le-

gible y lógico (Marsi y Krahmer, 2005). La evaluación de los resúmenes generados se realizó utilizando métricas automáticas como ROUGE y BertScore, estos métodos son ampliamente utilizados para medir la calidad en cuanto a semántica y se realizó la comparación los resultados con resúmenes de referencia (Zhang et al., 2020). En la tabla 6.2 se puede contrastar los resultados de estas métricas donde indicaron un alto grado de similitud semántica y coherencia entre los resúmenes generados automáticamente y los de referencia, lo que valida la eficacia del sistema.

Tabla 6.2. *Resultados experimentales para Rouge y BertScore*

Métrica	***Rouge-1***			***Rouge-2***			***Rouge-3***			***BertScore***		
Método de vectorización	R	P	F	R	P	F	R	P	F	R	P	F
Term Frequency-Inverse Document Frequency	0.35	0.17	0.23	0.12	0.05	0.07	0.29	0.15	0.20	0.54	0.51	0.52
Media	0.31	0.21	0.22	0.09	0.05	0.06	0.24	0.17	0.19	0.51	0.50	0.51
Desviación estándar	0.10	0.05	0.08	0.04	0.01	0.01	0.08	0.04	0.03	0.07	0.02	0.05

Notas: Las letras *R, P* y *F* corresponden a las métricas de *recall* (recuperación), *precision* (precisión) y *F1-score,* respectivamente.
Fuente: Elaboración propia.

Los resultados obtenidos mostraron un alta grado de similitud semántica y coherencia entre los resúmenes generados automáticamente y los de referencia, lo que valida la eficiencia del sistema en la generación de resúmenes precisos y relevantes. Además, el uso de TF-IDF en el proceso de selección de oraciones destacó la capacidad de identificar palabras clave relevantes en el texto de referencia, capturando coincidencias exactas con eficacia. Esto subraya su utilidad en contextos donde la precisión en la identificación de términos específicos es crucial.

Análisis de resultados en el contexto del desarrollo sostenible

El uso de resúmenes automáticos en el análisis financiero tiene implicaciones significativas para el desarrollo sostenible. La optimización de la extracción de información clave reduce el tiempo y los recursos necesarios para

analizar grandes volúmenes de datos financieros. Esto no solo mejora la eficiencia operativa en las empresas, sino que también promueve la conservación de recursos al minimizar el esfuerzo humano en tareas repetitivas, contribuyendo al ODS 9 (Industria, Innovación e Infraestructura).

Asimismo, la capacidad de generar resúmenes precisos a partir de grandes conjuntos de datos financieros fomenta la equidad económica, al permitir que tanto grandes empresas como pequeños inversionistas accedan rápidamente a información relevante, alineándose con el ODS 10 (Reducción de las desigualdades). Este acceso equitativo a información crítica puede ayudar a tomar decisiones financieras más informadas y sostenibles, reduciendo riesgos e inversiones insostenibles.

Conclusiones

La generación automática de resúmenes extractivos mediante el uso de grafos de oraciones ponderadas ha demostrado ser una herramienta eficiente y precisa para procesar grandes volúmenes de datos financieros. Este enfoque facilita la síntesis de información clave, permitiendo que los usuarios obtengan una visión clara y concisa del contenido de los documentos, optimizando el tiempo dedicado al análisis financiero. Además, el sistema contribuye significativamente a la transformación digital del sector, ya que ofrece soluciones innovadoras que mejoran la toma de decisiones y aumentan la competitividad empresarial.

El uso de técnicas avanzadas, como el algoritmo Maximal Marginal Relevance (MMR), permite equilibrar relevancia y diversidad en los resúmenes, asegurando que los resultados sean representativos del contenido original sin redundancias. Además, el sistema puede ser una herramienta valiosa no solo para grandes corporaciones, sino también para pequeños inversionistas, promoviendo la equidad en el acceso a la información y contribuyendo a un entorno financiero más transparente.

Este trabajo también tiene implicaciones directas para el desarrollo sostenible, ya que, al reducir el tiempo y los recursos necesarios para el análisis de datos, favorece la conservación de recursos y promueve prácticas más responsables y sostenibles en el ámbito empresarial. Finalmente, este estudio

establece un punto de partida para futuras investigaciones que podrían explorar la aplicación de este enfoque en otros idiomas y contextos, así como la integración de tecnologías más avanzadas, como el aprendizaje profundo, para mejorar la calidad de los resúmenes generados.

Referencias

Chen, J. (s/f). *Earnings call*. Investopedia. https://www.investopedia.com/terms/e/earnings-call.asp

Cormen, T. H., Leiserson, C. E., Rivest, R. L. y Stein, C. (2009). *Introduction to algorithms* (3ª ed.). MIT.

Fornero, R. A. (2017). *Fundamentos de análisis financiero*. Universidad Nacional del Cuyo. https://www.academia.edu/35162347/Fundamentos_de_análisis_financiero

Gómez Adorno, H. M. (2018). *Extracción de características de texto basada en grafos sintácticos integrados* [Tesis doctoral]. Instituto Politécnico Nacional, Centro de Investigación en Computación. https://tesis.ipn.mx/handle/123456789/25880

Marsi, E. C. y Krahmer, E. J. (2005). Explorations in sentence fusion. En *Proceedings of the 10th European Workshop on Natural Language Generation (ENLG)* (pp. 109–117).

Mihalcea, R. y Radev, D. (2011). Notations, properties, and representations. En *Graph-based natural language processing and information retrieval* (pp. 11–19). Cambridge University.

Moreiro González, J. A. (1988). El resumen. En J. Gimeno Perelló et al., *Operaciones de la cadena documental* (pp. 36–42). Instituto Oficial de Radio y Televisión. https://e-archivo.uc3m.es/bitstream/handle/10016/36085/resumen_moreiro_OC_1988.pdf

Nallapati, R., Zhai, F. y Zhou, B. (2016). *SummaRuNNer: A Recurrent Neural Network based sequence model for extractive summarization of documents.* Arxiv. https://doi.org/10.48550/arXiv.1611.04230

Thanaki, J. (2017). *Python natural language processing* (cap. 4, p. 79). Packt.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. y Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. En *International Conference on Learning Representations.* https://openreview.net/forum?id=SkeHuCVFDr