Between Ghosts and Graphs: A Data-Driven reading of The Canterville Ghost.

Análisis del estilo y la temática mediante metodología de análisis de datos

16 DE DICIEMBRE DE 2025

IT ACADEMY
TANIA ELIZAZBETH RODRÍGUEZ ZÁRATE

1. Resumen

El presente trabajo explora la posibilidad de analizar un texto literario mediante metodologías propias del análisis de datos, tomando como caso de estudio *The Canterville Ghost* de Oscar Wilde. El objetivo principal es examinar de qué manera las herramientas de análisis de datos permiten identificar patrones estructurales, estilísticos y temáticos sin sustituir la interpretación literaria tradicional.

El análisis se aborda desde dos ejes principales: el estilo y la temática. Para ello, el texto fue estructurado en distintos niveles (capítulos, escenas y frases) y procesado mediante técnicas de limpieza, tokenización y conteo léxico. Se emplearon visualizaciones para estudiar la longitud de las frases, la distribución temática, los sentimientos del léxico y la red narrativa.

Los resultados muestran que Wilde construye el relato a partir de un ritmo dinámico dominado por frases breves, una superposición constante de registros góticos, cómicos y culturales, y un predominio de la neutralidad emocional, que refuerza el tono irónico característico de la obra. Desde el punto de vista metodológico, el estudio demuestra que el análisis de datos puede aplicarse al ámbito literario como una herramienta complementaria, capaz de enriquecer la lectura y abrir nuevas perspectivas de interpretación.

2. Introducción

¿Es posible analizar un texto literario mediante metodología de análisis de datos? Esta ha sido la pregunta que ha guiado la realización del presente proyecto. Con el fin de explorar esta idea, he decidido trabajar con *The Canterville Ghost* de Oscar Wilde.

The Canterville Ghost, escrito por Oscar Wilde y publicado en 1887, narra la llegada de una familia estadounidense a una antigua mansión inglesa, donde descubren que habita un "temible" fantasma. La obra es una constante parodia entre lo cómico y gótico, lo moderno y la estructura vieja de la aristocracia inglesa.

El análisis que presento a continuación se divide en dos grandes ejes. En primer lugar, el análisis del estilo, centrado en aspectos como la distribución de palabras, frases, etc. En segundo lugar, el análisis temático, y como se va construyendo la obra. A partir de estos dos niveles, se pretende comprender qué nuevas perspectivas puede ofrecer un enfoque basado en datos dentro del campo de la literatura.

La motivación principal del proyecto es, en gran medida, personal. Con formación en humanidades, siempre he sentido fascinación por la capacidad humana de crear, imaginar y narrar. Al mismo tiempo, mi dificultad con el razonamiento lógico y analítico me ha llevado a buscar un punto medio que conecte ambos mundos. Este proyecto representa, así, un intento de encontrar un equilibrio entre dos formas de pensar: la sensibilidad literaria y la lectura cuantitativa de la información.

3. Metodología

3.1. Selección del corpus

El corpus utilizado para este análisis es *The Canterville Ghost* de Oscar Wilde. El texto fue obtenido de la biblioteca digital Project Gutenberg.

A partir de esta versión se realizó una limpieza inicial del documento. Este proceso consistió en eliminar elementos que no pertenecen a la obra en sí, como la tabla de ilustraciones, encabezados, notas editoriales y contenido adicional, con el fin de trabajar únicamente con el texto literario íntegro.

3.2. Herramientas y librerías utilizadas

Para el procesamiento, análisis y visualización de los datos textuales se emplearon las siguientes herramientas y librerías de Python:

- Pandas: organización del texto en dataframes y manipulación tabular.
- Matplotlib y Seaborn: generación de visualizaciones
- Regex (re): limpieza de caracteres, eliminación de marcas no
- NLTK: tokenización de oraciones, lo que permitió dividir el texto en unidades analizables.
- TextBlob: análisis de sentimiento para explorar la capa emocional del texto.
- Counter: conteo y frecuencia de palabras.
- Networkx: construcción de grafos para representar interacciones entre personajes.
- Itertools: manejo de combinaciones útiles para identificar coocurrencias narrativas.
- WordCloud: creación de nubes de palabras que permiten observar patrones de vocabulario.

El uso combinado de estas herramientas permitió transformar un texto literario en un conjunto de datos estructurados.

3.3. Procesamiento general del texto y construcción de datasets

Para este proyecto se trabajó con un DataFrame, que ha medida que el análisis se profundizaba se fue enriqueciendo con columnas adicionales. Principalmente, uno enfocado en el estilo, y el otro orientado al análisis temático.

3.3.1. Limpieza inicial y segmentación del texto

El texto limpio fue cargado en Python y posteriormente segmentado en capítulos, párrafos y oraciones. Esta segmentación permitió construir una estructura base, facilitando cálculos como el número de palabras por frase, la longitud en caracteres y la relación entre oraciones y unidades narrativas mayores.

3.3 Clasificación de oraciones

Las oraciones del corpus fueron clasificadas según su modalidad discursiva, distinguiendo entre:

Diálogo: intervenciones directas de los personajes, que permiten observar el tono del intercambio verbal.

Narración: fragmentos descriptivos o explicativos que construyen el contexto y modulan el ritmo del relato.

Acción del fantasma (ghost_action): oraciones en las que Sir Simon actúa como agente narrativo, ya sea mediante apariciones, intentos de intimidación o acciones con intención dramática o cómica.

Esta clasificación no busca analizar interacciones sociales en sentido relacional, sino identificar los recursos.

scene_id	chapter	paragraph	interaction_type	sir_simon_present	characters	sentence	sentence_length_words	sentence_length_chars	ghost_present	gothic_count	comic_count	cultural_count	dominant_theme
1	1	1	narration	False	[lordCanterville, mrOtis]	when mr. hiram b. otis, the american minister,...	65	348	0	2	0	4	cultural
2	1	2	dialogue	True	[duchess, lordCanterville, mrOtis, rector, sir...	"we have not cared to live in the place oursel...	80	434	1	2	0	2	gothic
3	1	2	narration	False	[ladyCanterville]	augustus dampier, who is a fellow of king's co...	50	303	0	1	0	1	gothic
4	1	3	dialogue	True	[mrOtis, sir_simon]	"my lord," answered the minister, "i will take...	16	90	1	1	0	2	cultural
5	1	3	ghost_mention	True	[sir_simon]	i have come from a modern country, where we ha...	72	357	1	1	0	4	cultural

Figura 1 DataFrame

3.4. Escenas y análisis de sentimientos

Para el análisis temático y de sentimientos se enriqueció el DataFrame basado en la noción de escena. A diferencia del capítulo, entendido como una unidad estructural fija, la escena se define aquí como un fragmento narrativo cohesionado por la concurrencia de uno o varios de los siguientes elementos: una misma localización, un evento o acción compartidos, una

intención narrativa común y, de manera específica para este estudio, la presencia del fantasma, Sir Simon.

Con el objetivo de abordar este segundo eje de análisis, se ha optado por una metodología basada en la construcción de diccionarios semánticos, entendidos como conjuntos de palabras representativas de campos temáticos aplicado a nivel de escena. La elección de los tres campos semánticos, gótico, cómico y cultural, responde a la tradición crítica en torno a la obra. Estos diccionarios permiten identificar la frecuencia y distribución de términos vinculados a cada temática, facilitando así la observación de patrones y relaciones entre el contenido temático y los rasgos estilísticos previamente analizados.

Para realizar una aproximación básica a los sentimientos del texto, se ha utilizado la librería TextBlob (Loria S., 2020). Esta herramienta permite clasificar el léxico, de acuerdo con su polaridad, en tres categorías generales: positivo, neutral y negativo. Ofreciendo una lectura cuantitativa de la polaridad emocional del lenguaje empleado.

Comentado [TR1]: ofreciendo una lectura cuantitativa de la polaridad emocional del lenguaje empleado.

4. Resultados

Nube de palabras. Una primera aproximación al texto

Como primer acercamiento visual al contenido del cuento The Canterville Ghost, se generó una nube de palabras a partir del texto completo. Esta técnica, de carácter exploratorio, permite identificar los términos con mayor frecuencia en la narración, ofreciendo una visión panorámica del universo léxico que compone el relato.

En la visualización obtenida, las palabras que aparecen con mayor tamaño son, indicando mayor frecuencia, son “Otis”, “ghost” y “Virginia”.

Figura 2. Nube de palabras

Distribución de párrafos por capítulo

Para comprender el estilo narrativo de The Canterville Ghost, se realizó una primera aproximación estructural basada en la organización formal del texto. El cuento se compone de siete capítulos cuya extensión varía significativamente.

Se elaboró un gráfico de barras que representa el número de párrafos por capítulo. Esta visualización permite observar cómo se distribuye el contenido narrativo a lo largo de la obra, revelando que el capítulo 5, con un total de 39 párrafos, destaca sobre el resto. En contraste, el capítulo 2, con solo 5 párrafos, presenta una estructura mucho más condensada.

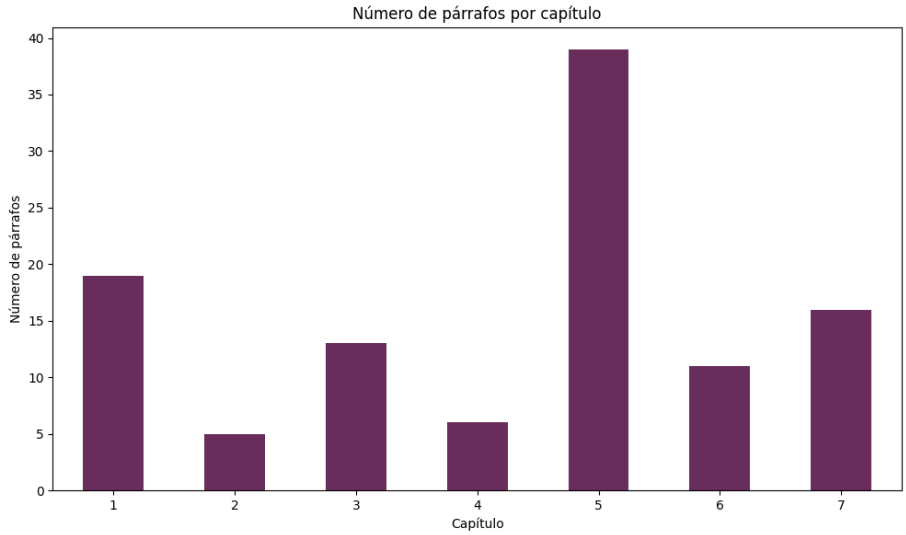

Figura 3. Número de párrafos por capítulo

Distribución de longitud de frases (en palabras) por capítulo

Para profundizar en el análisis del estilo, se examinó la distribución de la longitud de las frases (medida en número de palabras) por capítulo mediante un gráfico de tipo boxplot. Esta visualización permite observar la variabilidad en la construcción sintáctica a lo largo del texto, revelando diferencias significativas entre capítulos.

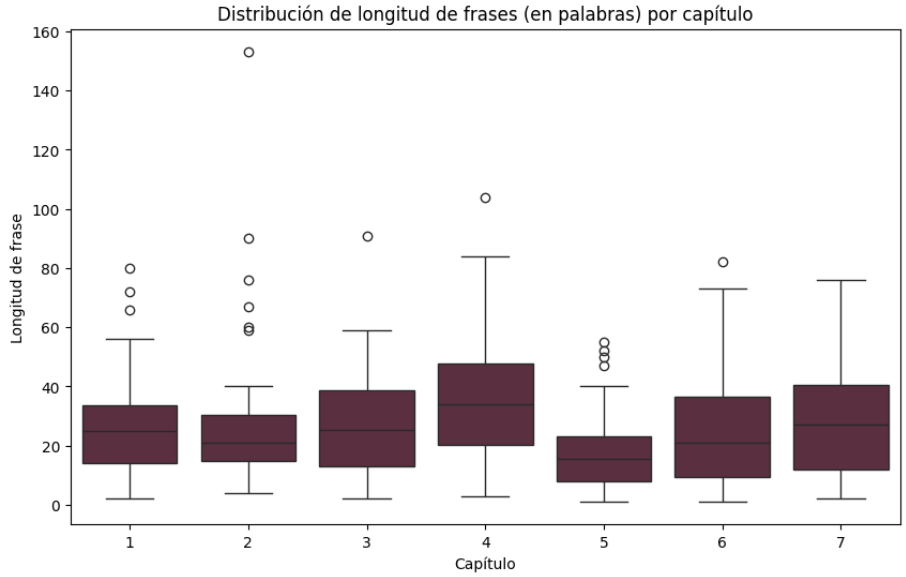

Figura 4. Boxplot Distribución de longitud de frase

El gráfico muestra que algunos capítulos presentan frases más extensas en promedio, mientras que otros se caracterizan por estructuras más breves. También se identifican outliers, es decir, frases excepcionalmente largas o cortas que se apartan del patrón general.

Relación entre longitud de frases (palabras vs caracteres)

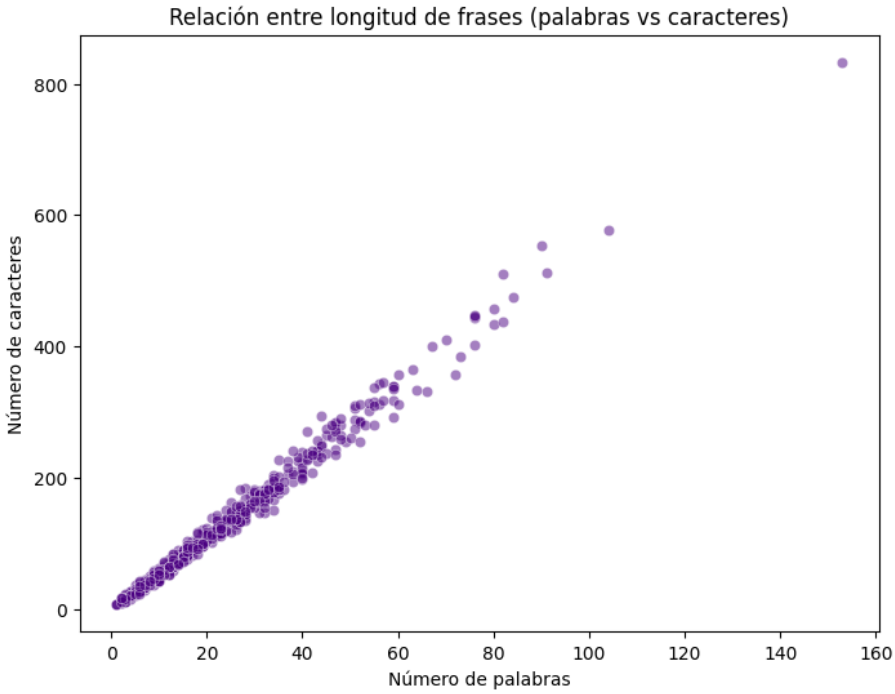

Figura 5. Scatterplot de relación entre longitud de frase

Para representar la relación entre el número de palabras y el número de caracteres por frase en el cuento se realizó un gráfico de dispersión (scatterplot). Como era esperable, se observa una correlación positiva: a mayor número de palabras, mayor número de caracteres.

La visualización muestra una nube de puntos alineada diagonalmente, lo que indica una relación proporcional entre ambas variables.

Comentado [TR2]: La dispersión nos habla de **regularidad estilística**, **flexibilidad sintáctica** y **control del ritmo**.

Controlado, no impulsivo
Flexible, pero no caótico
Rítmicamente consciente
Capaz de alternar entre sencillez y complejidad sin perder coherencia

Evolución de la longitud promedio de frases por capítulo

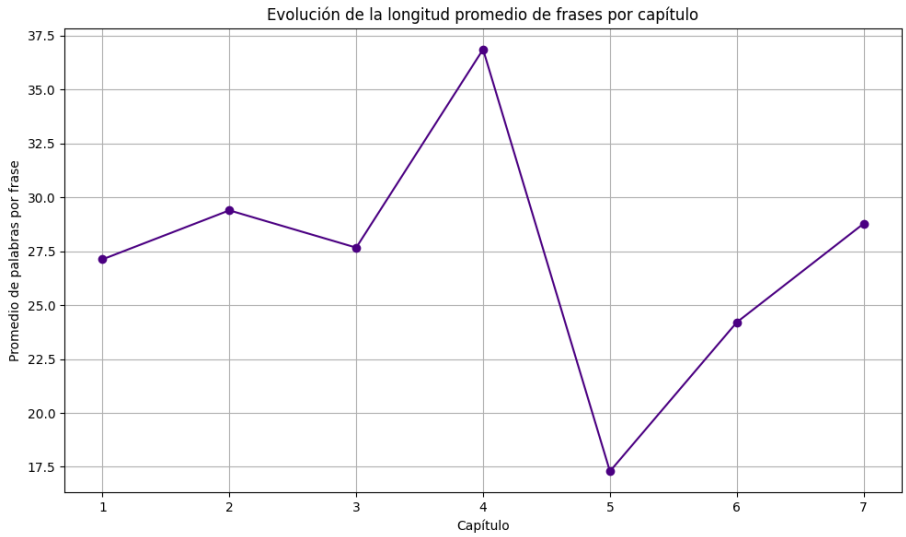

Figura 6. Evolución de la longitud promedio de frases por capítulo

Para concluir el apartado de estilo, se analizó con un gráfico de líneas la evolución de la longitud media de las frases —medida en número de palabras— a lo largo de los siete capítulos del cuento The Canterville Ghost. La visualización muestra una fluctuación estilística. Algunos capítulos presentan frases más extensas en promedio de 27.5 palabras por frase, mientras que otros se caracterizan por estructuras más breves.

El capítulo 4 alcanza el mayor promedio de longitud, mientras que el capítulo 5 registra el más bajo con 17.5 palabras.

4.1 Análisis emocional y relación con las temáticas

Este uso alternado del ritmo no es arbitrario. El estilo, entendido como la forma en que se construyen las frases y se organiza el discurso, permite observar cómo ciertas formas narrativas acompañan determinadas temáticas.

De este modo, el estilo funciona como una capa intermedia que modula el significado del texto y acompaña el desarrollo temático.

Distribución de temas por capítulo

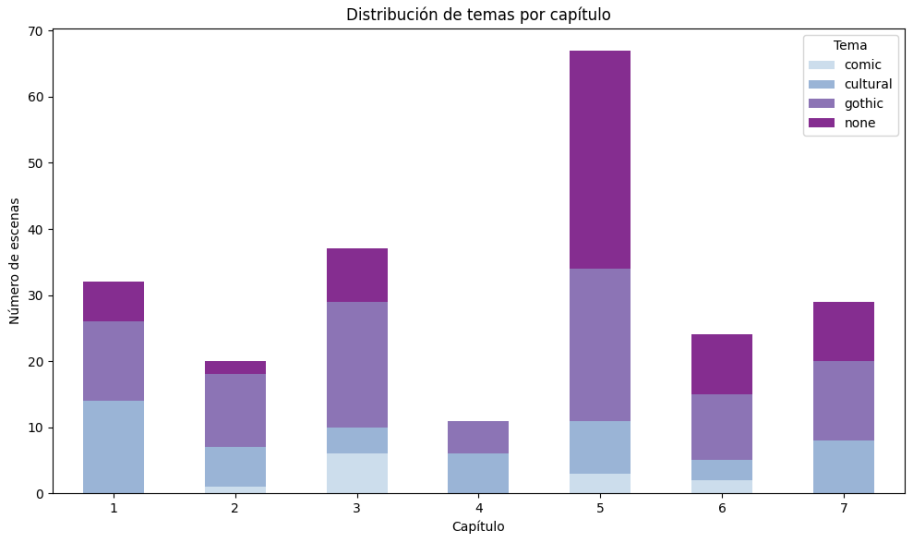

Figura 7. Distribución de temas por capítulo

Para visualizar la distribución de los temas, gótico, cómico y cultural, a lo largo de los distintos capítulos del cuento se proyecta una gráfica de barras. Esta visualización permite observar cómo los campos semánticos definidos no aparecen de forma aislada, sino que se entrelazan y coexisten a lo largo de la narración.

En la figura 7, se observa una predominancia del tema gótico, dejando de lado por el momento la categoría *none*, que más adelante se abordará, en contraste al tema de lo cultural.

Es importante señalar que este análisis no identifica emociones complejas ni estados psicológicos de los personajes, sino que proporciona una medida simplificada de la carga afectiva del lenguaje empleado. La clasificación en sentimientos positivos, negativos y neutrales se basa en la polaridad léxica del texto, y funciona como una capa complementaria al análisis temático, revelando el tono narrativo que subyace en cada escena.

Comentado [TR3]: Wilde parodia el gótico desde dentro, usando su léxico pero vaciándolo de intensidad emocional.

Su comicidad se basa en:
- la burla,
- el ridículo,
- la ironía,
- la incomodidad del fantasma.

Mapa de calor

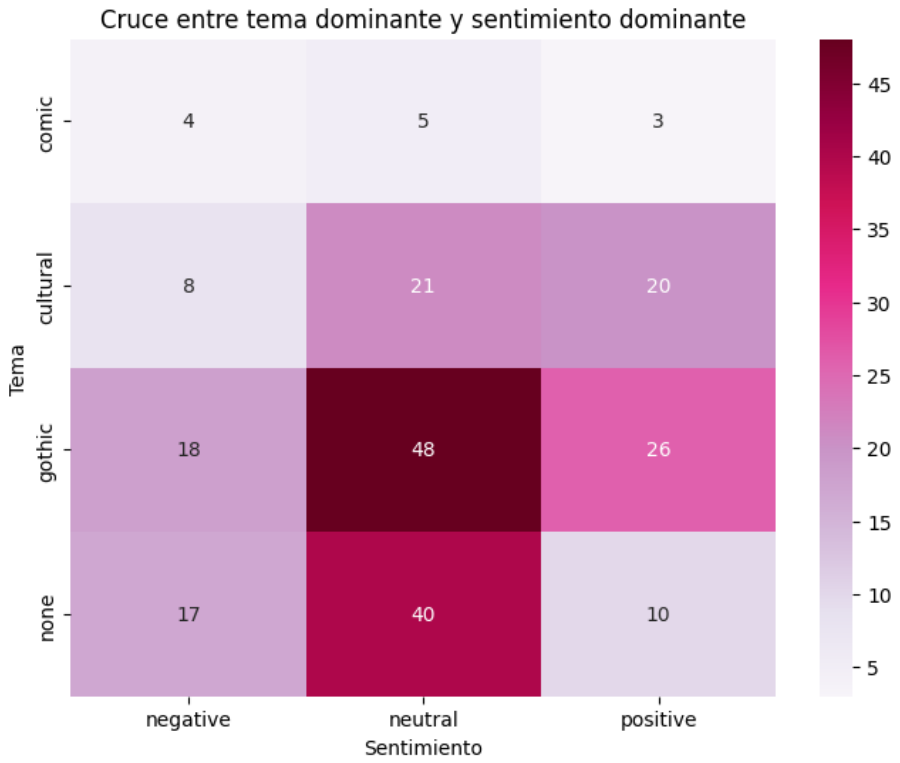

Figura 8.Cruce entre tema dominante y sentimiento dominante

El mapa de calor obtenido para visualizar la relación entre temas y sentimientos asignados a cada escena, muestra que, en todas las categorías temáticas, predomina ampliamente la neutralidad del tono.

El tema gótico presenta la mayor cantidad de escenas con sentimiento negativo, seguido por el cultural y el *none*. Sin embargo, en todos los casos, el sentimiento neutral supera en frecuencia a los demás. El ámbito cómico muestra una distribución más equilibrada, con presencia de sentimientos negativos y positivos en menor proporción.

Evolución de sentimientos por capítulo

Para finalizar esta etapa del análisis, se presenta un gráfico de líneas que muestra la evolución de la carga emocional a lo largo del relato, a partir de la clasificación léxica en categorías

positiva, neutral y negativa. Esta visualización permite observar tendencias y fluctuaciones de los sentimientos en el desarrollo narrativo.

Mostrando al capítulo 4 como el que posee menos escenas y por lo tanto las tres categorías de sentimientos tienen un descenso drástico. Mientras que en el capítulo 5 se observa una subida en el número de escenas y una carga significativa hacia los sentimientos neutrales.

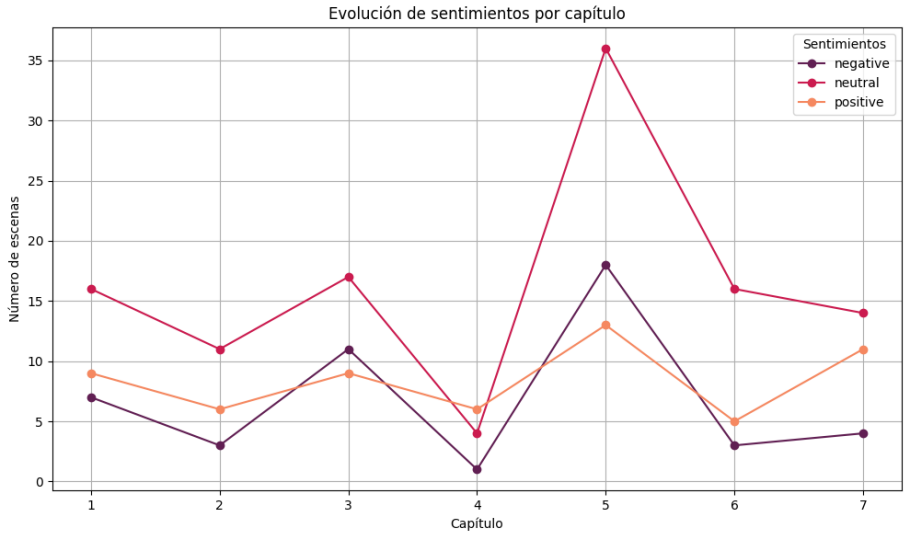

Figura 9. Evolución de emociones por capítulo

Red narrativa

Además del análisis estilístico, temático y emocional, resulta fundamental incorporar una dimensión relacional que permita observar el texto como una estructura de vínculos. La literatura también posee interacciones. Esta capa relacional revela cómo los personajes se articulan entre sí, y cómo esas conexiones configuran el entramado del relato.

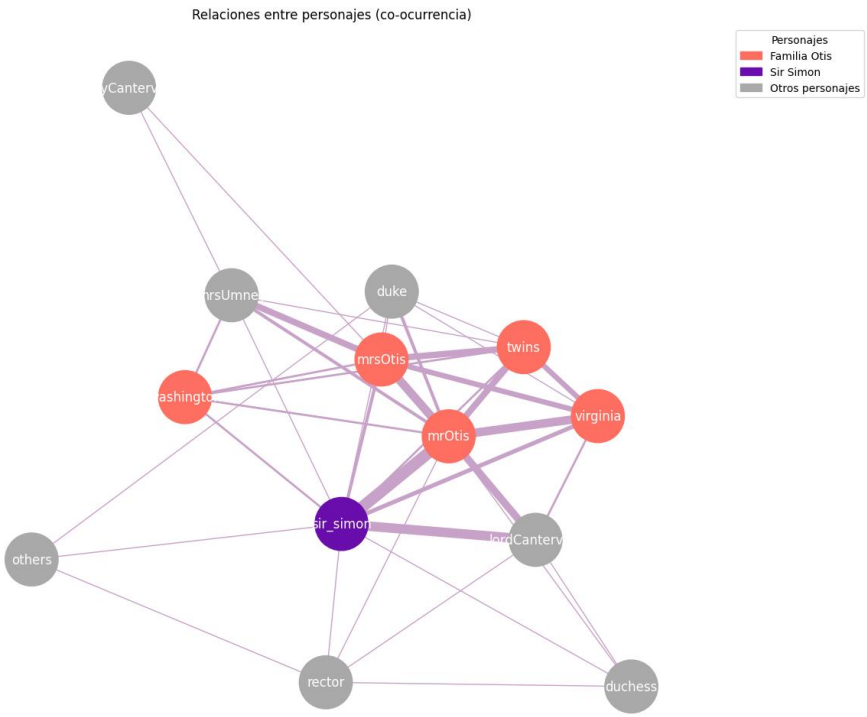

Figura 10. Relaciones entre personajes

En esta visualización, cada personaje se representa como un nodo, y cada interacción compartida se traduce en una conexión (arista) entre ellos. El grosor de las líneas indica la frecuencia de aparición conjunta, mientras que los colores distinguen grupos narrativos: la familia Otis, el fantasma Sir Simon, y los personajes secundarios.

La red revela que Sir Simon se encuentra en el centro de múltiples interacciones, especialmente con los miembros de la familia Otis. Los personajes secundarios aparecen más periféricos, con menor grado de conexión.

5. Discusiones

El análisis realizado permite interpretar El fantasma de Canterville como un texto cuidadosamente equilibrado entre forma, contenido y tono. Lejos de ofrecer una narración homogénea, los resultados sugieren que Oscar Wilde articula su relato a partir de contrastes deliberados, tanto a nivel estructural como temático y afectivo.

La nube de palabras (Figura 2) ofrece una primera clave interpretativa relevante. Si bien la prominencia de "Otis" confirma el papel estructural de la familia como eje de las acciones y reacciones frente al fantasma, resulta especialmente significativo que "Virginia" destaque. Este hallazgo anticipa su rol como mediadora dentro del relato, sugiriendo que el conflicto central no se limita a la oposición entre lo sobrenatural y lo racional, sino que incorpora una dimensión más. Así, una visualización basada en frecuencia léxica revela una jerarquía narrativa que no siempre resulta evidente en una lectura lineal.

Desde el punto de vista estructural, la variación en el número de párrafos por capítulo (Figura 3) y en la longitud de las frases (Figuras 4, 5 y 6) pone de manifiesto un uso consciente del ritmo narrativo. Los capítulos más extensos en párrafos y con mayor variabilidad sintáctica parecen coincidir con momentos de intensificación narrativa, mientras que los más breves cumplen funciones de transición o desarrollo. Esta alternancia refuerza la idea de que Wilde no construye su relato desde la uniformidad, sino desde la modulación constante de la narración.

El análisis de la longitud de las frases refuerza esta lectura. Las frases más largas, asociadas a estructuras sintácticas complejas, aparecen vinculadas a pasajes descriptivos o de construcción atmosférica, propios del imaginario gótico. En contraste, las frases cortas predominan en el diálogo, generando un ritmo ágil y teatral que refuerza el carácter satírico del texto. La dispersión observada en el scatterplot (Figura 5) no indica irregularidad, sino un control estilístico que ajusta la complejidad formal a la intención narrativa de cada momento.

En cuanto al análisis temático, la distribución de temas por capítulo (Figura 7) confirma que lo gótico, lo cómico y lo cultural no funcionan como categorías aisladas. La presencia dominante del léxico gótico, incluso dentro de escenas clasificadas como cómicas, refuerza la lectura de la obra como una parodia consciente del género. Wilde se apropia del imaginario gótico no para reproducir su solemnidad, sino para vaciarla de gravedad mediante el humor y la ironía. La categoría none, lejos de representar una ausencia de contenido, señala espacios de transición narrativa y ambigüedad temática que contribuyen a la cohesión del relato.

La incorporación del análisis de polaridad léxica aporta una dimensión afectiva que refuerza estas interpretaciones. El cruce entre tema dominante y sentimiento (Figura 8) muestra una clara predominancia de la neutralidad en todas las categorías temáticas. Este resultado sugiere una estrategia narrativa basada en la distancia irónica. Incluso en escenas potencialmente intensas, el lenguaje evita la carga emocional explícita. En el ámbito gótico, la coexistencia de términos con carga negativa y una neutralidad dominante refuerza la idea de que el terror es tratado como recurso estilístico más que un detonante de sentimientos. En el registro cómico, la neutralidad combinada con elementos negativos apunta a una comicidad crítica, sustentada en la burla y la ironía.

La evolución del tono afectivo por capítulo (Figura 9) confirma que no existe una progresión emocional lineal, sino un flujo oscilante entre tensión, comicidad y ambigüedad. El comportamiento del capítulo 5, con un pico notable de neutralidad, puede interpretarse como un punto álgido del relato, donde el conflicto se desplaza progresivamente hacia su resolución, manteniendo la distancia emocional característica de la obra.

Finalmente, la visualización de la red narrativa (Figura 10) permite integrar las distintas capas analizadas desde una perspectiva relacional. Al representar a los personajes como nodos y sus interacciones como conexiones, se observa que el núcleo del relato no reside exclusivamente en el fantasma, sino en la red de relaciones que se tejen en torno a la familia Otis. Este enfoque sintetiza los hallazgos previos, mostrando cómo estilo, temática y tono afectivo se articulan a través de las interacciones entre personajes, configurando la arquitectura narrativa del cuento.

En conjunto, la discusión evidencia que el uso de herramientas de análisis de datos no sustituye la interpretación literaria, sino que la amplía, permitiendo observar patrones estructurales, estilísticos y afectivos que enriquecen la comprensión del texto.

6. Conclusión y aprendizajes.

La pregunta que dio origen a este estudio fue clara y desafiante: ¿es posible analizar un texto literario desde una metodología propia del análisis de datos? Sí, es posible. Siempre que se reconozca que toda aproximación cuantitativa es una lectura situada, dependiente de las dimensiones que se decida explorar. En este caso, se optó por dos ejes: el estilo y la temática.

Al cerrar este recorrido, emergen dos planos de aprendizaje que se entrelazan, el literario y el metodológico.

Desde la perspectiva literaria, el análisis revela que *The Canterville Ghost* se construye sobre una tensión constante entre agilidad y densidad, entre comicidad y mesura. El estudio muestra que el autor privilegia frases breves, favoreciendo un ritmo ágil, en momentos de atmósfera gótica o introspección narrativa, la prosa se alarga.

En cuanto a la temática, el texto se sostiene principalmente sobre tres registros lo gótico, lo cultural y lo cómico. Una superposición que no busca equilibrio, sino sátira. El fantasma, lejos de ser una figura de horror, se convierte en un cuerpo paródico. El análisis léxico revela un predominio de palabras de polaridad neutral, incluso en escenas cargadas de tensión

Desde el enfoque del análisis de datos, la construcción de distintos DataFrames permitió descomponer el texto en capas y observar su arquitectura interna: estilo, temas, polaridad emocional, relaciones entre personajes. Esta vía, no ofrece respuestas definitivas, sino nuevas preguntas, rutas de interpretación y posibilidades de contraste.

Aunque el objeto de estudio fue un relato literario, la metodología desarrollada es extrapolable a contextos empresariales y de inteligencia de negocio. Las técnicas aplicadas permiten detectar patrones de contenido, analizar tono y percepción, identificar tendencias narrativas, comparar discursos y modelar interacciones entre entidades.

6. Referencias

Wilde, O. (1887/2023). The Canterville Ghost. Project Gutenberg. https://www.gutenberg.org/ebooks/14522

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.

Hagberg, A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95.

Loria, S. (2020). TextBlob: Simplified text processing. https://textblob.readthedocs.io/

McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. (Opcional, solo si lo mencionas)

Seaborn Documentation. (2024). Statistical data visualization. https://seaborn.pydata.org/

WordCloud Documentation. (2024). Word cloud generation in Python. https://amueller.github.io/word_cloud/