¿QUÉ NOS DICE UNA MÁQUINA SOBRE *FORTUNATA Y JACINTA*?

WHAT DOES A MACHINE TELL US ABOUT *FORTUNATA AND JACINTA*?

Pedro Juan Baquero Pérez

Universidad de La Laguna

RESUMEN

En este trabajo se han aplicado diversas técnicas encuadradas dentro del procesamiento del lenguaje natural, el cual es una rama de la inteligencia artificial, para obtener información semántica de la obra *Fortunata y Jacinta*. Entre los resultados que se muestran tenemos la representación resumida a lo largo del texto, una propuesta de temas de la obra, un mapa de su sentimiento y la descripción de los personajes. Por otra parte, otro de los objetivos de este trabajo es promover el uso de estas técnicas para adquirir mayor conocimiento sobre la obra de Galdós. En este sentido, sobre estas técnicas, y con el fin de que se puedan conocer sus posibilidades y limitaciones, se hace una exposición breve y conceptual.

PALABRAS CLAVE: Galdós, *Fortunata y Jacinta*, Procesamiento del lenguaje natural, Inteligencia Artificial.

ABSTRACT

In this work several techniques framed within natural language processing, which is a branch of artificial intelligence, have been applied to obtain semantic information from the work *Fortunata y Jacinta*. Among the results shown we have the summarized representation throughout the text, a proposal of themes of this work, a map of its sentiment and the description of the characters. On the other hand, another of the aims of this work is to promote the use of these techniques to acquire greater knowledge of Galdós' work. In this sense, a brief and conceptual presentation of these techniques is given so that their possibilities and limitations can be understood.

KEYWORDS: Galdós, *Fortunata y Jacinta*, Natural language processing, Artificial Intelligence.

INTRODUCCIÓN

Dentro de la inteligencia artificial se entiende por una máquina como un ente artificial que hace un tratamiento de datos y que termina resolviendo algún tipo de problema, para lo que utiliza un conjunto de algoritmos. Una máquina no comprende un texto y menos aún lo disfruta, la palabra 'inteligencia' dentro de 'inteligencia artificial' solo se queda en resolver problemas, pero una máquina no llega a formular problemas en la medida que lo podemos hacer nosotros, los humanos. Así, cuando planteo la pregunta ¿qué nos dice una máquina sobre *Fortunata y Jacinta*?, no podemos esperar que la máquina vaya más allá de los datos que extrae de un texto. Una máquina trabaja muy bien analizando números. De esta forma, puede extraer información numérica y estadística que permite obtener o descubrir información valiosa, como puede ser el conocer qué probabilidad tiene un texto de pertenecer a determinado autor. Sin embargo, si queremos extraer información relacionada con la comprensión de un texto, actualmente no

Galván González, Victoria (coord.) (2023): *Coda a un centenario. Galdós, miradas y perspectivas.*
Casa-Museo Pérez Galdós, Cabildo de Gran Canaria. Las Palmas de Gran Canaria. ISBN: 978-84-1353-120-5.

podemos esperar grandes resultados. Entonces, ¿qué podríamos esperar?: una máquina puede descubrir información semántica dentro del texto, donde parte de esta información nos puede resultar obvia, pero otra puede estar oculta. En otras palabras, lo que podemos esperar es que una máquina nos descubra información semántica oculta, lo que en la jerga de la inteligencia artificial se denomina minería de textos.

La minería de textos aplicada a los textos literarios se pueden encuadrar dentro de las humanidades digitales. Se trata de aplicar métodos cuantitativos utilizando las nuevas tecnologías que complementan, y que no excluyen, a la tradicional lectura atenta de los textos literarios. En este trabajo lo concretamos en la aplicación de técnicas de procesamiento del lenguaje natural, que es un campo dentro de la inteligencia artificial. Estas técnicas permiten procesar los textos literarios de una manera muy diferente a lo que podría hacer un humano. De esta forma, facilitan acercarse a una obra literaria con otras perspectivas y obtener resultados que con su lectura tradicional solo se podrían especular. En este sentido, hay investigadores, como Moretti (2013) y Jockers (2016), que exigen un nuevo tipo de análisis que enriquezca al analista de los textos literarios con la recogida de las evidencias que aportan estas técnicas computacionales. Este trabajo trata de aplicar estas técnicas a la obra *Fortunata y Jacinta* y al mismo tiempo exponer conceptualmente estas técnicas.

En este sentido, el objetivo de este trabajo tiene dos vertientes. En primer lugar, se introducirá las técnicas que son utilizadas en el procesamiento del lenguaje natural que nos pueden ayudar al análisis de textos literarios. Principalmente, se expondrá cómo una máquina reduce un texto a números y cómo estos se procesan. En segundo lugar, se presentarán los resultados que nos proveen estas técnicas para la obra de *Fortunata y Jacinta*. Son cuatro los resultados que se muestran: el resumen de la obra a lo largo del texto, los temas que descubre la máquina, un mapa del sentimiento que transmite la obra y la descripción de los personajes. Hay que indicar que estos resultados deben ser interpretados, no siendo el objetivo principal de este trabajo realizar su interpretación, para lo que hay que dar paso al analista o crítico literario.

¿Cómo las máquinas analizan los textos?

Una máquina no comprende un texto, tan solo lo procesa y entrega resultados. Una máquina no disfruta de un texto: reduce y traduce un texto a números, hace operaciones con ellos y muestra unos resultados que la misma u otra máquina no comprende. Será el humano quien, en tal caso, los interprete. Entonces, ¿cómo traducir un texto a números? Básicamente, se trabaja con palabras a las que se le asocia un vector a cada una. Un vector no deja de ser una secuencia

de números que nos indican algo. Ese algo tiene un significado distinto dependiendo del método que se vaya a utilizar para representar el texto. Así, lo primero que se tiene que hacer es convertir un texto, en nuestro caso, la obra *Fortunata y Jacinta*, a vectores, es decir, a números, los cuales nos tienen que dar algún tipo de información sobre el texto.

Pongamos un ejemplo muy sencillo: supongamos que nuestro texto solo contiene dos capítulos, y cada capítulo contiene solo una frase:

- Capítulo I: 'Las noticias más remotas que tengo me las ha dado Jacinto.'
- Capítulo II: 'Tuve que darle mil noticias del asilo.'

Entre ambos capítulos existen un total de 15 palabras únicas, entonces, el texto, ordenando las palabras alfabéticamente, se puede representar en algo como lo siguiente:

	asilo	dado	darle	del	ha	Jacinto	Las	más	Me	Mil	noticias	Que	remotas	tengo	Tuve
I	0	1	0	0	1	1	2	1	1	0	1	1	1	1	0
II	1	0	1	1	0	0	0	0	0	1	1	1	0	0	1

A esto se le denomina una matriz, que en este caso tiene 2 filas para cada capítulo, y 15 columnas para cada palabra, es decir, tenemos 15 palabras únicas. Cada columna es un vector que define a cada palabra donde, en este ejemplo simplificado, un vector contiene dos números, por ejemplo: para la palabra 'asilo' el vector es [0,1], que nos indica que esta palabra aparece 0 veces en el capítulo I y 1 vez en el capítulo II; y para la palabra 'las' el vector es [2,0], es decir, aparece 2 veces en el capítulo I y 0 veces en el capítulo II.

Para el caso de *Fortunata y Jacinta* tenemos un texto con 4 partes, divididas en capítulos, y cada capítulo dividido en subcapítulos[1]. En total tenemos 198 subcapítulos, con unas 470.000 palabras, existiendo unas 32.000 palabras únicas. Si dividimos el texto en sus 198 subcapítulos, tendríamos unos 32.000 vectores de palabras, donde cada vector, es decir, lo que define una palabra, tendría 198 números, uno por cada subcapítulo. En definitiva, tenemos, en total, cerca de seis millones de números, donde muchos toman el valor cero, que indica que tal palabra no aparece en un subcapítulo. Aunque es una cifra importante de números, una máquina no tiene demasiados problemas en tratarlos.

Este modelo, basado en una matriz, es una representación simplificada de un texto. Realmente, este modelo nos da tan solo información de las palabras que aparecen en cada

[1] En este trabajo se ha trabajado a partir del texto en formato libre y digital de *Fortunata y Jacinta* (Pérez Galdós: 1887) disponible en Proyecto Gutenberg.

capítulo. Se pierde el orden y el contexto inmediato de cada palabra en el texto. Por ejemplo, se pierde la relación de las palabras que acompañan a la palabra 'las' —que aparece dos veces en nuestro ejemplo simplificado—, es decir, se pierde que esta palabra acompaña a 'noticias' en un caso, y a 'me' y 'ha' en el otro caso. También, frases como 'un hombre se come un perro' y 'un perro se come un hombre', en este modelo, son equivalentes, dado que tienen las mismas palabras, a pesar de que cada frase tiene un significado distinto.

Con el anterior párrafo se quiere mostrar que se trata de un modelo simplificado, que realmente está lejos de representar al completo el texto original. En cualquier caso, la experiencia ha dado buenos resultados para descubrir información de los textos usando este tipo de modelos. De todas formas, como veremos más adelante, cuando se describan los personajes, existen modelos de representación más complejos donde el vector para cada palabra dispone de información contextual de las otras palabras que le acompañan.

A partir de modelos de representación como el que se ha expuesto, las máquinas procesan los datos y obtienen resultados. Y es cuando empiezan las dificultades. En nuestro caso, solo nos interesa la información semántica, donde un problema inicial es que para una máquina estas matrices tienen mucho ruido e información redundante. Por ejemplo, existen muchas palabras que no aportan mucha información semántica, por ejemplo, las palabras 'del', 'las', 'ha' y 'que' producen ruido. También, existen palabras que tienen la misma información semántica, por ejemplo, 'comí' y 'come', que, quitando la información del tiempo verbal, una máquina las considera palabras distintas, pero que tienen el mismo significado o lema.

Un primer proceso para adquirir mejor información semántica del texto es eliminar ruido e información redundante. Estamos hablando de que la máquina trabaje con las palabras que nos aporten significado semántico. En el español existen nueve tipos de palabras —artículo, sustantivo, pronombre, adjetivo, verbo, adverbio, preposición, conjunción e interjección—, de estos, el sustantivo, adjetivo y verbo son los que tienen más información semántica. Por ello, en este trabajo, se ha seleccionado solo estos tres tipos de palabras. Por otra parte, dentro de los sustantivos podemos diferenciar los comunes y los propios. Sobre estos últimos tenemos nombres de personas —que en nuestro caso serían los personajes—, lugares, y otros tipos de entidades. En este trabajo se está más interesado en los personajes y no se tendrán en cuenta los otros tipos de entidades, como los lugares. Por tanto, de todas las palabras nos quedaremos con los nombres —los sustantivos comunes—, los adjetivos, los verbos y los personajes.

Además, dentro de los nombres, adjetivos y verbos, existen muchas palabras que tienen una alta frecuencia de repetición en cualquier texto, por ejemplo, 'decir', 'cosa', 'manera', etc. Si no se eliminan estas palabras, la máquina las puede considerar como relevantes por su alta

frecuencia de aparición, cuando en realidad son palabras que no son específicas del texto bajo análisis, con lo que no aportan mucha información del texto. Por tanto, el humano debe indicar a la máquina qué palabras no hay que tener en cuenta[2]. Hay que ser consciente que esto último aporta cierto sesgo por parte del analista del texto. Por otra parte, la selección de estas palabras hay que considerarlas siempre dentro del contexto del análisis que se quiere hacer del texto. Por ejemplo, si se quiere analizar aspectos relacionados con los diálogos, palabras como 'decir', se podrían considerar como relevantes.

Una vez que se ha determinado qué tipos de palabras se van a seleccionar, a la máquina hay que decirle cómo categorizarlas. Un humano no tiene muchos problemas para categorizar las palabras por tipos y saber qué palabras tienen el mismo lema. Sin embargo, para una máquina no es tan evidente. Por tanto, este es el momento cuando a una máquina se le tenga que incorporar lo que se denomina inteligencia. Realmente, por muchas redes neuronales que se utilicen, esta inteligencia no deja de ser un tratamiento matemático o estadístico que puede dar resultados 'milagrosos', pero que actualmente nunca llegan al nivel de aciertos que haría un humano. En este trabajo se han aplicado técnicas basadas en redes neuronales para seleccionar solo los nombres, adjetivos, verbos y el nombre de los personajes[3]. Se ha estimado un porcentaje de aciertos del orden del 98%[4], lo cual, aunque puede parecer un porcentaje alto, sin embargo, sobre un corpus de 470.000 palabras, supone que hay cerca de 10.000 palabras mal categorizadas. Si un humano se dedicase durante varias semanas o meses a categorizar todas las palabras el porcentaje de errores sería muy inferior. Eso sí, una máquina lo hace en segundos. De esta forma, la matriz inicial, que como vimos tenía 198 filas y unas 32.000 columnas con mucho ruido y redundancia, se convierte en pocos segundos en una matriz de 198 filas y unas 8.000 columnas, es decir, con 8.000 palabras, solo con información semántica. En la Figura 1 podemos ver el resultado de cómo sería esta representación semántica de *Fortunata y Jacinta* a través de una matriz.

[2] A estas palabras se les denomina palabras vacías, más comúnmente conocidas con la terminología inglesa como *stopwords.*

[3] En este trabajo se ha utilizado para la categorización de los tipos de palabras la herramienta Stanza (Qi et al.: 2020). Stanza dispone de un conjunto de herramientas de procesamiento del lenguaje natural en el lenguaje de programación Python. Stanza se basa en redes neuronales para el análisis de texto, incluyendo la lematización, y el reconocimiento de entidades con nombre (NER), donde están incluidos los nombres propios de personas.

[4] La estimación del 98% se ha realizado a través de la corrección manual de los resultados obtenidos con la herramienta Stanza sobre el primer subcapítulo de *Fortunata y Jacinta*.

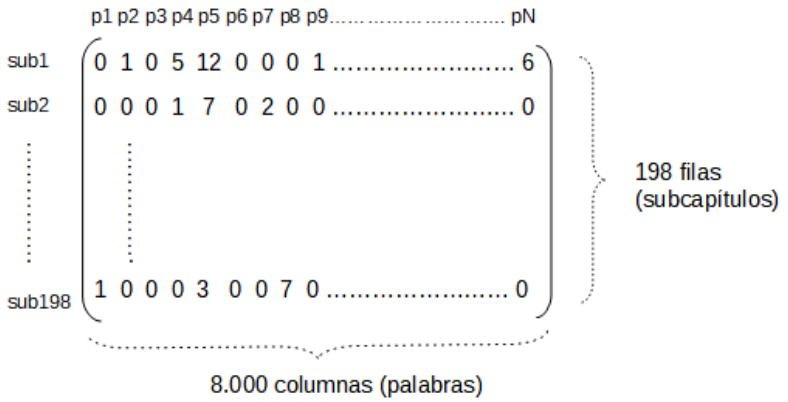

Figura 1. Matriz que representa un texto con 198 subcapítulos y 8.000 palabras.

En definitiva, el texto original se ha simplificado en una matriz, donde esta matriz nos da información matemática del texto, la cual tiene *N* palabras únicas y *M* subcapítulos. Básicamente, la información que nos provee esta matriz es la frecuencia de aparición de cada palabra (*p1, ...pN*) para cada subcapítulo de la obra (*sub1, ..., subM*), y donde las *N* palabras únicas se refieren a solo los nombres, adjetivos, verbos y personajes. A partir de aquí, se aplican diversas técnicas estadísticas para sacar más información semántica del texto.

Otro problema es la selección de los personajes. Sobre este problema se abundará más adelante cuando se describan los personajes. Ahora solo se indicará que para cada personaje se ha asociado una sola palabra, por ejemplo, la menciones en el texto a Fortunata como 'la de Rubín', 'la Pitusa' o 'la señora de Rubín' se engloban todas sobre el nombre 'Fortunata'.

A continuación, veremos cómo se puede extractar la información semántica por capítulos o subcapítulos, para luego descubrir los temas generales de la obra, el sentimiento del texto y describir los personajes. Para ello, se expondrá breve y conceptualmente las técnicas que se aplican. Empezaremos por lo más sencillo, es decir, extraer o resumir información por capítulos o subcapítulos.

Resumir la información semántica a lo largo del texto

Una vez que nos hemos quedado solo con los nombres, adjetivos, verbos y los nombres de los personajes, con las dificultades que ello supone, es directo disponer de una representación resumida para cada subcapítulo. En este caso, cuando asociamos a cada palabra un número, esto es, su frecuencia de aparición, en la jerga informática se le denomina pesado. Este pesado nos indica la frecuencia de cada palabra que realmente está en cada subcapítulo. A este tipo de

pesado se le denomina local, y es lo que más se aproxima a la distribución de palabras del texto original.

Una de las formas más comunes de resumir un texto visualmente[5] cuando se utilizan las tecnologías del lenguaje es el uso de las nubes de palabras teniendo en cuenta la frecuencia de aparición de cada palabra [6]. Aunque la forma más corriente de esta representación es mostrando todas las palabras en una sola nube, en nuestro caso, se ha optado en mostrar por separado cada uno de los tipos de palabras, es decir, se muestra para cada subcapítulo cuatro nubes de palabras para: los personajes, los nombres, los adjetivos y los verbos.

Esta representación tiene como ventaja una mejor interpretación sobre una obra compleja como *Fortunata y Jacinta*, donde existe una abundancia de personajes que van teniendo distinta importancia junto con la trama que evoluciona a lo largo de la obra. Así, con una representación secuencial de cada subcapítulo: con la nube de personajes se puede ir viendo el peso que tiene cada personaje; con los nombres se puede asociar con el tema del subcapítulo, con los adjetivos de califica el tema; y con los verbos se puede asociar al tipo de acción.

Como ejemplo, en la Figura 2 podemos ver los resultados obtenidos usando nubes de palabras para un subcapítulo concreto de la obra.

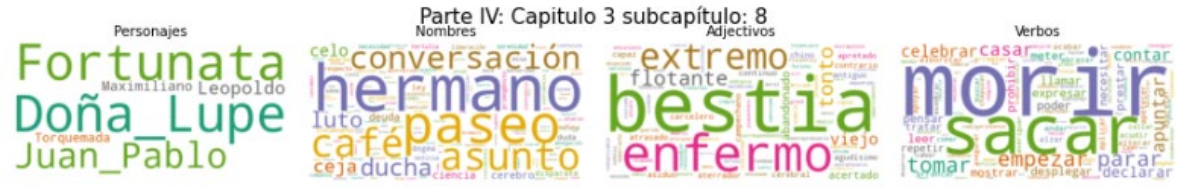

Figura 2. Nubes de palabra para el capítulo 3 y subcapítulo 8 de la parte IV, utilizando pesado local.

Esta nube de palabras nos provee información sobre quiénes son los personajes más activos en este subcapítulo —Fortunata, Doña Lupe y Juan Pablo— y además sobre qué tema actúan y tipo de acción dentro de este subcapítulo y tipo de acción.

Pero este pesado local no pondera lo suficiente determinadas palabras que, aunque sean poco frecuentes dentro de un subcapítulo, sí lo son con respecto a la obra completa, por lo que pueden ser importantes para representar el contenido de un subcapítulo, dado que son más específicas en este subcapítulo. Para solventar este problema, se puede aplicar otro tipo de pesado a la matriz, que penaliza aquellas palabras frecuentes que aparecen en muchos subcapítulos

[5] Realmente, detrás de esta representación visual existe una representación más cuantitativa, consistente en conjuntos de palabras con su peso correspondiente.

[6] Para indicar este proceso de resumen de un texto, en el ámbito tecnológico se utiliza la palabra sumarización, lo cual es una mala castellanización de la palabra inglesa *summarization*.

distintos y que de más peso a palabras que tienen inferior frecuencia de aparición en el conjunto de la obra. A este tipo de pesado se le denomina global[7]. En la Figura 3 se puede ver el resultado de aplicar este pesado para el mismo subcapítulo del texto.

Figura 3. Nubes de palabra para el capítulo 3 y subcapítulo 8 de la parte IV, utilizando pesado global.

Con este pesado global se nos muestra, entre otros aspectos, que los personajes Torquemada y Leopoldo tienen un peso más importante dentro de este subcapítulo en comparación con el total de la obra. Hay que considerar que, aunque parezca una contradicción terminológica, mientras con el pesado local se nos muestra información más global dentro del contexto de la obra, con el pesado global se nos da información más específica dentro del contexto local del subcapítulo.

Aplicando estas técnicas se ha realizado una representación visual de cada parte, de todos los capítulos y los subcapítulos de *Fortunata y Jacinta*. Dado que la inclusión de los resultados de los 31 capítulos o 198 subcapítulos sobrepasaría con creces la longitud de este artículo, a título de ejemplo, se muestran en la Figura 4 solo el resultado de aplicar ambos pesados al capítulo 2, donde se puede observar cómo la trama principal —pesado local— de este capítulo gira principalmente entorno a los personajes de Bárbara y Baldomero sobre el negocio textil de esta familia —representado con las palabras 'tienda', 'genero' o 'color')—. En cambio, con el pesado global se nos representan otros personajes que adquieren un protagonismo más importante dentro de este capítulo sobre el conjunto de la obra, como son Bonifacio y Gumersindo, en donde la trama de este capítulo se especifica mayormente los aspectos relativos al negocio textil —ejemplificado con palabras como 'pañolería'—.

Por último, también podemos disponer de una visión general de la obra en base a representar un resumen para cada una de las partes. En este sentido, la Figura 5 representa el resumen de cada una de las cuatro partes, donde, entre otras cosas, se puede ver la distinta importancia que tienen los personajes en cada parte: por ejemplo, mientras en la primera parte la trama transcurre principalmente sobre los personajes de Jacinta y Juanito, en las siguientes partes

[7] El pesado global que se ha utilizado es TF-IDF (*Term frequency – Inverse document frequency*), el cual es un pesado ampliamente utilizado en el procesamiento de lenguaje natural (Salton y Bucley: 1988).

Fortunata es la principal protagonista junto a Maximiliano en la parte II y IV y Evaristo y Doña Lupe en la parte III. Por otra parte, se observa como la temática de cada parte gira sobre aspectos diferentes: mientras que para la parte I tenemos que predomina la familia —de Santa Cruz— y para las otras partes la esfera de acción gira en base a tramas distintas.

Figura 4. Nubes de palabra, con pesado local y global, para el capítulo 2 de Fortunata y Jacinta.

Figura 5. Nubes de palabras para cada una de las cuatro partes de Fortunata y Jacinta.

En verdad, lo que nos muestra una máquina con los resultados de este apartado no deja de ser un resumen representado visualmente en forma de nubes de palabras. En cualquier caso, la aplicación de técnicas de pesado nos puede proveer de una mayor información. En nuestro

caso, hemos aplicado un pesado global que nos muestra información relevante del capítulo o subcapítulo que no sigue la trama general de la obra. Pero, se pueden aplicar otros pesados, por ejemplo, ponderando mayormente alguno de los tipos de nombres, adjetivos o verbos, de forma que se resalte algún tipo de aspecto semántico que este bajo el interés del analista.

DESCUBRIMIENTO DE LOS TEMAS

Una máquina no sabe qué es un tema tal como lo entendemos los humanos. Lo que sí puede llegar a obtener son qué palabras guardan algún tipo de relación entre sí, que agrupadas forman un conjunto de palabras. En el contexto del procesamiento del lenguaje natural, un tema se puede definir como una distribución de palabras (Blei, et al.: 2003, 994) junto con un peso para cada palabra. En otras palabras, una máquina considera un tema a cada uno de los conjuntos de palabras que encuentra.

Existen diversos de algoritmos que localizan temas en base a agrupar palabras que guardan relación entre sí, pudiendo cada uno obtener resultados diferentes, es decir, nos propondrán temas diferentes. En nuestro caso, se ha seleccionado un método que es fácilmente interpretable y que da buenos resultados cuando se procesa un texto, el cual es un algoritmo matemático ampliamente conocido en el mundo de la ciencia de datos que se basa en separar la matriz inicial que representa el texto en tres matrices[8]. En este sentido, la matriz que vimos anteriormente en la Figura 1 se descompone en tres matrices, que multiplicadas nos da esa misma matriz.

En la Figura 6 se puede ver estas tres matrices, donde cada una nos da información en tres aspectos distintos. Se parte de un número de temas, *N*, donde este número lo fija la persona que quiere analizar el texto. El significado de cada matriz es el siguiente: en la matriz A los números de cada columna nos indican la importancia de cada tema para cada subcapítulo; la matriz B nos da información de la importancia de cada tema en el texto completo; y, en la matriz C, los números de cada fila nos dan información del peso de cada palabra para cada tema.

[8] Este método se denomina SVD, en español, *Descomposición de Valores Singulares* (Golub y Reinsch: 1971).

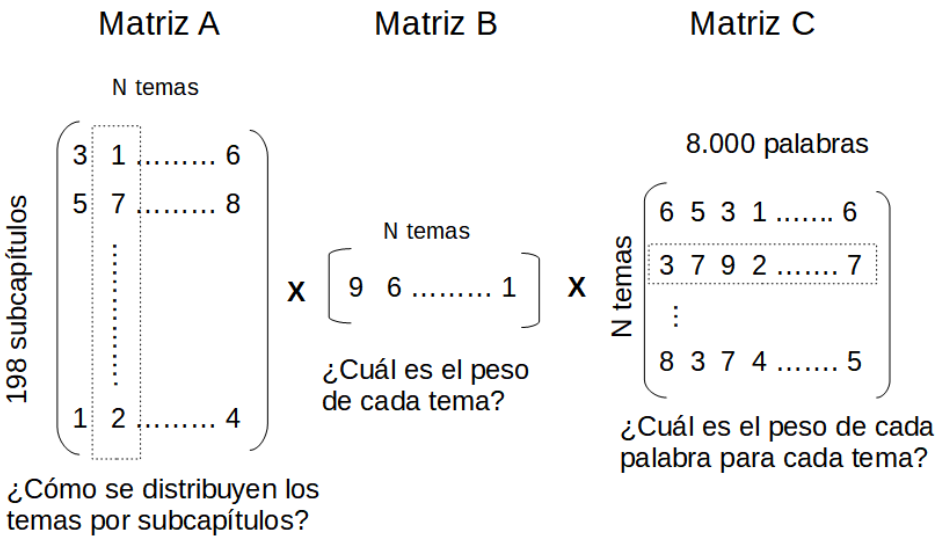

Figura 6. Descomposición de la matriz original en tres matrices.

Como se ha dicho, el número de temas es un valor que lo fija un humano. En verdad, no existe un método automático que nos diga cuántos temas existen. De esta forma, es una elección arbitraria por parte de un humano. Para nuestro caso, se han seleccionado 10 temas. Los temas se han etiquetado del 1 al 10, sin darles un nombre específico. En la Figura 7 se representa los valores de la matriz B, es decir, el peso que tiene cada uno de los 10 temas, donde se observa que el tema 10 es el que pesa más y el tema 1 el que menos.

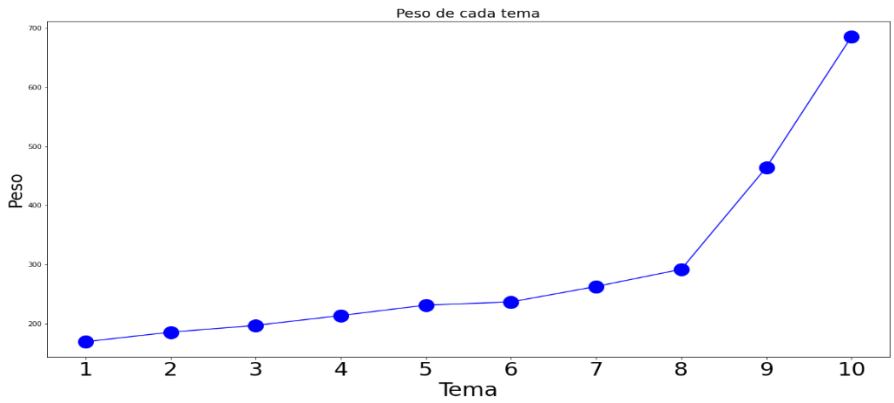

Figura 7. Valores de la matriz C: Importancia o peso de cada tema.

Por cuestiones de espacio, solo se muestran en la Figura 8 y Figura 9 los seis temas con más peso que la máquina ha descubierto, esto es, los temas 10, 9, 8, 7, 6 y 5. Para cada uno de los temas se muestran las nubes de palabras que nos define el tema. Estas nubes de palabras son el resultado de representar los números de cada fila de la matriz C: en otras palabras, el peso que tiene cada palabra en el tema. Cada tema viene acompañado por una gráfica, la cual nos

Galván González, Victoria (coord.) (2023): *Coda a un centenario. Galdós, miradas y perspectivas.*
Casa-Museo Pérez Galdós, Cabildo de Gran Canaria. Las Palmas de Gran Canaria. ISBN: 978-84-1353-120-5.

representa los números de cada columna de la matriz A, esto es, nos indican el peso que tiene cada subcapítulo en el tema. Por tanto, es la matriz C la que utiliza la máquina para descubrir la información de los temas que obra en base a conjuntos de palabras y que le corresponderá a un humano interpretarla.

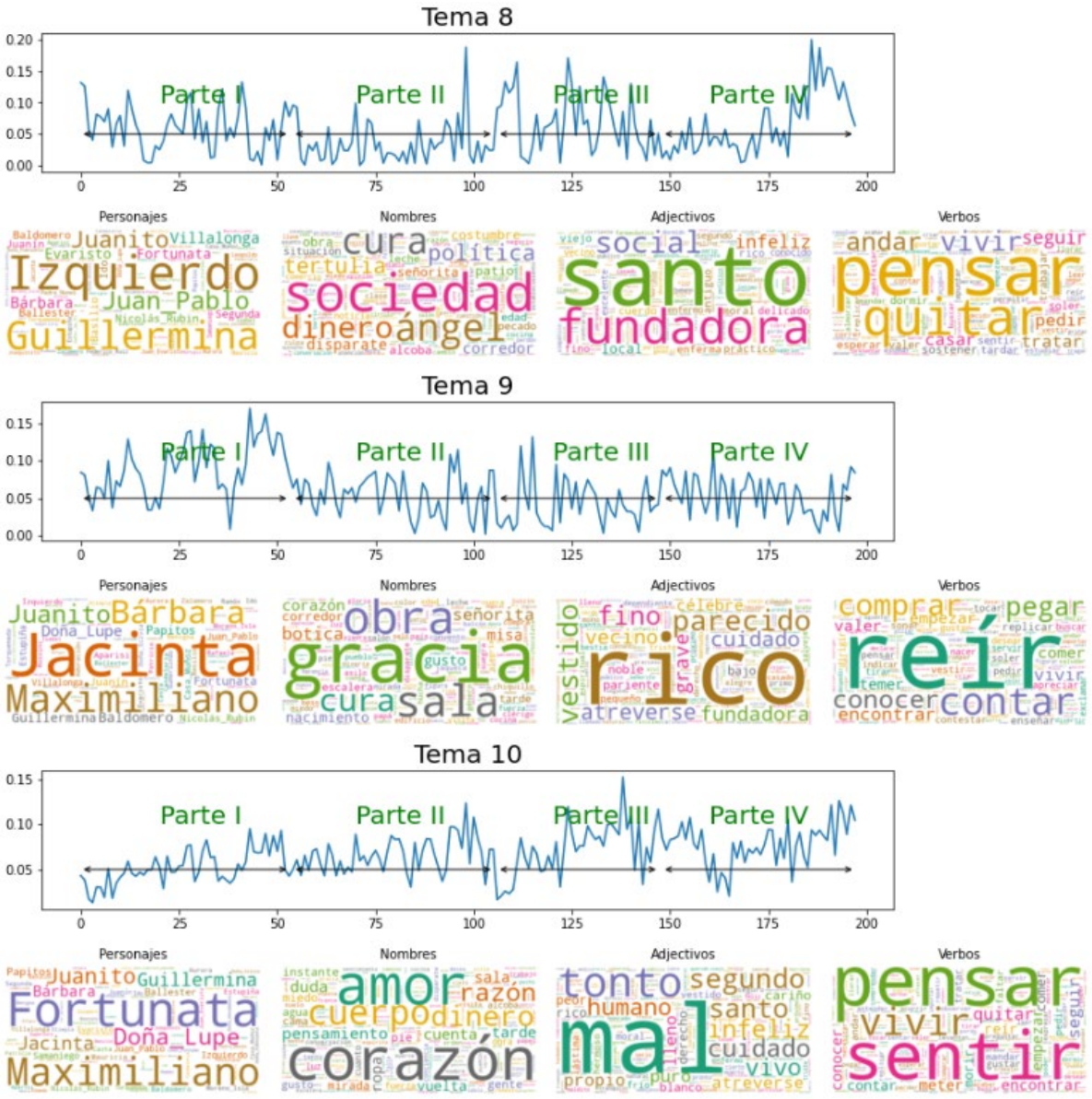

Figura 8. Representación de los temas 10, 9 y 8 junto con su peso en cada subcapítulo. Con pesado local.

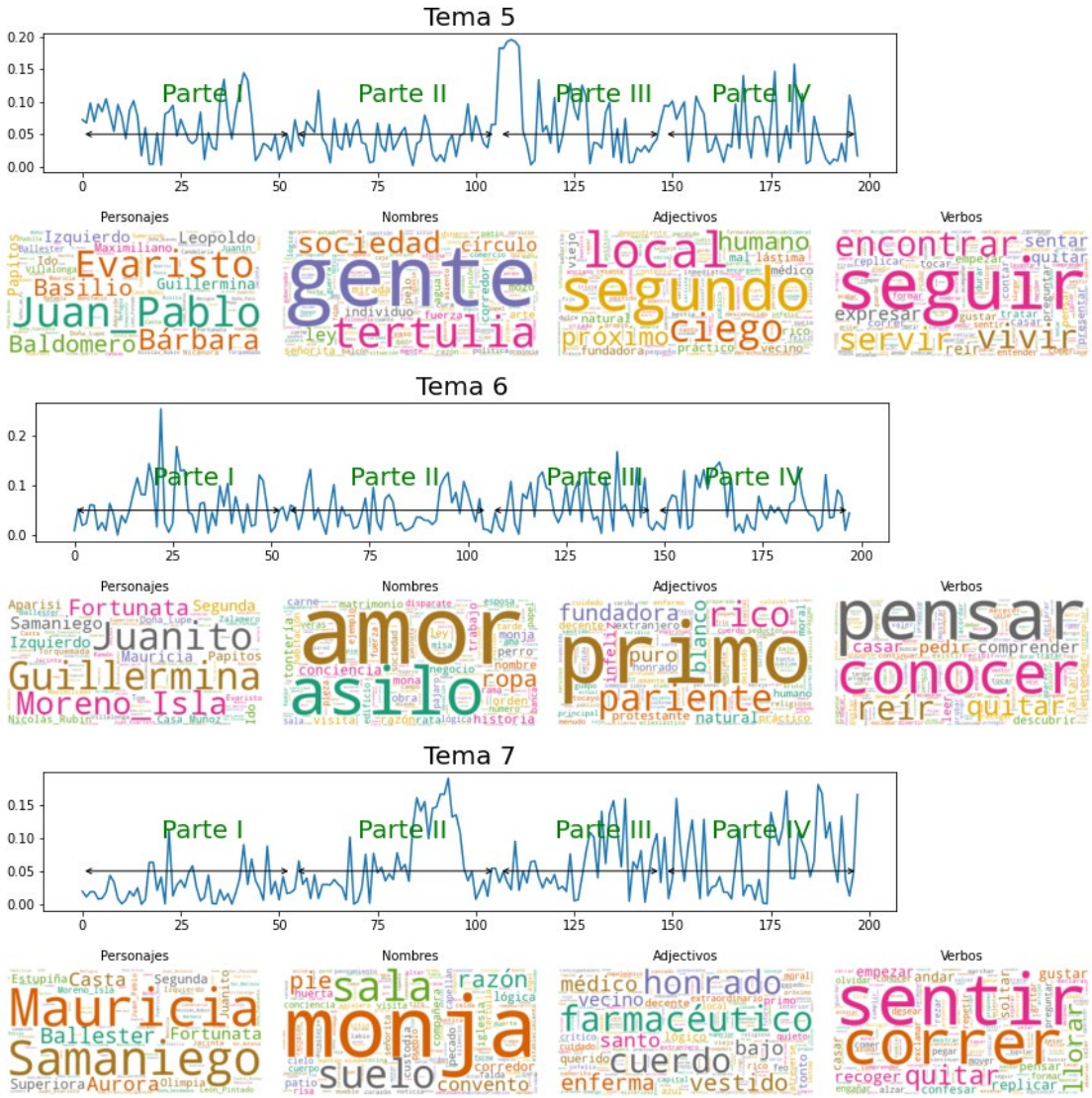

Figura 9. Representación de los temas 7, 6 y 5 junto con su peso en cada subcapítulo. Con pesado local.

Desde que la lógica de una máquina es diferente a la de un humano, no todos los temas que encuentra una máquina tienen que poder ser interpretados directamente siguiendo los cánones de expertos humanos. Por ejemplo, Sherstinova et al. (2022: 307-309) para poder evaluar distintos modelos recurrieron a la evaluación por expertos, tres filólogos con formación literaria y lingüística, los cuales evaluaron de forma independiente la interpretabilidad de todos los temas sobre obras literarias obtenidos por cada modelo, determinando para cada tema si es interpretable o no interpretable. Un tema lo consideraban interpretable si la mayoría de las palabras que formaban el tópico se relacionaban dentro de un mismo o similar campo semántico, o si era posible construir una historia plausible o un fragmento de una historia, basándose en las palabras que definían cada tema. Los resultados no fueron uniformes para todos los expertos, llegando a obtener porcentajes de interpretabilidad de los temas para el

mejor de los casos del orden del 50%. Por tanto, los temas que encuentra una máquina deben ser considerados como propuestas de temas que hace la máquina al analista experto, el cual será el que en mayor medida pueda interpretarlos.

Desde que el objetivo principal de este trabajo es mostrar los resultados que nos da una máquina, por tanto, no se va a realizar un análisis exhaustivo o interpretación de los resultados obtenidos. Sin embargo, y a título de ejemplo, a partir de las Figura 8, se puede realizar un simple análisis de los temas principales que ha descubierto la máquina. En el tema 10, el personaje de Fortunata, junto a Maximiliano, es el que tiene más peso, girando el tema sobre el 'amor' y el 'corazón'; y, según la gráfica que lo acompaña, viene a ser más predominante durante la parte III y IV. En el tema 9, Jacinta es el personaje con más peso, girando sobre el tema religioso, tomando más importancia durante el final de la parte I. En el tema 8, son Guillermina e Izquierdo los principales protagonistas, donde parece tener más peso la parte económica, con mayor peso al final de la parte IV.

Por otra parte, estos temas se han obtenido a partir de la matriz con pesado local. Sin embargo, se puede partir de la matriz con pesado global. En la Figura 10 se muestran los tres temas con más peso. Al utilizar el pesado global, los temas que la máquina ha descubierto hay que interpretarlos como temas que se salen de la trama general de la obra, pero que tienen importancia dentro de esta. En este sentido, se puede observar en la Figura 10 como los temas 8 y 9 se concentran en unos pocos capítulos muy específicos de la obra. Por ejemplo, el tema 8, con pesado global, gira sobre el comercio y se concentra principalmente a principio de la parte I.

Finalmente, hay que tener en cuenta que el descubrimiento de los temas que nos propone la máquina no se tiene que considerar como una forma de contrastar temas encontrados a partir de un análisis con una lectura atenta. En verdad, los resultados que nos proporciona una máquina son propuestas de temas que buscan guiar al analista a profundizar en temas que pueden haber pasado desapercibidos a través de una lectura atenta. Por otra parte, se puede afirmar que partiendo de pesados distintos obtenemos temas distintos, con lo que el analista puede aplicar sus propios sesgos ponderando aspectos específicos, por ejemplo, incluyendo lugares y ponderándolos, de forma que los resultados giren en función de los lugares donde se desarrolla la obra.

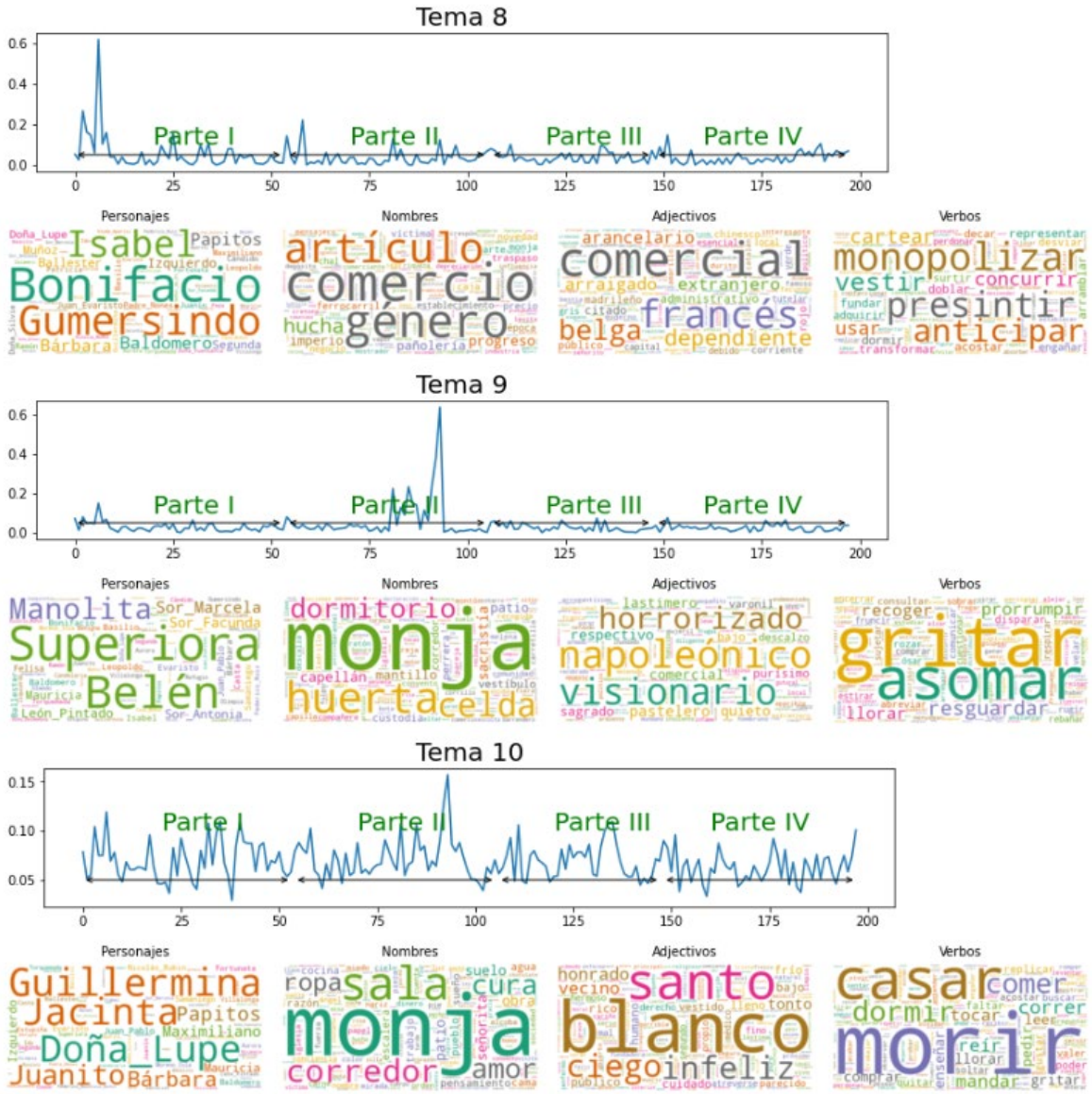

Figura 10. Representación de los temas 10, 9 y 8 junto con su peso en cada subcapítulo. Con pesado global.

ANÁLISIS DE SENTIMIENTO

Otro aspecto que puede ser objetivo de análisis es el descubrimiento del sentimiento que nos transmite un texto. Para descubrir el sentimiento existen diversas técnicas. Estas técnicas determinan la polaridad positiva, neutra o negativa de un texto. Por ejemplo, la frase 'estoy caminando' tiene un sentimiento neutro; 'estoy contento' tiene polaridad positiva, y 'estoy muy contento', todavía es más positiva; por contra, 'estoy triste' tiene un sentimiento negativo. Existe una abundante literatura sobre el análisis de sentimiento (Birjali, et al.: 2021), teniendo un importante campo de aplicación dentro de las redes sociales y que, en nuestro caso, se ha

Galván González, Victoria (coord.) (2023): *Coda a un centenario. Galdós, miradas y perspectivas.*
Casa-Museo Pérez Galdós, Cabildo de Gran Canaria. Las Palmas de Gran Canaria. ISBN: 978-84-1353-120-5.

aplicado sobre un texto literario. Hay que decir que no existe un único enfoque o técnicas para abordar el análisis del sentimiento.

Un enfoque para abordar el análisis del sentimiento está basado en un paradigma simbólico o de reglas que hace uso de una serie de diccionarios, teniendo en cuenta además el papel de la negación, la intensificación y las expresiones, es decir, se analiza la sintaxis de una expresión con el contenido de ciertas palabras que nos indican distintos grados de positividad o negatividad, y se le da un valor. Esto, en sí mismo, tienen un sesgo de quién ha definido estos diccionarios con los valores asociados pero que, en cualquier caso, nos indica valores numéricos y continuos que permiten guiar al analista a comparar diferentes partes de un texto en base a la tendencia del sentimiento.

Otro enfoque se basa en la aplicación de un paradigma conexionista —es decir, aplicando redes neuronales— sobre corpus entrenados, es decir, se etiqueta un corpus manualmente por un humano indicando para cada elemento si la polaridad es positiva, negativa o neutra, y posteriormente se entrena una red neuronal con este corpus etiquetado, para a continuación comparar en qué medida una expresión es similar a estas expresiones etiquetadas. Para nuestro caso, este enfoque tiene el inconveniente de que los corpus entrenados en español se basan en textos actuales, lo cual pueden no ser fiel reflejo sobre un texto de época de *Fortunata y Jacinta*. Por otra parte, los resultados no suelen tomar valores continuos, es decir, nos indican solo tres posibles valores, si es positivo, neutro o negativo, sin cuantificarlos.

Por ello, se ha optado por el uso del primer enfoque, es decir, se utilizará herramientas que hagan uso de reglas para determinar la polaridad del sentimiento de un texto[9]. Si aplicamos estas técnicas a nuestro texto podemos ver el resultado en la Figura 11, que nos muestra cómo va variando el sentimiento a lo largo de la obra[10].

[9] En nuestro trabajo, para el análisis de sentimiento se ha utilizado la herramienta Textblob (Loria: 2020), la cual es una librería basada en el lenguaje de programación Python.

[10] Hay que indicar que sobre este análisis de sentimiento se puede realizar un estudio más profundo. Por ejemplo, un subcapítulo puede ser etiquetado como neutro ya que tiene partes positivas y negativas que se contrarrestan. En este sentido, se puede analizar cada subcapítulo con más profundidad, determinando qué partes, párrafos o frases son más negativas o positivas.

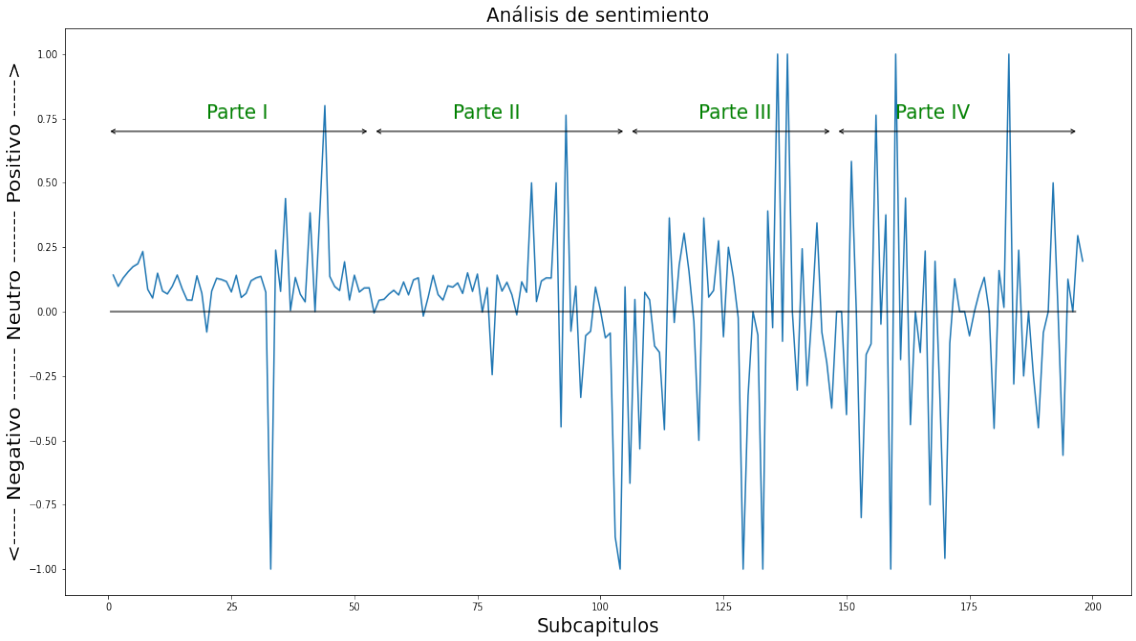

Figura 11. Distribución por subcapítulos del sentimiento en Fortunata y Jacinta.

Tal como se observa en la Figura 11, desde el inicio de la obra hasta aproximadamente la mitad de la segunda parte, se transmite un sentimiento más uniforme y ligeramente positivo, donde se observa que existen dos subcapítulos, uno claramente negativo (cap. 8, sub. 4) y otro claramente positivo (cap. 10, sub. 1). En el primer caso, se trata del episodio donde, estando Jacinta y Juanito conversando, entra en escena el infeliz don José Ido, representándose como hambriento y tocado de la locura por los celos; Juanito se mofa del infeliz, mientras a Jacinta le produce miedo y pena. En el segundo caso, don Baldomero anuncia que tocó la lotería, con la correspondiente alegría en el reparto del premio a los agraciados. En resumidas cuentas, se trata de dos episodios que rompen con la tendencia más neutral, en cuanto al sentimiento transmitido, de la obra y que las técnicas de análisis del sentimiento permiten detectar.

También, se puede observar que las partes III y VI son más caóticas, en el sentido, que van variando entre subcapítulos negativos y positivos. Así, se puede observar que entre el final de la parte II y principio de la parte III predomina un sentimiento más negativo, con independencia que existen tramos ligeramente positivos. Por otra parte, las partes III y IV son menos neutras, con más subcapítulos positivos y negativos, que van alternándose, lo que nos indica una trama a lo largo de estas dos partes más agitada.

En resumen, estos resultados nos muestran un mapa de la obra, que permite localizar cómo el sentimiento va variando a lo largo de la obra y, de esta forma, el crítico dispone de evidencias que le permiten dirigirse hacia partes específicas y realizar un análisis más detallado sobre aquellos aspectos que hacen del texto tenga una orientación más positiva o negativa.

Galván González, Victoria (coord.) (2023): *Coda a un centenario. Galdós, miradas y perspectivas.*
Casa-Museo Pérez Galdós, Cabildo de Gran Canaria. Las Palmas de Gran Canaria. ISBN: 978-84-1353-120-5.

DESCUBRIMIENTO DE INFORMACIÓN SOBRE LOS PERSONAJES

En verdad, no existe una única forma de abordar el análisis de los personajes. Inferir el personaje no deja de ser un reto desde una perspectiva literaria desde que no se ha llegado a un consenso sobre el significado del término. Lo primero que hay que tener en cuenta es que no está claro qué es un personaje en un texto literario. Pensemos en un personaje de *Fortunata y Jacinta*, Barbarita. Veamos el siguiente extracto de esta novela:

> Barbarita no se trataba con todos los individuos que aparecen en esta complicada enredadera. A muchos les esquivaba por hallarse demasiado altos; a otros apenas les distinguía por hallarse muy bajos. Sus amistades verdaderas, como los parentescos reconocidos, no eran en gran número, aunque sí abarcaban un círculo muy extenso, en el cual se entremezclaban todas las jerarquías. En un mismo día, al salir de paseo o de compras, cambiaba saludos más o menos afectuosos con la de Ruiz Ochoa, con la generala Minio, con Adela Trujillo, con un Villuendas rico, con un Villuendas pobre, con el pescadero pariente de Samaniego, con la duquesa de Gravelinas, con un Moreno Vallejo magistrado, con un Moreno Rubio médico, con un Moreno Jáuregui sombrerero, con un Aparisi canónigo, con varios horteras, con tan diversa gente, en fin, que otra persona de menos tino habría trocado los nombres y tratamientos. (Pérez Galdós: 1887, parte I, cap. IV, sub. 2).

A parte de Barbarita, en este extracto podemos ver diversas personas[11]. Está claro para un humano que sobre Barbarita giran el resto de las personas, por ello, aunque solo aparezca solo una vez de forma explícita, en verdad, Barbarita aparece más de una vez en este texto, a través de correferencias como 'cambiaba saludos con…'. Es por ello, que un humano infiere que Barbarita es el eje central de este texto. Por eso, tenemos claro que Barbarita es un personaje de este texto. Pero ¿qué podemos decir que el resto de las personas son realmente personajes? Desde un punto de vista extenso podemos afirmar que sí, ya que realmente cada uno de estos personajes nos dicen algo.

En definitiva, una máquina se encuentra con un problema: descubrir cuáles son las personas clave, es decir, los personajes y que serán objeto de análisis. Por otra parte, aparece un segundo problema: descubrir las menciones a una misma persona con referencias distintas.

Para entender el primer problema, antes debemos concretar qué entendemos por un personaje. Se parte de la definición de Chatman (2006: 219) sobre lo que se considera un personaje en una narración, siempre buscando una definición que una máquina pueda tratar. De esta forma, desde una concepción de los formalistas y de algunos estructuralistas los

[11] Entre las referencias a personas tenemos: *'todos los individuos que aparecen en esta complicada enredadera', 'la de Ruiz Ochoa', 'la generala Minio', 'Adela Trujillo', 'un Villuendas rico', 'un Villuendas pobre', 'el pescadero pariente de Samaniego', 'la duquesa de Gravelinas', 'un Moreno Vallejo magistrado', 'un Moreno Rubio médico', 'un Moreno Jáuregui sombrerero', 'un Aparisi canónigo', 'varios horteras'* y *'diversa gente'*.

personajes son productos de las tramas, que su estatus es funcional, que estos son participantes. En esta concepción, se debe evitar los aspectos psicológicos, ya que lo que define un personaje son sus funciones. En otras palabras, se tiene que analizar sólo lo que los personajes hacen en una historia, no lo que son, y analizar sus esferas de acción. Así, un personaje es algo que desempeña cualquiera acción sobre un conjunto de papeles activos en una narración y la trama denota los principales acontecimientos de una historia. De esta forma, un personaje es un ser animado que es importante para la trama que, en el caso de nuestro ejemplo, Barbarita se puede definir como personaje, pero no a las diversas entidades que encuentran casualmente en su paseo.

Por tanto, una máquina debería inferir cuál es la función de todas las personas que aparecen en la obra, cuáles son sus esferas de acción y determinar si es importante en la trama. Realmente, estamos hablando de que la máquina tome decisiones subjetivas para seleccionar aquellas personas que puedan considerarse personajes. Aunque este tipo de decisiones sean implementadas en una máquina, siempre serán definidas por un humano con sus propios sesgos. Por ello, en este trabajo, se ha facilitado el trabajo a la máquina seleccionando los personajes manualmente. En este sentido, se han seleccionado 86 personajes, los cuales se les ha indicado a la máquina que son los que debe analizar.

El segundo problema, es decir, el descubrir las menciones a una misma persona con referencias distintas, realmente trata de resolver retos distintos desde el punto de vista de una máquina. Tenemos las distintas formas que se mencionan a un mismo personaje, por ejemplo, en el caso de Barbarita se la menciona de distintas formas a lo largo de la obra: ‘Bárbara’, ‘Bárbara Arnáiz’, ‘Barbarita Arnáiz’, ‘esposa de Santa Cruz’, ‘la señora de Baldomero Santa Cruz’. Tenemos, además, apelativos sin ningún referente claro para una máquina, por ejemplo: Fortunata con ‘la Pitusa’ o Jacinta con ‘la Delfina’. ¿Cómo sabe la máquina que realmente el texto está hablando de la misma persona?

Por tanto, en este caso también la máquina debería poder inferir que, de todas las entidades de nombres de personas, muchas realmente se refieren al mismo personaje. Este es un problema que se ha abordado a través de distintas técnicas suponiendo importantes esfuerzos de implementación y no con buenos resultados. Por ello, en muchos casos, es el humano el que facilita manualmente la tarea a la máquina a través de definir diccionarios de personajes, cada uno definido con un conjunto de las distintas menciones que aparecen en el texto. En este sentido, en este trabajo se ha definido un diccionario que abarca los 86 personajes donde a cada uno se relaciona con las distintas formas que aparecen en el texto.

Una vez que se ataja los dos problemas anteriores, es decir, el descubrimiento de los personajes en el texto, a través de una ayuda manual a la máquina, tenemos otro problema: ¿cómo describir un personaje? En verdad, un personaje se puede describir con distintas orientaciones: puede ser descrito en base a un perfil psicológico (Flekova. y Gurevych: 2015)[12], en base a la función que desempeña en un texto (Valls-Vargas, et al.: 2015)[13], o sobre el contexto en que se mueve dentro de la obra (Bamman, et al.: 2014). En nuestro caso, se ha basado en esta última aproximación, para lo que se describe un personaje en base a su contexto.

Hasta ahora, hemos visto una representación del texto literario basado en una matriz donde figuran las palabras que aparecen en los subcapítulos de la obra. Como se ha expuesto más arriba, en esta representación no se tiene información sobre las palabras en función de su orden de aparición en el texto, con lo que se pierde en gran medida su contexto. Existen otras representaciones que sí nos dan información sobre el contexto de cada palabra con el resto. Son representaciones mucho más complejas y muchas de ellas basadas en redes neuronales.

Sin entrar a describir estas redes neuronales, los resultados que nos dan estos modelos también son vectores para cada palabra pero que, en lugar de darnos información de su aparición en cada subcapítulo, nos dan información del contexto con otras palabras[14]. Por ejemplo, la palabra 'Fortunata' nos produce un vector como el siguiente: [0,2341, 0.3445, …, 0.7899]. Estos vectores tienen longitudes que elige el humano, pero que suelen ser superiores a los 300 números. Si observamos este vector, realmente un humano tiene muchas dificultades para interpretarlo. Este es uno de los grandes problemas del uso de las redes neuronales: es muy complicado interpretar los resultados.

Por ello, lo mejor es no entrar a entender los números de estos vectores, sino tan solo saber que nos dan información sobre la cercanía de cada palabra con el resto. A esto se le denomina similitud. Las medidas de similitud basadas en el corpus consisten en realizar cálculos de similitud entre palabras. En este tipo de enfoques se extrae información valiosa del análisis de un corpus que, en nuestro caso, es la obra *Fortunata y Jacinta*. El principio de la hipótesis distributiva establece que las palabras con significados similares aparecen en contextos

[12] La forma más ampliamente utilizada para abordar la descripción de los personajes, o la predicción del rasgo psicológico de los personajes, utilizando herramientas de procesamiento de lenguaje natural se hace procurando encontrar correlaciones entre los aspectos léxicos y estilísticos del texto y ciertos rasgos de personalidad que se definen en el campo de la psicología.

[13] Estas funciones del personaje se basan en la obra de Vladimir Propp, el cual desarrolla una teoría narrativa estructuralista, donde clasificó a los personajes de los cuentos populares rusos en varios papeles funcionales básicos o funciones del personaje: Héroe, Villano, Despachador, Donante, Ayudante (Mágico), Buscador y Falso Héroe. Cada rol cumple funciones narrativas específicas y desempeña su esfera de acción concreta.

[14] En el procesamiento del lenguaje natural esta forma de representar a las palabras se le denomina *word embedding*, incrustación de palabras, que es una técnica que asigna un vector a cada palabra, el cual guarda información semántica en función de diferentes contextos gramaticales.

similares en los documentos (Gorman y Curran: 2006, 361) y este principio constituye la base de la mayoría de los métodos basados en corpus, que tiene como consecuencia que estos métodos no tienen en cuenta el significado real de las palabras individuales, sino el contexto que las acompañan. Otra consecuencia de la aplicación de esta hipótesis es que palabras que suelen aparecer juntas serán también similares.

Aplicando funciones matemáticas sencillas se puede saber lo similares que son dos palabras[15]. Por ejemplo, si tenemos dos frases como 'me tomo una taza de café' y 'me tomo una taza de té', 'café' y 'té' se consideran similares ya que tienden a aparecer en el mismo contexto. Sin embargo, esto no nos indica que sean sinónimas. En otras palabras, la similitud no indica en qué medida dos palabras están próximas en el texto entre sí, sino en qué medida dos palabras tienen las mismas palabras próximas entre sí o tienen contextos similares.

En este sentido, si elegimos una palabra, por ejemplo, 'Fortunata', podemos ver qué otras palabras son más similares. Para ello, previamente hay que entrenar redes neuronales para calcular los vectores de cada palabra. Las posibilidades de sacar resultados de similitud son muy amplias y estas posibilidades las decide un humano. Básicamente, se pueden aplicar dos enfoques para analizar esta similitud. Uno es basarse en valores de vectores obtenidos de corpus ya entrenados[16] y otro en basarse en el mismo contexto del texto, para lo que hay que entrenar el mismo texto. El primer caso se ha descartado, dado que es necesario disponer de un corpus amplio de la época de *Fortunata y Jacinta*, y utilizar corpus actuales puede generar significados inconsistentes[17]; y, por otra parte, los nombres de los personajes son característicos de la misma obra.

Por ello, en nuestro caso, se ha incorporado el texto de *Fortunata y Jacinta* en una red neuronal[18], lo que se denomina, 'entrenar una red neuronal', y se han obtenido los vectores para todas las palabras. En esta parte solo se ha centrado en los 86 personajes seleccionados. De esta forma, una máquina cuantifica la similitud de cada personaje en función de otros personajes, nombres, adjetivos y verbos.

[15] Existen diversas funciones para calcular la similitud entre dos palabras, sin embargo, la función coseno es la que más ampliamente ha sido utilizada entre los investigadores del procesamiento del lenguaje natural (Mohammad y Hirst: 2012), y es la que se ha utilizado en este trabajo.

[16] Por ejemplo, para el idioma español existen colecciones de vectores de palabras basados en corpus que abarcan del orden de 1,5 billones de palabras obtenidos de multitud de fuentes (Cardellino: 2016).

[17] A título de ejemplo, la palabra 'remotas' que aparece en la primera frase en *Fortunata y Jacinta*, utilizando la colección de Cardellino (2016) nos da palabras similares como 'máquina' y 'servidor', las cuales están asociadas con el contexto actual y no con el contexto de la época de Benito Pérez Galdós.

[18] En este trabajo, se ha utilizado el algoritmo de word2vec (Mikolov et al: 2013), el cual es uno de los más ampliamente utilizados para el *word embedding*.

Cada tipo de palabra nos podrá indicar aspectos distintos de un personaje. Así, un personaje similar nos indicaría que existe una relación entre ellos, que pueden ser tanto por su interacción mutua en el texto, como por que ambos interaccionan en los mismos contextos. Los nombres nos darían más información sobre en qué temas se mueven en la obra. Los adjetivos nos definirían cómo es el personaje. Por último, los verbos nos describirán cuáles son las acciones principales del personaje en la obra.

De los 86 personajes analizados, solo se mostrarán, en la Figura 12 y Figura 13 los resultados para los personajes Fortunata y Jacinta, respectivamente. En primer lugar, nos aparece su descripción a nivel de toda la obra. Por otra parte, en una obra como *Fortunata y Jacinta* los personajes evolucionan, por ello, un análisis interesante es cómo un personaje va variando a lo largo de una obra. Por ejemplo, podemos ver cómo la máquina describe un personaje para cada parte. Para ello, entrenamos nuestra red neuronal para cada parte de la obra por separado.

Figura 12. Descripción y evolución del personaje Fortunata por cada parte de la obra.

Tampoco, en este caso, se va a entrar en un análisis detallado de estos resultados. En cualquier caso, se observa cómo el personaje Fortunata va cambiando a lo largo del texto en los diferentes aspectos del análisis. Si nos centramos en los adjetivos, mientras en el texto completo a Fortunata se le asocia con 'inútil' y 'honrada', se observa que va cambiando en cuando se considera cada una de las partes por separado: en la parte I se puede ver una Fortunata 'abierta' y 'pura'; 'fina' en la parte II; 'rica' en la parte III; en cambio, en la parte IV se le asocia con 'propio', que, si se analiza el texto de esta parte, este adjetivo se puede encuadrar dentro de 'amor propio'. Otra información que se puede extraer está asociada con las relaciones o afinidad de Fortunata con otros personajes. Por ejemplo, se puede ver que, mientras a nivel de toda la novela prima la afinidad de Fortunata con Maximiliano, Jacinta y Guillermina, en las diversas partes los pesos de las relaciones con otros personajes van cambiando. En este caso, llama la atención que no se resalta la afinidad con Juanito. En cuanto al análisis de los nombres y verbos, en el primer caso, esto nos da más información de Fortunata sobre la esfera de temas sobre los que prevalece la acción y, en el segundo caso, sobre su tipo de acción.

Figura 13. Descripción y evolución del personaje Jacinta por cada parte de la obra.

Galván González, Victoria (coord.) (2023): *Coda a un centenario. Galdós, miradas y perspectivas.*
Casa-Museo Pérez Galdós, Cabildo de Gran Canaria. Las Palmas de Gran Canaria. ISBN: 978-84-1353-120-5.

En el caso se Jacinta se puede ver también una evolución del personaje. En este caso, se observa que en el conjunto de la obra Jacinta se tiene afinidad con Juanito, mientras que esta se diluye para cada una de las partes. Fijándonos en los adjetivos, vemos cómo tenemos una Jacinta 'dulce' en la primera parte; y 'espiritual' en la segunda parte. Mientras que en la parte III y IV la máquina la define de una forma similar con los adjetivos 'hermosa' y 'loca'.

En definitiva, la descripción de los personajes usando este tipo de algoritmos nos proporciona una visión más cuantitativa sobre los personajes. No hay que olvidar que los personajes son una parte esencial del análisis de una obra, sobre todo en una obra compleja como *Fortunata y Jacinta*, por tanto, una herramienta de este tipo abre al analista con nuevas perspectivas para su estudio basado en las evidencias encontradas por una máquina. Por último, hay que recordar que existen otras formas de abordar la descripción de un personaje, por ejemplo, asignándoles características psicológicas, lo que abre en mayor medida las posibilidades de estas técnicas.

CONCLUSIONES

En la conclusión toca contestar a la pregunta: ¿Qué nos dice una máquina de *Fortunata y Jacinta*? En primer lugar, la máquina no nos dice lo que ella quiere, en el sentido que sus respuestas van guiadas por los intereses de un humano. Un humano le dice a la máquina qué algoritmos utilizar: se ha visto que para el mismo problema existen una variedad de algoritmos que no tienen que dar los mismos resultados. Además, sobre estos algoritmos el humano le indica a la máquina que aplique una serie de sesgos. En el contexto de este trabajo, el uso de sesgos no debe considerarse que tenga una connotación negativa, dado que son estos sesgos lo que dicen a la máquina sobre qué aspectos debe centrarse. En otras palabras, es el humano el que guía a la máquina en base a sus intereses, y sobre estos, la máquina nos muestra una serie de resultados.

A lo largo de este trabajo se ha ido mostrando parte de estos resultados que, dada la limitación de espacio, no dejan de ser una parte reducida de lo que realmente se ha generado. En todo caso, podemos decir que una máquina nos puede decir muchas cosas. Se ha visto que, aparte de resumir visualmente cada subcapítulo —o capítulo—, a través del pesado global nos da información sobre distintos aspectos, como es el caso de personajes o temas, que son relevantes en la trama concreta del subcapítulo y no en la trama global. La máquina nos propone como resultados un conjunto de temas, ya sea sobre la trama global de la obra, o como temas que son muy específicos en partes concretas de esta. Nos da un mapa detallado del sentimiento

de la obra, informándonos dónde existen subcapítulos con sentimiento positivo, neutro o negativo. Por último, nos provee de una descripción de los personajes en base al contexto sobre el que se mueven a lo largo del texto.

En verdad, estos resultados no dejan de ser la punta del iceberg de toda la información que nos pueden aportar este tipo de técnicas utilizadas dentro de la inteligencia artificial. Se puede decir que para una obra literaria tan compleja tan compleja como *Fortunata y Jacinta* las posibilidades de extraer información son ilimitadas, donde la interacción con el experto literario y lingüista será clave para que guiar a la máquina y para la interpretación de los resultados. En otras palabras, una máquina nos podrá decir muchas más cosas, pero será el analista quién realmente interprete e indague en los resultados.

En cualquier caso, lo qué nos da una máquina es información que, en base a esta, el humano tiene que generar mayor conocimiento. Básicamente, la máquina, procesando numérica y rápidamente gran cantidad de datos, nos provee de resultados que el humano difícilmente puede obtener a través de la lectura atenta. Estos resultados no solo tienen que ser interpretados por un humano, sino que también la máquina le indica dónde buscar o sobre qué aspectos nuevos puede indagar a través de una lectura atenta. No es, por tanto, el objetivo el utilizar estas técnicas para generar conocimiento a través del tradicional contraste de hipótesis de las ciencias experimentales, sino generar mayor conocimiento a través de la obtención de evidencias cuantitativas y objetivas que están de alguna manera ocultas en los textos y que guíen al analista en sus investigaciones. En definitiva, este trabajo no deja, además, de ser una propuesta para un planteamiento más científico sobre la obra de *Fortunata y Jacinta* que, sin duda alguna, puede extenderse a toda la obra de Galdós.

Galván González, Victoria (coord.) (2023): *Coda a un centenario. Galdós, miradas y perspectivas.*
Casa-Museo Pérez Galdós, Cabildo de Gran Canaria. Las Palmas de Gran Canaria. ISBN: 978-84-1353-120-5.

BIBLIOGRAFÍA

BAMMAN, D., UNDERWOOD, T., y SMITH, N. A., "A bayesian mixed effects model of literary character", en *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, Vol. 1, pp. 370–379.

BIRJALI, M., KASRI, M., BENI-HSSANE, A., "A comprehensive survey on sentiment analysis: Approaches, challenges and trends", *Knowledge-Based Systems*, 2021, vol. 226, p. 107-134.

BLEI, D. M., NG, A. Y., y JORDAN, M. I., "Latent Dirichlet allocation", *Journal of Machine Learning Research*, 2003, 3 (Jan), pp. 993-1022.

CARDELLINO, C., "Spanish Billion Words Corpus and Embeddings" (March 2016), https://crscardellino.github.io/SBWCE/

CHATMAN, S. "Story and discourse: Narrative structure". In Hale, D. J. (ed.). *The novel: An anthology of criticism and theory 1900-2000*. John Wiley & Sons, 2006, 219.

FLEKOVA, L., y GUREVYCH, I., "Personality profiling of fictional characters using sense-level links between lexical resources", en *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1805–1816.

GORMAN, J. y CURRAN, J. R., "Scaling distributional similarity to large corpora", en *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 2006, pp. 361-368.

GOLUB, G. H., y REINSCH, C., "Singular value decomposition and least squares solutions", en *Handbook Series Linear algebra*, 1971, (pp. 134-151). Springer, Berlin, Heidelberg.

JOCKERS, M. L., *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

QI, P., ZHANG, Y., ZHANG, Y., BOLTON, J. y MANNING, C.D., "Stanza: A python natural language processing toolkit for many human languages", en *Association for Computational Linguistics (ACL) System Demonstrations*, 2020.

LORIA, S., *textblob Documentation. Release 0.16.0*., 2020.

MIKOLOV, T., CHEN, K., CORRADO, G., y DEAN, J. "Efficient estimation of word representations in vector space", *arXiv preprint arXiv*:1301.3781, 2013.

MOHAMMAD, S.M., y HIRST, G., "Distributional measures of semantic distance: A survey". A*rxiv Preprint Arxiv*:1203.1858, 2012.

MORETTI, F.; *Distant reading*, Verso, 2013

PÉREZ GALDÓS, B., *Fortunata y Jacinta: Dos historias de casadas*, 1887, .accesible en https://www.gutenberg.org/ebooks/17013

SALTON, G., y BUCKLEY, C., "Term-Weighting approaches in Automatic Text Retrieval". *Information Processing and Management*, 1988, 24(5), pp. 513–523.

SHERSTINOVA, T., MOSKVINA, A., KIRINA, M., ZAVYALOVA, I., KARYSHEVA, A., KOLPASHCHIKOVA, E., MAKSIMENKO, P. y MOSKALENKO, A. "Topic Modeling of Literary Texts Using LDA: on the Influence of Linguistic Preprocessing on Model Interpretability", en *31st Conference of Open Innovations Association (FRUCT). IEEE*, 2022, pp. 305-312.

VALLS-VARGAS, J., ZHU, J., y ONTANÓN, S., "Narrative hermeneutic circle: Improving character role identification from natural language text via feedback loops", en *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015, pp. 2517-2523.