SELECCIÓN DE CUENTOS DE LOS HERMANOS GRIMM

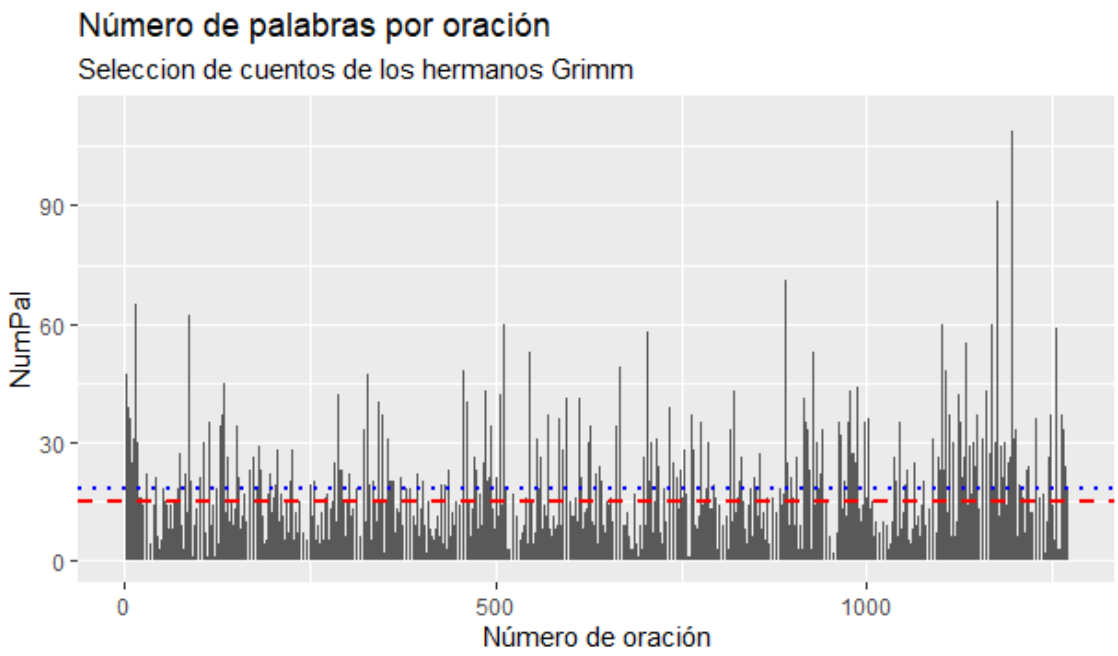
El corpus de textos con el que he decidido trabajar ha consistido en una selección de los cuentos de los hermanos Grimm. En un primer momento pensé en hacerlo con un corpus que siguiera de manera más o menos lineal una historia, algo parecido a episodios nacionales de Galdós, pero me llamaba más la atención la idea de analizar un corpus en el que los textos fueran de un mismo autor, pero independientes unos de otros y ver qué resultados se podían obtener. Con lo cual, me decanté por un pequeño corpus literario más individualizado, y opté por quince de los cuentos de los hermanos Grimm, ya que cada cuento es independiente.

El primer paso fue cosecharlos. En cuanto al formato, mis dos primeras opciones eran o en pdf o txt. Me decanté por el formato de texto plano. Descargué los cuentos de Wikisource uno por uno en formato *.txt*, acto seguido, los limpié y borré el texto relativo a la edición y que no formaba parte de los cuentos.

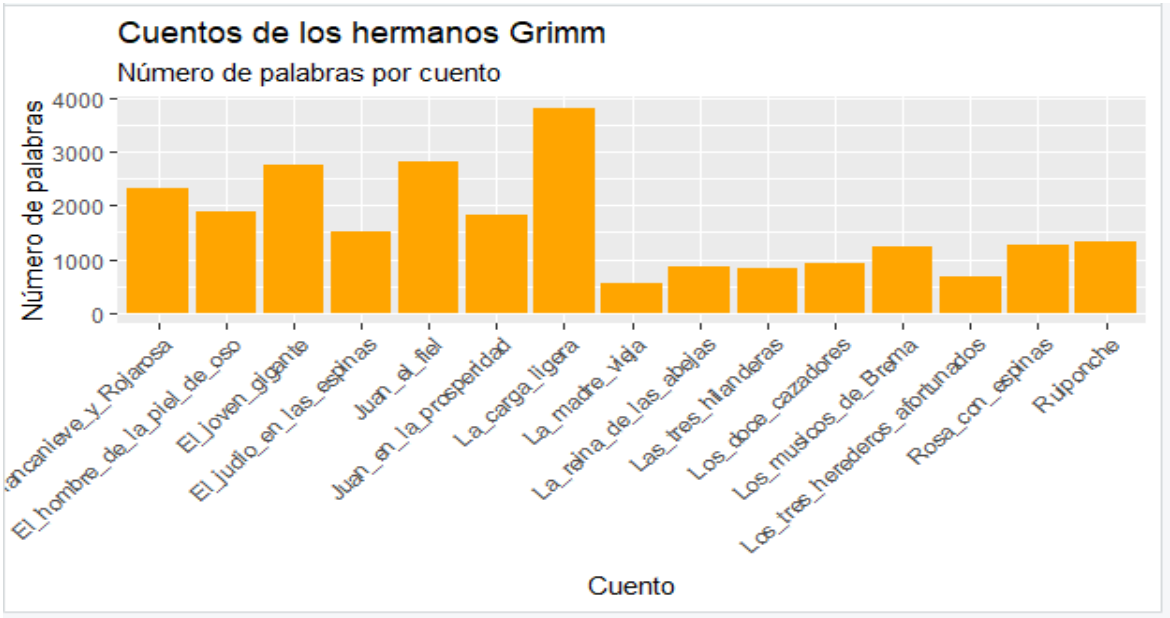
Seguidamente creé un nuevo directorio de trabajo *cuentosgrimm*, con los subdirectorios, datos/cuentos donde guardé los quince archivos de texto de los cuentos. Tenía como preferencia trabajar con los textos de manera local y no que tuviera que acceder a ellos indicando una ruta externa.

Dicho esto, cargué los textos y, en primer lugar, comencé a trabajar las estadísticas básicas. Las dos primeras gráficas muestran las estadísticas principales tras haber dividido los textos en palabras, con el número de palabras por cuento y el número de palabras por oración a lo largo de los 15 cuentos. Dentro de los gráficos básicos, lo que resultaba más interesante era saber cuáles eran las palabras que de verdad tenían un valor semántico relevante, por lo que después de realizar el vaciado de palabras, obtuve la gráfica 1.3, que muestra los términos más frecuentes de los cuentos.

Gráfica 1.1



Gráfica 1.2



Gráfica 1.3

Después de haberme hecho con la información básica de los textos, las dos técnicas que más me llamaban la atención eran el *Topic-Modeling* y sobre todo el análisis de sentimientos.

TOPIC-MODELING

Tras ver los resultados de la gráfica 1.2, decidí realizar el análisis semántico de los 7 primeros cuentos con la técnica de Topic-Modeling, ya que los últimos resultaron ser muy cortos en comparación. Tengo que decir que los resultados que he obtenido me han resultado algo confusos y difíciles de interpretar. En primer lugar, al realizar el código, en vez de dividir los textos por páginas, ya que se tratan de historias cortas, hice la división por cuentos. De esta manera, calculé la media de palabras de los cuentos y era de 2.299 por cuento.

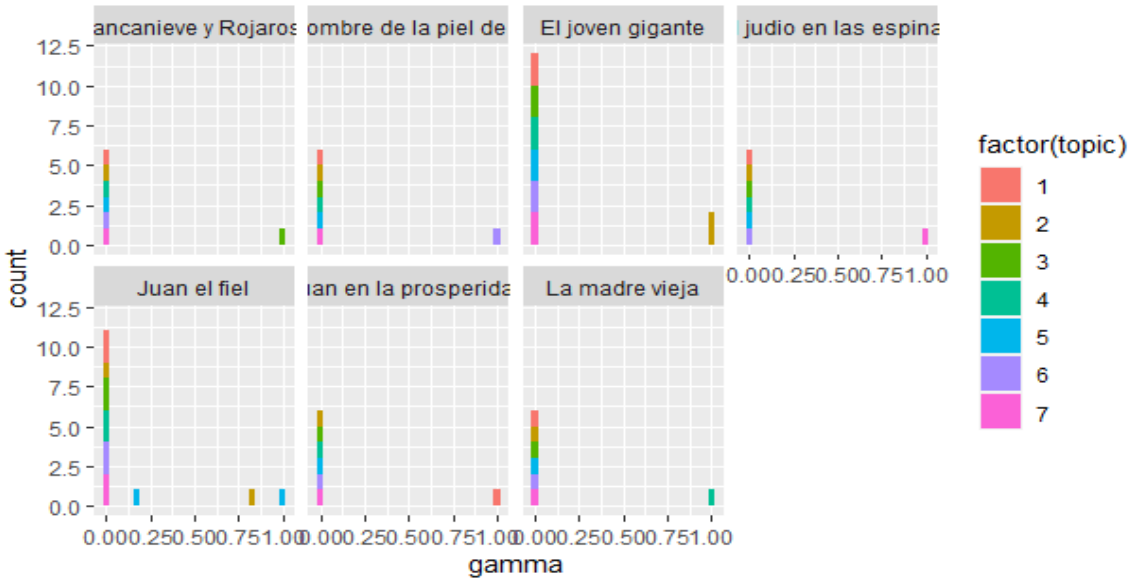
En la primera gráfica podemos observar los términos que se han relacionado con los diferentes tópicos de los cuentos. Algo que me ha llamado mucho la atención y que no entiendo de dónde viene es el resultado de la tabla número 4 dentro de la gráfica 2.1, ya que los textos no contienen nada más que no sea del cuento. Me encargué de borrar todo lo relativo a la edición y al sitio de la descarga antes de empezar a trabajar con los textos,

por lo que las palabras *http*, *wikisource* y *licencia* no tendrían que aparecer por ningún lado.

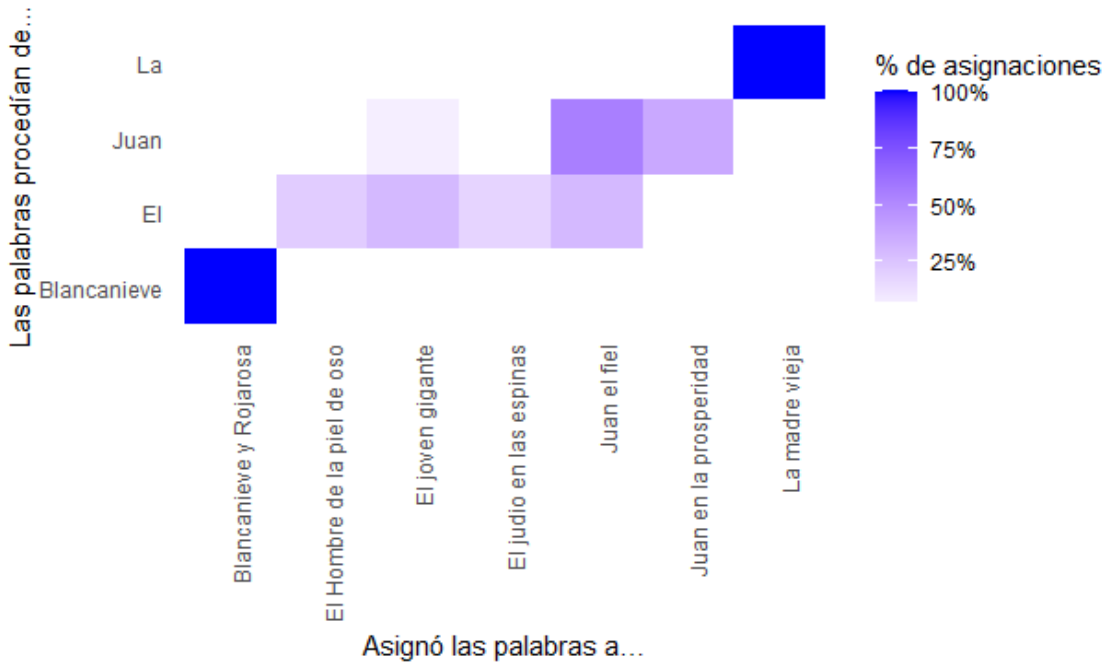
En cuanto a las gráficas restantes, nos encontramos con la 2.2, que asocia las partes de los cuentos con diferentes tópicos, y 2.3, que responde a la asignación correcta de palabras. La última no es exactamente el resultado esperado, ya que en el eje y no salen correctamente los 7 cuentos de los que proceden las palabras.

Gráfica 2.1

Gráfica 2.2



Gráfica 2.3



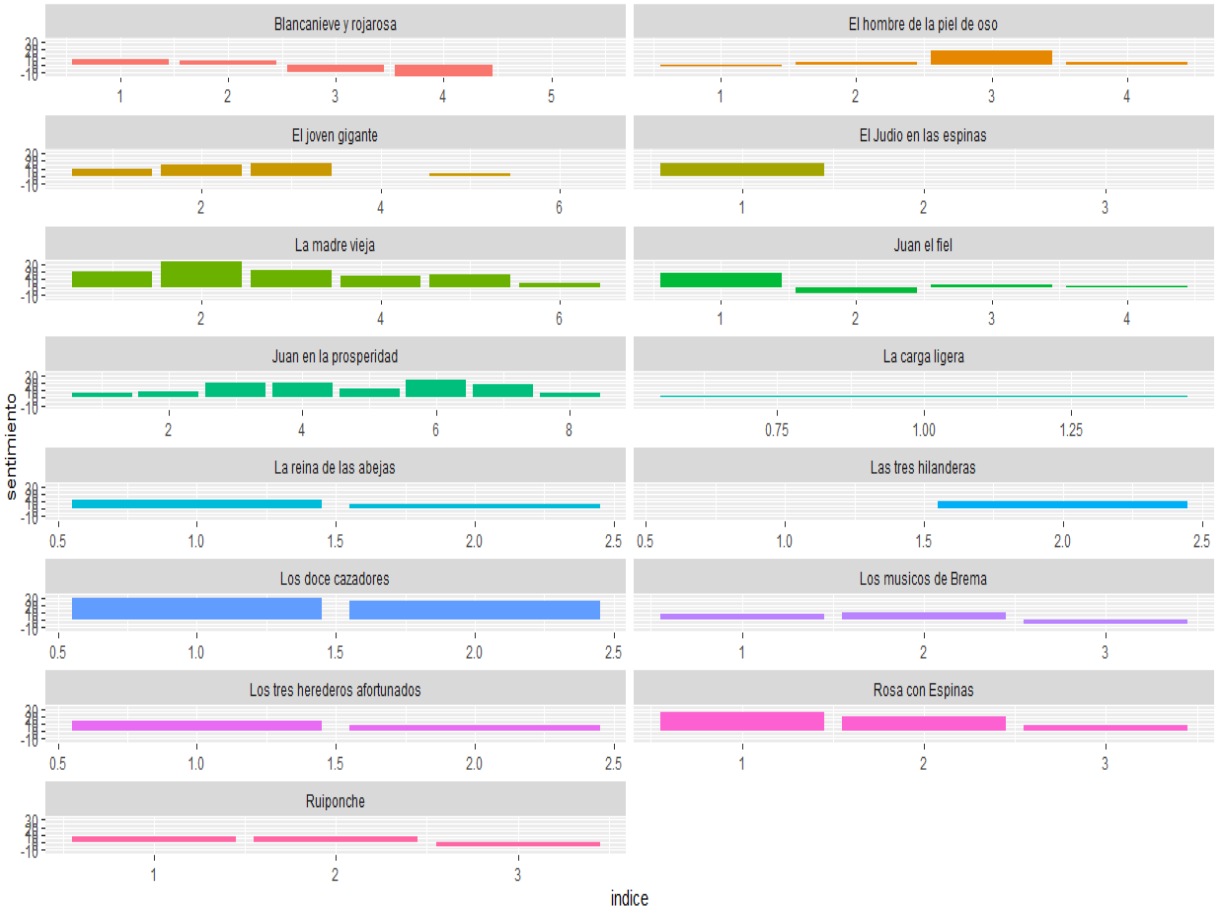
ANÁLISIS DE SENTIMIENTOS

Con la técnica de análisis de sentimientos, podemos ver los resultados de los cuentos uno por uno y una evolución y valencia positiva o negativa en general de los quince juntos.

De los resultados obtenidos que se pueden ver a continuación en las gráficas, la primera (3.1) muestra la valoración del sentimiento durante la trayectoria narrativa de cada uno de los cuentos. Hay más cantidad de valores positivos, no solo por las valencias que se muestran en los gráficos sino también por la extensión de los cuentos, ya que no todos son igual de largos. Los cuentos *La madre vieja* y *Juan en la prosperidad* son los que destacan en cuanto a sentimiento positivo en extensión. Por el contrario, *Blancanieve y rosaroja* sale a la luz como el que encierra más valores de sentimiento negativo. Llama la atención el resultado de *La carga ligera*, ya que se mantiene con la misma valencia durante todo el cuento, aunque también hay que tener en cuenta que es el más corto de todos los cuentos. Otro de los resultados a destacar es el declive que pasa paulatinamente de positivo a negativo en *Blancanieve y rosaroja* y en *Rosa con espinas*.

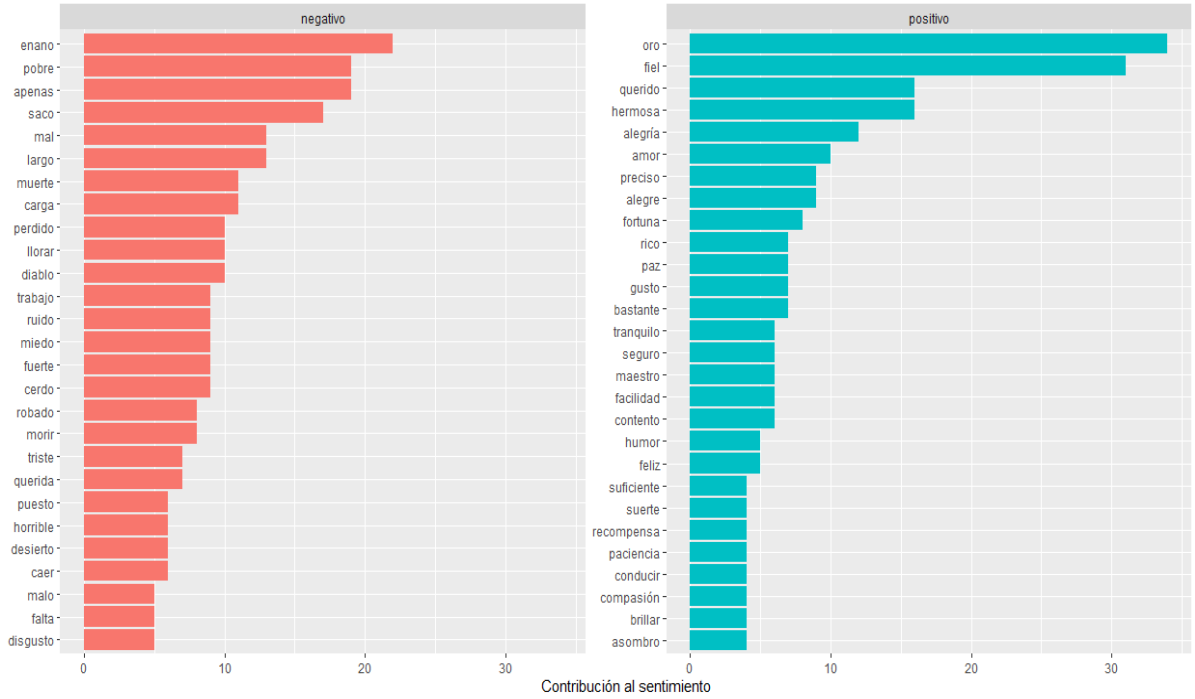
Con relación a cada uno de los 15 cuentos, aquí podemos apreciar las valencias positivas y negativas de cada uno.

Gráfica 3.1

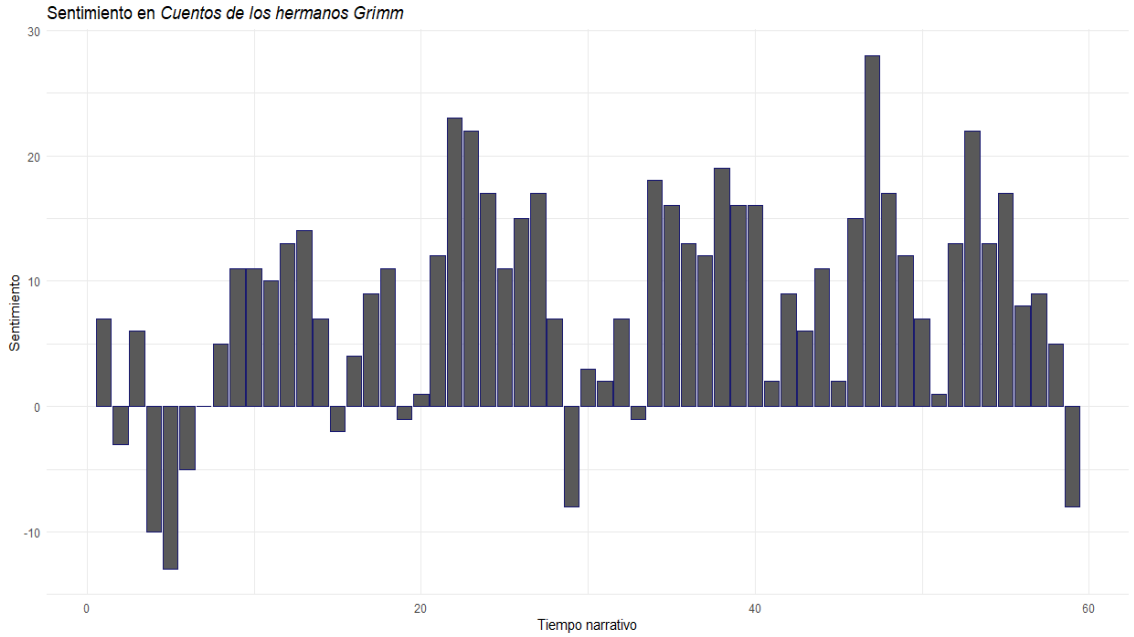


Las siguientes gráficas, a diferencia de la primera, presentan una valoración en conjunto de los 15 cuentos. En la 3.2 aparecen las 25 palabras que se consideran negativas y positivas debido a la frecuencia de aparición. En cuanto a 3.3, 3.4 y 3.5 ambas representan la misma información con gráficos ligeramente diferentes, que es el trascurso y la evolución sentimental a lo largo de todos los cuentos, mostrando qué partes son más positivas y cuáles con un tono más negativo.

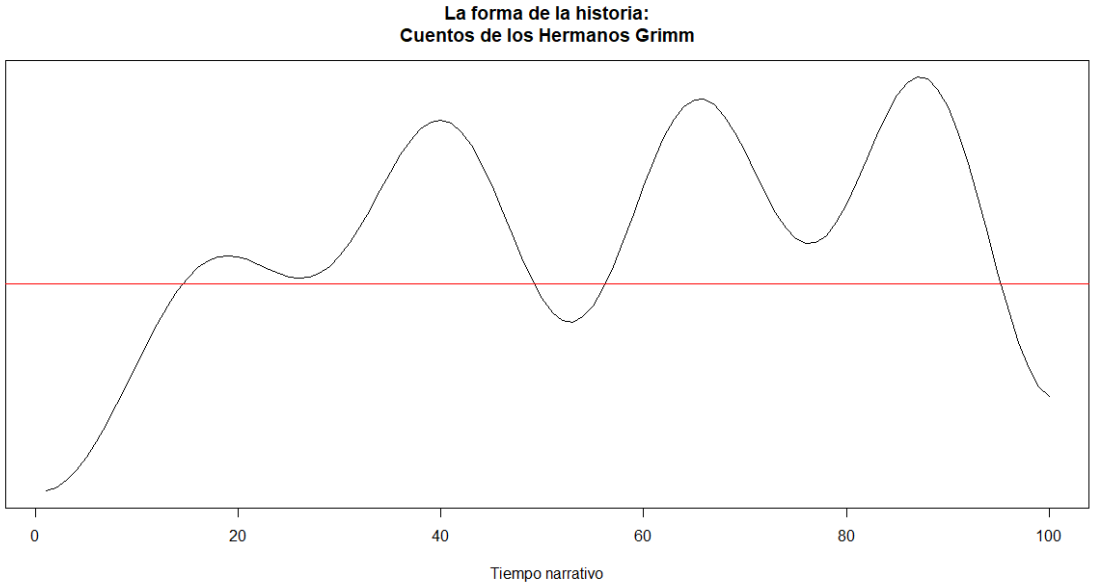
Gráfica 3.2



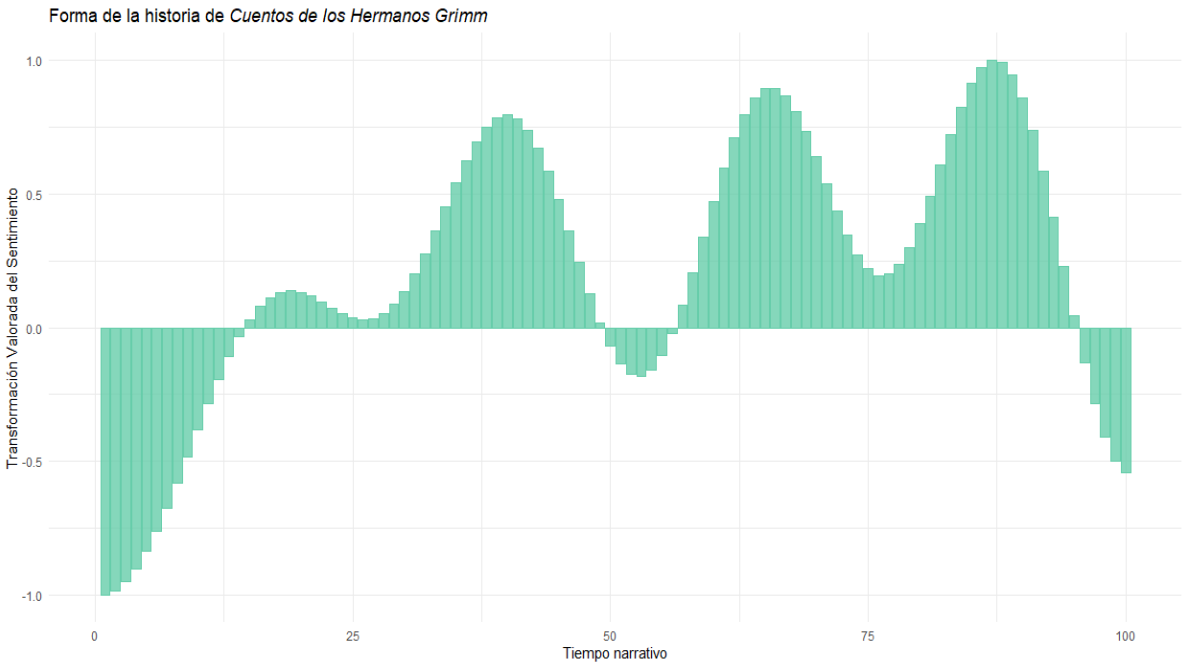
Gráfica 3.3



Gráfica 3.4



Gráfica 3.5



En general se puede apreciar una valoración general más positiva que negativa de los cuentos, tanto de manera general como particular. Todo esto sin olvidarse de que los diccionarios usados operan de una manera específica y que en la realidad pueda diferir en algo de los resultados obtenidos.