Universidad de Valladolid

Facultad de Filosofía y Letras

Departamento de Lengua Española

TRABAJO FIN DE MÁSTER:

Análisis de sentimientos de *Cien años de soledad* y *El amor en los tiempos del cólera* de Gabriel García Márquez

TITULACIÓN A LA QUE SE OPTA:

Máster en Estudios Filológicos Superiores: investigación y aplicaciones profesionales

AUTOR:

Danny Fernando Murillo Lanza

TUTOR:

Dr. José Manuel Fradejas Rueda

Valladolid, 12 de julio de 2017

Índice

Introducción

A las alturas de la segunda década de este nuevo milenio hay una verdad que no es desconocida: Una nueva sociedad y civilización se está construyendo. Esta nueva sociedad se está forjando sobre la base de las nuevas tecnologías de la información y la comunicación. En la actualidad y en un futuro todas estas tecnologías, para ejemplificar: un móvil o un ordenador, estarán presentes en todas las actividades y ámbitos tanto personales, académicos y laborales de la humanidad. Todo ello implica, forzosamente, una transformación de nuestra cosmovisión, de nuestras maneras de pensar, de hacer las cosas, de trabajar y de estudiar el mundo.

Las humanidades, conocidas también como las ciencias humanas se han dedicado al estudio del lenguaje, la literatura, la educación, la historia, la música, la historia del arte, el cine, la lógica, la filosofía y la religión, entre muchos otros campos. Así, pues, se ha tenido la imagen de que los que nos dedicamos al estudio de cualquiera de estas disciplinas, por ejemplo: un filólogo o historiador, nos pasamos la vida encerrados en una biblioteca, leyendo vastas cantidades de libros, documentos o manuscritos; meditando y escribiendo. No obstante, hoy por hoy, parece ser que esa idea se está esfumando (o debería de ser así), puesto que cada vez somos más los humanistas que nos interesamos en incorporar de forma activa a los ordenadores como parte de nuestra actividad creativa, investigativa y docente.

El filólogo, es decir, el que siente pasión e interés por el estudio de las letras, ya sea desde el ámbito de la Lingüística o de la Literatura, también tenemos que fijar la mirada en los ordenadores y aliarnos con ellos para facilitar nuestra labor. Desde nuestra percepción, tarde o temprano, ya sea en la faceta docente o en la investigadora, el filólogo no podrá prescindir de los ordenadores. Si bien es cierto, será capaz de efectuar muchas actividades que ha venido realizando, pero, de una u otra forma los necesitará como medio o como fin. Lo importante es que los emplee, aunque, el sueño, para nosotros, es que los use de forma activa.

A partir de todas estas cuestiones que hemos expuesto hasta aquí y de la imperiosa necesidad de integrar a los ordenadores, en general, a las Humanidades y, en particular, a Lingüística y a la Literatura, se han originado una serie de disciplinas encaminadas al estudio y el trabajo de la fusión de estas áreas: El lenguaje, la literatura y los ordenadores.

Una de esas disciplinas tan interesantes, novedosas y desafiantes que aúna el estudio de las relaciones entre lenguaje y ordenador es el *Procesamiento del Lenguaje Natural (PLN)*. En términos sencillos el PLN se encarga de diseñar programas informáticos que puedan *emular* la

capacidad lingüística humana (Lavid, 2005). En el marco de esta disciplina han surgido una serie de líneas de investigación o aplicaciones que intentan alcanzar tal fin. El *Análisis de Sentimientos (AS)* o Minería de Opiniones (MO) es una de esas aplicaciones o líneas sugestivas que, en términos generales, se encarga de la detección automática de los sentimientos expresados en textos y su clasificación según la polaridad/orientación que tienen (normalmente *positiva, negativa o neutra*) (Balahur Dobrescu, 2011)

¡Eureka! Si las obras literarias (la literatura) son, por una parte, textos y, por otro lado, uno de los grandes propósitos que se plantean sus esculpidores, es decir, los escritores es producir en los lectores un abanico de sentimientos o emociones a través de las letras y de lo que nos cuentan; como también, si es necesario que integremos los ordenadores a nuestras actividades filológicas… Entonces, ¿por qué no atreverse y analizar los sentimientos o emociones de algunas obras literarias mediante una herramienta o un sistema de PLN? ¿Un humano lo podría hacer de manera manual? Por supuesto que sí. Pero ¿lo podríamos hacer de manera rápida, es decir en poco tiempo, con muchas obras literarias? Sería muy difícil y nos podría llevar una vida.

Por todo ello, nos hemos planteado este Trabajo Fin de Máster, cuyo título es: **«Análisis de sentimientos de *Cien años de soledad* y *El amor en los tiempos del cólera* de Gabriel García Márquez»** cuyo principal objetivo es, justamente, analizar, mediante un sistema informático de PLN, los sentimientos o las emociones que figuran en las obras de García Márquez

Para lograr este objetivo nos hemos planteado una serie de objetivos específicos. Estos son: primero, conocer las bases conceptuales y terminológicas; y trazar un panorama histórico del PLN y el AS, a partir de una revisión y selección bibliográfica. Segundo, establecer y conocer la relación, la aplicabilidad y los estudios que se han elaborado en torno al AS en la literatura. Tercero, explicar cuáles son los sentimientos (negativos, positivos o neutros) y las emociones (alegría, confianza, expectativa, sorpresa, miedo, disgusto, ira y tristeza) que figuran en las dos novelas de García Márquez.

Finalmente, determinar si existe correspondencia entre los datos ofrecidos por el sistema de PLN y la narración de los lances más importantes de las novelas en lo que respecta a los sentimientos. O sea, determinar si, por ejemplo: un hecho negativo o positivo que es contado en la novela es detectado en términos del discurso narrativo por el sistema de PLN.

A partir de estos objetivos, hemos organizado y estructurado el presente trabajo en los siguientes apartados:

Capítulo I. El Procesamiento de Lenguaje Natural (PLN): ¿Un ordenador es capaz de emular nuestro lenguaje humano?: En este se construye una base conceptual, terminológica y se traza una evolución del PLN. Además, detallamos cuáles es la arquitectura de un sistema de PLN, su delimitación, las diversas aplicaciones o áreas de esta disciplina.

Capítulo II. Análisis de sentimientos (AS) o Minería de Opinión (MO): ¿El ordenador es capaz detectar los sentimientos o las emociones de un texto?: En este igualmente se construye una base conceptual, terminológica y se esboza la evolución histórica del AS. Asimismo, se presentan cuáles han sido las tareas que se ha planteado realizar el AS, mediante qué enfoque o métodos se ha hecho, cuáles han sido los niveles de análisis o clasificación de los sentimientos y, por último, qué herramientas informáticas se han desarrollado para efectuarlo y cuál ha sido el idioma de base de estas herramientas.

Capítulo III. Análisis de sentimientos aplicado a la literatura: ¿El ordenador es capaz de detectar los sentimientos de la literatura?: En este, básicamente, intentamos establecer la relación y la aplicabilidad que tiene el AS en la literatura. Además, describimos algunos trabajos que se han hecho de AS aplicados a la literatura, los cuales nos impulsaron a realizar este trabajo y se erigen como auténticos modelos para efectuar nuestro AS.

Capítulo IV. Abriendo caminos: Análisis de sentimientos de *Cien años de soledad* (1967) y *El amor en los tiempos del cólera* (1985): Este es el apartado central, aquí nos planteamos desarrollar lo concerniente al autor y las obras literarias que analizamos, el sistema de PLN que hemos empleado para efectuarlo y, finalmente, explicamos el proceso e interpretamos los datos generados del análisis. Esto es, en definitiva, el análisis de las novelas.

Este trabajo, sin lugar a dudas, aporta tanto a esta nueva disciplina, es decir, el Procesamiento del Lenguaje Natural (PLN) y específicamente, al área de Análisis de Sentimientos (AS), puesto que al ser pocos los trabajos que hallamos sobre AS aplicado a la literatura, con este se podría determinar hasta qué punto el AS se puede aplicar a la literatura, como también, se podría dar paso a la creación de sistemas o herramientas informáticas de PLN específicas y funcionales para el Análisis de Sentimientos Literarios. Así mismo, el presente le permitirá al filólogo, especialmente al que se dedica al estudio de la literatura, a valorar la capacidad que tiene un ordenador y un sistema de PLN como herramienta efectiva para la investigación literaria.

Finalmente, queremos acabar este prólogo guiñando a los lectores para que abramos nuestras mentes y empecemos a creer plenamente que los ordenadores pueden ser grandes aliados en aras del desarrollo del conocimiento de las letras y el lenguaje. ¿Habrá errores en este proceso? Por supuesto que sí, lo malo sería que no los hubiera porque la disciplina no avanzaría y no se consolidaría. Por tal razón, debemos entender el presente como un estudio experimental que podría poner los primeros ladrillos de un área que aúna al lenguaje, a la literatura y a los ordenadores.

I. El Procesamiento de Lenguaje Natural (PLN): ¿Un ordenador es capaz de emular nuestro lenguaje humano?

El lenguaje constituye uno de los aspectos y rasgos fundamentales del comportamiento humano y, por consiguiente, de su naturaleza. El lenguaje es, quizá, la herramienta más importante del ser humano, puesto que le permite comunicar a sus semejantes todos sus pensamientos, sus ideas, sus gustos, sus sentimientos y sus emociones. Ningún logro de la humanidad –en la ciencia, la tecnología, la historia y el arte– se hubiera conseguido, si el hombre no hubiera tenido esta capacidad tan potente.

El lenguaje natural, propiamente dicho, en forma escrita sirve para transmitir el conocimiento de una generación a la siguiente durante un largo tiempo. En forma hablada sirve como vehículo de comunicación principal en el comportamiento cotidiano con los demás. (Moreno Boronat, 1999)

¿Qué sería de la humanidad sin el lenguaje? No podríamos realizar ninguna de las actividades personales, económicas, sociales, culturales, académicas, profesionales, ni políticas que llevamos a cabo día con día. He ahí la gran trascendencia de este.

Ahora bien, dado que el área que estamos abordando se centra en procesar el *Lenguaje Natural*, cabe, entonces, definirlo y diferenciarlo frente a otros lenguajes conocidos como los Lenguajes Formales.

1.1.Lenguaje Natural frente a Lenguaje Formal

Aunque pareciera complejo diferenciar a un lenguaje del otro, desde nuestra perspectiva, es más fácil de lo que se cree. A partir del trabajo de Augusto Cortez, Hugo Vega y Jaime Pariona (2009), podemos distinguirlos así:

Por un parte, nos referimos a Lenguaje Natural (LN) a aquel que ha evolucionado con el tiempo para fines de comunicación humana, como el español o alemán (Brookshear, 1993) Estos lenguajes van evolucionando sin tomar en cuenta de forma intencionada la gramática, cualquier regla se establece después de sucedido el hecho. El lenguaje natural es ese lenguaje del que hablábamos al comienzo de este primer capítulo, es decir, el que posibilita a las personas comunicarse entre sí día a día. ¿Por qué naturales? Simplemente porque su evolución se produce de forma natural y porque son propios de la comunicación humana.

Por otro lado, los Lenguajes Formales (LF) son aquellos que el hombre ha desarrollado para expresar las situaciones que se dan en específico en cada área del conocimiento científico. (Cortez Vásquez *et al.*, 2009) A diferencia de los Lenguajes Naturales, en estos, los símbolos y las reglas para unir estos símbolos ya están formalmente preestablecidas y, por tal motivo, se rigen a ellas sin flexibilidad. Mediante estos se puede modelar cualquier teoría de las ciencias exactas con la ventaja de que las ambigüedades no tendrían cabida. Algunos ejemplos de estos son: los lenguajes de programación, los matemáticos y los lógicos.

En conclusión, diremos que un LN es aquel cuyo principal uso y fin es la comunicación entre los humanos y, por lo tanto, las reglas para unir sus símbolos se establecen mediante un proceso de retroalimentación continua. Estos son más flexibles y puede existir en ellos cierta ambigüedad. Ahora bien, un LF es aquel que se emplea para expresar situaciones particulares de algunas áreas del conocimiento científico y, por tal motivo, sus símbolos y las reglas para su unión son preestablecidas, a partir de unos principios teóricos y formales. Por ello, estos son más rígidos y cualquier ambigüedad es eliminada. La distinción de ambos lenguajes nos permite abordar las cuestiones propias del área en la que estamos enmarcando nuestro estudio, es decir, el Procesamiento del Lenguaje Natural.

1.2. ¿Qué es el Procesamiento de Lenguaje Natural (PLN)? ¿El PLN es una disciplina equivalente a la Lingüística Computacional?

Hemos apuntado al inicio de este capítulo que el lenguaje natural es la capacidad más importante del ser humano y, a su vez, es el rasgo más diferenciador frente a otros seres vivos. El lenguaje natural es tan extraordinario, pero a la vez tan complejo que, entender la manera en la que se producen sus procesos de realización es una tarea que, en la actualidad, todavía se está estudiando. El lenguaje natural tiene dos principales mecanismos: el de producción (hablar y escribir) y el de la comprensión (escuchar y leer). Los humanos realizamos estas actividades con mucha facilidad día a día y el grueso de sus usuarios ni siquiera se dan cuenta de lo enigmático y potente que puede serlo.

Ese es el reto que se plantea el Procesamiento del Lenguaje Natural y muchas otras disciplinas científicas, tales como la Lingüística Computacional, la Inteligencia Artificial, la Lingüística Informática, la Ingeniería Lingüística, entre muchas otras.

¿Por qué tantas disciplinas persiguen (casi) un mismo fin? Simplemente porque la relación que conlleva el Lenguaje-Ordenador genera, naturalmente, un vínculo interdisciplinar sobre todo entre la Lingüística y la Informática, desde el cual, según la perspectiva, se estudia el mismo objeto.

Ahora bien, ¿Cuál es el reto que se plantea el PLN? La respuesta es muy sencilla, pero con un fondo, ciertamente, complicado (valga la paradoja). El Procesamiento del Lenguaje Natural tiene el desafío de hacer que un ordenador sea capaz de realizar (o al menos emular, hasta la actualidad) los diferentes procesos y actividades que efectúa el lenguaje natural, es decir, la comprensión (escuchar, leer) y la producción o generación (hablar, escribir). Sí, aunque parezca descabellado, lo que se intenta es que la computadora pueda, por ejemplo, analizarnos sintácticamente una oración, pueda resumirnos el contenido un texto, nos los pueda traducir automáticamente, nos pueda proporcionar una respuesta ante una pregunta y, entre otras cuestiones, que pueda detectar la subjetivad, la actitud o los sentimientos (positivos, negativos o neutros) de un texto. Si bien es cierto, muchas de estas tareas son fáciles de ejecutar para los seres humanos, en cambio, para los ordenadores, dada su naturaleza, se torna mucho más difícil, pero no por eso, imposible.

Para tener una noción conceptual de la disciplina en la que estamos enmarcando nuestro estudio, es decir, el Procesamiento del Lenguaje Natural (PLN) es necesario que la establezcamos a partir de la delimitación con un área muy cercana, nos referimos a la Lingüística Computacional. Por tal motivo, a continuación, expondremos las definiciones que han vertido diversos autores sobre cada una de estas disciplinas (LC y PLN) y, a partir de ellas, determinaremos si existe alguna diferencia o si, por el contrario, son dos áreas de estudio equivalentes.

La *Lingüística Computacional* (LC) se puede definir como: «el estudio de los sistemas de computación utilizados para la comprensión y la generación de las lenguas naturales» (Grishman y Moreno Sandoval, 1991), por su parte, Tordera Yllescas (2012) , señala que la LC «es la disciplina cuyo objetivo persigue la simulación de la competencia comunicativa del hombre a nivel escrito y/ o a nivel oral o, al menos, la simulación de alguna subcompetencia de esta», por otro lado, Lavid (2005) explica que la LC «es un área interdisciplinaria entre la Lingüística y la Informática que se ocupa de la construcción de sistemas informáticos capaces de procesar el lenguaje humano» Finalmente, Gómez Guinovart (1998), apunta que la LC «es un campo científico interdisciplinar relativamente reciente -cerca de cincuenta años de investigación y desarrollo- cuyo objetivo radica en incorporar en los ordenadores la habilidad en el manejo del lenguaje humano.»

Ahora bien, para el área denominada *Procesamiento del Lenguaje Natural* (PLN): «el objetivo de esta investigación es crear modelos computacionales del lenguaje lo suficientemente detallados que permitan escribir programas informáticos que realicen las diferentes tareas donde intervienen el lenguaje natural» (Allen, 1995) , asimismo, el PLN «como disciplina busca desarrollar programas computacionales que sean capaces de ejecutar actividades relacionadas con la comprensión, análisis y producción de textos o discursos escritos en lenguaje natural, de una manera similar a como lo hace el ser humano» (Gelbukh, 2010), por otra parte, el PLN consiste en:

> […] la utilización de un lenguaje natural para comunicarnos con la computadora, debiendo ésta entender las oraciones que le sean proporcionadas, el uso de estos lenguajes naturales facilita el desarrollo de programas que realicen tareas relacionadas con el lenguaje o bien, desarrollar modelos que ayuden a comprender los mecanismos humanos relacionados con el lenguaje (Cortez Vásquez *et al.*).

En otro orden, Teresa Moure y Joaquim Llisterri en su capítulo titulado: *«Lenguaje y Nuevas Tecnologías: El campo de la Lingüística Computacional»,* explican detalladamente que el Procesamiento de Lenguaje Natural es una disciplina que tiene por objetivo:

> Realizar automáticamente transformaciones entre distintas representaciones u objetos lingüísticos: pasar de un texto a una representación con información sobre la categoría gramatical de cada palabra, su estructura de constituyentes o sus significados, traducir de una lengua a otra, resumir el contenido de un texto, extraer la información necesaria para recuperarlo después de haberlo introducido en un sistema de archivos, o escribirlo a partir de los conceptos básicos que forman su estructura. Por ello se habla de *procesamiento o tratamiento del lenguaje natural* (Natural Language Processing, NLP), usándose el término 'natural' para distinguir el lenguaje humano de los lenguajes de programación comunes en informática (Moure y Llisterri, 1996).

Finalmente, otros autores consideran que el Procesamiento del Lenguaje Natural (PLN): «es una parte esencial de la Inteligencia Artificial que investiga y formula mecanismos computacionalmente efectivos que faciliten la interrelación hombre/máquina y permitan una comunicación mucho más fluida y menos rígida que los lenguajes formales» (Moreno Boronat, 1999) y, por lo tanto, «todo sistema de PLN intenta simular un comportamiento lingüístico humano; para ello debe tomar conciencia tanto de las estructuras propias del lenguaje, como del conocimiento general acerca del universo del discurso» (Moreno Boronat, 1999).

Luego de todo este repaso conceptual que hemos hecho sobre las dos disciplinas en materia, o sea, la Lingüística Computacional y el Procesamiento del Lenguaje Natural, podemos decir que ambas persiguen el mismo objetivo, es decir: «diseñar programas informáticos que puedan *emular* la capacidad lingüística humana» (Lavid, 2005).

Por otro lado, también existen algunos autores de las materias que estamos abordando que consideran que hay ciertos rasgos diferenciadores entre la LC y el PLN. Estos son:

> […] mientras la lingüística computacional se centra más en la modelización del conocimiento lingüístico para posibilitar la construcción de sistemas computacionales que analicen y/o generen textos en lenguaje natural, el PLN hace un mayor énfasis en la búsqueda de soluciones a los problemas que plantea la lingüística computacional, pero en el marco de aplicaciones concretas: p.ej. recuperación y extracción de información, resúmenes automáticos, traducción mecánica, etc. (Martí Antonín, María Antonia, 2003).

¿Cómo se traduce esta última percepción? Pues, básicamente, en que la Lingüística Computacional manifiesta una orientación más teórica y el Procesamiento del Lenguaje Natural una faceta más práctica y funcional. Esto es, mientras la primera se encarga de elaborar modelos formales sobre el lenguaje y evaluar las teorías que suministra la Lingüística Teórica para la construcción de los sistemas informáticos que procesan el LN; la segunda, intenta desarrollar aplicaciones concretas en las que el lenguaje humano desempeña un papel central y que tengan trascendencia en la sociedad.

En todo caso, lo más sensato y dado que la distinción solo varía dependiendo del prisma desde el cual se observe, por tal motivo: «actualmente los términos lingüística computacional, PLN y tecnologías lingüísticas suelen ser utilizados de forma indiscriminada tanto por lingüistas como informáticos para hacer referencia básicamente a una misma disciplina» (Periñán Pascual, 2005). Así, concluye Antonio Moreno Sandoval: «Por tanto, Lingüística Computacional y Procesamiento del Lenguaje Natural tratan de lo mismo: del desarrollo de programas de ordenador que simulan la capacidad lingüística humana» (Grishman y Moreno Sandoval, 1991).

En conclusión, hemos de deducir dos cuestiones sustanciales: la primera, es que tanto LC como el PLN deben concebirse como sinónimos o como disciplinas equivalentes. La segunda, es que podemos definir al Procesamiento del Lenguaje Natural (PLN) como: el área de estudio interdisciplinar que trata básicamente a las lenguas naturales (Lingüística) y a los ordenadores (Informática). ¿De qué forma? Diseñando o construyendo modelos, sistemas, mecanismos y sobre todo programas o aplicaciones informáticas que sean capaces de emular o realizar las actividades que efectúa el comportamiento o la capacidad lingüística humana, tales como: la comprensión, el análisis y la generación de textos o discursos orales o escritos en lenguaje natural, entendiendo al LN como aquel que es empleado por los seres humanos para comunicarse entre sí.

1.3. Antecedentes históricos del Procesamiento del Lenguaje Natural (PLN)

El Procesamiento del Lenguaje Natural es una disciplina relativamente reciente en contraste con otras áreas de estudios que tienen centenares de años de haberse consolidado y, por lo tanto, estudiado. A continuación, desarrollaremos una sucinta reseña de los antecedentes históricos del PLN, enfatizando en los proyectos y las aplicaciones más importantes que se crearon, algunas teorías formales del lenguaje que influyeron en su desarrollo y algunos problemas que fue experimentando a través de las décadas. Hemos estructurado este panorama en tres etapas: 1. Desde los orígenes hasta los ochenta. 2. La década de los noventa. 3. Desde los noventa hasta la actualidad. Esta reseña la hemos elaborado a partir de los trabajos de Lidia Moreno, Manuel Palomar, Antonio Molina y Antonio Fernández (1999) y de Isidoro Gil Leiva y José Vicente Rodríguez (1996).

1.3.1. Desde los orígenes hasta los ochenta

Los genes del PLN se remontan aproximadamente a las décadas de los años 40 y 50 del siglo pasado. Con la aparición de las primeras computadoras se emprendieron los primeros proyectos para procesar el lenguaje, es así, como se crea en los años cincuenta, el primer sistema de Traducción Automática de inglés-ruso con un enfoque basado en la equivalencia de palabras a partir de grandes diccionarios. A este, le siguió el GAT (Georgetown Automatic Translator), y el CETA (Centre d'études pour la Traduction Automatique). Ambos proyectos no cumplieron con las expectativas originales, por tal motivo, en el año de 1964 se cancelaron los fondos para los proyectos de Traducción Automática en EE.UU., esto supuso un freno para el desarrollo, pero no un obstáculo para la creación de otros sistemas.

Es así, como en los años sesenta, aparecen nuevos sistemas como el BASEBALL de Green, SIR de Raphael y STUDENT de Bobrow, estos buscaban información a través de patrones o expresiones regulares dejando a un lado el resto de información del texto. El ejemplo más conocido que implementó esta tecnología fue el sistema de dialogo ELIZA de Weizenbaum. Adicionalmente a la creación de todos estos sistemas, los aportes de los trabajos de la gramática transformacional y de los lenguajes formales elaborados por Chomsky proporcionaron la maquinaria para la siguiente generación de investigadores del PLN. En los años setenta se crean las primeras interfaces en LN a Base de Datos como el sistema LUNAR de Woods. Asimismo, aparecen un abanico de analizadores sintácticos que emplean gramáticas incontextuales como SAD-SAM de Lindsay. En otro orden, Halliday propuso un formalismo ("systemic grammar") que codificaba las

relaciones funcionales en una oración. ¿En qué sistema se pudo ver la aplicación de esta teoría? En el conocido mundo de los bloques ("blocks-world") de Winograd. En estos mismos años, Woods mejoró la potencia de las expresiones regulares y de las gramáticas incontextuales agregando a un autómata de estados finitos variables y restricciones funcionales.

¿Qué ocurrió en los años ochenta? Básicamente, en reacción a la naturaleza de las Redes de Transición Aumentadas desarrolladas por Woods en la década de los setenta, surgieron una serie de formalismos que se basaban en estructuras teóricas más formales. Para el caso, en el año de 1983, Noam Chomsky propuso su Teoría de Rección y Ligadura en la que se da mayor importancia al léxico. Igualmente, en esta línea aparecieron una variedad de gramáticas como las Gramáticas de Estructura Sintagmática, las Gramáticas Léxico-Funcionales de Bresnan y muchas otras más. Ahora bien, en lo concerniente a las aplicaciones se construyeron, lógicamente, sistemas más sofisticados y cada vez menos robustos, tales como Ariane-78, EUROTRA o ATLAS (en el área de la Traducción Automática), y TEAM, CHAT-80 y ORBU en el campo de las interfaces con Bases de Datos.

1.3.2. La década de los noventa

En los años noventa se recuperaron los formalismos que fueron introducidos en la década anterior y se desarrollaron extensiones de estos. ¿En qué consistieron esas extensiones? Pues, en representaciones de las dependencias a larga distancia y las estrategias requeridas para el análisis y eliminación de la ambigüedad del texto. No obstante, estas no fueron resueltas completamente por la variedad que, hasta la actualidad, representa el lenguaje natural.

Por tales cuestiones, los métodos basados en la definición de reglas y la codificación manual del conocimiento fueron perdiendo fuerza y dieron paso a la creación de métodos estadísticos y de aprendizaje automático, cuyas bases fundamentales fueron grandes corpus de información.

1.3.3. Desde los noventa hasta la actualidad

La disciplina se fue consolidando y cada vez más se han ido sumando interesados en estudiarla. Actualmente, se está buscando soluciones a problemas parciales como: revisión lingüística de textos, recuperación de información, extracción de información, resúmenes y clasificación; reconocimiento y síntesis de voz, traducción automática y generación automática de textos.

1.4. Arquitectura de un sistema de Procesamiento del Lenguaje Natural (PLN)

Iniciaremos este apartado recordando que el objetivo de cualquier sistema de Procesamiento del Lenguaje Natural es realizar o emular las tareas de comprensión y producción del lenguaje de forma semejante a un humano. Ahora bien, como es lógico, alcanzar este cometido no es nada fácil, todo lo contrario, es una tarea difícil y compleja.

Hasta el día de hoy, ningún ente natural o artificial ha superado las capacidades de los seres humanos. Si bien es cierto, las tecnologías, en la actualidad, son tan inteligentes y pueden realizar un gran número de actividades, todas estas han sido creadas gracias a las capacidades, el estudio y el trabajo de la humanidad. El lenguaje, por ejemplo, es una capacidad humana que realiza labores increíbles e importantes: somos capaces de generar un sinfín de expresiones cargadas de significados teniendo en cuenta el contexto situacional de donde las producimos. Igualmente, somos capaces de comprenderlas, mediante un proceso de abstracción de los significados propiamente lingüísticos y sociales. Todos estos mecanismos de expresión y producción nos mantienen, naturalmente, en un constante proceso de actualización de nuestros conocimientos lingüísticos y contextuales. Por eso es que estudiar y comprender al lenguaje y a su funcionamiento no es tan fácil.

Ahora bien, ¿cómo es que somos capaces de realizar todas las tareas descritas anteriormente? Pues, gracias a una vasta cantidad de conocimiento, parte del cual es innato, y parte del cual hemos ido adquiriendo en el transcurso de la vida, estos son: conocimiento del lenguaje, conocimiento sociocultural y conocimiento del mundo. En el proceso de producción y comprensión del lenguaje, también hacemos uso de otras facultades cognitivas como el raciocinio.

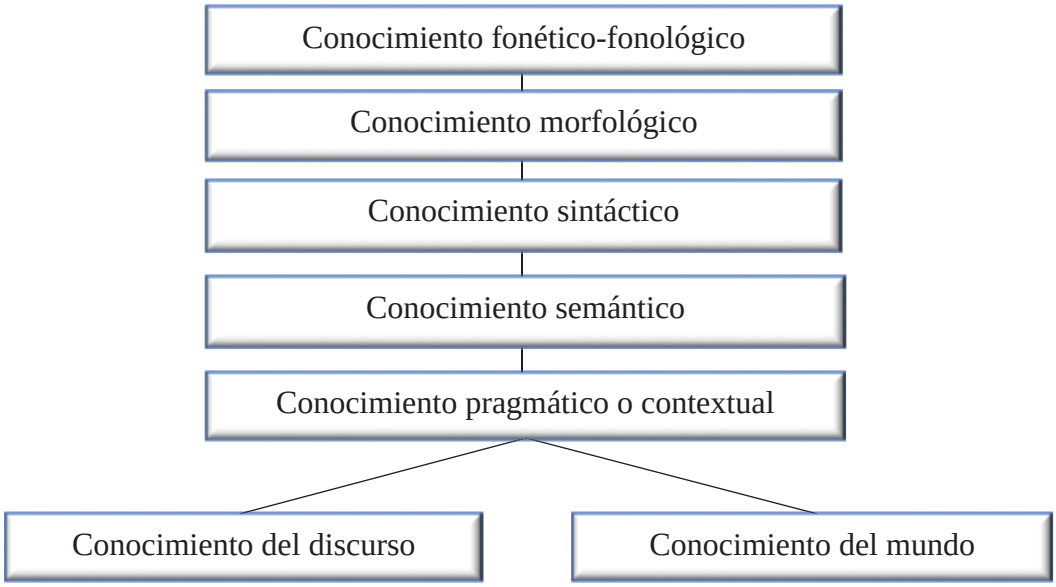
Por todas las cuestiones que hemos expuesto hasta aquí, compartimos la percepción de Julia Lavid (2005), quien considera que: «emular la capacidad lingüística es una tarea de gran envergadura: tenemos que manejar el razonamiento, comprender cómo funciona el lenguaje y proporcionar al sistema informático conocimiento enciclopédico». Por esto último, dada la complejidad del fenómeno lingüístico, es necesario dividirlo en diferentes partes más simples que posibiliten su estudio de forma individual. La Lingüística Descriptiva coadyuvó en este cometido, puesto que ya había divido la lengua en diversos niveles de análisis: fonético, morfológico, sintáctico, semántico y pragmático.

En PLN se le conoce a este proceso como *modularidad «*el cual consiste en dividir el sistema en componentes relativamente independientes» (Moreno Sandoval, 1998), también,

aparece la *estratificación* que son «los diferentes tipos de organización de los niveles lingüísticos» (Lavid, 2005). Tanto uno como otro son de gran importancia porque hacen que los sistemas sean más flexibles, ampliables y modificables, puesto que aseguran su integridad, aunque se produzcan cambios o adiciones en algunos de sus componentes. Por otra parte, esto permite un estudio más específico de las áreas y, por consiguiente, les posibilita a los investigadores enfocarse en los problemas que se originan en cada uno de los componentes.

Por todo esto, para que un sistema computacional de PLN cumpla con su objetivo principal, según diversos autores tales como Moreno Boronat (1999), Lavid (2005), Moreno Sandoval (1998), Gil Leiva y Rodríguez Muñoz (1996), Fernández Gavilanes (2012) y otros más, tiene que disponer de los diferentes niveles, formas o conocimientos de la lengua que se almacenan en diversos módulos como ilustra la figura 1.

FIGURA 1: Tipos de conocimientos para un sistema de PLN



Como podemos observar, cada uno de los módulos guarda diferentes tipos de conocimiento:

- **Conocimiento fonético-fonológico:** se ocupa de las realizaciones acústicas, así como de su transcripción. Es necesario en los sistemas de reconocimiento y síntesis de habla.
- **Conocimiento morfológico:** estudia cómo las palabras son construidas a partir de unidades más pequeñas, que son los monemas (morfemas y lexemas).
- **Conocimiento sintáctico:** estudia las combinaciones de las palabras para formar sentencias correctas y las relaciones de las mismas unas con otras. Este es un componente básico, puesto que se encarga de reconocer las estructuras de las oraciones.

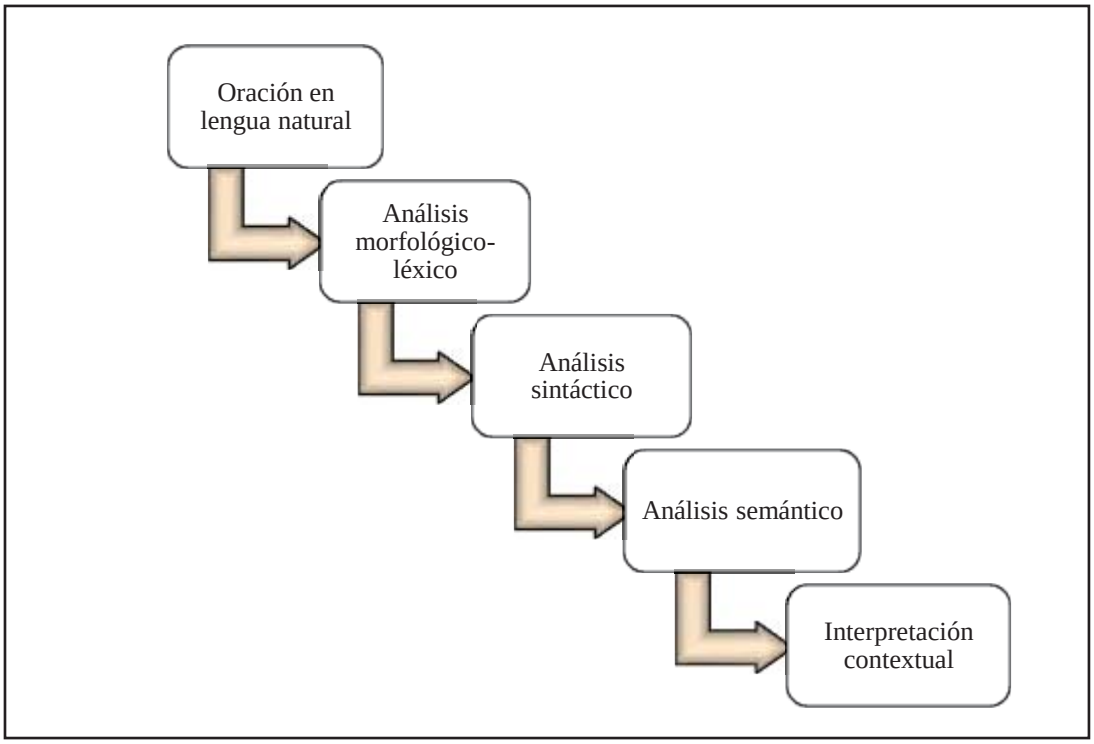
- **Conocimiento semántico:** se ocupa del significado de las palabras y de cómo estos significados se combinan en oraciones para conformar significados oracionales. Otro de los componentes básicos, dado que asigna el significado a las estructuras.
- **Conocimiento pragmático o contextual:** se ocupa de cómo las oraciones se usan en diferentes contextos situacionales y de cómo estos contextos influyen o afectan en el procesamiento e interpretación de las emisiones lingüísticas. Algunos de los autores que hemos señalado, como Moreno Sandoval (1998), y Cortez Vásquez, Vega Huerta y Pariona Quispe (2009) consideran que dentro de este conocimiento hay dos subniveles más:
    a) **Conocimiento del discurso:** aquí se abordan los aspectos de interpretación afectados por las oraciones emitidas anteriormente. En síntesis, este conocimiento se emplea para interpretar los pronombres anafóricos y los aspectos temporales.
    b) **Conocimiento del mundo:** incluye todos los conocimientos conceptuales del mundo que tienen en cuenta los hablantes cuando se comunican en una lengua determinada. Este es fundamental para entender mucha información sobrentendida, pero no proferida explícitamente en las oraciones.

A partir de la clasificación modular que hemos descrito anteriormente y siguiendo a Moreno Boronat (1999) y a Vilares Ferro (2005), un sistema computacional de PLN divide las fases o niveles de análisis de una oración en:

- **Análisis morfológico-léxico:** convierte la cadena de caracteres de entrada en una secuencia de unidades léxicas, o sea, determina las palabras de un texto y su etiqueta morfosintáctica mediante el uso de diccionarios y reglas morfológicas.
- **Análisis sintáctico:** analiza la secuencia de unidades léxicas y produce representaciones de su estructura (árbol, red, etc.), es decir, realiza el agrupamiento de las palabras en sintagmas y frases.
- **Análisis semántico:** a partir de la estructura anterior (sintáctica) genera otra que representa el significado o sentido de la oración, dicho de otra forma, determina el significado de las frases de acuerdo con el significado de los sintagmas, palabras y morfemas que lo conforman.
- **Análisis contextual o de función pragmático:** utilizada la estructura semántica de la frase anterior para desarrollar la interpretación final de la oración, en función de las circunstancias del contexto, es decir, establece la identidad de las personas y objetos determina la estructura del discurso y gestiona el diálogo en un entorno conversacional.

Para apreciar de mejor manera todo este proceso y siguiendo a Moreno Sandoval (Moreno 1998), mostramos las cuatro fases de análisis en el esquema de la Figura 2.

FIGURA 2: Fases o etapas en el procesamiento computacional del lenguaje natural



Ahora bien, Antonín Martí y Castellón (2000) señalan que cualquier sistema computacional que trate el lenguaje natural ha de constar de tres componentes básicos:

1. ***Textos* en lengua natural**: que constituirán como el material o la sustancia que queremos procesar o generar.
2. ***Datos lingüísticos:*** usualmente gramáticas y lexicones, que nos indican cómo procesar o generar los textos:
    a) El diccionario o lexicón: es el encargado de recoger las unidades léxicas que pertenecen a la lengua en cuestión junto con la información necesaria para su procesamiento morfológico, sintáctico y semántico.
    b) La gramática: es la que recoge las reglas necesarias para, por un lado, determinar la gramaticalidad de los textos, y, por otro lado, para efectuar el procesamiento de los mismos, es decir, mostrar su estructura morfológica, sintáctica, semántica, etc.
3. ***Programas informáticos*:** que son los responsables de llevar a cabo el procesamiento o la generación de los textos de acuerdo con la información que le proporcionan los datos lingüísticos

Hasta aquí, hemos visto qué conocimientos y componentes tiene que tener un sistema informático de PLN, como también el proceso que debe seguir para realizarlo. ¿En dónde hallamos estos sistemas? Básicamente en una gama de aplicaciones que se encargan de resolver problemas o desarrollar o emular las tareas de producción y comprensión del lenguaje natural.

1.5. Aplicaciones o áreas del Procesamiento de Lenguaje Natural (PLN)

Para acabar de perfilar esta disciplina, es necesario hacer un repaso suscito por las principales aplicaciones o áreas de estudio y trabajo que se han venido desarrollando desde sus orígenes y otras que han surgido en los últimos años. Antes de ello, tenemos que acotar dos consideraciones importantes: la primera, dado que tanto el PLN como la LC son disciplinas equivalentes, entonces, aunque algunos autores las propongan como aplicaciones para el PLN y otros para la LC, nosotros las tomaremos en cuenta de forma indiscriminada como si se trataran de la misma área. La segunda, puesto que no existe una clasificación única de todas las aplicaciones (aunque sí hay un consenso en casi todas), por tal motivo, presentamos algunas de estas y al final explicaremos cuál es el objeto de las aplicaciones más importantes o de las que interesan para el desarrollo de este estudio.

La primera clasificación la propone Ralph Grishman (1986), quien establece tres aplicaciones principales:

1. Traducción automática.
2. Recuperación de información.
3. Interfaces hombre-máquina.

La segunda tipología nos la ofrece el propio Antonio Moreno Sandoval (1998). Esta clasificación es mucho más completa y específica, puesto que las agrupa en varios bloques:

1. Sistemas que tratan de emular la capacidad humana de procesar lenguas naturales:
    a) Traducción automática.
    b) Recuperación y extracción de información.
    c) Interfaces hombre-máquina.
2. Sistemas que ayudan en las tareas lingüísticas. En este grupo se agrupan las herramientas que pueden ser utilizadas por los lingüistas para facilitarles ciertas tareas complejas:
    a) Herramientas de análisis textual.

b) Herramientas para manejo de corpus.
c) Bases de datos lexicográficas.

3. Sistemas de ayuda a la escritura y composición textual. Este grupo lo integran una serie de aplicaciones ampliamente desarrolladas y cualquier usuario habitual de procesadores de textos está familiarizado con ellas:
    a) Correctores ortográficos.
    b) Correctores sintácticos y de estilo.
4. Enseñanza asistida por ordenadores.

La tercera clasificación nos la brinda Javier Gómez Guinovart (1998)*:*

1. Tecnologías del habla:
    a) Reconocimiento del habla.
    b) Síntesis del habla.
2. Traducción automática:
    a) Traducción totalmente automática.
    b) Traducción asistida por ordenador.
3. Extracción de información.

La cuarta tipología ha sido elaborada por Julia Lavid (2005). Esta ha sido agrupada en diferentes bloques, según sus funcionalidades:

1. Sistemas que permiten al usuario comunicarse con el ordenador:
    a) Recuperación y extracción de la información.
    b) Interfaces hombre-máquina.
2. Sistemas que permiten a los humanos comunicarse entre sí en diferentes lenguas:
    a) Traducción Automática.
    b) Enseñanza de Lenguas Asistida por Ordenadores (ELAO)
3. Sistemas de ayuda en las tareas lingüísticas:
    a) Las herramientas de análisis textual y de corpus.
    b) Las herramientas de ayuda a la escritura.
    c) Las bases de datos lexicográficas.
    d) Las bases de datos terminológicas.

La quinta clasificación nos la dan Cortez Vásquez, Vega Huerta y Pariona Quispe (2009):

1. Traducción automática.
2. Recuperación de información.

3. Extracción de información y resúmenes.
4. Resolución cooperativa de problemas.
5. Tutores inteligentes.
6. Reconocimiento de voz.

En la tesis de Milka Villayandre Llamanzanares (2010) hallamos la sexta tipología:

1. Aplicaciones que tratan de reproducir la capacidad humana de procesar el lenguaje:
    a) Traducción automática.
    b) Interfaces en lenguaje natural.
    c) Recuperación y extracción de la información.
2. Aplicaciones basadas en el tratamiento de información textual:
    a) Herramientas de ayuda a la escritura.
    b) Creación automática de resúmenes.
    c) Extracción de terminología.
    d) Indexación automática.
    e) "*Data mining*" textual o descubrimiento datos en textos.
3. Tecnologías del habla:
    a) Síntesis del habla.
    b) Reconocimiento del habla.
    c) Sistemas de diálogos.

La última de estas clasificaciones que queremos destacar es la que ofrece Hernández y Gómez (2013). La resaltamos, sobre todo, porque incluye al área en específico en el que se enmarca nuestro estudio. Esta es:

1. Recuperación y extracción de información.
2. Minería de datos.
3. Traducción automática
4. Sistemas de búsqueda de respuesta.
5. Generación de resúmenes automáticos.
6. Análisis de sentimientos.

Para cerrar este primer capítulo de nuestro trabajo y luego de haber hecho un repaso por las diferentes clasificaciones de las aplicaciones que buscan cumplir los objetivos del Procesamiento de Lenguaje Natural, hemos de conocer en qué consisten algunas de las más importantes o de las que más se han estudiado desde sus orígenes hasta la actualidad.

1. **Traducción automática:** El sueño de construir máquinas traductoras ha sido desde sus orígenes uno de los grandes retos de la Informática, y la Traducción Automática (TA) pasó a posicionarse como una de las aplicaciones estrellas desde el momento en que una computadora estuvo disponible en los años cuarenta. Esta asimismo es, hoy por hoy, una de las tareas más complejas. ¿En qué consiste la TA? Seguramente, tanto usted como yo hemos usado algún sistema de TA y sabemos en qué consiste, pero vale la pena puntualizarlo: «Se trata de tomar oraciones (o textos completos) en una lengua natural, que se denomina como lengua fuente, y producir automáticamente una traducción a otra lengua, llamada la lengua meta» (Moreno Sandoval, 1998).
2. **Recuperación y extracción de información:** Estas dos aplicaciones están estrechamente ligadas, pero no son idénticas, el principal objetivo que intentan alcanzar es tratar a la información almacenada en grandes bases de datos textuales. *La recuperación de información* se ocupa de tomar las consultas de un usuario a una base o banco de datos textuales y «el sistema se encarga de proporcionar los materiales que se ajustan a los criterios de búsqueda, no solo basándose en la detección de palabras clave sino también llevando a cabo una labor de comprensión lingüística de la consulta» (Llamazares, 2010). *La extracción de información* «pretende 'leer' grandes cantidades de texto reconocer la información importante contenida en ellos y trasladarla a un formato predefinido para que pueda ser tratada y recuperada con mayor facilidad» (Moreno Sandoval, 1998).
3. **Reconocimiento y síntesis del habla:** Estas aplicaciones se centran en el tratamiento de la lengua oral. *El reconocimiento del habla* «de forma inversa a la síntesis, transforma un enunciado oral en su contrapartida escrita». Por su parte, *la síntesis del habla* «o generación de habla artificial, sobre todo, conversión de textos escritos en su equivalente oral» (Llamazares, 2010).
4. **Interfaces hombre-máquina: «**son sistemas que permiten la interacción del usuario con el ordenador en su lengua natural en vez de utilizar lenguajes informáticos o menús complejos» (Lavid, 2005).
5. **Enseñanza de Lenguas Asistida por Ordenador (ELAO):** En los últimos tiempos la enseñanza de lenguas ha sido una actividad de gran predicamento por todo el mundo. Es así que la ELAO se ha encargado de ofrecer a los usuarios «una serie de programas educativos que permiten el aprendizaje de las lenguas combinando

recursos (voz, texto, imágenes) y llevando a cabo, en algunos casos, auténticos análisis sintácticos» (Lavid, 2005).

Desde nuestra perspectiva todas estas aplicaciones que hemos señalado y que hemos descrito someramente son muy interesantes, apasionantes, pero también, muy desafiantes. Nos hubiese gustado explicar con demora en qué consiste cada una, pero no es objetivo del presente trabajo llevar a cabo tal labor.

Hasta aquí hemos realizado un largo, pero sustancial recorrido por todos los aspectos más importantes del Procesamiento de Lenguaje Natural (o, al menos los ineludibles para entender este trabajo). Ahora, estamos listos para que nuestro cohete descienda a otra de las capas (campo, área o aplicación) en la que hemos enmarcado el presente, nos referimos al Análisis de Sentimientos (AS). Por ello, no hemos definido en este capítulo en qué consiste esta área, porque le dedicaremos el capítulo siguiente.

II. Análisis de Sentimientos (AS) o Minería de Opinión (MO): ¿El ordenador es capaz detectar los sentimientos, las emociones o las opiniones de un texto?

Desde tiempos ancestrales, los seres humanos siempre hemos querido saber qué piensan o cuáles son los sentimientos, opiniones o percepciones que tienen los demás sobre nosotros mismos, sobre alguna situación, tópico, objeto, etc. ¿Para qué? Pues, simplemente para saberlo o bien para tomar decisiones. A esta propiedad de la humanidad se le conoce como ***subjetividad.*** El lenguaje, por ejemplo, ha sido una de las principales herramientas para expresar esa subjetividad en forma de lo que se denomina *opinión*.

En la actualidad, gracias al acelerado desarrollo, expansión y uso del internet y de sitios Web 2.0, tales como: blogs, foros, redes sociales, servicios de mensajerías, wikis, prensa en línea con participación de lectores, plataformas de comparativa y evaluación de productos y servicios, se genera una cantidad impresionante de datos que expresan esa subjetivad o, dicho concretamente, que expresan las opiniones de millones de usuarios sobre un determinado tema, persona, actuación, objeto, producto, servicio, etc. Estas fuentes de información son tan peculiares porque cualquiera puede verter su percepción, opinión o sentimiento sin ser un profesional en la materia u objeto, como también, porque todos podemos acceder a ellas y valorarlas para nuestros propósitos.

Todo este cúmulo de información es de gran valor para diversos ámbitos de la sociedad, sobre todo para los campos industriales, comerciales, académicos y, hasta, políticos. Sin embargo, esta información tiene una serie de características que hacen que sea complejo su procesamiento de forma manual, por ejemplo: la cantidad que se genera es muchísima y la codificación de esta está en texto, o sea, datos sin estructura aparente para la máquina. Por todo esto, según Vilares, Alonso y Gómez Rodríguez (2013): «desde hace tiempo se ha venido demostrando la necesidad de crear herramientas informáticas de PLN capaces de analizar estas grandes cantidades de texto y extraer y sintetizar información de forma automatizada, surgiendo así disciplinas como la *minería de textos*», dentro de este marco, procesar de forma automática todo tipo de información subjetiva (opiniones, sentimientos) han capturado el interés y ha venido recibiendo mayor atención. De este proceso se ha ocupado una nueva área de estudio conocida como ***Análisis de Sentimiento*** **(AS)** o ***Minería de Opinión*** **(MO)**.

Es en el marco de esta disciplina en el que centramos nuestro estudio, por tal motivo, a continuación, pretendemos desarrollar algunas cuestiones importantes para entenderla, tales como: su definición, las diferentes denominaciones que ha recibido, algunas razones de su surgimiento, su evolución histórica, sus principales tareas, métodos, niveles de análisis, y las herramientas que se han usado para realizar AS tanto para el inglés como para el español.

2.1. ¿Qué es el Análisis de Sentimiento (AS) o Minería de Opinión (MO)?

Antes de definir, como tal, el área o la disciplina en la que estamos enmarcando nuestro trabajo y también antes de presentar un breve panorama histórico de la misma, hay una cuestión que conviene abordar previamente y que tiene que ver con las denominaciones que ha recibido este campo.

Esta disciplina, tal y como lo veremos más adelante, ha adquirido un mayor estudio y popularidad sobre todo en este siglo XXI y eso ha implicado que todavía no se termine de consolidar, ni homogenizar en algunos de sus aspectos más importantes, tales como el nombre o la denominación que la definen. Por tal razón, es común en la literatura referirse a este campo como *Análisis de Sentimiento (Sentiment Analysis), Minería de Opinión (Opinion Mining), Análisis de Subjetividad (Subjectivity Analysis), Extracción de Opinión (Opinion Extraction), Análisis de Afectos (Affect Analysis), Análisis de Emoción (Emotion Analysis),* etc. Sin embargo, de toda esta lista terminológica, los dos primeros, es decir, ***Minería de Opinión*** *(Opinion Mining)* y ***Análisis de Sentimiento*** *(Sentiment Analysis),* han sido los nombres en torno a los cuales se ha ido reorganizando la disciplina. Estos, a su vez, se han usado indistintamente.

La primera vez que se empleó el término ***Opinion Mining*** fue en el trabajo de Dave, Lawrence y Pennock (2003). Según sus autores un sistema ideal de MO debería actuar como un buscador de características de productos, y mostrar junto a las características de cada producto la orientación general de la opinión existente sobre ellas. Por su parte, la primera vez que se usó el segundo término, es decir, ***Sentiment Analysis*** fue en el trabajo de Nasukawa y Yi (2003), quienes describen a un sistema de AS como un sistema de clasificación de la opinión a nivel de entidad, es decir, el sistema que presentan se atreve a obtener la orientación de cada opinión vertida sobre cada uno de los sujetos que aparecen en los documentos.

Aunque ambos términos se usan en la actualidad de forma equivalente, en un sentido estricto, realmente, no lo son. Es más, existen algunos trabajos en los que se empieza a marcar una diferenciación conceptual (Cambria *et al.*, 2013). No obstante, la utilización indiscriminada de cualquiera de ambos términos para referirse a la detección o identificación de opiniones y la orientación positiva o negativa de estas en un documento es ampliamente aceptado (Liu, 2010).

Teniendo en cuenta las diferentes denominaciones que ha recibido esta área de estudio de PLN, pasamos a abordar uno de los elementos centrales de todo el trabajo, nos referimos a explicar de qué se encarga o en qué consiste el Análisis de Sentimientos (AS) o Minería de Opinión (MO). Haremos un repaso por las definiciones más interesantes que han dado algunos de los autores más representativos del campo.

Una de las definiciones más generales y difundidas en esta área ha sido la vertida por Pang y Lee (2008), quienes explican que la MO: «se centra en tratar automáticamente información con opinión, lo que permite, entre otras cosas, extraer la polaridad (positiva, negativa, neutra o mixta) de un texto». Por otro lado, «la minería de opiniones permite valorar cuantitativamente expresiones subjetivas como sentimientos o sensaciones» (Chen y Zimbra, 2010). En otro orden, «la MO busca analizar las opiniones, sentimientos, valoraciones, actitudes y emociones de las personas hacia entidades como productos, servicios, organizaciones, individuos, problemas, sucesos, temas y sus atributos» (Liu, 2012).

Existen otros autores como Ortiz, Castillo y García (2010) quienes señalan que la MO o AS es un área «consistente en la valoración, clasificación del componente axiológico del lenguaje, es decir, aquellos aspectos lingüísticos que codifican la subjetividad, en términos de positividad o negatividad, del hablante con respecto a aquello de lo que habla». Asimismo, en la tesis de Alexandra Dobrescu (2011) se acota que el Análisis de Sentimiento es una tarea del PLN cuyo principal objetivo es:

> […] la detección automática de los sentimientos expresados en textos (normalmente por una fuente sobre un 'objeto', que puede ser una persona, un evento, un producto, una organización, etc.) y su clasificación según su polaridad/orientación que tienen (normalmente *positiva, negativa* o *neutra*, aunque distintos autores han propuesto escalas más finas de sentimientos, incluyendo por ejemplo las clases *muy positivo* o *muy negativo*).

Por otro lado, Jorge Carrillo de Albornoz Cuadrado (2011) en su tesis, aunque se refiere a la misma como *Análisis Sentimental*, considera que este término «hace referencia a la tarea

de análisis, identificación y clasificación de todo tipo de contenido emocional, subjetivo u opinado». También, existen otros autores que marcan una división conceptual entre AS y MO:

> *El análisis de sentimiento* se centra en determinar la actitud del autor de un texto con respecto a un determinado tema. *La minería de opiniones*, por su parte, analiza los textos a un nivel de granularidad más fino y se plantea identificar qué opina el autor del texto sobre aspectos concretos del tema sobre el que escribe (un producto, una institución, una persona, un partido político…) (Troyano Jiménez *et al.*, 2015)

Finalmente, autores más recientes explican de forma sintética que el Análisis de Sentimientos es una tarea particular del Procesamiento de Lenguaje Natural (PLN) que: «trata de determinar la polaridad que un texto pretende transmitir. Generalmente, esta tarea se ha enfocado desde el PLN como una tarea de clasificación automática de un texto en tres clases de polaridad: positiva, negativa o neutra» (Ortiz, Castillo y García, 2010)

En el contexto español, explica Eugenio Martínez Cámara (2015) que «aunque en la bibliografía internacional se emplea indistintamente *Sentiment Analysis* y *Opinion Mining*, en español hay una cierta tendencia al uso de *Análisis de Sentimientos*». Para efectos del presente estudio y dado el tipo de tarea que aquí realizaremos, preferimos ser parte de esta tendencia, dicho de otro modo, nos referiremos al área como ***Análisis de Sentimientos (AS***).

Desde nuestro enfoque y siguiendo la definición de Alexandra Dobrescu (2011) y la de Bing Liu (2012), podemos definir al Análisis de Sentimientos (AS) como la disciplina o área del Procesamiento del Lenguaje Natural (PLN) que se ocupa de detectar, identificar o analizar automáticamente los sentimientos, emociones u opiniones expresadas en textos (normalmente por una fuente sobre un 'objeto', que puede ser una persona, un evento, un producto, una organización, un problema, un suceso, un tema y sus atributos etc.) y su clasificación según su polaridad (que puede ser *positiva*, *negativa o neutra*).

2.2. Antecedentes históricos del Análisis de Sentimientos (AS)

Hemos expuesto en el apartado anterior que el mayor auge, estudio y popularidad del Análisis de Sentimientos se produjo en los primeros años del presente siglo XXI, sin embargo, hay una serie de trabajos que indican que las raíces se remontan a las últimas dos décadas del siglo pasado, sobre todo, en la década de los años noventa. Los estudios realizados por Albornoz Cuadrado (2011), Martínez Cámara (2015) y Peñalver Martínez (2015) son los principales materiales de apoyo para esbozar el subsiguiente panorama histórico del AS.

Los tres autores confeccionan un recorrido histórico por la disciplina partiendo de los distintos trabajos que se han venido elaborando y que han supuesto un avance significativo para la materia.

Los estudios que elaboraron Carbonell (1979) y Wilks y Bien (1983) se han considerado como los precursores de esta disciplina. El primero, propone en su tesis un modelo computacional para representar el pensamiento subjetivo de las personas. El segundo, por su parte, en años posteriores propone un estudio sobre la opinión o creencia que tiene un sujeto sobre otro a partir de un diálogo entre ambos.

Si bien es cierto, ambos trabajos fueron importantes en el desarrollo de esta disciplina, los diversos estudios elaborados en la última década del siglo XX, fueron todavía mucho más trascendentales porque se enfocaron en el análisis de textos y su interpretación subjetiva. Es así, como la investigación da un giro y se centra, por ejemplo, en la interpretación de metáforas (Hearst, 1992), la clasificación de lenguaje afectivo y emocional (Kantrowitz, 2003), el reconocimiento de bloques textuales subjetivos con el punto de vista del autor sobre un determinado agente (Wiebe y Bruce, 1995)

Ahora bien, si los años noventa fueron determinantes para la evolución del AS, lo ha sido mucho más este siglo XXI. En el transcurso de casi estas dos décadas es cuando la disciplina ha adquirido gran popularidad y estudio en diferentes países del mundo. El gran abanico de publicaciones, libros, artículos, revistas, tesis, congresos, asociaciones, talleres y experimentos que se han venido acumulando en este breve espacio de tiempo son innumerables y dan muestra del nivel de interés y estudio que ha ganado el campo. Ahora bien, ante esto es lógico preguntarnos a qué se debe este interés por estudiar las opiniones y los sentimientos que vierten las personas sobre un objeto, un individuo, un tópico, un producto, un servicio, etc. Al respecto, Pang y Lee (2008) consideran que este notable avance se debe a tres posibles factores:

1. La proliferación de métodos de aprendizaje automático aplicables a problemas de PLN.
2. La disponibilidad de conjuntos de datos etiquetados, prestos y dispuestos para su uso en sistemas basados en aprendizaje automático. La posibilidad de compilar y usar conjuntos de datos para el análisis de la opinión, vino precedida por el florecimiento que Internet estaba experimentando en los últimos años del siglo XX. En los primeros años del siglo XXI, comenzaron a brotar las primeras plataformas

web donde se publicaban opiniones, lo que contribuía, en gran manera, a la preparación de esos primeros corpus de opiniones.

3. El inicio, por parte de la comunidad investigadora, de la toma de conciencia sobre el reto intelectual que supone el extraer la posición de una persona respecto a un agente con el simple hecho de escudriñar automáticamente lo que ha escrito, así como las posibles aplicaciones en las que esa capacidad podría derivar.

Los distintos trabajos que se han venido publicando en este siglo, han dado paso a que se empleen los términos: *Análisis de Sentimientos* y *Minería de Opinión* para denominar a la disciplina (como lo vimos en el apartado anterior).

Asimismo, podemos hallarnos con otras cuestiones que se han venido discutiendo y estudiando en los últimos años. Por ejemplo, una de estas cuestiones es la relacionada con las técnicas que se han construido para realizar la tarea que conlleva la MO y el AS. Las técnicas más importantes que se han desarrollado se han venido agrupando en tres grupos: medidas probabilísticas de asociaciones de palabras (Kamps y Marx, 2002), técnicas que usan información sobre relaciones léxicas (Kamps y Marx, 2002; Kim, Jeong y Ryul Jeong, 2014), y técnicas que usan bases de datos léxicas (Esuli y Sebastiani, 2005; Andreevskaia y Bergler, 2006). En resumen, todos estos trabajos clasifican las palabras en dos niveles o categorías: *positivas o negativas*, y además asignan a cada texto una puntuación positiva o negativa.

Poco tiempo después, surgieron otros trabajos que no solo se centraron en determinar la polaridad (positivo/negativo) de un texto, sino también se enfocaron en determinar el nivel o la intensidad de la polaridad del mismo como alto/medio/bajo, positivo/negativo (Dave, Lawrence y Pennock, 2003; Goldberg y Zhu, 2006). Por otro lado, mediante un conjunto de métodos estadísticos de selección de características y aplicando técnicas de aprendizaje computacional, algunos autores como Pang, Lee y Vaithyanathan (2002), Pang y Lee (2004), Gamon (2004) y Baccianella, Esuli y Sebastiani (2009) explotaron una gran cantidad de datos que provenían de opiniones sobre sobre películas, servicios y productos. No obstante, «las investigaciones llevadas a cabo por este conjunto de investigadores concluyeron que las técnicas de aprendizaje computacional, a nivel general, no alcanzan un rendimiento satisfactorio en la tarea de análisis de sentimientos» (Peñalver Martínez, 2015).

Partiendo de esta situación, otros autores como Zhou y Chaovalit (2008) y Zhao y Li (2009) comenzaron a emplear nuevas técnicas para el análisis de sentimientos distintas a las que tienen como eje central el aprendizaje computacional, nos referimos a las técnicas de la

Web semántica, específicamente las ontologías. ¿Cuál es el principal objetivo de estas? La determinación de la polaridad partiendo de las características de un concepto. Ahora bien, según, Peñalver Martínez (2015) «En los últimos años se han impuesto dos métodos principales para clasificar las opiniones de los usuarios a nivel de características: los métodos basados en modelos y los métodos estadísticos».

Finalizamos este recorrido histórico afirmando que en la actualidad las diversas técnicas, métodos y herramientas están en un constante proceso de validación y cada vez son más los talleres que se organizan, las asociaciones que se crean y los trabajos que se publican en torno al Análisis de Sentimientos. Los próximos años serán fundamentales para la disciplina.

2.3. Tareas del Análisis de Sentimientos (AS)

El Análisis de Sentimientos se encarga de la realización, en mayor o en menor medida y según el propósito, de una serie de tareas. Actualmente, siguiendo lo expuesto en la tesis de Jorge Carrillo Albornoz Cuadrado (2011), son tres las principales tareas que se pueden englobar en el marco de esta nueva disciplina de PLN y, por consiguiente, de la Minería de Textos:

1. **Detección o clasificación de la polaridad:** Es, quizá, la principal tarea o a la que se le ha dedicado más tiempo porque puede estar ligada tanto a la MO como al AS. En la literatura se le conoce como clasificación sentimental. ¿En qué consiste? Es el proceso mediante el cual los textos de entrada son analizados, procesados y clasificados en ***positivos o negativos*** según la carga emotiva que presenten. Los niveles en los que se puede efectuar este análisis pueden ser a nivel de documento, frase o característica. (Más adelante detallaremos sobre cada uno de estos niveles).
2. **Detección de la subjetividad:** Esta tarea está más vinculada a lo que se ha denominado Análisis de Subjetividad. La detección de la subjetividad se efectúa mediante un proceso en el cual el texto es analizado y diferenciado atendiendo a la carga subjetiva expresada por parte del autor. En definitiva, lo que busca es determinar cuándo un texto es subjetivo u objetivo. Su principal impulsora ha sido Janyce M. Wiebe.

3. **Detección de fragmentos del texto que contienen opiniones:** El nexo entre la MO y esta tarea es estrecho. Aquí, el texto es analizado para extraer aquellos fragmentos que expresan opiniones por parte del autor.

Albornoz Cuadrado (2011), adiciona a estas tres tareas centrales, otras de mayor complejidad:

1. **Detección de la intensidad emocional:** En este caso, la clasificación se ramifica y ya no solo se intenta reconocer la polaridad de un texto, sino también la intensidad de su orientación. Aquí, el texto de entrada es clasificado atendiendo a la intensidad de su polaridad en diferentes clases (*fuertemente negativo, negativo, neutro, positivo* y *fuertemente positivo*). Como es razonable, a mayor número de clases a las que se enfrenta un sistema, menor es su porcentaje de acierto.
2. **Clasificación sentimental mediante tópicos o características del texto:** donde los parámetros que ponderan la clasificación están basados en los tópicos o características de los temas tratados en los textos. Usualmente, este tipo de sistemas suelen evaluar documentos que recogen opiniones sobre productos o servicios donde ciertos aspectos de esos productos o servicios condicionan más que otros la carga afectiva global de la opinión

Por nuestra parte, añadiríamos una tarea mucho más avanzada que ha ganado mucho interés en los últimos años y que no solo se limita a la clasificación del texto según la polaridad:

1. **Detección de estados emocionales:** Aquí se intenta detectar o determinar, desde un plano lingüístico, los estados emocionales (ira, tristeza, alegría, disgusto, sorpresa, etc.) que subyacen en el texto. Indudablemente, esta tarea está relacionada con lo que se ha denominado como Análisis de Afectos, Análisis de Emociones y, desde luego, el Análisis de Sentimientos.

2.4. Clasificación de los enfoques o métodos para realizar Análisis de Sentimientos (AS)

Enfoques, métodos, técnicas, paradigmas, líneas o aproximaciones son algunos de los términos que emplean diversos autores en la literatura para referirse a las formas o a los caminos que se seleccionan para ejecutar las tareas propias del AS (descritas en el apartado anterior), aunque más propiamente para efectuar la clasificación de la polaridad (positivo, negativo o

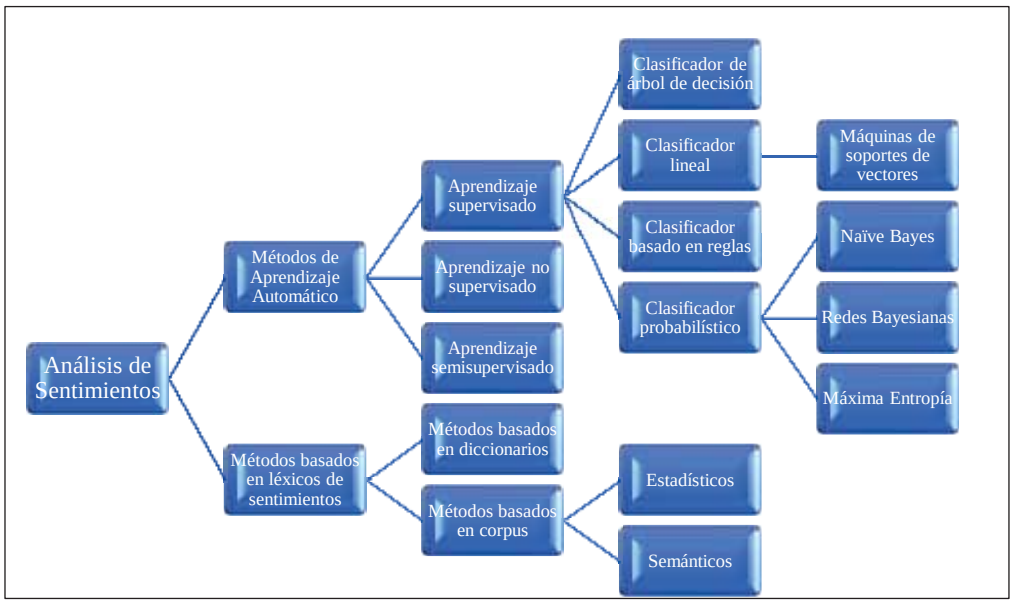
neutro) de un texto. Aunque no hay unificación de criterios para su denominación, sí la hay, en torno a cuáles son los enfoques, métodos o técnicas más utilizados.

Los trabajos de (Martínez Cámara, Martín Maldivia y Ureña López, 2011), (Vilares, Alonso y Gómez Rodríguez, 2013), (Molina González *et al.*, 2015), (Peñalver Martínez, 2015) y López Barbosa (2015) coinciden categóricamente que son dos los principales enfoques o métodos que se pueden elegir para clasificar los sentimientos: ***aplicando aprendizaje computacional o automático*** (Machine Learning ML) o ***aplicando un enfoque semántico o basado en diccionarios o léxicos de sentimientos.*** A continuación, siguiendo a los autores descritos anteriormente, explicaremos de manera sintética en qué consisten cada uno de estos:

a) **Enfoques o métodos que aplican aprendizaje computacional o automático:** son los que han sido más empleados y «se basan en entrenar unos modelos a partir de una colección de datos etiquetados a priori, con el objetivo de predecir el valor de salida correspondiente a cualquier dato de entrada válido» (Molina González *et al.*, 2015). O sea, estos sistemas o modelos están entrenados con una serie de ejemplos que han sido analizados previamente por personas y que han determinado su valor negativo/positivo. ¿Cuál es el principal inconveniente que manifiesta este enfoque? La falta de datos etiquetados, es decir, si hay pocos ejemplos anotados, entonces el campo o área de análisis se limita a un determinado dominio o canal, puesto que, al cambiar de uno a otro, el lenguaje puede variar y, por lo tanto, los datos con los que ha sido entrenada la máquina no responden plenamente.
b) **Enfoques o métodos con una orientación semántica o basado en diccionarios o léxicos de sentimientos:** estos se apoyan «en la orientación semántica (OS) de las palabras, donde cada término que expresa opinión es anotado con un valor que representa su polaridad» (Vilares, Alonso y Gómez Rodríguez, 2013). Básicamente, su estrategia es buscar en el texto, que queremos procesar, todas las palabras que tienen fijadas una polaridad (positiva, negativa o neutra) y luego realizar una métrica de lo hallado. Una de las propiedades fundamentales de estos métodos o técnicas es que se basan en diccionarios o lexicones de sentimientos, en estos se almacenan las palabras en conjunto con la asignación de la polaridad que se le ha dado a cada una. Un inconveniente es que al momento de procesar el texto solo se analizarán las palabras que estén en el diccionario y, por lo tanto, se dejarán por fuera otros términos que, quizá, sean relevantes. No obstante, esta dificultad se puede superar invirtiendo tiempo y recursos para construir lexicones con un gran número de palabras.

En la tesis de Rutilio Rodolfo López Barbosa (2015) se propone un esquema en el que se ilustra de manera perfecta la clasificación de los métodos o técnicas que se emplean para realizar análisis de sentimientos, según los dos enfoques que hemos abordado anteriormente, es decir, los que aplican aprendizaje computacional o automático (Machine Learning ML) y los de orientación semántica o basados en lexicones de sentimientos. Este esquema que presenta el autor se ha elaborado a partir de un esquema previo que es propuesto en el estado de arte del trabajo de (Medhat, Hassan y Korashy, 2014). El esquema que presenta López Barbosa (2015) se adaptó para enfatizar que la investigación es sobre Análisis de Sentimientos.

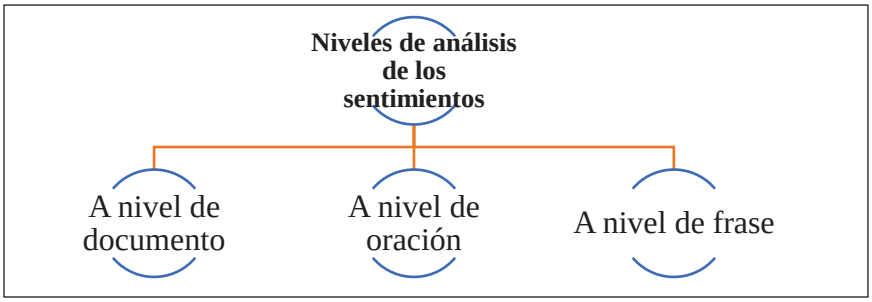
FIGURA 3: Clasificación de métodos de análisis de sentimientos



2.5. Niveles de análisis o de clasificación de los sentimientos

El alcance, la amplitud, el detallismo o la clasificación del análisis de los sentimientos de un texto se puede realizar desde diferentes niveles de análisis. Según los estudios de Liu (2012), Pang y Lee (2008), López Barbosa (2015) y las tesis de Vilares, Alonso y Gómez Rodríguez (2013), Martínez Cámara (2015) y Peñalver Martínez (2015) se pueden distinguir tres niveles de análisis principales, tal y como lo ilustra la Figura 4.

FIGURA 4: Niveles de análisis o de clasificación de los sentimientos



Como observamos los niveles que proponen los autores son a nivel de documento, a nivel de oración y a nivel de frase. A continuación, explicaremos en qué consiste cada uno de ellos:

a) Análisis a nivel de documento

En este nivel se concibe al texto en su totalidad o globalidad, es decir, como una única unidad básica de información. Por lo tanto, **«**se asume que los documentos (por ejemplo, las críticas a productos y servicios) contienen opiniones. De esta forma, la tarea principal es la identificación del sentimiento o actitud positiva o negativa global en todo el documento» (López Barbosa, 2015). La mayoría de los enfoques o sistemas de Análisis de Sentimientos existentes están basados en la clasificación de los sentimientos a nivel de documento (Peñalver Martínez, 2015).

b) Análisis a nivel de oración

En este nivel de análisis como ya su nombre lo indica la unidad que se toma como referencia para analizarse es la oración, por lo tanto, este nivel considera que no todas las oraciones tienen opiniones o sentimientos, **«**por lo que es común que en una primera etapa se separen aquellas oraciones que expresan opiniones de aquellas que expresan hechos, es decir, que se realice una clasificación de subjetividad» (Liu, 2010). ¿Para qué? Básicamente, para intentar identificar la orientación en términos de polaridad (positivo, negativo o neutro) que tienen.

c) Análisis a nivel de frase

A este nivel también se la ha denominado «análisis a nivel de cláusula». Wilson, Wiebe y Hoffmann (2005) han sido los investigadores que lo han usado principalmente y, por lo tanto, se erigen como los impulsadores de este tipo de análisis. Estos investigadores presentaron un modelo para identificar la subjetividad/objetivad de una gama de oraciones pertenecientes a conjunto de documentos. ¿De qué manera lo lograron? Etiquetando o anotando manualmente la polaridad contextual (positivo/negativo) de aproximadamente 16.000 expresiones o frases subjetivas. Todas las oraciones de los documentos (material de prueba) se compararon con los patrones aprendidos (las expresiones o frases), primero, para determinar si las oraciones eran subjetivas u objetivas (Análisis de Subjetividad), segundo, y luego de realizar el paso anterior, para detectar la polaridad u orientación semántica de aquellas que fueron consideradas como subjetivas. Este tipo de sistemas nos daría, en definitiva, una clasificación de subjetividad.

2.6. Herramientas o recursos para realizar Análisis de Sentimientos (AS)

Acabaremos este segundo capítulo, exponiendo que desde los orígenes del Análisis de Sentimientos (AS) se han venido desarrollando, naturalmente, una gama de herramientas, recursos o aplicaciones para realizar las diferentes tareas que se propone esta disciplina. En la actualidad, todo este conjunto de herramientas pueden ser clasificadas según distintos criterios: de acuerdo a los recursos iniciales que necesitan, en relación con los métodos o técnicas de clasificación de sentimientos que emplean, de acuerdo al grado de automatización que tienen, según el dominio para el que han sido validadas, en relación con el idioma para las que han sido aplicadas sus metodologías y sus corpus, de acuerdo con el nivel de clasificación o análisis de los sentimientos y, finalmente, en función del acceso que tienen los usuarios con las mismas.

Por todo ello, consideramos pertinente señalar y describir brevemente cuáles son esas herramientas y recursos que se han creado y empleado para efectuar Análisis de Sentimientos. Un punto que queremos mencionar es que, aunque nos gustaría, por cuestiones de interés y espacio, no podemos hacer referencia a la totalidad de herramientas, sino que, presentamos aquellas que han tenido un mayor uso en los últimos años y, por consiguiente, que han supuesto un avance significativo para el área.

Por una parte, Peñalver Martínez (2015) habla de que en los últimos años se han ido construyendo una serie de recursos lingüísticos para poder utilizarse en las soluciones ideadas en esta rama. Estos son los ejemplos de los recursos lingüísticos más significativos para el idioma inglés:

a) ***WordNet*** (Miller, 1995): Es una gran base de datos léxica de inglés. Los sustantivos, los verbos, los adjetivos y los adverbios se agrupan en conjuntos de sinónimos cognitivos (*synsets*), cada uno expresando un concepto distinto. Los *synsets* están interrelacionados por medio de relaciones conceptuales-semánticas y léxicas.
b) ***WordNet Affect*** (Strapparava y Valitutti, 2004): Este recurso se construyó partiendo de *WordNet* a través de la selección y etiquetado de subconjuntos de *synsets* que representaban el conocimiento afectivo o emocional. Se partió de un conjunto inicial de palabras afectivas semilla, clasificadas de acuerdo con las seis categorías básicas de emociones (alegría, tristeza, miedo, sorpresa, ira y disgusto) y se expandió el léxico utilizando las relaciones de *WordNet*.

c) ***SentiWordNet*** (Esuli y Sebastiani, 2006): Es un recurso léxico para la MO. SentiWordNet asigna a cada *synsets* de WordNet tres puntuaciones de sentimiento: positivo, negativo, objetivo.

d) ***Emotion triggers*** (Balahur Dobrescu y Montoyo Guijarro, 2008): Es un método de identificación y clasificación de la valencia y las emociones presentes en un texto. Con este método se introduce un nuevo concepto denominado disparador de emoción "*Emotion trigger*". Inicialmente, se construyó de forma incremental una base de datos léxica de disparadores de emoción asociados a la cultura con la que se quiere trabajar, luego esta fue alimentada con otras bases de datos como *WordNet, NomLex* y otros dominios relevantes. Un disparador de emoción es una palabra o concepto que expresa una idea que, dependiendo del mundo de interés del lector, de los factores culturales, educativos y sociales, conduce a una interpretación emocional del contenido del texto o no.

e) ***MicroWNOp*** (Cerini *et al.*, 2007): Es un corpus que está compuesto por 1105 conjuntos de *synsets* de *WordNet*. El corpus se divide en tres partes: Una parte común que está integrada por 110 *synsets* que 5 evaluadores han evaluado conjuntamente para alinear sus criterios de valoración. El grupo1, que lo conforman 496 *synsets*, pero en este caso estos han sido evaluados por tres evaluadores. El grupo 2 que lo componen 499 *synsets* que han sido evaluados por los dos evaluadores restantes. Tanto en el grupo 1 como el 2, los evaluadores realizaron su evaluación de forma independiente a la parte común.

Por otra parte, López Barbosa (2015) en su tesis hace una clasificación de las herramientas y recursos en función del acceso que puede tener el usuario a las mismas, esto es, las divide en herramientas comerciales y de uso libre, ambas son dirigidas para el inglés:

a) **Herramientas comerciales:** estas ofrecen diferentes servicios, tales como el análisis de tendencias, la extracción de características de los objetos de opinión más comentados, la agrupación de sentimientos, emociones, actitudes e incluso el análisis de ironía. Algunas de las herramientas más populares son: *AlchemyAPI, Lymbix, openAmplify, Repustate, Semantria, SentimentAnalizer, SentiRate, Sentimetrix.*

b) **Herramientas de uso libre:** En este caso, las herramientas que propone el autor son específicas para el Análisis de Sentimientos de Tweets. Estas son: *OpinionCrawl, Sentimentor, Sentiment140, StreamCrab, TweetFeel* y *Twitrratr.*

Hasta aquí, hemos descrito algunas de las principales herramientas y recursos que se han creado y usado para el Análisis de Sentimientos. Evidentemente, hay una cuestión que podemos deducir y que difícilmente podríamos pasar por alto y es que todas estas herramientas y recursos han sido creadas con metodologías y corpus en inglés.

Al respecto, en la literatura son muchos los investigadores, tales como Dobrescu (2011), Martínez Cámara, Martínez Maldivia y Ureña López (2011), Vilares, Alonso y Gómez Rodríguez (2013) Molina González *et al.* (2015) y Henriquez, Guzmán y Salcedo (2016), que comparten dos percepciones: la primera es que la mayoría de trabajos, herramientas e investigaciones se han centrado en el análisis de textos o corpus escritos en inglés, así, lo señalan Martínez Cámara, Martí-Maldivia y Ureña López (2011): «existen muchos trabajos en el campo del análisis de sentimientos, habiéndose aplicado en multitud de dominios, pero la mayor parte de ellos han sido realizados sobre corpus de documentos en inglés». La segunda, es que reconocen que «si bien el inglés es la lengua predominante en Internet, hay otros idiomas como el chino o el español que cada vez tienen más presencia en la red» (Martínez Cámara *et al.*, 2011), y por este motivo, «es que la generación y uso de recursos propios en el idioma de los documentos a tratar se esté convirtiendo en un tema crucial para realizar la clasificación de opiniones mediante orientación semántica» (Molina González *et al.*, 2015).

Para el caso del español, en lo relacionado a la investigación de AS, los grupos que más sobresalen son el ITALICA de la Universidad de Sevilla y el grupo SINAÍ de la Universidad de Jaen, pero enfocándose en el AS de la lengua árabe (Rushdi Saleh *et al.*, 2009). Ahora bien, en cuanto a las herramientas y recursos, podemos afirmar, a partir de la tesis de Martínez Cámara (2015), que se han creado y usado algunas herramientas y recursos para efectuar Análisis de Sentimientos en español. A continuación, presentamos las herramientas y recursos que ha recopilado el autor antes mencionado:

1. **Corpus de opiniones disponibles en español:**
   - *Spanish Movie Reviews*[1]*:* Es el primer corpus de opiniones que se puso a disposición de la comunidad investigadora. El corpus está formado por 3878 opiniones circunscritas al dominio del cine, y están etiquetadas en una escala de intensidad de opinión que oscila entre el nivel 1 y 5.

[1] http://www.lsi.us.es/~fermin/corpusCine.zip

- *Corpus General de TASS*[2]*: TASS* es el primer taller enfocado a la promoción de la investigación de técnicas de AS en español aplicadas a textos publicados en Twitter. En los años que se viene celebrando el TASS se han desarrollado cuatro corpus de tweets. De esos cuatro corpus el principal es el conocido como Corpus General de TASS. Este tiene dos versiones: una en la que los tweets están etiquetados en una escala de 6 niveles de polaridad, y otra en la que la escala se reduce a 4 niveles. Los 6 niveles de polaridad son: P+ (muy positivo), P (positivo), NEU (neutro), N (negativo), N+ (muy negativo) y NONE (sin polaridad). Los 4 estratos de polaridad se corresponden con: P (positivo), NEU (neutro), N (negativo) y NONE (sin polaridad).
- *OCA*[3]*:* Corpus elaborado durante el año 2010 sobre críticas de cine extraídas de varias páginas web especializadas escritas en árabe. El corpus está etiquetado a nivel de documento, y está formado por un total de 500 críticas, 250 positivas y otras 250 negativas.
- *EVOCA*[4]*:* Este corpus es la traducción automática del corpus OCA a inglés, por lo que está conformado por el mismo número de opiniones que OCA.
- *COAH*[5]*:* Se trata de un corpus de opiniones de hoteles de Andalucía. El corpus está conformado por 1816 opiniones etiquetadas en una escala de 5 niveles de opinión.
- *Corpus MCE*[6]*:* El corpus MCE es el resultado de la traducción automática del corpus *Spanish Movie Reviews*, y fue desarrollado para el estudio de la combinación de clasificadores especializados en distintos idiomas, con el fin de mejorar la clasificación de la polaridad en español.
- *COST*[7]*:* es un corpus de tweets en español con etiquetas de polaridad impuras, es decir, su etiqueta de opinión está determinada por los emoticonos que aparecen en los tweets.

2. Bases de conocimiento de opinión orientadas al tratamiento del español:

- *Léxico de Pérez Rosas*[8]*:* la denominación "Léxico de Pérez Rosas" se ha utilizado para diferenciar a la lista de palabras en el trabajo de Martínez Cámara (2015), ya que sus autores no le han asignado ningún nombre a su léxico. El léxico de Pérez Rosas es una

[2] http://www.daedalus.es/TASS2015/private/general-tweets-train-tagged.xml
[3] http://sinai.ujaen.es/oca-corpus/
[4] http://sinai.ujaen.es/evoca-corpus/
[5] http://sinai.ujaen.es/coah/
[6] http://sinai.ujaen.es/mce-corpus/
[7] http://sinai.ujaen.es/cost-2/
[8] http://web.eecs.umich.edu/~mihalcea/downloads/SpanishSentimentLexicons.tar.gz

lista de palabras de opinión en dos niveles que los autores denominan como: *strength lexicon* (lexicón preciso) y *medium strength lexicon* (lexicón menos preciso). Los elementos que los distinguen a cada uno, además del número de palabras que incluyen (strength: 1347 y medium: 2496), estriba en los términos semilla que se han empleado para su generación

- *ML-SentiCon*[9]*:* Lista de lemas de opinión estratificados en 8 niveles de precisión. Los autores aplican una versión mejorada del algoritmo de generación de *SentiWordNet* para crear una lista de *synsets* con su respectiva puntuación de polaridad.
- *Spanish Emotion Lexicon*[10]*:* Se trata de una lista de 2036 palabras clasificadas en 6 estados de ánimo diferentes: alegría, enfado, miedo, tristeza, sorpresa y disgusto. Las palabras tienen asociado un valor de probabilidad, que los autores llaman PFA, de pertenencia a una de las 6 categorías.
- *ElhPolar*[11]*:* Lista de palabras de opinión construida a partir de varias fuentes de datos. Primero los autores tradujeron al español el *MPQA Subjectivity Lexicon.* Seguidamente, mediante la aplicación de un método basado en la ratio de verosimilitud (*log likehood ratio*) los autores seleccionaron las palabras positivas y negativas más prominentes del conjunto de entrenamiento del Corpus General de TASS. Finalmente, complementaron el lexicón con expresiones coloquiales.
- *iSOL*[12]*:* Lista de palabras de opinión en español desarrollada durante el transcurso de la investigación que realizó Martínez Cámara (2015). El proceso de generación de la lista constó de dos fases, una primera que se centró en la traducción automática al español de la lista de palabras de opinión en inglés *BLOL*, y una segunda que se circunscribió a la corrección manual de los errores de traducción y a la inclusión de más términos indicadores de opinión.
- *eSOLDomainGlobal*[13]*:* Las experimentaciones relacionadas con la adaptación al dominio de iSOL dieron lugar a distintas versiones de iSOL adaptadas a los 8 dominios del corpus *SFU*: coches, hoteles, lavadoras, libros, teléfonos móviles, música, ordenadores y películas.

[9] http://www.lsi.us.es/~fermin/ML-SentiCon.zip
[10] http://www.cic.ipn.mx/~sidorov/SEL.zip
[11] http://komunitatea.elhuyar.org/ig/files/2013/10/ElhPolar_esV1.lex
[12] http://sinai.ujaen.es/isol/
[13] http://sinai.ujaen.es/esoldomainglobal/

III. Análisis de Sentimientos aplicado a la literatura: ¿El ordenador es capaz de detectar los sentimientos de la literatura?

El Análisis de Sentimientos se ha aplicado, desde sus orígenes, primordialmente a textos que contienen explícitamente información subjetiva o en las que hallamos opiniones, tales como: críticas o reseñas de productos, servicios, temas, personajes, etc. Los sitios web 2.0 han sido las principales fuentes o plataformas de donde se han extraído todos estos textos. Como habíamos expuesto en apartados anteriores, desde sus inicios, los campos más interesados en crear sistemas de AS y usarlos fueron el ámbito comercial, empresarial y político. En nuestro caso, para el área que queremos realizar AS es para el ámbito literario.

Dado que los textos literarios no han sido los principales textos con los que se ha efectuado AS, entonces, cabe preguntarnos: ¿si es viable realizar Análisis de Sentimientos con obras literarias? o ¿los textos literarios contienen información subjetiva, expresan emociones o sentimientos? o, quizá, si ¿los escritores al escribir una obra literaria intentan, mediante el discurso narrativo, el narrador y las experiencias vividas por los personajes, plasmar una serie de emociones y suscitarlas, seguidamente, en los lectores? Las respuestas se pueden resumir en una sola palabra: sí.

Todas las personas que alguna vez hemos establecido una comunicación o interacción con una novela, un cuento, un microrrelato, un poema, una tragedia, una comedia, un drama, etc., o dicho de forma sencilla, todas las personas que hemos leído alguna obra literaria, sin lugar a dudas, su lectura nos tuvo que haber causado risas, lagrimas, disgustos, sorpresas, enfado, miedo, tristeza… en fin, tuvimos que haber sentido lo que se denomina una emoción o un sentimiento. Si no nos hubiera producido alguno de estos sentimientos, seguramente, nos hubiéramos aburrido y hubiésemos dejado de leer, porque son las emociones las que, desde nuestra percepción, mantienen el interés por lo que se nos está relatando.

Todo esto se conecta perfectamente con una apreciación que tenemos y es que los seres humanos cuando nos enfrentamos a una obra de ficción (sea una novela, película, cuento, etc.), queremos sentir y pensar de la manera más cercana a la realidad, o sea, intentamos adoptar ese mundo ficcional como un mundo verosímil (al menos mientras estamos leyendo). Al respecto, Bermúdez Antúnez (2010) señala que:

> […] la respuesta emocional es una realidad determinante durante el proceso de lectura o recepción estética. La interrelación (leer literatura y disfrutar de mundos ficcionales) tiene como meta central producir en los receptores respuestas emocionales que lo conecten (acercándolo o resistiéndose) al mundo ficcional propuesto.

Por tal razón, nosotros somos del parecer que cuando un escritor se plantea el objetivo de escribir una obra literaria, indudablemente uno de los propósitos que persigue es suscitar una serie de emociones en los lectores. Este efecto, lo puede lograr con la propia historia que está relatando o bien a través de las formas, técnicas o estructuras de organizar esa historia.

Solo para ejemplificar nos vamos a ceñir a los inicios de algunos textos de Márquez, el primero, es el de *Crónica de una muerte anunciada:* «El día en que lo iban **a *matar***, Santiago Nasar se levantó a las 5.30 de la mañana para esperar el buque en que llegaba el obispo» (García Márquez, 1981). El segundo, es el de *Cien años de soledad:* «Muchos años después, ***frente al pelotón de fusilamiento***, el coronel Aureliano Buendía había de recordar aquella tarde remota en que su padre lo llevó a conocer el hielo» (García Márquez, 1967), o bien, al segundo párrafo de *El amor en los tiempos del cólera:* «***Encontró el cadáver*** cubierto con una manta en el catre donde había dormido siempre, cerca de un taburete con la cubeta que había servido ***para vaporizar el veneno***» (García Márquez, 1985). Como podemos ver los tres inicios de los tres textos nos suscitan ciertas emociones, por ejemplo, miedo o sorpresa, pero lo más interesante es que la trasmisión de estas emociones nos estimula o nos motiva a seguir leyendo. De este proceso es el que estamos hablando.

Según Bermúdez Antúnez (2010), este proceso de la construcción emocional a través de los textos literarios puede ser vehiculizada solo por el hecho de que los escritores presenten etiquetas léxicas, dicho de forma sencilla, palabras o términos que hagan referencia a esos estados o procesos emocionales, por ejemplo, en el inicio de *Crónica…* la etiqueta léxica o la palabra *«matar»*, trae consigo una carga emocional impresionante, en este caso negativa. El autor antes mencionado, explica que además de emplear una serie de palabras, los escritores también recurren a una serie de procedimientos retóricos para terminar de impregnar emocionalmente el mundo ficcional.

A partir de todo esto, entonces, podemos afirmar que sí se podría realizar el Análisis de Sentimientos a textos literarios, o al menos, sí podríamos experimentar, porque los sistemas (por lo menos el que nosotros emplearemos) podrían detectar esas etiquetas léxicas con cargas emocionales.

Si bien es cierto, hay algunos trabajos como el de Silge (2016) o bien las diversas publicaciones que ha cargado en su blog,[14] Matthew L. Jockers, creador del paquete que utilizaremos para efectuar nuestro AS, todavía, son muy pocos los investigadores o los estudiosos que está interesados en experimentar con el AS aplicado a literatura. En nuestro caso, este interés nació a partir de la lectura del trabajo de Silge (2016), quien, a su vez, fue motivada por los trabajos que ha elaborado Jockers (2015). Por tal razón, parece acertado que expongamos de forma sucinta el experimento que llevó a cabo la autora mencionada anteriormente. ¿Para qué? Para conocer el modelo que utilizaremos y para valorar la viabilidad de la aplicación del AS a obras literarias.

3.1. You must allow me to tell you how ardently I admire and love Natural Language Processing[15] de Julia Silge

El 08 de marzo de 2016, Julia Silge[16] publicó en su blog el *post «You must allow me to tell you how ardently I admire and love Natural Language Processing»*, cuya traducción al español podría ser: *«Permíteme decirte cómo admiro y amo apasionadamente el Procesamiento de Lenguaje Natural»*, en este *post*, la autora nos presenta un trabajo experimental de Análisis de Sentimientos de una novela, específicamente, de *Orgullo y Prejuicio* de Jane Austen. A continuación, exponemos su experimento. No sin antes puntualizar que ciertos términos como librería, texto plano, etc., serán definidos en el siguiente capítulo, por tal motivo, dirigimos el interés más a los resultados que al proceso.

El primer objetivo que se planteó la autora fue obtener la versión digital de la novela, para ello acudió al *Proyecto Gutenberg*[17], de donde la obtuvo en formato UTF-8, es decir, como texto plano. El segundo paso que realizó fue modificar los datos para que el AS fuera lo más confiable posible. A través de *R*, el sistema, primero leyó el texto plano con la librería *readr*, seguidamente, eliminó la información del encabezado y del pie de página que traen consigo los ficheros que se descargan del Proyecto Gutenberg, esto se hizo para tener exclusivamente el texto real de la novela.

[14] http://www.matthewjockers.net/
[15] https://juliasilge.com/blog/you-must-allow-me/
[16] Es una científica de datos en Stack Overflow. Estudió Física y Astronomía y presentó su tesis doctoral en 2005. Es amante de la literatura, sobre todo de Jane Austen, y le gusta trabajar con el lenguaje de programación *R*.
[17] https://www.gutenberg.org/

El texto plano estaba compuesto por líneas de 70 caracteres y, puesto que, estas no constituyen un fragmento de texto ideal para los propósitos que se proponía la investigadora, por ello, usó la librería *stringr* para concatenarlas en fragmentos de 10 líneas, más o menos del tamaño de un párrafo. Básicamente, lo que hizo fue dividir todo el texto en fragmentos de 10 líneas.

Luego de haber hecho esta división, Silge (2016) señaló qué diccionario utilizó para llevar a cabo su AS. El diccionario o lexicón de sentimientos que usó fue el *NRC Word-Emotion Association Lexicon* de Saif Mohammad y Peter Turkey, el cual es implementado en *R* mediante el paquete *syuzhet.* Es así, como mostró algunos ejemplos de la funcionalidad que tiene este diccionario con textos literarios, es decir, nos enseñó cómo las puntuaciones que asigna a cada fragmento el *NRC* se acercan a la valoración que haría una persona. En la Figura 5, 6 y 7., se presentan dichos ejemplos.

FIGURA 5: Ejemplo de la valoración del NRC con fragmentos de *Orgullo y Prejuicio*

```
> get_nrc_sentiment("Nobody can tell what I suffer! But it is always so. Those who do no
t complain are never pitied.")
  anger anticipation disgust fear joy sadness surprise trust negative positive
1     1            0       0    0   0       1        0     0        2        0
```

La traducción del fragmento sería: “Nadie puede decir lo que sufro! Pero siempre es así. Los que no se quejan nunca tienen piedad.”. Las puntuaciones que le asignó el NRC son: enfado/ira = 1, tristeza = 1, negativo = 2.

FIGURA 6: Ejemplo de la valoración del NRC con fragmentos de *Orgullo y Prejuicio*

```
> get_nrc_sentiment("And your defect is to hate everybody.")
  anger anticipation disgust fear joy sadness surprise trust negative positive
1     2            0       1    1   0       1        0     0        2        0
```

La traducción del fragmento sería: “Y tu defecto es odiar a todo el mundo.”. Las puntuaciones que le asigno el NRC son: enfado/ira = 2, disgusto = 1, miedo = 1, negativo = 2.

FIGURA 7: Ejemplo de la valoración del NRC con fragmentos de *Orgullo y Prejuicio*

```
> get_nrc_sentiment("You must allow me to tell you how ardently I admire and love you.")
  anger anticipation disgust fear joy sadness surprise trust negative positive
1     0            0       0    0   1       0        0     1        0        2
```

La traducción del fragmento sería: “Debes permitirme decirte cuán apasionadamente te admiro y te amo.”. Las puntuaciones que le asignó el NRC son: alegría = 1, confianza = 1, positivo = 2.

Después de habernos mostrado la forma de trabajo y la funcionalidad que tiene el diccionario NRC analizando fragmentos de textos literarios, Silge (2016) continuó realizando el tercer paso que consistió en dividir todo el texto en los volúmenes correspondientes que tiene la novela. Así, pues, explicó la autora que *Orgullo y Prejuicio* contiene 61 capítulos divididos en tres volúmenes: el volumen I está conformado por los capítulos 1-23, el volumen II por los capítulos 24-42 y el volumen III por los capítulos 43-61. A partir de ello, se buscó y se indicó dónde estaban estas divisiones en el texto.

Ahora que se han modificado los datos, parece que todo está listo para iniciar con el AS. Por ello y en cuarto paso se muestra cuál es el sentimiento tanto positivo como negativo de la novela. ¿Cómo se hizo esto? A cada fragmento del texto se le asignó una puntuación tanto para el sentimiento positivo como para el negativo. Por ello, un fragmento de texto dado podría tener puntuaciones altas para ambas, puntuaciones bajas para los dos, o bien, cualquier combinación de los mismos. Asimismo, se identificaron y se señalaron los eventos o pasajes más importantes de la novela para anotar las tramas. Finalmente, los resultados se trazaron en dos gráficos: uno de puntos (Figura 8.) y otro de barras (Figura 9.)

FIGURA 8: Gráfico de puntos de los sentimientos positivos y negativos de *Orgullo y Prejuicio*

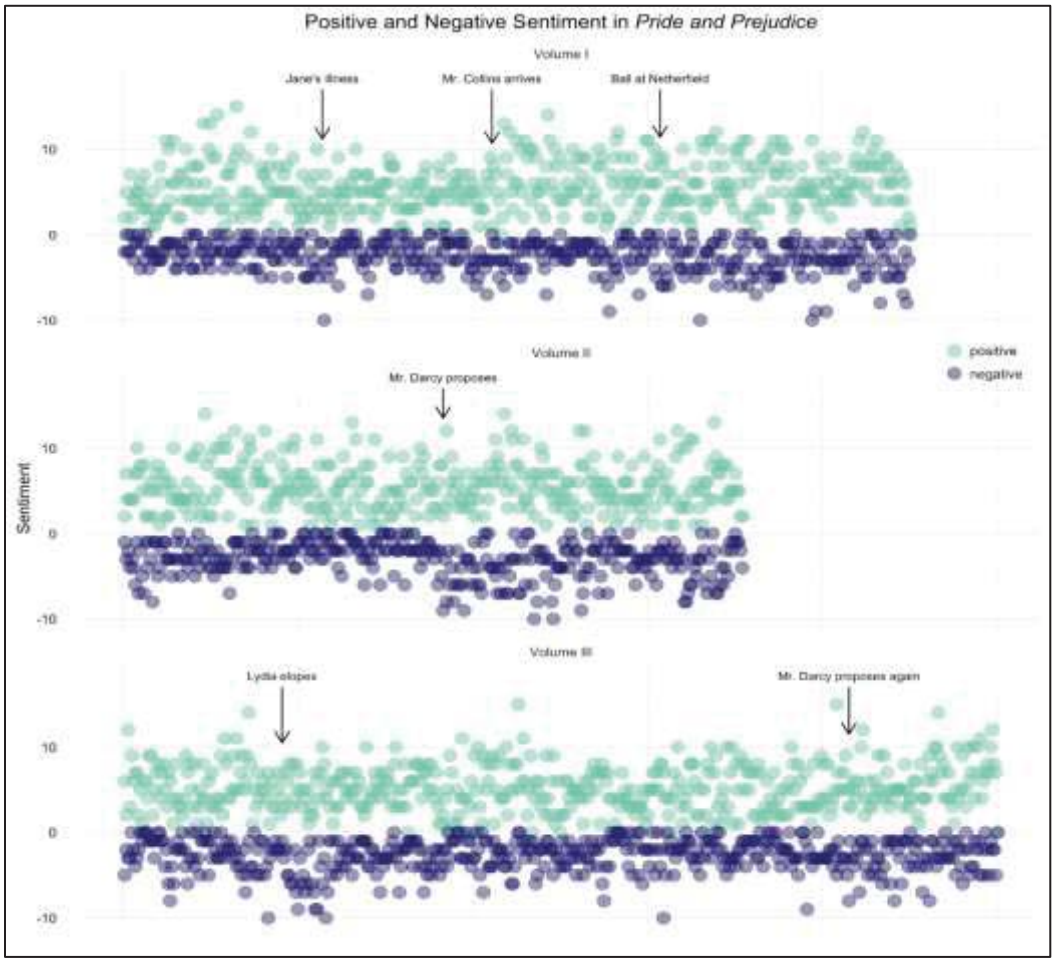
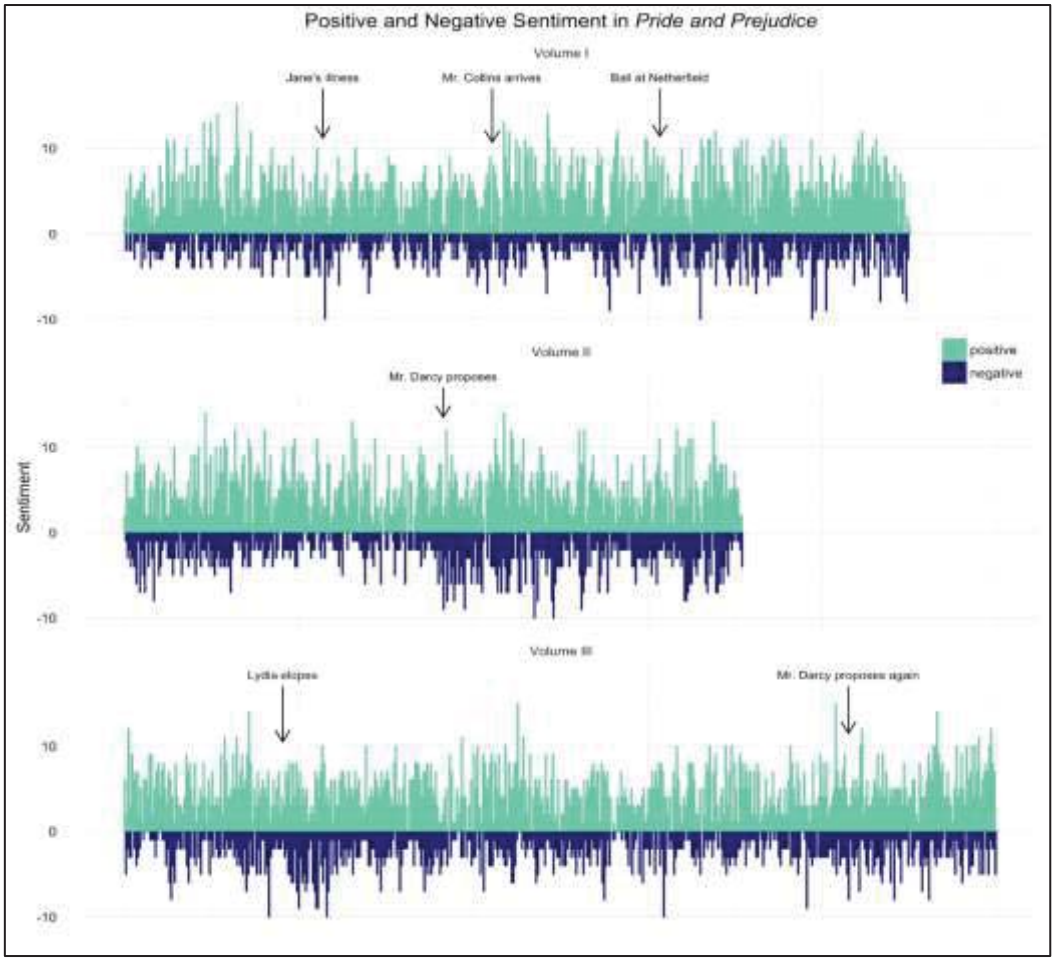


FIGURA 9: Gráfico de barras de los sentimientos positivos y negativos de *Orgullo y Prejuicio*

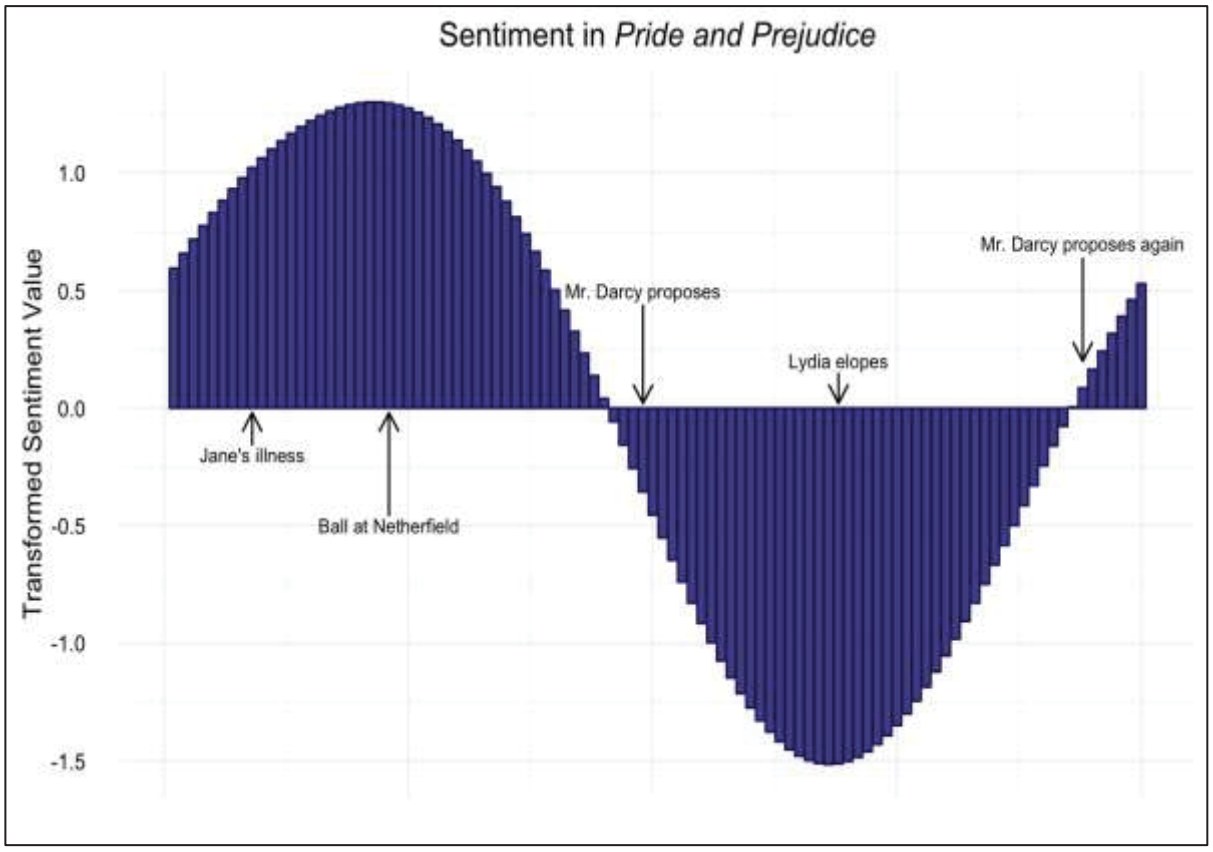


Explica Silge (2016) que, como se puede ver, las puntuaciones del sentimiento positivo son en términos generales mucho más altas que las del sentimiento negativo, y esto tiene mucho sentido si lo relacionamos con el estilo de escritura que tiene Jane Austen. Por otro lado, apunta que se puede ver un sentimiento más fuertemente negativo en dos pasajes de la novela: 1. Cuando el Sr. Darcy hace la primera propuesta y 2. Cuando Lydia se fuga con su amante.

En los gráficos anteriores se mostraron tanto los sentimientos positivos como negativos que tiene la novela, pero todavía no sabemos qué sentimiento predomina más en cada fragmento de texto y por ende en la novela. Por esto y en el quinto paso se le asignó un valor a cada fragmento, es decir, el sentimiento positivo menos el sentimiento negativo, todo esto, para tener una idea, sentido o valor general del contenido emocional del texto. Seguidamente, para visualizar los resultados y para ver mucho mejor la trayectoria global de la narrativa de la novela, se filtraron y transformaron las puntuaciones que se las ha dado a estos sentimientos ¿Cómo? Usando una Transformación de Furier de paso bajo.

La Transformación de Fourier nos muestra dónde el sentimiento narrativo es positivo o negativo; y el filtro de paso bajo nos permite ver la estructura general de estos sentimientos en la narración. Veamos, entonces este gráfico en la Figura 10.

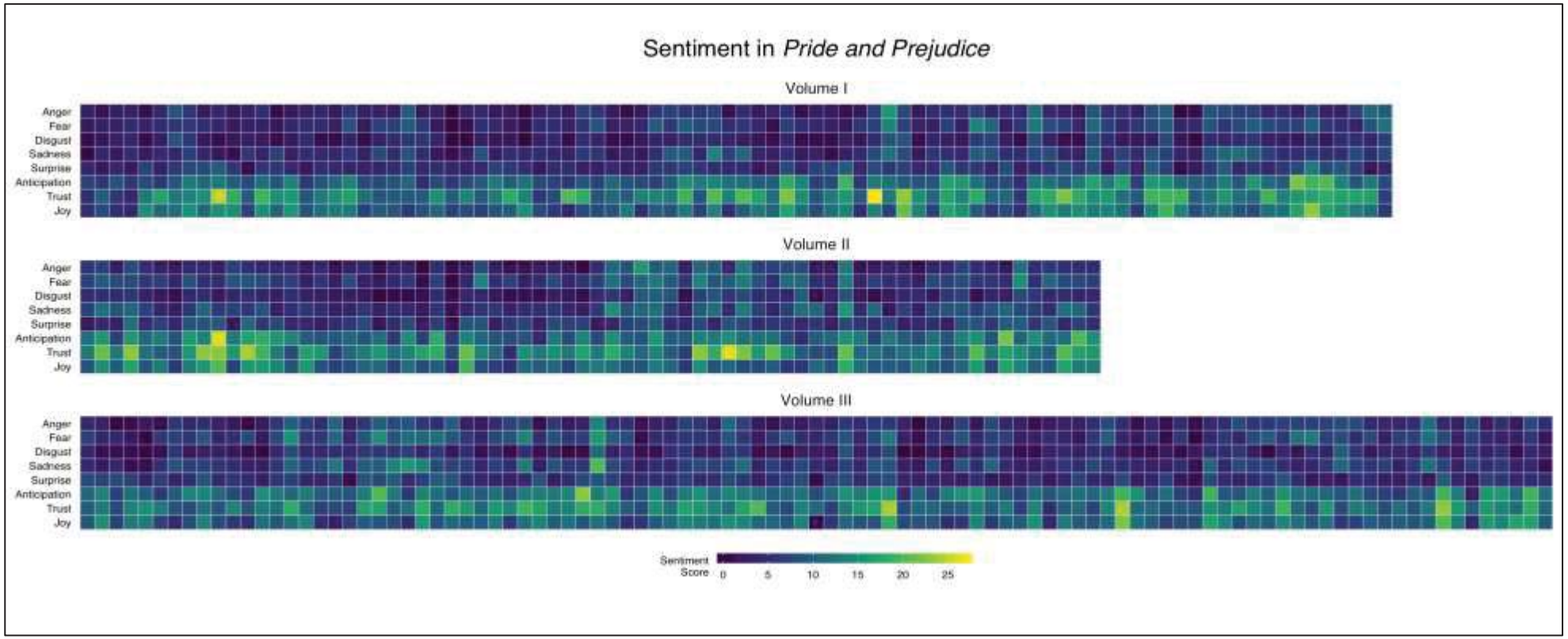
FIGURA 10: Gráfico de la Transformación Valorada de los Sentimientos en *Orgullo y*



Al respecto de este gráfico, expone Silge (2016) que las puntuaciones de los sentimientos en bruto fueron todas positivas en *Orgullo y Prejuicio*, pero las puntuaciones filtradas y transformadas de los sentimientos se han escalado y centrado para visualizar de forma general la estructura narrativa. Hay que observar que los eventos importantes se corresponden al máximo con el puntaje transformado y filtrado del sentimiento.

En el último paso que se efectuó en este trabajo, lo que se intentó es mostrarnos la valoración que hace el NRC para ocho emociones. Estas son: ira, miedo, anticipación, confianza, sorpresa, tristeza, alegría y el disgusto (*anger, fear, anticipation, trust, surprise, sadness, joy y disgust*). En consecuencia, lo que se intenta es ver las puntuaciones de estas emociones durante la novela. Los resultados, tal y como lo muestra la Figura 11., se graficaron en un mapa de calor al estilo de Bob Rudis.

FIGURA 11: Gráfico de las emociones en *Orgullo y Prejuicio*



Finalmente, y para cerrar este capítulo, Julia Silge (2016) explica que, como se puede apreciar en el mapa de calor, las emociones positivas son más fuertes que las negativas, lo cual es sensato dado el brillante estilo de escritura de Austen. Por su parte, señala que las emociones negativas son más fuertes en el medio del volumen II, cuando el Sr. Darcy hace la primera propuesta y cerca del comienzo del volumen III cuando Lydia se fuga con su amante.

IV. Abriendo caminos: Análisis de sentimientos de *Cien años de soledad* (1967) y *El amor en los tiempos del cólera* (1985)

Este capítulo se erige como el último de este trabajo, pero no por eso el menos importante, todo lo contrario. Con este capítulo vamos cerrando nuestro estudio, pero, a la vez, vamos abriendo un nuevo camino dentro del Procesamiento del Lenguaje Natural (PLN) y, propiamente, dentro del Análisis de Sentimientos, disciplina que, como hemos visto hasta aquí, es reciente y tiene frente a ella muchos desafíos que superar.

A partir de aquí, nos empezamos a sumergir en un mundo en el que casi todo es nuevo y, por lo tanto, hay más cosas por descubrir, hay más obstáculos por sobreponer y más logros por contar. Todo lo que presentemos se instituye como un avance significativo para los ordenadores, el lenguaje y la literatura, para el PLN y, especialmente, para el AS. ¿Cómo podemos llamar a este nuevo cosmos? Queremos ser valientes y llamarle (por lo menos para efectos de este estudio) como ***Análisis de Sentimientos Literarios (ASL)***. Esa es, justamente, la tarea que desarrollaremos o, dicho de forma sensata, con la que experimentaremos: analizar los sentimientos o las emociones de forma automática de dos importantes novelas contemporáneas: *Cien años de soledad* (1967*)* y *El amor en los tiempos del cólera* (1985) escritas por Gabriel García Márquez, uno de los grandes, uno de los mejores escritores de los últimos tiempos y de toda la historia de la literatura universal, uno de los máximos representantes del Realismo mágico, movimiento literario revolucionario del siglo pasado, nacido en las entrañas hispanoamericanas.

Empezaremos nuestro análisis recordando lo expuesto por Antonín Martí y Castellón (2000) quienes señalan que cualquier sistema computacional que trate el lenguaje natural, tal y como lo hacen todo sistema de Análisis de Sentimientos, ha de constar de tres componentes básicos:

a) Textos en lengua natural.
b) Datos lingüísticos: en forma de diccionarios y gramáticas.
c) Programas informáticos.

Dado que en este trabajo realizamos en sí un tratamiento o procesamiento del lenguaje natural, entonces, parece razonable que desglosemos y describamos en su plenitud

todo el sistema que estamos empleando para efectuar el AS de las novelas, a partir de los tres componentes referidos por el autor antes mencionado.

4.1. ¿Qué textos en lengua natural analizaremos?

El Análisis de Sentimientos puede ser aplicado a cualquier tipo de documento de texto (Pang y Lee, 2008), no obstante, desde los primeros experimentos que realizaron diversos investigadores, el AS se ha empleado casi de forma exclusiva al análisis de los distintos tipos de textos publicados por usuarios de los servicios Web, más concretamente medios sociales (Tsytsarau y Palpanas, 2012). Por su parte, nosotros nos ubicamos dentro de ese *«casi»*, puesto que nuestros textos en lengua natural son ***obras literarias***. Desde nuestra percepción, creemos que los impulsadores de esta área no se imaginaron que algún día se iba a realizar AS a textos literarios (o al menos no eran esas las intenciones). Sin embargo, al ser esta disciplina tan interesante y, además, al ser los textos literarios una fuente en la que, constantemente, los escritores plasman ideas, pero, sobre todo sentimientos y emociones mediante las historias que nos van relatando, entonces, ¿por qué no experimentar? ¿por qué no realizar AS a obras literarias y, exclusivamente a novelas? y ver cuáles son los resultados de este experimento.

Tal y como lo expusimos en el comienzo de este capítulo, las dos novelas que seleccionamos son: *Cien años de soledad* (1967*)* y *El amor en los tiempos del cólera* (1985) del escritor latinoamericano Gabriel García Márquez. El idioma original en las que estas obras están escritas es, naturalmente, el español. No obstante, dado que, tanto el modelo y, específicamente, el diccionario o lexicón de sentimientos que estamos utilizando (o al menos, al que tenemos acceso) lo conforman un corpus de palabras en inglés[18], por tal razón, los textos que son analizados en este estudio son las versiones traducidas al inglés de las tres novelas originales en español. Un punto a señalar es que hemos revisado las traducciones para verificar que sean de calidad y que reflejen en su máximo nivel el lenguaje de las novelas originales de Márquez.

Nuestro sueño es realizar el AS con textos en español, cometido que no está lejos de cumplirse, porque cada vez más son los investigadores que crean corpus y bases de datos en

[18] Nos referíamos en el segundo capítulo de este trabajo que la mayoría de herramientas accesibles para efectuar AS están concebidas para el inglés.

español. Pero, no todo está perdido, aunque aquí nos valgamos de textos en inglés, lo importante es que podremos comprobar si es viable realizar AS con textos literarios.

A continuación, para aquellos que desconocen estas tres importantes novelas, señalaremos sus datos más interesantes, sus fechas de publicación, sus argumentos y sus estructuras externas. Asimismo, nos referiremos de forma sucinta a su autor.

4.1.1. Gabriel García Márquez

Los que somos amantes de la literatura y la estudiamos, seguramente, no desconocemos este nombre, todo lo contrario. Hoy por hoy, los apellidos: *García Márquez,* tanto para apasionados como para los estudiosos de las letras hispánicas y universales nos sonarán de algún viento. No es necesario ser lector ni académico de las letras hispanoamericanas contemporáneas para conocer estos apellidos, basta con leer o estudiar literatura, como tal, para saber que García Márquez es uno de los más ilustres escritores de los últimos tiempos. Por nuestra parte, por ejemplo, al leer los dos apellidos nos acordamos de algunos de sus mejores cuentos y peculiares novelas, los asociamos con el *realismo mágico*, uno de los movimientos literarios más revolucionarios de la mitad del siglo pasado y los relacionamos con el fenómeno literario denominado *Boom latinoamericano,* que hizo que las letras hispanoamericanas se extendieran por todo el mundo. La trascendencia de estos apellidos es tanta que, nosotros creemos, que hasta los que no tienen un interés obstinado por la literatura, seguro, los habrán escuchado de algún sitio.

Gabriel José de la Concordia García Márquez es el nombre completo de uno de los grandes mitos y escritores de la literatura universal y, desde luego, de las letras hispánicas. García Márquez es el autor genuino de los textos que aquí analizaremos. Su nacimiento se remonta a una fecha: 6 de marzo de 1927 y a un municipio colombiano: Aracataca; asimismo, (aunque más tristemente), aparece la fecha de su fallecimiento: 17 de abril de 2014 en la Ciudad de México.

Gabriel García Márquez fue un excelso escritor, periodista, guionista y editor. En el ámbito literario fue apodado por su familia, amigos y colegas como *Gabito o Gabo.* Tanto su hipocorístico como su apocope han sido tan reconocidos, entre otras razones, gracias al enorme legado literario que nos heredó, legado que se traduce en un abanico extraordinario de novelas, cuentos, guiones de cine, reportajes novelados, discursos, entrevistas, trabajos periodísticos,

entre otros. No por nada, en el año de 1982 recibió el Premio Nobel de Literatura, máximo laurel al que puede aspirar un escritor en el mundo académico-literario. Por otro lado, y adicionalmente a esta presea, García Márquez se ganó (y todavía lo sigue haciendo) la afición de millones de lectores por todo el mundo (quizá, premio de mayor valor para un escritor).

Hablar de García Márquez es hablar forzosamente tanto de *Realismo mágico* como de *Boom latinoamericano.* Aunque no vamos a desarrollar una explicación detallada y exhaustiva de ambas cuestiones, diremos en qué consisten para tener una noción básica al respecto.

Por una parte, el *Realismo mágico*, tal y como lo habíamos expuesto anteriormente, es una corriente o movimiento literario que nació a mediados del siglo XX en Hispanoamérica. Su principal característica es que la realidad de los hechos narrativos que se nos relatan se ven asaltados de forma sigilosa, discreta y natural por acciones, ciertamente, irreales, extrañas, fantásticas o mágicas a tal punto que las percibimos como comunes, normales, cotidianas y reales dentro de la narrativa. Por otro lado, el llamado *Boom latinoamericano* fue un fenómeno literario o, más propiamente dicho, editorial y comercial que surgió entre los años 1960 y 1970 del siglo pasado, y cuya principal labor fue distribuir ampliamente por Europa y por todo el mundo un conjunto de trabajos (obras literarias) escritos por extraordinarios novelistas hispanoamericanos. Algunos de estos escritores más relevantes de este grupo son: Gabriel García Márquez (lógicamente), Mario Vargas Llosa, Carlos Fuentes y Julio Cortázar.

Aunque podríamos seguir tratando una infinidad de aspectos interesantes sobre el autor y su obra no es objetivo de este trabajo hacerlo, por tal motivo, a continuación, nos referimos puntualmente a las tres novelas que analizaremos y las cuales se constituyen como los textos en lengua natural que procesara nuestro sistema de Análisis de Sentimientos.

4.1.2. *Cien años de soledad* (1967)

«He leído el 'Quijote' americano» fueron las palabras que le escribió Carlos Fuentes, –escritor representativo del *Boom* y *del Realismo mágico*– a Julio Cortázar en una carta que le dirigió tras leer el manuscrito de la novela que pronto se convirtió en obra cumbre del *Realismo mágico*, es decir, *Cien años de soledad.* Esas cinco palabras, bastan y sobran, para tomar conciencia de la magnitud del texto literario que analizaremos.

Cien años de soledad es considerada la obra maestra del escritor colombiano y la novela más representativa del *Realismo mágico*. Asimismo, es considerada una de las obras más renombradas de la literatura hispanoamericana y universal, como también, uno de los textos literarios más traducidos y leídos por todo el mundo. Este año, casualmente, se está celebrando el cincuenta aniversario de la publicación de su primera edición, la cual fue publicada un 5 de junio de 1967, en la ciudad de Buenos Aires, por la editorial Sudamericana.

Antes de la aparición de esta novela, el escritor colombiano ya había elaborado algunas otras obras literarias más breves como *La hojarasca* (1955), *El coronel no tiene quien le escriba* (1961) y *La mala hora* (1962), además de un volumen de cuentos, *Los funerales de la Mamá Grande* (1962), los cuales nos iban sumergiendo en el mundo trágico que caracteriza a Macondo, mundo ficticio que florece plenamente en *Cien años.…*

Si bien es cierto, hasta la publicación de todas estas obras, Márquez era un conocido escritor en el ámbito hispanoamericano, fue con *Cien años de soledad* cuando se consagró en todo el mundo y para la inmortalidad.

En esta novela parece que todo se fusionó de forma perfecta, tanto la historia como la forma en la que se nos relata la misma son extraordinarias y quizá sea esta unión una de las razones por las que ha trascendido. Al respecto, para nuestro análisis es fundamental conocer el fondo del texto, es decir, la historia que se no está contando porque, desde nuestra percepción, un discurso narrativo variará (en lo relacionado al plano lingüístico y, propiamente, al léxico) dependiendo las historias que se nos narren (aunque esto no sea una ley). Por tal motivo, a continuación, exponemos brevemente cuál es el argumento de la novela.

Cien años de soledad relata de forma fabulosa la historia de una aldea imaginaria, Macondo, y de la estirpe de sus fundadores, la familia Buendía. Los Buendía son siete generaciones que viven en Macondo, lugar de donde nunca salen, hasta que un ciclón bíblico arranca los cimientos de su casona y con esto no dejar rastro de toda la estirpe.

Es así, como la historia nos empieza contando cómo José Arcadio Buendía –el patriarca de la estirpe– se casó con Úrsula Iguarán, quien verdaderamente es su prima. Esta relación, le genera un intenso miedo a ella, pues, Úrsula cree que cuando dos familiares se unen tanto amorosa, pero, sobre todo sexualmente, sus hijos pueden nacer con cola de cerdo. Tanto es el temor que le genera esto a Úrsula que se niega a cumplir sus obligaciones de mujer y priva a José Arcadio de consumar su matrimonio (tener relaciones sexuales). Si bien, esto no le agrada a José Arcadio, la gota que derramó el vaso fue el insulto que le dirigió Prudencio Aguilar,

luego de que José Arcadio le ganara una pelea de gallos. El perdedor le comentó que quizás su gallo podría ser más eficaz con su mujer Úrsula. La vergüenza, la humillación, la ofensa y la ira que le produjo este insulto a José Arcadio fue tanto que termina matando en honor a Prudencio Aguilar, y ese mismo día consuma su matrimonio con Úrsula.

José Arcadio creía haber acabado con Prudencio, pero esto no fue así, porque a los días de su muerte se le aparecía como un alma en pena, como un fantasma con un rostro lleno de tristeza y soledad. A José Arcadio esta situación le causó un gran temor a tal punto que mata a todos sus gallos y decide partir con su mujer a un nuevo lugar para vivir. Este viaje no lo realizaron solos, sino de la mano de otros amigos, quienes los acompañaron en toda la travesía durante varios meses de difícil camino por la ciénaga. En el transcurso del viaje, nació el hijo mayor de José Arcadio y Úrsula. A este lo llamaron como al padre, o sea: José Arcadio.

Una tarde, mientras José Arcadio estaba descansando a la orilla de un río, estaba soñando con un pueblo lleno de espejos y que se llamaba Macondo, al despertar y, motivado por el sueño, decide terminar el viaje y fundar Macondo sobre ese sitio. Naturalmente, fue él quien asumió la dirección, la organización y la fundación del nuevo pueblo. Luego de haberse establecido y con José Arcadio y Aureliano (hijos mayores de Úrsula y José Arcadio Buendía), el patriarca de la familia empezó a mostrar interés por la invención de artefactos. Un día llegaron los gitanos a Macondo y junto con ellos venía Melquíades un gitano –que tiene la habilidad de ir y regresar de la muerte–, quien se convierte en amigo de José Arcadio y quien escribe una serie de pergaminos en los que se relata la historia de los Buendía, pero que se revela hasta cien años después. Tanta es la obsesión que tiene José Arcadio con algunos temas y objetos que acaba atado a un patio, pues, creían que la locura lo había capturado.

¿Qué ocurre con Úrsula? Ella, asume las responsabilidades de la casa por más de cien años, además de cuidar a sus tres hijos bilógicos: José Arcadio, Aureliano y Amaranta, como también a Rebeca, una hija adoptiva que come tierra y que trajo la enfermedad del insomnio a Macondo. Los cuatro hijos conforman la segunda generación de los Buendía y gran parte de la historia se desarrolla gracias a los hechos que le suscitan a cada uno. José Arcadio (el hijo mayor) al enterarse que está embaraza Pilar Ternera (mujer con la que tuvo relaciones sexuales), huye del pueblo con los gitanos. Amaranta y Rebeca entran en una disputa por el amor de Pietro Crespi, rivalidad que se acaba cuando Rebeca rechaza a Pietro por José Arcadio (el hijo mayor) quien regresa a Macondo, como todo un semental y quien muere de forma misteriosa. Por su parte, Amaranta elige a la soltería como su forma de vida. Finalmente, Aureliano –quien

siempre fue silencioso y taciturno- termina convirtiéndose en el coronel Aureliano Buendía, y comandando el Ejército de resistencia de la Guerra Civil.

En esta misma línea narrativa, se continúa relatando la historia de las siguientes cinco generaciones. La tercera generación nace del vientre de Pilar Ternera, quien procrea dos hijos junto a los dos hermanos mayores de los Buendía. Junto a José Arcadio tiene a Arcadio y con Aureliano tiene a Aureliano José. La cuarta generación de los Buendía es engendrada por Arcadio, quien junto a Santa Sofía de la Piedad tiene cuatro hijos: Remedios, José Arcadio Segundo y Aureliano Segundo. La quinta generación es reproducida por Aureliano Segundo, quien junto a Fernanda del Carpio tiene tres hijos: Amaranta Úrsula, José Arcadio y Renata Remedios. La penúltima generación es procreada por Renata Remedios en conjunto con Mauricio Babilonia, quienes solamente tienen un hijo: Aureliano Babilonia, personaje que finalmente, sin que estuviese viva Úrsula Iguarán para impedirlo, acaba concibiendo con su tía, Amaranta Úrsula (quien muere desangrada luego del parto), al último Aureliano, quien nace con cola de cerdo, y que, como profetizaron los manuscritos de Melquíades, acaba siendo arrastrado por todas las hormigas del mundo.

Finalmente, a Aureliano Babilonia esta imagen del bebé, le hace recordar los pergaminos que escribió Melquíades y que él estaba intentado descifrar. Al darse cuenta que esa escena se encontraba en los manuscritos, se dispone a descifrar toda la historia que ahí estaba descrita con anticipación en medio de unos vientos huracanados que estaban asediando al sitio en donde se hallaba y, desde luego, a Macondo. ¿Qué encontró Aureliano al terminar de leerlos? Pues, se dio cuenta que, al terminar la lectura, también, finalizaría su propia historia y con él, la historia de Macondo, aldea que sería arrasada por el viento y borrada de cualquier memoria humana... «porque las estirpes condenadas a cien años de soledad no tenían una segunda oportunidad sobre la tierra».

La historia de la familia Buendía se nos relata, de forma externa, en una estructura lineal de veinte capítulos no enumerados. En esta estructura es usual encontrarnos con anacronías tanto de analepsis como de prolepsis. No sabemos si fue a causa del realismo mágico o qué, pero cuando seleccionamos este texto, lo hicimos sin ningún criterio en específico, más que por la trascendencia, la calidad, la autenticidad y, sobre todo, el gusto por la novela de Márquez. En fin, lo cierto es que sea cual sea el motivo analizar los sentimientos de forma automatizada de uno de los textos clásicos de la literatura, siempre será sugestivo para las letras hispánicas y universales.

4.1.3. *El amor en los tiempos del cólera (1985)*

Concluimos la presentación del conjunto de textos que analizaremos con la novela que fue publicada por García Márquez después que en 1982 le otorgaran el Premio Nobel. *El amor en los tiempos del cólera* es una novela mucho más extensa que *Crónica*…, y casi de la misma extensión de *Cien años.*.. fue publicada en 1985 por la editorial Oveja Negra en Colombia y por la Sudamericana en Argentina.

Las novelas de Márquez, aunque tienen muchas similitudes, son como una caja de Pandora: nunca sabemos que habrá luego de que empezamos a leer sus primeras páginas… este es el caso de *El amor en los tiempos del cólera*, aunque aparentemente, el título nos anuncia el tema principal de la historia, a medida que avanzamos en la lectura, otros tópicos van apareciendo y se van enredando perfecta, pero a la vez, ambiguamente con el eje temático principal: el amor (¿o la obsesión?).

La historia que se nos relata en *El amor en los tiempos del cólera* es la historia que Márquez siempre quiso narrar y fue pensada por el escritor colombiano para consagrar el amor. Básicamente, se nos cuenta la historia de amor –aparente– entre dos jóvenes: Fermina Daza y Florentino Ariza, relación que fue interrumpida de forma sorprendente por Daza, al regresar de un viaje que le obligó hacer su padre para que tratara de olvidar a Florentino. Este amor contrariado tuvo que esperar cincuenta y un años, nueve meses y cuatro días para que se pudiera consumar.

Se inicia relatando la muerte de dos personajes: la de Jeremiah de Saint-Amour, un refugiado antillano inválido de guerra, y la de Juvenal Urbino –médico, hombre rico, de buena familia y quien logra casarse, luego de varios rechazos, con Fermina Daza– este, al regresar de hacer las labores de médico forense de la casa de su amigo suicida, en su intento de recuperar a un loro que tenía en su casa y que se le quería escapar, se cae desde lo alto de una escalera y se mata. ¡Vaya, comienzo!

Ambos sucesos trágicos ocurrieron un domingo de Pentecostés de principios de la década de los años treinta, en un pueblecito portuario colombiano del litoral Caribe que se sitúa cerca del río Magdalena.

La repentina muerte del doctor Juvenal Urbino fue el momento que tanto había esperado Florentino Ariza dado que, aunque su carrera en los negocios había florecido, y aunque había sostenido 622 pequeños romances, su corazón todavía pertenecía a Fermina. Por ello, cuando

Urbino muere, Florentino acude al funeral para volverle a declarar su amor a Fermina Daza, petición que nuevamente fue rechazada por Daza. Ante esto, Florentino, así como lo hizo en la adolescencia, vuelve a escribirle cartas e intenta conquistarla, tanta es su obsesión que Fermina acaba aceptando primero su amistad y luego hacer un viaje por el río Magdalena en uno de los barcos de la compañía de Ariza, sin saber, hasta el último momento, que Florentino la acompañará.

Es en el río Magdalena donde estos viejos, que ya pasan de los setenta, se entregan a su amor, con tanto apasionamiento que, para librarse de testigos y permanecer a solas en el barco, Florentino planea y hace que en el viaje de vuelta se ice la bandera amarilla que anuncia que el barco está infectado por el cólera, y, una vez llegados a la desembocadura, y por lo tanto a la ciudad, vuelva a remontar el río para permanecer juntos.

Mientras se nos va relatando esta historia de amor contrariado y otros lances más, también se nos narra de forma colateral cómo el cólera hace estragos y se suceden las guerras entre liberales y conservadores, sin que por ello se resienta demasiado la vida de la ciudad caribeña.

La estructura externa del texto está organizada en seis capítulos no enumerados o seis secuencias narrativas. Por otro lado, estos seis capítulos se distribuyen en las 490 páginas que tiene la novela.

Hasta aquí, hemos hecho un repaso descriptivo de los tres textos que utilizaremos para realizar nuestro AS, pero desde un punto de vista literario. Sin embargo, también tenemos que explicar las características estructurales de los textos genuinos que nuestro sistema procesará. Hay que recordar que los textos son traducciones al inglés de las versiones originales en español, por lo tanto, en a continuación nos referimos a tales características.

4.2. Formato y extensión de los textos en lengua natural que analizaremos

Para que nuestro sistema de AS realice el respectivo procesamiento, los textos deben estar como archivos de *texto simple*, *texto sencillo*, *texto sin formato* o *texto plano*. Estos son archivos informáticos que contiene únicamente texto formado solo por caracteres que pueden ser leídos por los humanos, por tal razón, no tienen elementos o tipos de formatos tipográficos. En la figura 12., para ilustrar mostramos una imagen de cómo luce uno de los textos en este tipo de formato.

FIGURA 12: Ejemplo de texto plano del *Amor en los tiempos del cólera*

```
Love in the Time of Cholera

CHAPTER ONE

IT WAS INEVITABLE: the scent of bitter almonds always reminded him of
the fate of unrequited love. Dr. Juvenal Urbino noticed it as soon as he
entered the still darkened house where he had hurried on an urgent call
to attend a case that for him had lost all urgency many years before.
The Antillean refugee Jeremiah de Saint-Amour, disabled war veteran,
photographer of children, and his most sympathetic opponent in chess,
had escaped the torments of memory with the aromatic fumes of gold
cyanide.
```

Por otro lado, en lo que respecta a la extensión de los textos, en la Tabla 1., se detalla la cantidad de líneas que tiene cada uno de estos.

TABLA 1: Extensión (en líneas) de los textos

N.º	Novela (texto)	Extensión (en líneas)
1	*Cien años de soledad* (1967)	12,835
3	*El amor en los tiempos del cólera* (1985)	13, 005

4.3. ¿Qué gramáticas o diccionarios nos indicarán cómo analizar los sentimientos de las novelas?

Exponíamos en el capítulo II de este trabajo que uno de los métodos o enfoques para realizar Análisis de Sentimientos es empleando un diccionario o lexicón de sentimientos en el que se almacenan una lista de palabras a las que les ha asignado previamente una orientación semántica: negativa o positiva. Por nuestra parte, haremos uso de uno de los muchos diccionarios o lexicones de sentimientos que se han creado.

En lo que respecta al AS del lenguaje figurado, el *NRC Word-Emotion Association Lexicon (EmoLex)*, ha demostrado dar buenos resultados, por tal razón, hemos decido utilizarlo en este trabajo. Por tal motivo, consideramos acertado detallar ciertas particularidades que tiene el diccionario NRC, no sin antes, ofrecer una definición de lo qué es un diccionario o lexicón de sentimientos.

4.3.1. ¿Qué es un diccionario o un lexicón de sentimientos?

Un diccionario o lexicón de sentimientos se puede concebir como una base datos que almacena una serie de entradas, que pueden ser términos o conceptos. Estas tienen una propiedad y es que han sido etiquetadas previamente con algún tipo de información emocional, sentimental o afectiva. Según, Albornoz Cuadrado (2011) estos diccionarios o lexicones se pueden clasificar en dos grandes grupos:

a) Lexicones o diccionarios de sentimientos basados en términos o palabras.
b) Lexicones o diccionarios de sentimientos basados en conceptos o significados.

Los diccionarios o lexicones de sentimientos basados en términos «tienen como unidad primitiva el término o palabra, el cual es etiquetado con cierta información emocional», mientras que los diccionarios o lexicones de sentimientos basados en conceptos «utilizan el significado de las palabras como unidad primitiva, siendo el concepto el que es etiquetado emocionalmente» (Albornoz Cuadrado, 2011).

Igualmente, tanto unos como los otros pueden adquirir dos orientaciones: la primera orientación, etiqueta las entradas con categorías emocionales que representan las emociones básicas, tales como: *sadness* (tristeza), *fear* (miedo), *anger (*ira o enfado), etc. La segunda, etiqueta las entradas en función de diferentes intensidades emocionales, como, por ejemplo: la *polaridad* (negativa, positiva o neutra) o la *activación* que tiene cada una.

El diccionario *NRC Word-Emotion Association Lexicon* (*EmoLex*), pertenece a los diccionarios de sentimientos basados en términos o palabras y manifiesta una orientación mixta dado que ha anotado las entradas tanto por la intensidad emocional, específicamente, la polaridad que tienen (negativa o positiva), como por categorías emocionales.

4.3.2. Diccionario NRC Word-Emotion Association Lexicon (*EmoLex*)

El *NRC Word-Emotion Association Lexicon (EmoLex)*[19] es un diccionario o lexicón de sentimientos que fue creado de forma colaborativa por Saif Mohammad y Peter Turney. Este lexicón contiene una lista de palabras en inglés con sus correspondientes asociaciones con las ocho emociones básicas de Plutchik (Plutchik y Kellerman, 1980). Estas son: ira, miedo,

[19] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

anticipación, confianza, sorpresa, tristeza, alegría y el disgusto (*anger, fear, anticipation, trust, surprise, sadness, joy y disgust*) y dos sentimientos: positivo y negativo (*negative y positive*).

¿Cuál fue el proceso que se siguió para asignar tanto las emociones como la polaridad? Pues, básicamente mediante un proceso de etiquetación o anotación manual vía *crowdsourcing*, cuya traducción en español sería *colaboración abierta distribuida* o *externalización abierta de tareas.* ¿En qué consistió esta anotación manual? Se contrató a un grupo de personas por convocatoria abierta, se les dio un corpus de textos y su tarea consistió en leer todos estos textos, identificar aquellas palabras que expresan información subjetiva, afectiva, emocional o sentimental y luego anotar, primero, qué orientación semántica tiene cada término, es decir, la polaridad (negativa o positiva), y, segundo, determinar qué palabras expresan una emoción y etiquetar qué tipo de emoción expresan.

Si la palabra pertenece a la categoría se indica con un 1, en caso contrario con un 0. En este recurso podemos encontrar 14,182 palabras etiquetadas. En su trabajo, Escortell Pérez y Paolo Rosso (2017) nos muestran cómo se codifica la información en el *NRC.* La Tabla 2., es la misma tabla que presentan las autoras:

TABLA 2: *EmoLex:* representación de la palabra *dark* (oscuro)

Palabra	Categoría	Asociación
dark	Anger	0
dark	anticipation	0
dark	disgust	0
dark	fear	0
dark	joy	0
dark	negative	0
dark	positive	0
dark	sadness	1
dark	surprise	0
dark	trust	0

Como hemos visto las anotaciones afectivas o sentimentales que se han hecho en el *NRC Word-Emotion Association Lexicon (EmoLex)* han sido para palabras en inglés. No obstante, el trabajo ha ido un poco más allá y sus autores vía el *Traductor de Google* ofrecen versiones del léxico en más de veinte idiomas. La Tabla 3., nos ilustra el número de entradas que contiene el diccionario *NRC* según el idioma.[20]

[20] Recientemente, hemos obtenido la versión en español, sin embargo, por cuestiones de tiempo no hemos podido emplearla. Pero, no tenemos la menor duda que haremos el experimento en español en próximos trabajos.

TABLA 3: Número de entradas que tiene *EmoLex*, según el idioma

Idioma	N.º de entradas	Idioma	N.º de entradas
Inglés	14, 182	Hebrero	7, 828
Ruso	14, 182	Griego	7, 198
Japonés	14, 182	Somalí	7, 031
Español	14, 182	Urdu	6, 035
Chino (simple)	14, 182	Latín	5, 871
Portugués	14, 182	Finés	5, 785
Francés	14, 182	Tamil	5, 488
Persa	13, 618	Gujarati	5, 385
Chino (tradicional)	13, 037	Euskera	5, 344
Vietnamita	12, 351	Sueco	5, 266
Alemán	11, 812	Suajili	5, 230
Italiano	11, 114	Telugu	4, 782
Turco	9, 725	Danés	4, 671
Bengalí	9, 453	Catalán	4, 617
Ucraniano	8, 903	Marathi	4, 476
Rumano	8, 581	Irlandés	4, 460
Tailandés	8, 562	Galés	4, 214
Hindi	8, 116	Esperanto	4, 208
Neerlandés	7, 850	Zulu	4, 174

4.4. ¿Qué herramientas o programas informáticos empleamos para efectuar el AS de las novelas?

Finalmente, luego que sabemos qué textos analizaremos, como también, cuáles son los datos lingüísticos (diccionarios o gramáticas) que le dirán al sistema cómo procesar el material, solamente nos falta saber qué programa o herramienta informática será la responsable de llevar a cabo el procesamiento de esos textos de acuerdo con esa información lingüística.

En nuestro caso, los programas o, dicho de forma puntual, los paquetes y el entorno de programación que emplearemos para llevar a cabo nuestro AS son los que utilizó Silge (2016) en el trabajo que nos impulsó a realizar el presente estudio y que, desde luego, nos sirvió de modelo. Por una parte, el entorno o lenguaje de programación que emplearemos es *R* y, por otro lado, las principales librerías o paquetes informáticos que usaremos son: *syuzhet*, *readr*, *stringr*, *dplyr*, *reshape2*, *ggplot2* y *ggthemes.* Un informático, seguramente, sabrá qué es *R* o qué es un librería o paquete informático y con un poco más de suerte conocerá alguno de los que hemos hecho referencia, no obstante, a un filólogo le sonará mucho menos. Por tal motivo, a continuación, explicaremos cada una de estas cuestiones.

Queremos hacer un paréntesis y señalar que la curiosidad, las habilidades y los conocimientos que adquirimos en la clase de *Tecnologías de la información aplicadas a la docencia e investigación en lengua española* sobre el entorno de programación *R*, también nos hicieron atrevernos a experimentar con este trabajo.

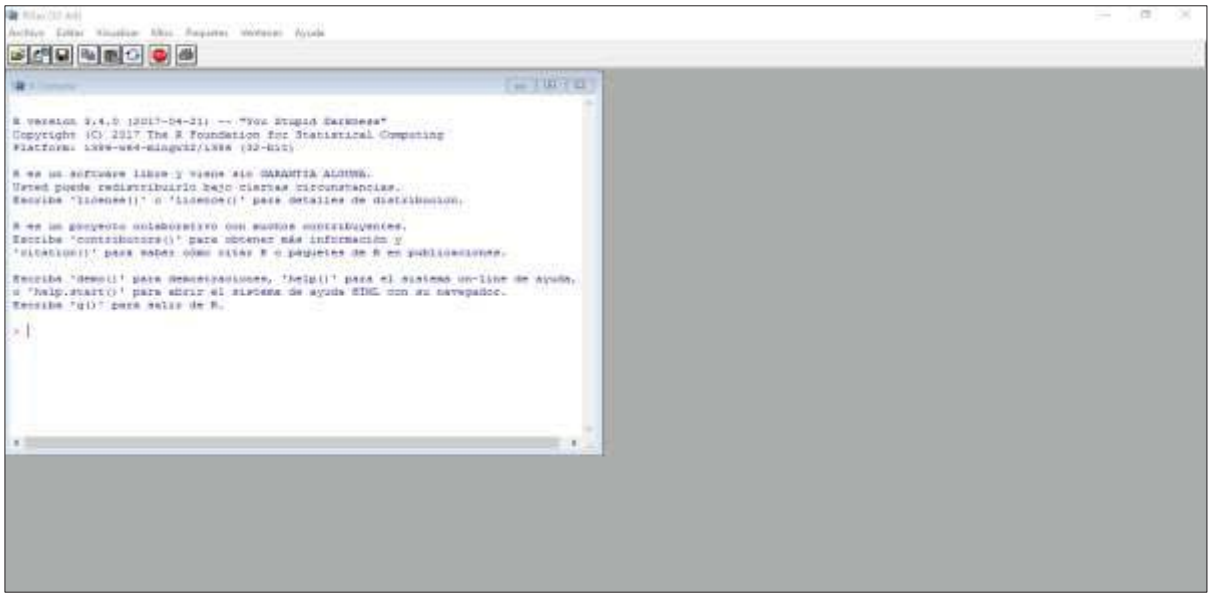
4.4.1. ¿Qué es *R*[21]*?* ¿Qué es *RStudio?*

Iniciaremos acotando que lenguajes de programación hay muchos. Algunos de los más usados en la actualidad son, por ejemplo: *Python, Java, Perl, PHP*. Algunos de estos se pueden descargar de forma gratuita en nuestros ordenadores. ¿Saber utilizar un lenguaje de programación le puede abrir puertas a un filólogo? Sin lugar a dudas.

En nuestro caso y como hemos dicho, emplearemos *R* un lenguaje y entorno de programación un lenguaje de uso extendido en los análisis estadísticos y ha demostrado que es muy útil en los estudios filológicos, de hecho, el presente trabajo es un ejemplo de ello.

R es un conjunto integrado de programas para manipulación de datos, cálculo y gráficos. En la Figura 13., ilustramos cómo luce la consola de trabajo de *R.*

FIGURA 13: Consola de trabajo de *R*



Entre otras características, *R* dispone de:

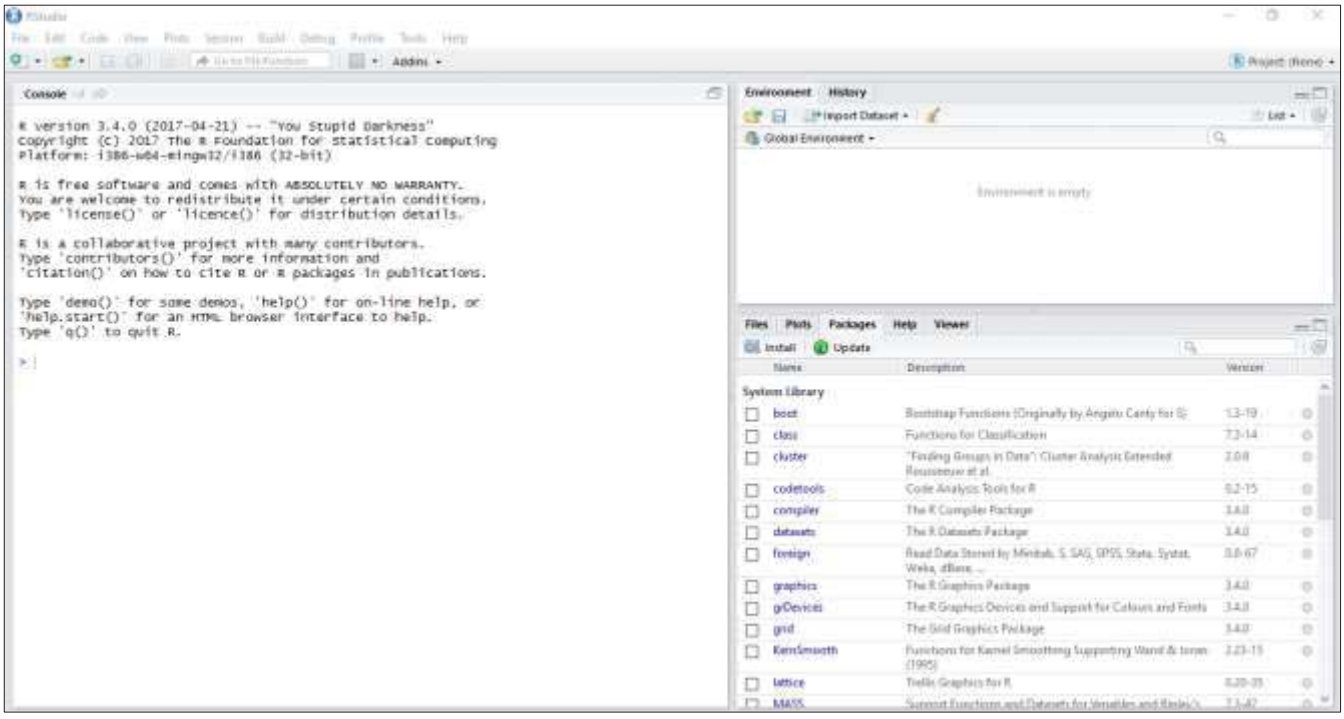
- almacenamiento y manipulación efectiva de datos,
- operadores para cálculo sobre variables indexadas (*Arrays*), en particular matrices,
- una amplia, coherente e integrada colección de herramientas para análisis de datos,

[21] https://cran.r-project.org/

- posibilidades gráficas para análisis de datos, que funcionan directamente sobre pantalla o impresora, y
- un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas y salidas. (Debe destacarse que muchas de las funciones suministradas con el sistema están escritas en el lenguaje R) (González y González, 2000)

Para hacer mucho más sencillo el manejo de *R* es recomendable usar un programa que se llama *RStudio*[22], por nuestra parte es el programa que estamos utilizando para efectuar el AS. *RStudio* es un entorno de desarrollo integrado (IDE) para R. Es software libre con licencia GPLv3 y se puede ejecutar sobre distintas plataformas (Windows, Mac, o Linux) o incluso desde la web usando *RStudio Server*. En la Figura 14., igualmente mostramos cómo luce el entorno de trabajo de *RStudio.*

FIGURA 14: Entorno de trabajo de *RStudio*



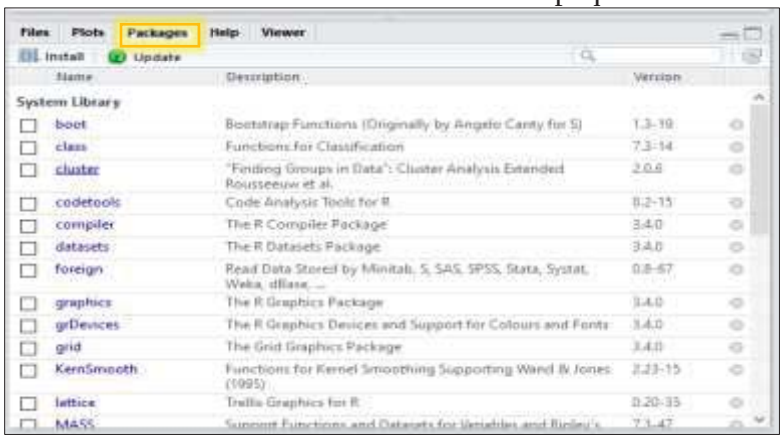
[22] https://www.rstudio.com/products/rstudio/download/

4.4.2. ¿Qué es una librería, una biblioteca o un paquete?

Una de las grandes ventajas y, quizá, la característica más potente que tiene *R* es que, además, de las múltiples funciones que trae consigo, cualquier usuario puede crear, incorporar y usar nuevas funciones que sean capaces de realizar tareas específicas. Estas múltiples funciones se integran en lo que se denomina como *librería, biblioteca o paquete.*

«Un paquete (*package*) es una colección de funciones, datos y código R que se almacenan en una carpeta conforme a una estructura bien definida, fácilmente accesible para *R*» (Santana y Nieves Hernández, 2016). En la Figura 15., mostramos la sección del entorno de trabajo de *RStudio,* donde se encuentran los paquetes y donde los podemos instalar.

FIGURA 15: Sección de ubicación de los paquetes en *RStudio*



4.4.3. ¿Qué paquetes o librerías utilizamos para realizar nuestro AS?

Silge (2016) en su trabajo experimental ha empleado una serie de paquetes o librerías que le permitieran llevar a cabo de forma efectiva y ordenada las tareas que se planteó y, sobre todo, para alcanzar el principal objetivo que era analizar los sentimientos de forma automatizada de las novelas de Jane Austen. En nuestro caso, como lo habíamos señalado anteriormente, hemos seguido fielmente el modelo de trabajo de la autora, antes mencionada, esto incluye los diversos paquetes.

Los paquetes o librerías que emplearemos son los siguientes: *syuzhet*, *readr*, *stringr*, *dplyr*, *reshape2*, *ggplot2* y *ggthemes*, como ya lo habíamos expuesto en apartados anteriores. Sin embargo, no hemos explicado qué funciones puede llevar a cabo cada uno de estos.

Por tal razón, a continuación, nos referiremos de forma sucinta sobre cada uno, pero, haremos un especial énfasis en *syuzhet,* puesto que este es el que realiza la principal tarea, es decir, analizar los sentimientos. En la Tabla 4., se ha sintetizado toda la información referente a cada uno de estos.

TABLA 3: Paquetes utilizados para realizar el AS

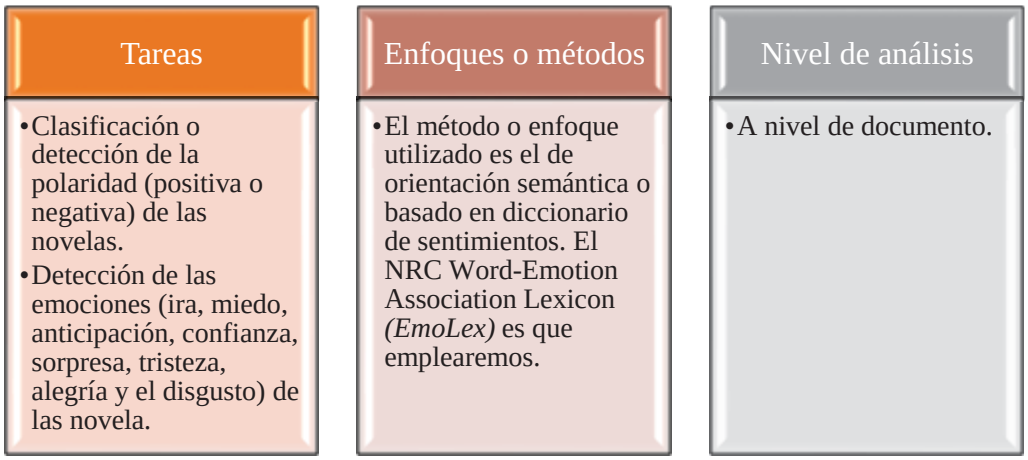
Paquete	Creador/es	Funciones
syuzhet	Matthew Jockers	Las principales funciones que se pueden realizar con este extraordinario paquete desarrollado por Jockers, miembro del grupo de PLN de Stanford son la extracción de sentimientos y de arcos narrativos derivados de los sentimientos del texto. La manera de efectuar ambas funciones es mediante cuatro diccionarios o lexicones de sentimientos que trae consigo. Estos son: – *syuzhet:* que es el predeterminado y que es desarrollado por el Laboratorio Literario de Nebraska. – *afinn:* desarrollado por Finn Årup Nielsen. – *bing:* desarrollado por Minqing Hu y Bing Liu. – *nrc:* desarrollado por Saif Mohammad y Peter Turney. El paquete también proporciona un *hack*[23], para implementar el analizador de sentimientos *coreNLP* de Stanford. El paquete proporciona varios métodos para la normalización del arco narrativo.
readr	Hadley Wickham, Jim Hester y Romain Francois	El objetivo de *readr* es proporcionar una forma rápida y amigable de leer una tabla cuadrada o una rectangular de datos (como csv, tsv y fwf).
stringr	Hadley Wickham	Puesto que las funciones básicas de *R* para tratar cadenas pueden resultar ciertamente incomodas. Por ello, con este paquete se podrá realizar de manera más fácil las operaciones con cadenas de textos.
dplyr	Hadley Wickham	Proporciona una forma bastante ágil de manejar los ficheros de datos de *R.*
reshape2	Hadley Wickham	Esta biblioteca, dispone fundamentalmente de dos funciones, *melt* y *cast*, muy útiles para determinado tipo de transformaciones de datos.

[23] Término usado en informática para denominar a ciertas modificaciones.

Paquete	Creador/es	Funciones
ggplot2	Hadley Wickham y Winston Chang	En lo que se refiere a la visualización de la información mediante gráficos, este paquete es eficaz. Esta librería proporciona un poderoso sistema que hace que sea menos difícil la tarea de producir gráficos complejos de varias capas.
ggthemes	Jeffrey B. Arnold	Es una función se encarga de proporcionarle a ggplot2 una serie de objetos geométricos, escalas, misceláneas y temas adicionales.

Después de este recorrido descriptivo y explicativo de los tres componentes que conforman nuestro sistema de Procesamiento de Lenguaje Natural y, sobre todo, de Análisis de Sentimientos solo resta definir, en definitiva, cuáles son las principales tareas en las que se enfoca nuestro análisis, el método o enfoque que estamos utilizando y, finalmente, el nivel de análisis o de clasificación de sentimientos en el que centramos nuestro trabajo. En el esquema de la Figura. 16, ilustramos estos tres aspectos.

FIGURA 16: Tareas, enfoques o métodos y niveles de análisis del trabajo



4.5. Análisis de Sentimientos automatizados de *Cien años de soledad* (1967)

Iniciamos este experimento, justamente, con un cúmulo de sentimientos y emociones mezcladas, puesto que no sabíamos qué resultados íbamos a obtener, puesto que el lenguaje figurado, la literatura y todavía más el estilo de escribir de Márquez en ciertos momentos se tornan complejos. No obstante, los resultados y hallazgos que, a continuación, presentaremos nos han dejado impresionados (más de lo que creíamos).

Este análisis intentará describir el proceso que se llevó a cabo para efectuar el AS a cada uno de los textos literarios, en este caso, el de *Cien años…*, visualizar los resultados encontrados mediante una serie de gráficos de distintos diseños y, por nuestra parte, intentaremos interpretar y explicar esos resultados. Sin más, nos sumergimos en los sentimientos y las emociones que figuran en la historia de las siete generaciones de la familia Buendía.

Para que nuestro sistema de AS nos realizara el procesamiento de *Cien años…*, hubo una fase previa de preparación del material. El primer paso que llevamos a cabo fue buscar una versión digital de la novela en su versión traducida al inglés, el formato que hallamos de *Cien años…* fue en PDF, por tal motivo, lo convertimos a texto plano o simple y eliminamos información innecesaria (para este estudio), por ejemplo, la del encabezado y la de pie de página.

En un segundo paso, modificamos los datos del texto plano de *Cien años*, básicamente, lo que hicimos fue dividir la novela en fragmentos de 10 líneas de, aproximadamente, 70 caracteres. Este es, más o menos, el tamaño de un párrafo y es el tamaño ideal para que el sistema de AS le asigne una puntuación de los sentimientos y las emociones. Así, pues, la novela se dividió en 1284 fragmentos.

Al tener modificados los datos, entonces, ejecutamos el tercer paso y, quizá, el que más interesante. A través del paquete *syuzhet* que, como habíamos expuesto, trae consigo el NRC se les asignó a los 1284 segmentos una puntuación, de la polaridad de los sentimientos, tanto positivo como negativos. En otras palabras, el sistema de AS identificó y puntuó cuántos términos de cada fragmento expresaban un sentimiento positivo o negativo. Por lo tanto, podría ser que alguno de estos tenga puntuación alta o baja tanto en una orientación como en la otra. Además, identificamos 25 pasajes importantes de la historia para reconocer la trama de la novela. Pues bien, visualicemos los sentimientos positivos y negativos de *Cien años de soledad* en los gráficos de puntos (Figura 17.) y de barras (Figura 18.)

FIGURA 17: Gráfico de puntos de los sentimientos positivos y negativos de *Cien años de soledad*

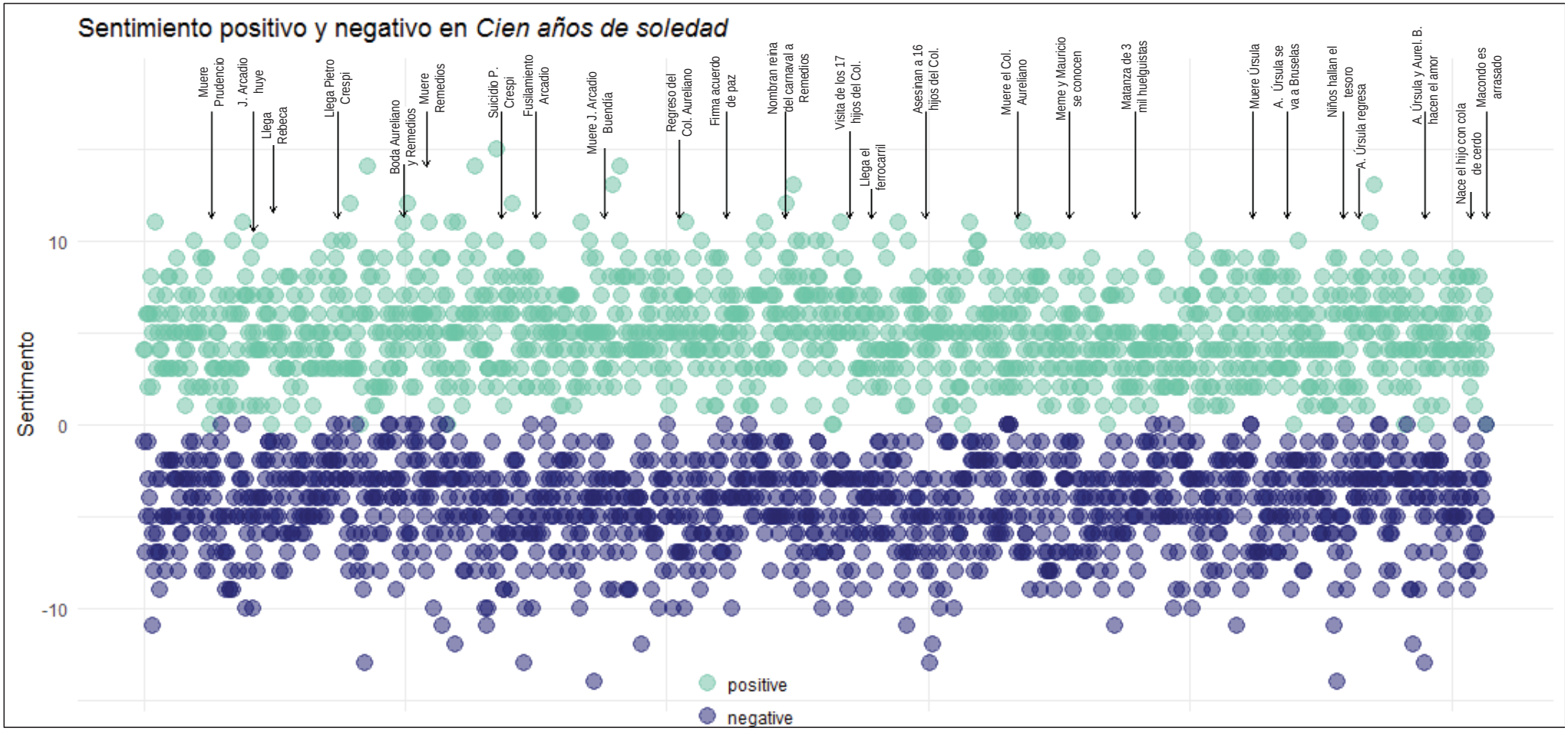
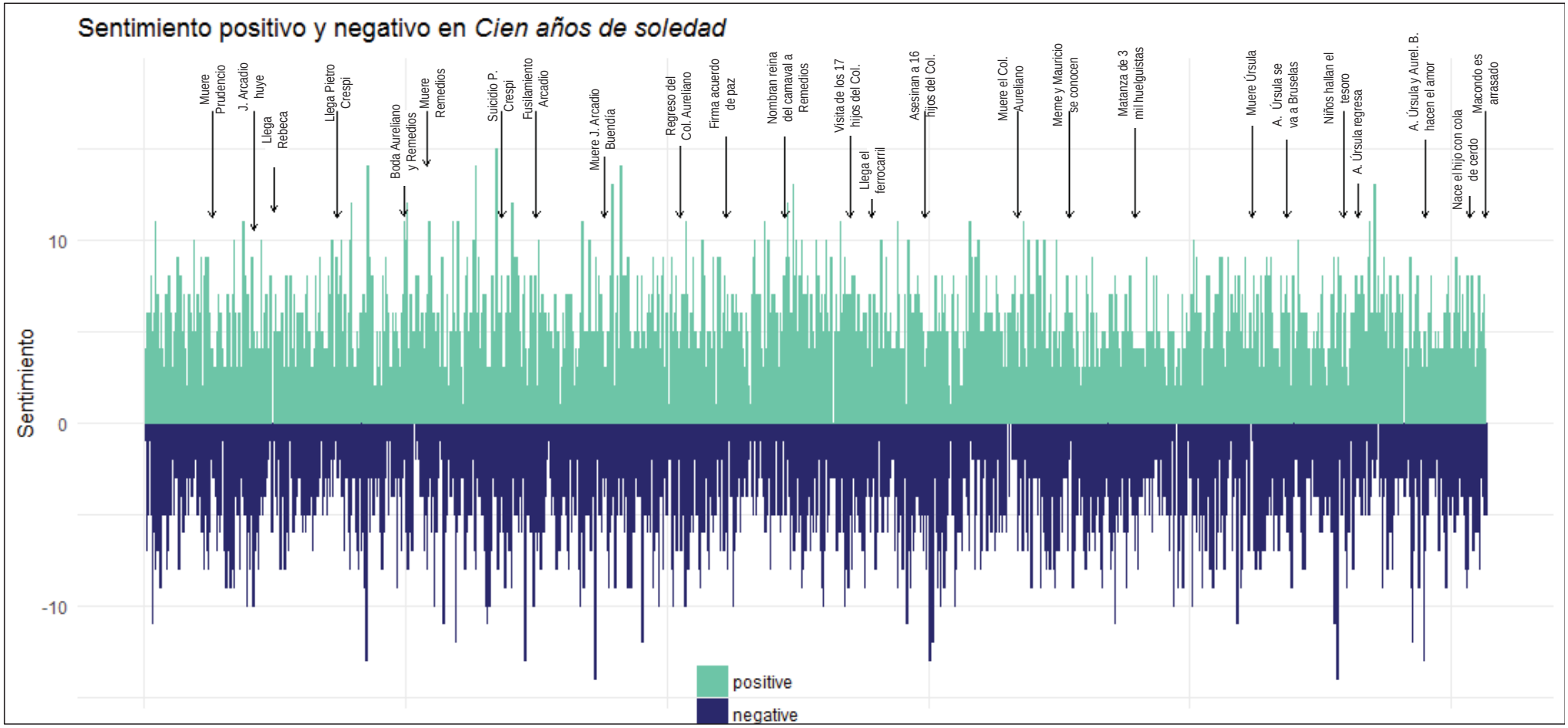


FIGURA 18: Gráfico de barras de los sentimientos positivos y negativos de *Cien años de soledad*



A partir de los gráficos presentados anteriormente podemos deducir dos cuestiones importantes: la primera, es que, en términos generales, las puntuaciones del sentimiento positivo son más que las del sentimiento negativo. Esto significa que en la novela el sentimiento positivo está presente, en mayor o menor medida, durante toda la historia de la familia Buendía. Por otro lado, la segunda cuestión que queremos señalar es que, si bien es cierto, el sentimiento positivo tiene presencia en toda la novela, sin embargo, las máximas puntuaciones para cada sentimiento las hallamos en el negativo. ¿Qué significa esto? Significa que hay momentos, hechos o sucesos de la historia que se relatan con un léxico que expresa fuertemente los sentimientos negativos en comparación, quizá, con aquellas partes de la narración en las que el sentimiento es positivo, pero que no se expresa con términos palabras marcadamente positivas.

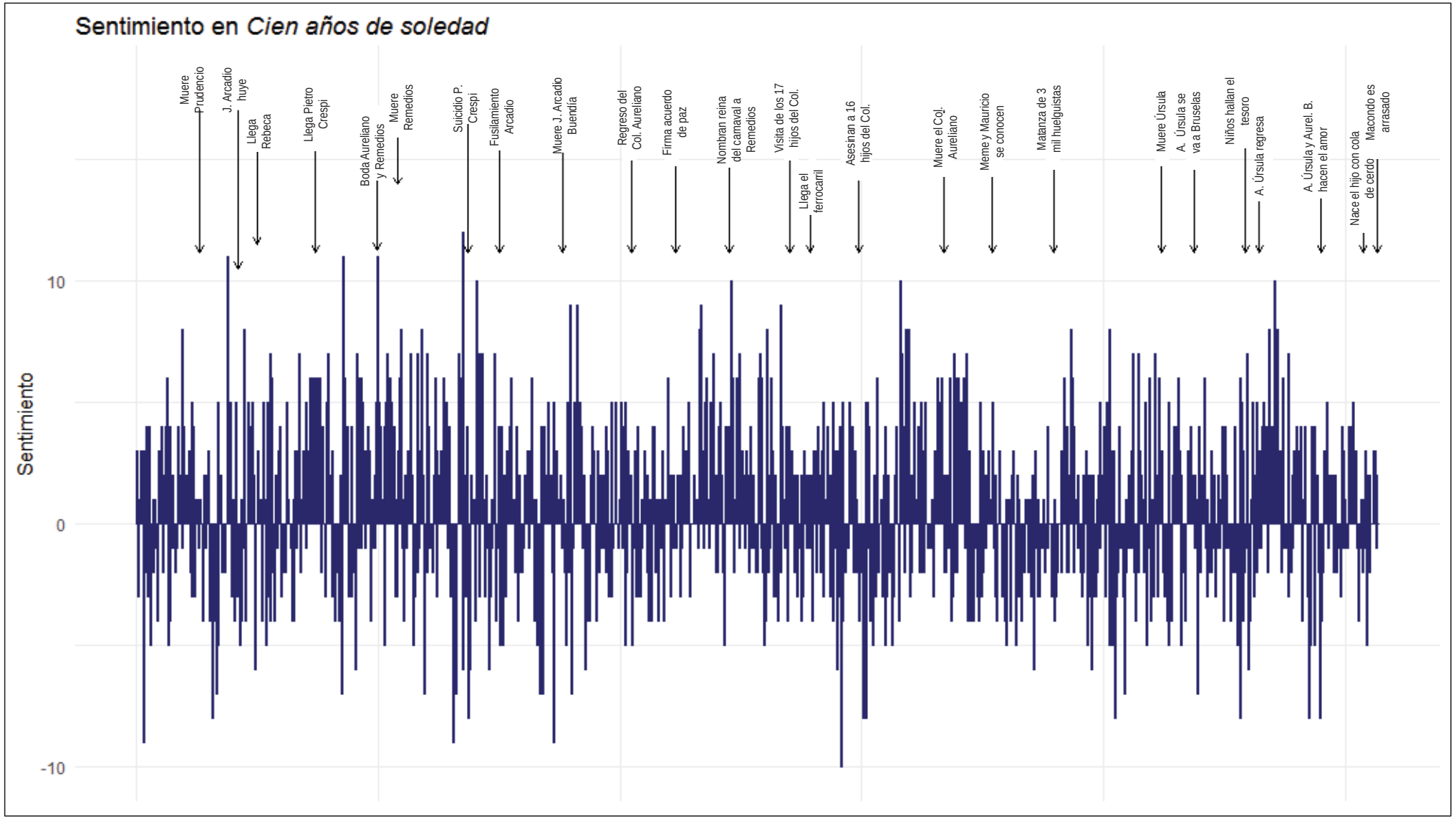
En conclusión, en función de los resultados obtenidos por el SA, diremos que el sentimiento positivo lo hallamos, en mayor o menor medida, durante toda la historia de *Cien años…*, pero, en los momentos que aparece el sentimiento negativo, este se expresa con un léxico vehementemente negativo lo que podría causar mayor impacto durante la lectura de la novela, puesto que, se nos mantendría en un estado de lectura relativamente positiva y de forma repentina y punzante los sentimientos negativos se apoderarían de nosotros.

Hasta aquí, el sistema de AS le asignó una puntuación tanto positiva como negativa a cada fragmento. Ahora bien, lo interesante es saber qué sentimiento es predominante en cada uno de los segmentos y así tener otro panorama de los sentimientos y emociones de la novela. ¿Cómo se logró esto? Básicamente, haciendo un procedimiento de resta sencillo: a la puntuación del sentimiento positiva de cada fragmento se le restó la puntuación del sentimiento negativo. En la Tabla 4., se encuentra el resultado de los 10 primero fragmento de *Cien años…*, luego de realizar este procedimiento. Por otro lado, en la Figura 19., podemos visualizar lo que exponíamos anteriores: que hay más presencia del sentimiento positivo en *Cien años de soledad.*

TABLA 4: Puntuación del sentimiento en *Cien años de soledad*

	linenumber	sentiment
1	1	3
2	2	-3
3	3	1
4	4	3
5	5	1
6	6	2
7	7	3
8	8	-9
9	9	-1
10	10	-2

FIGURA 19: Gráfico del sentimiento en *Cien años de soledad*



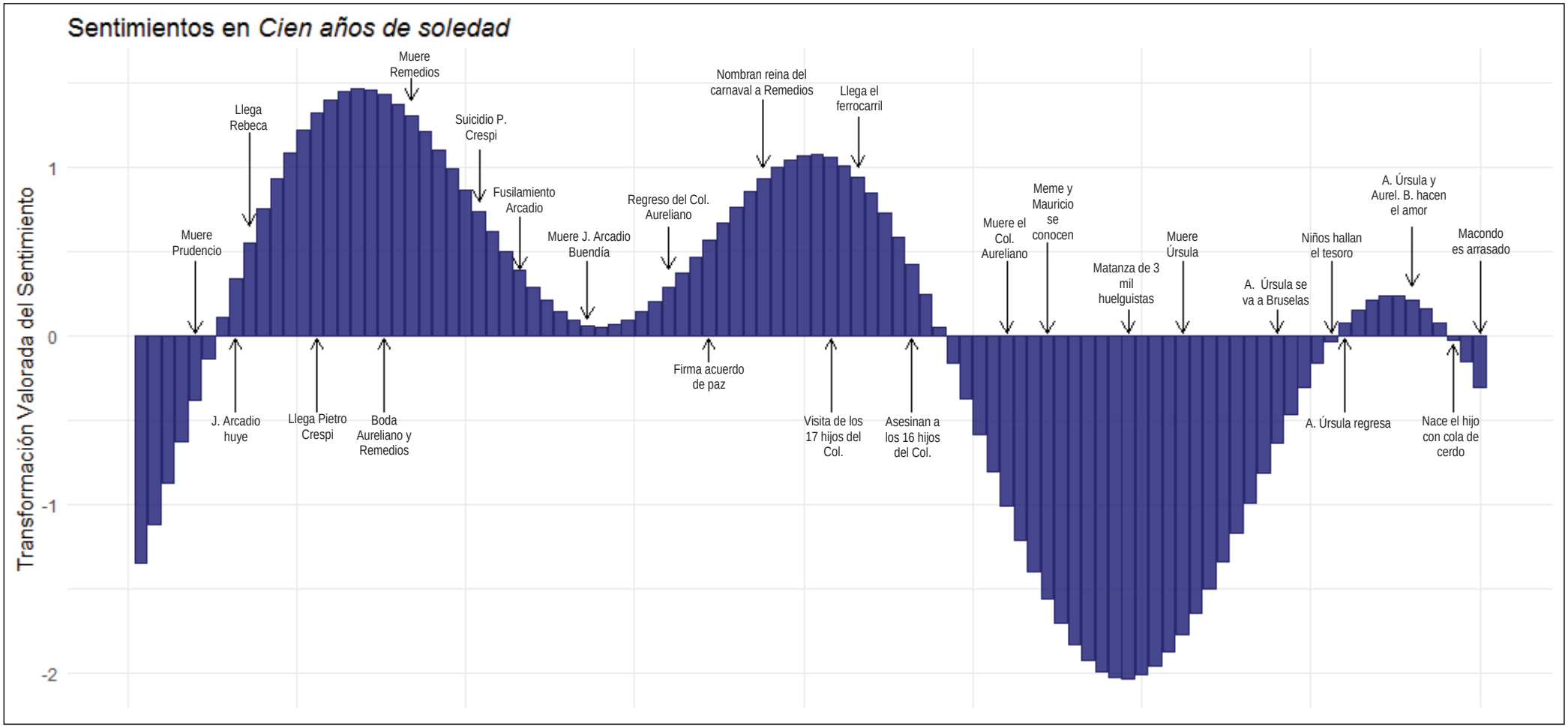
Ahora realizaremos la parte más interesante que es ver la trayectoria global de los sentimientos en toda la historia y relacionarla con algunos de los pasajes más importantes de la novela. ¿Qué queremos lograr con esto? Pues, ver si existe una correlación entre los sentimientos y los hechos, sucesos o acontecimientos que se nos relatan en *Cien años de soledad.* Dicho de forma sencillo sí, por ejemplo, cuando José Arcadio Buendía mata a Prudencio Aguilar o cuando llega el ferrocarril a Macondo los sentimientos son positivos o negativos.

El gráfico que presentamos en la Figura 20., está compuesto por 100 barras, cada una de esas barras contiene la puntuación colectiva de los sentimientos de, aproximadamente, 13 fragmentos del texto, a partir de esa puntuación, se determinó si el sentimiento predominante en ese trayecto de la novela es positivo o negativo. En la Tabla 5., a manera de ejemplo, mostramos la puntuación que asignó el sistema a las primeras 40 barras.

TABLA 5: Puntuación colectiva del sentimiento por cada 13 fragmentos de *Cien años de soledad*

Barra	Puntuación	Barra	Puntuación
1	-1.34909757	21	1.30159493
2	-1.11790112	22	1.20989708
3	-0.87691475	23	1.10428181
4	-0.62995768	24	0.98840931
5	-0.38102744	25	0.86610618
6	-0.13420625	26	0.74126498
7	0.10643569	27	0.61774210
8	0.33693745	28	0.49925654
9	0.55354279	29	0.38929224
10	0.75279112	30	0.29100673
11	0.93160159	31	0.20714822
12	1.08734825	32	0.13998359
13	1.21792404	33	0.09123912
14	1.32179207	34	0.06205551
15	1.39802262	35	0.05295847
16	1.44631493	36	0.06384567
17	1.46700305	37	0.09399052
18	1.46104555	38	0.14206254
19	1.42999926	39	0.20616404
20	1.37597766	40	0.28388201

FIGURA 20: Gráfico de la trayectoria global de los sentimientos en *Cien años de soledad*



¡Impresionante! —: fue la palabra que dijimos cuando visualizamos el gráfico anterior. Los resultados que nos ha generado el sistema de AS son extraordinarios y muy interesantes. A continuación, haremos una interpretación de los mismos.

Lo primero que hay que acotar es que, según nuestro sistema de AS y según el diccionario NRC, *Cien años de soledad* es una novela que inicia y finaliza con sentimientos negativos. Ahora bien, la trayectoria de los sentimientos en toda la novela la podemos estructurar en cinco trayectos importantes: en el primero, de la barra 1 a la 6, los sentimientos son negativos. En el segundo, de la barra 7 a la 60, los sentimientos son, en mayor o menor medida, positivos. En el tercero, de la barra 61 a la 89, los sentimientos son fuertemente negativos. En el cuarto, de la barra 90 a la 97, los sentimientos son levemente positivos. En el quinto, de la barra 98 a la 100, los sentimientos son levemente negativos. Ahora bien, ¿qué hechos o acontecimientos pudieron haber desencadenado estos sentimientos? A continuación, lo veremos.

En el primer tramo, es decir, en el inicio de la novela los sentimientos son negativos. Esto es pertinente, si relacionamos, por ejemplo, que aquí el sistema de AS ubica la muerte de Prudencio Aguilar a manos de José Arcadio Buendía, luego de que el patriarca de los Buendía le ganara una pelea de gallos y Aguilar le dijera que ojalá su gallo le hiciera el favor a su mujer.

En el segundo tramo, los sentimientos dejan de ser negativos y empiezan a subir lentamente de forma positiva. Este ascenso de los sentimientos se inicia, por ejemplo, con la huida de José Arcadio con una gitana, al enterarse que Pilar Ternera estaba embarazada. Asimismo, con la llegada de Rebeca a la casa de los Buendía. Los sentimientos positivos crecen intensamente tras la llegada del italiano Pietro Crespi a la fiesta que se organizó por las reformas de la casa y tras la celebración de la boda entre Aureliano y Remedios. Luego, los sentimientos, aunque positivos siempre, empiezan a descender con la muerte de Remedios, el suicido de Pietro Crespi, el Fusilamiento de Arcadio y llegan a su punto más bajo, dentro de los sentimientos positivos, con la muerte del patriarca de los Buendía, es decir, José Arcadio Buendía. El ascenso de los sentimientos positivos inicia de nuevo con el regreso al pueblo del coronel Aureliano Buendía, la firma del acuerdo de paz que hace el coronel para parar la guerra civil que se estaba viviendo entre liberales y conservadores, también, con el nombramiento de Remedios, la bella, como reina del carnaval; y la visita que le hacen sus 17 hijos al coronel Aureliano. El descenso, dentro de los sentimientos positivos, empieza de nuevo con la llegada del Ferrocarril a Macondo y con el asesinato de los 16 hijos del coronel Aureliano.

El tercer tramo, se concatena con la bajada que se venía generando en los sentimientos positivos, puesto que, en este trayecto los sentimientos son fuertemente negativos. ¿Qué pudo producir este descenso? Pues, seguramente, la muerte del coronel Aureliano Buendía y el punto más bajo, es decir, el momento en el que la novela tiene los sentimientos negativos a su máximo nivel es cuando ocurre la matanza de los tres mil huelguistas en la plaza de Macondo y después con la muerte de Úrsula Iguarán. ¿Coincidencia? ¿*Realismo mágico*? No, esto es: Procesamiento del Lenguaje Natural (PLN) y, especialmente, Análisis de Sentimientos automatizados. Seguidamente, los sentimientos empiezan a alzarse, aunque siempre dentro de los sentimientos negativos. Por ejemplo: con el viaje de Amaranta Úrsula a Bruselas y con el hallazgo, que hicieron los cuatro niños, del tesoro que escondía Úrsula Iguarán en la casa.

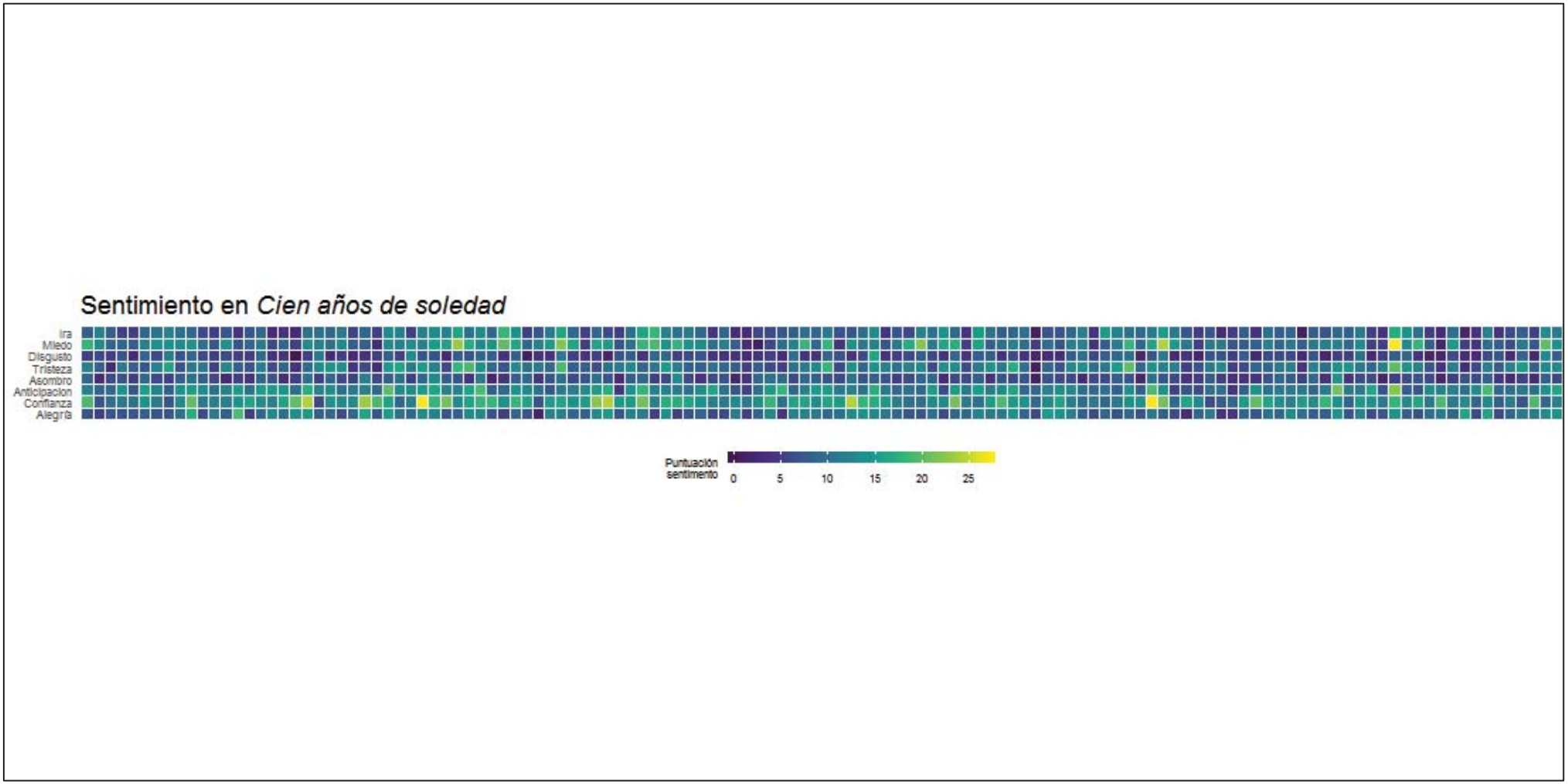
El cuarto tramo de la novela, se encadena con la ascensión que se venía produciendo en el trayecto anterior, aunque aquí con una presencia de sentimientos positivos muy leves. Esto, se conecta perfectamente con algunos de los sucesos que ocurrieron en este intervalo narrativo, por ejemplo, el regreso de Amaranta Úrsula de Bélgica y el momento en el que Aureliano Babilonia consuma el amor (sexualmente hablando) con Amaranta Úrsula, justamente, cuando el esposo de esta, o sea, Gastón, escribía una carta en la misma casa.

Finalmente, en el quinto y último trayecto de la novela, los sentimientos negativos se hacen presente. Seguramente, todos los que hemos leído la novela coincidiremos con esto, puesto que la historia, en verdad, termina mal. Algunos de los hechos más importantes que pudieron haber generado este descenso negativo de los sentimientos fueron: el nacimiento con cola de cerdo del último hijo y miembro de la familia Buendía (el gran temor de Úrsula), el momento en el que las hormigas se comen a este bebé y, naturalmente, porque:

> […] estaba previsto que la ciudad de los espejos (o los espejismos) sería arrasada por el viento y desterrada de la memoria de los hombres en el instante en que Aureliano Babilonia acabara de descifrar los pergaminos, y que todo lo escrito en ellos era irrepetible desde siempre y para siempre porque las estirpes condenadas a cien años de soledad no tenían una segunda oportunidad sobre la tierra (García Márquez, 1967).

Hasta aquí, hemos analizado los sentimientos en lo que respecta a su orientación semántica o a su polaridad: negativa o positiva. Ahora, pasamos a realizar una tarea más avanzada y compleja, dentro del AS, es decir: detectar o identificar estados emocionales en *Cien años de soledad.* En la Figura 21., mediante un mapa de calor, ilustramos la puntuación que le ha asignado al texto de las 8 emociones que tiene el diccionario NRC (ira, miedo, anticipación, confianza, sorpresa, tristeza, alegría y disgusto).

FIGURA 21: Mapa de calor de las emociones en *Cien años de soledad*



Para interpretar de mejor manera el mapa de calor tenemos que tener en cuenta las siguientes cuestiones: por una parte, hay que señalar que se han estructurado las emociones en dos grupos, para efectos de mejor visualización, el primero (de arriba hacia abajo) lo componen las emociones de tendencia negativa: ira, miedo, disgusto y tristeza; el segundo, lo integran las emociones con una orientación más positiva: asombro, anticipación, confianza y alegría. En otro orden, hay que destacar que la puntuación de cada emoción se relaciona con los colores (azul y amarillo), por tal razón, si la emoción tiene mayor intensidad del color azul, significa que la emoción tiene poca puntuación o poca presencia en la novela; por otro lado, si la emoción tiene mayor intensidad del color amarillo, significa que la emoción tiene mucha puntuación o presencia en el texto.

¿Qué emociones tienen poca o mucha puntuación y, por lo tanto, presencia en *Cien años de soledad?* Dar respuesta a tal pregunta no es tan fácil, en este caso, hay que recordar que esta tarea es más avanzada y, ciertamente, más compleja. Sin embargo, parece ser que en nuestro mapa de calor de *Cien años…*, las emociones de orientación positiva tienen, ligeramente, una mayor presencia durante todo el hilo narrativo. ¿Cómo explicamos esto? Pues, lo deducimos porque en los intervalos narrativos en los que se manifiesta la presencia intensa de las emociones negativas, curiosamente, las emociones de tendencia positiva también tienen una puntuación alta.

A partir de ello, concluimos que las emociones de orientación positiva, como el asombro, la anticipación, la confianza y la alegría están presentes, en mayor o menor medida, durante toda la novela. En los trayectos en los que se manifiestan las emociones de orientación negativa como la ira, el miedo, el disgusto y la tristeza son, sutilmente, atenuadas por las emociones positivas. Sin embargo, parece sensato afirmar que, aunque las emociones de tendencia positiva están presentes en casi toda la novela, su presencia o predominio no es rotundo, aplastante o integro. O sea, no hay una marcada diferencia, sino, sutil.

Los resultados hallados en lo que respecta a la primera tarea: identificar o clasificar los sentimientos de la novela según su polaridad (negativa o positiva) fue mucho más fácil, o al menos se podía interpretar con menos dificultad, en contraste con la segunda tarea: identificar las emociones. Así, finalizamos el Análisis de Sentimientos automatizado de *Cien años de soledad* y continuamos con la otra obra de Márquez.

4.6. Análisis de sentimientos de *El amor en los tiempos del cólera* (1985)

Cerramos nuestro Trabajo de Fin de Máster exponiendo el proceso y los resultados del Análisis de Sentimientos de la novela que García Márquez publicó tres años más tarde de ganar su Premio Nobel. Es decir, *El amor en los tiempos del cólera.* Un elemento que queremos destacar y que será determinante en este análisis, es que, a diferencia de *Cien años…*, la temática principal de *El amor…* es totalmente distinta (la visión particular que tiene el escritor colombiano del amor).

En lo que respecta al proceso previo de preparación del material y modificación de los datos, hay que resaltar que se realizaron los mismos pasos o etapas que se llevaron a cabo en *Cien años…*, los principales cambios que implicó esta novela son los siguientes: Primero, la división de la novela, aunque fue igualmente en fragmentos de 10 líneas, en este caso el número total de segmentos que se obtuvieron fueron: 1300, tan solo 16 fragmentos más que la primera novela. Segundo, los pasajes, acontecimientos o sucesos más importantes de la trama central de *El amor…*, que se anotaron o etiquetaron fueron 31, tan solo 6 más que el primer texto.

Por tal razón, damos paso a efectuar nuestra primera tarea de AS, es decir, identificar o clasificar la polaridad de los sentimientos: negativos o positivos en esta segunda obra literaria. Hay que recordar que, para ello, nuestro sistema de AS, a través del paquete *syuzhet* que trae consigo el diccionario de sentimientos NRC, asignó una puntuación de los sentimientos de positivos como negativos a cada fragmento. En la Tabla 6. y la Tabla 7., se muestra esta puntuación de los últimos 10 segmentos del texto. Seguidamente, presentamos los sentimientos positivos y negativos de toda la novela: en el gráfico de puntos (Figura 22.) y en el gráfico de barras (Figura 23.)

TABLA 6: Puntuación del sentimiento positivo de *En el amor en los tiempos del cólera*

linenumber	sentiment	value
1291	positive	2
1292	positive	5
1293	positive	9
1294	positive	6
1295	positive	6
1296	positive	4
1297	positive	3
1298	positive	3
1299	positive	5
1300	positive	4

TABLA 7: Puntuación del sentimiento negativo de *En el amor en los tiempos del*

linenumber	sentiment	value
1291	negative	-5
1292	negative	-3
1293	negative	-4
1294	negative	-5
1295	negative	-3
1296	negative	0
1297	negative	-7
1298	negative	-4
1299	negative	0
1300	negative	-4

FIGURA 22: Gráfico de puntos de los sentimientos positivos y negativos de *El amor en los tiempos del cólera*

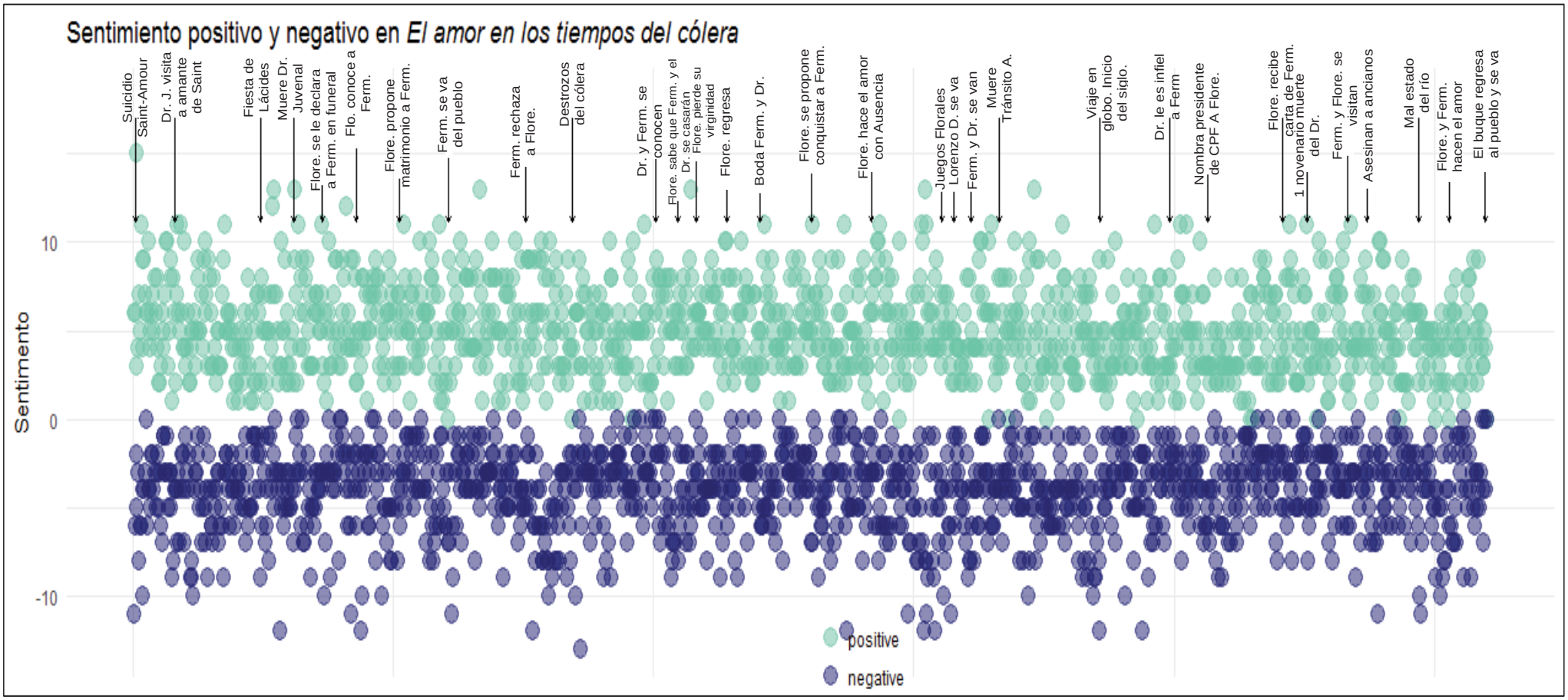
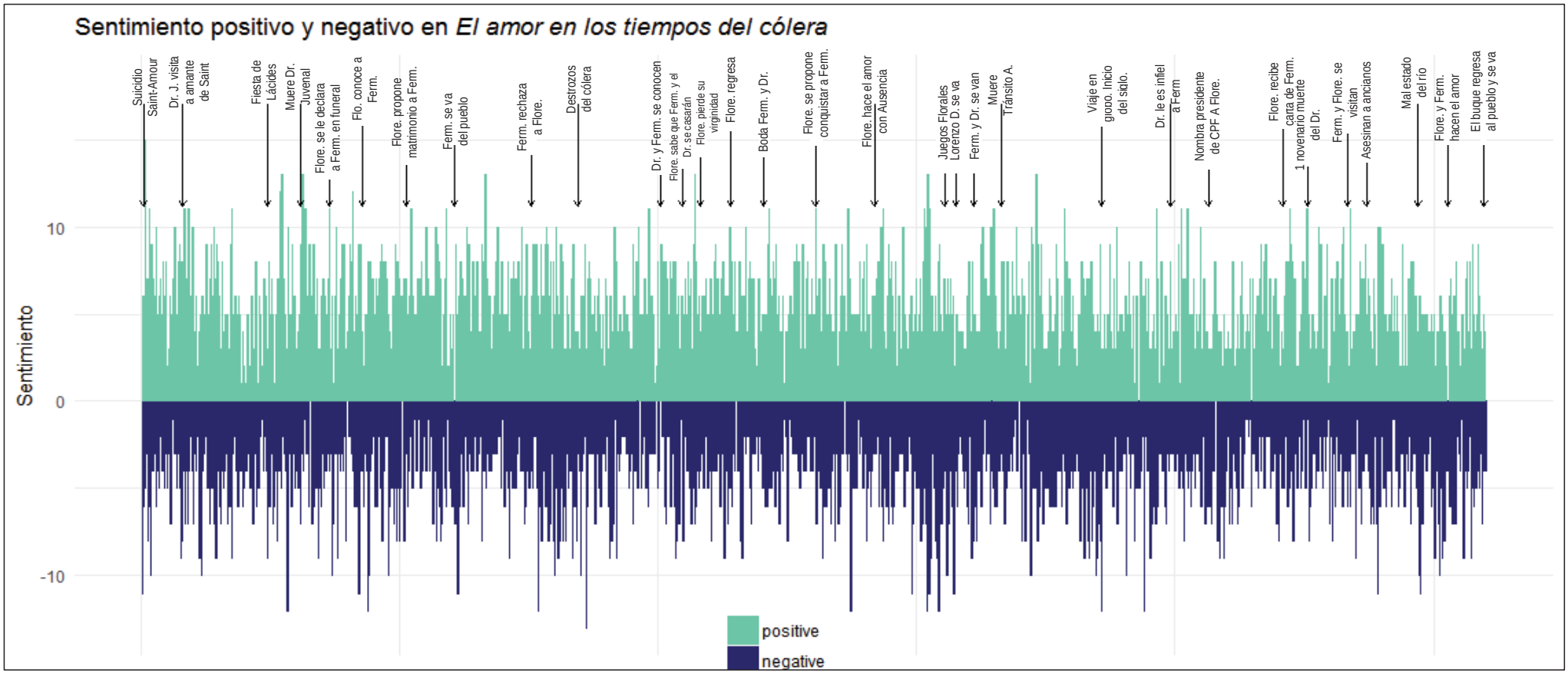


FIGURA 23: Gráfico de barras de los sentimientos positivos y negativos de *El amor en los tiempos del cólera*



Los dos gráficos presentados anteriormente de las puntuaciones tanto para los sentimientos positivos como negativos de *El amor en los tiempos del cólera*, nos reflejan resultados y diferencias mucho más claras, si lo ponemos en contraste con los gráficos que se generaron de *Cien años…* Aquí, las puntuaciones para el sentimiento positivo, desde un panorama general, son mucho mayores que las del sentimiento negativo. Así, pues, diremos que el sentimiento positivo permanece, en mayor o menor medida, durante la narración de los diferentes sucesos de la novela.

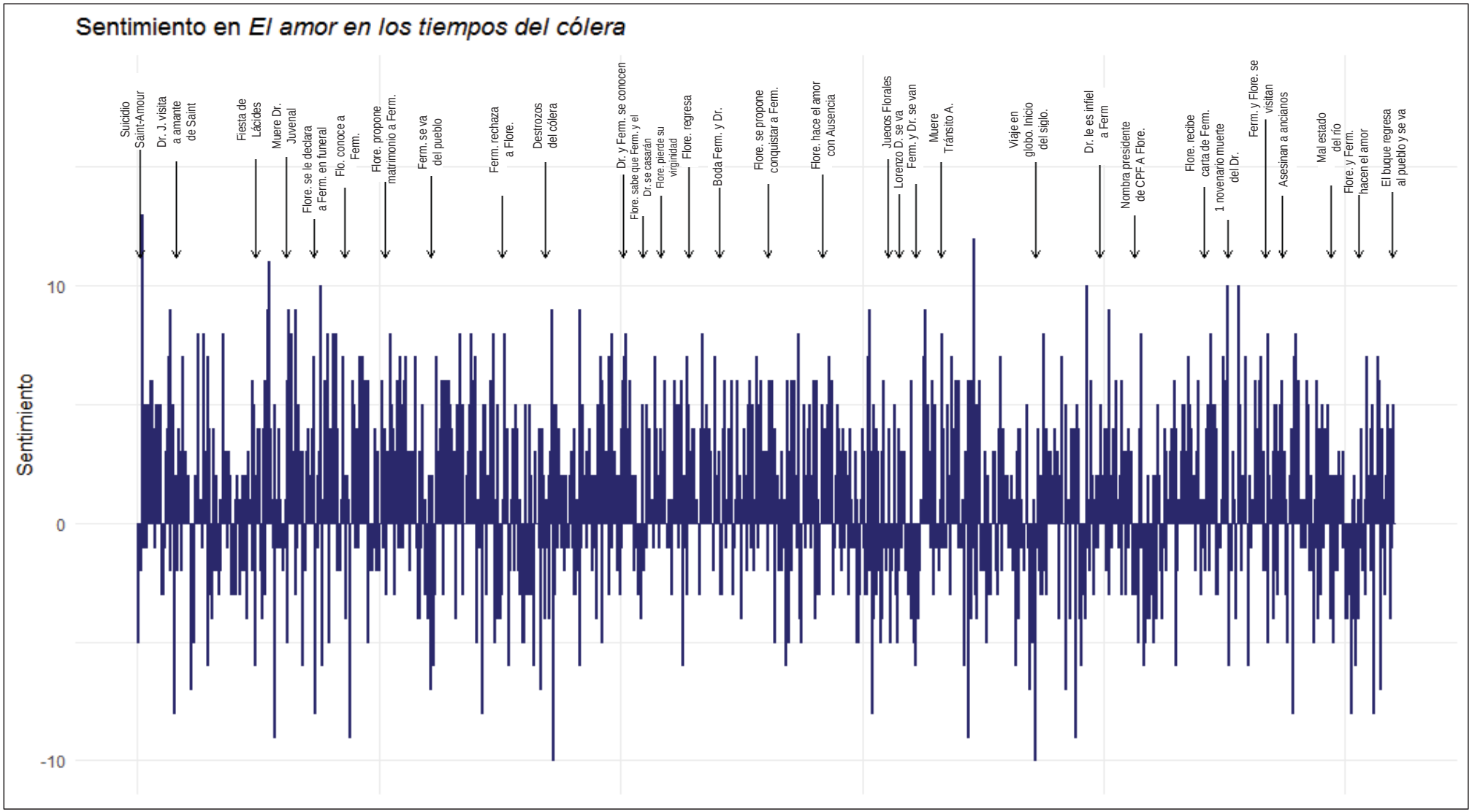
Por otro lado, notamos que, en algunos pasajes narrativos, las puntuaciones del sentimiento negativo se despuntan, pero no lo hacen de forma repentina, sino que vienen precedidas, generalmente, por un descenso procesual del sentimiento. En términos sencillos, esto significa que cuando nos hallamos en la lectura con un suceso o acontecimiento en el que predomina el sentimiento negativo, nosotros, ya podríamos haber sido avisados de lo que va ocurrir gracias a una serie de hechos previos que tienen presencia de sentimiento negativo. Esto, quizá, podría hacer que el lector se prepare emocionalmente para recibir tales acciones, todo depende, desde luego, de su intuición o su lógica.

A continuación, comprobaremos que el sentimiento positivo es predominante o tiene mayor presencia en la narración de esta novela. Al igual que se hizo en *Cien años…*, en este texto se efectúo el proceso de resta de los sentimientos positivos menos los sentimientos negativos de cada fragmento. En la Tabla 8., se muestra la puntuación del sentimiento de los fragmentos 650 al 659, que representan, tentativamente, la mitad de la novela. Asimismo, en la Figura 24., podemos notar que el sentimiento predominante de la novela es el positivo.

TABLA 8: Puntuación del sentimiento en *El amor en los tiempos del cólera*

linenumber	sentiment
650	3
651	2
652	-2
653	4
654	6
655	3
656	6
657	1
658	1
659	-5

FIGURA 24: Gráfico de los sentimientos de *El amor en los tiempos del cólera*



Después de haber determinado o identificado que el sentimiento positivo es el sentimiento que tiene mayor presencia en la novela, damos paso a la parte que, quizá, se torna más interesante y curiosa, desde nuestra perspectiva, es decir, ver la trayectoria global de los sentimientos en toda la historia y vincularla con algunos de los pasajes más importantes de la obra literaria. En este caso, reconoceremos si, por ejemplo, el sentimiento es negativo o positivo cuando Fermina Daza es obligada por su padre a dejar el pueblo, o, cuando Florentino Ariza conoce y habla por primera vez con Fermina.

Este trayecto global de los sentimientos de la novela es ilustrado por el gráfico que presentamos en la Figura 25., así como el de *Cien años…*, este está compuesto por 100 barras y también cada una de las barras agrupa, aproximadamente, 13 fragmentos del texto. Hay que recordar que, a cada uno de estos grupos, nuestro sistema de AS le asignó una puntuación colectiva de los sentimientos y determinó cuál sentimiento es predominante en cada trayecto, si el positivo o el negativo. A manera de ejemplo, mostramos en la Tabla 9., la puntuación que se le asignó a las primeras 30 barras.

TABLA 9: Puntuación colectiva del sentimiento por cada 13 fragmentos de *El amor en los tiempos del cólera*

Barra	Puntuación	Barra	Puntuación
1	1.00080251	21	1.04185698
2	-0.73810930	22	0.95647044
3	-0.47843329	23	0.86865042
4	-0.22553446	24	0.78174995
5	0.01696201	25	0.69896210
6	0.24565109	26	0.62323450
7	0.45742832	27	0.55719066
8	0.64956313	28	0.50305998
9	0.81976304	29	0.46261860
10	0.96622713	30	0.43714248
11	1.08768724	31	0.42737404
12	1.18343580	32	0.43350346
13	1.25333947	33	0.45516487
14	1.29783825	34	0.49144786
15	1.31792986	35	0.54092388
16	1.31513995	36	0.60168693
17	1.29147866	37	0.67140743
18	1.24938479	38	0.74739798
19	1.19165885	39	0.82668915
20	1.12138690	40	0.90611340

FIGURA 25: Gráfico de la trayectoria global de los sentimientos en *El amor en los tiempos del cólera*