UNIVERSIDAD DE JAÉN

ESCUELA POLITÉCNICA SUPERIOR DE JAÉN
DEPARTAMENTO DE INFORMÁTICA

TESIS DOCTORAL

SENTIMENT ANALYSIS IN SPANISH

PRESENTADA POR:
EUGENIO MARTÍNEZ CÁMARA

DIRIGIDA POR:
DR. D. ALFONSO UREÑA LÓPEZ
DRA. DÑA. M. TERESA MARTÍN VALDIVIA

JAÉN, 26 DE OCTUBRE DE 2015

ISBN 978-84-16819-02-7

PhD Thesis

Sentiment Analysis in Spanish

Eugenio Martínez Cámara

Supervisors: PhD L. Alfonso Ureña López and PhD M. Teresa Martín Valdivia

Universidad de Jaén
Departamento de Informática

Tesis Doctoral

Análisis de Opiniones en Español

Sentiment Analysis in Spanish

Autor:
Eugenio Martínez Cámara

Dirigida por:
Dr. L. Alfonso Ureña López
Dra. María Teresa Martín Valdivia

Octubre 2015

A mi madre, mi padre y mi hermana

Agradecimientos[1]

Es fácil caer en el error de pensar que una tesis doctoral es el resultado del trabajo individual de una persona, pero la exigencia que requiere su elaboración obliga al concurso de varias más además de su autor. El objetivo de estas líneas es agradecer a esas personas que me han acompañado en todos los aspectos de mi vida durante los años en los que he estado preparándome y redactando mi tesis doctoral.

Quisiera comenzar por mis directores de tesis L. Alfonso Ureña López y M. Teresa Martín Valdivia, a los cuales siempre les agradeceré la gran confianza que han tenido en mi. A mi mente vienen ahora dos imágenes, el día en que Alfonso, con mi Proyecto Fin de Carrera recién presentado, me dijo que si había pensado en hacer una tesis doctoral; y el momento en que Maite me preguntó "¿alguna vez has escrito un artículo?", a lo que yo respondí "no", y casi antes de que yo terminara de pronunciar la corta partícula negativa, Maite ya me estaba diciendo, "pues para todo hay una primera vez". Esos fueron los dos momentos iniciales y empujones necesarios para comenzar el estudio que ha desembocado en la presente tesis, y que supusieron la primera muestra de confianza de mis directores.

Parece que de unos directores de tesis exclusivamente se aprende los conocimientos propios del campo de investigación en el que están especializados, pero la teoría relativa al Procesamiento del Lenguaje Natural es una pequeña muestra de lo que he aprendido de ellos. De Alfonso me llevo la gran cantidad de consejos que me ha ido dando a lo largo de los años que llevo trabajando con él, los cuales no se han limitado al ámbito científico o laboral, sino que han llegado hasta el nivel personal. Alfonso me ha enseñado a valorar la perseverancia, la cual siempre da su fruto aunque las circunstancias no sean las adecuadas, así como a no buscar atajos en el trabajo, sino a tratar de desarrollar adecuadamente y con esmero cualquier tarea. De Maite espero haberme llevado una pizca de su capacidad de trabajo, la cual es envidiable; un poco de su pasión por la investigación, la cual es inabarcable; y una migaja de su fuerza vital. Gracias a los dos.

Mi estudio no ha estado sólo impulsado por Alfonso y por Maite, sino también por cada uno de los miembros del grupo Sistemas Inteligentes de Acceso a la Información. No puedo olvidar mi primer trabajo sobre *tweets*

[1]Este trabajo de investigación ha sido parcialmente financiado por el Fondo Europeo de Desarrollo Regional (FEDER), el proyecto FIRST FP7-287607 del Séptimo Programa Marco para el Desarrollo de la Investigación y la Tecnología de la Comisión Europea; el proyecto ATTOS TIN2012-38536-C03-0 del Ministerio de Economía y Competitividad y el proyecto AROESCU P11-TIC-7684 MO de Excelencia de la Junta de Andalucía.

con Miguel Ángel, ni tampoco cada una de las ediciones de TASS que hemos organizado junto a Julio Villena. Arturo ha sido en muchas ocasiones un referente por su hambre insaciable de aprender cada día algo nuevo. Nunca me ha faltado la ayuda de Manuel Carlos en todo aquello que le he pedido, por lo cual siempre le estaré agradecido. La mención de la palabra ayuda me obliga a recordar a Jose, el cual siempre paraba de hacer cualquier cosa que estuviera haciendo para escucharme en el momento que me sentaba en la silla que había al lado de su mesa. El humor y los consejos de Manolo también permanecerán en mi memoria, porque muchos han sido los días en el rato del café, en los que ha sabido arrancar una sonrisa a todos los que estábamos a su alrededor. Con Fernando he trabajado poco, pero al igual que el resto siempre me he ayudado cuando se lo he pedido. Loles se merecería un párrafo entero, pero no quiero que haya más agradecimientos que tesis. Una parte importante de mi investigación la he desarrollado con Loles, juntos hemos aprendido, juntos hemos trabajado muchas horas, y juntos hemos sacado muchos experimentos adelante antes de que acabaran con nosotros mismos. Salud y Eladio llegaron al final, pero eso no quiere decir que no hayan aportado su granito de arena para que saliera esta tesis, porque ambos me han echado un capote cuando se lo he pedido.

No quiero dejar de mencionar a todos los que han pasado por el despacho 154 del A3 porque son muchas las horas que hemos pasado juntos. Las largas conversaciones con Cristobal será complicado dejar de recordarlas; las risas y confidencias con Javi fueron una gran compañía en mis primeros años; la alegría de Poncho será muy difícil encontrarla en otro compañero de trabajo; la manera de ser de Antonio Galiano nunca se me olvidará; fue un lujo que pasara por el despacho la profesionalidad de Víctor; así como fue una suerte haber podido disfrutar de la enorme cultura de Eduard, su conocimiento sobre lingüística computacional y su peculiar personalidad.

El apoyo de los compañeros de la Universidad ha sido muy importante, pero creo que más aún ha sido el que he encontrado fuera. Sin el apoyo y confianza de mis padres y mi hermana no hubiera podido completar mi tesis, porque ellos han sido mi gran sustento durante estos años de esfuerzo. Ellos han sabido entender mi ausencia y mi dedicación al trabajo, lo cual nunca se lo podré agradecer como se merecen. Mi madre ha sabido siempre como darme la mano en los momentos complicados que he pasado, así como a animarme a seguir hacia adelante. El apoyo de mi padre ha llegado hasta el punto de convertirse en mi corrector oficial, y con el mérito añadido de tener un conocimiento reducido de informática. A mi hermana siempre le agradeceré el haber sabido entender el hecho de tener un hermano continuamente trabajando, ocupado y en ocasiones ausente. Gracias papi,

mami y niña.

Carmen llegó en la recta final, en esos momentos en los que falta el empujón necesario para sentarse a escribir. Pues Carmen, sin que nadie se lo dijera, supo que tenía que hacer eso, y encontró la manera de hacer que me sentara a escribir. Muy importante ha sido el ánimo que me ha dado cada día, meritorio el haberme escuchado con comprensión cada vez que lo he necesitado, y admirable el haber aguantado a un novio en ocasiones ausente. Gracias por estar a mi lado.

El apoyo de los amigos es muy importante y todos los que tengo, desde Antonio Turnes, hasta Jorge, Jesús, Chemita, Marta Amate, Elia, Cristina, Lucía, Marta Cámara, Teresa, Ignacio, Víctor, Antonio, Alberto Montoro, Pedrito, María, Chari, Manuel Caro, JuanSe, Manolín..., han sabido entenderme cuando lo he necesitado. Pero es a dos a los que debo desde aquí darles un gran abrazo, y son mi gran amiga Regina y mi eterno compañero de prácticas y amigo Luisk. Ellos dos son mis confidentes, los que nunca me han fallado, los que siempre que los he necesitado han estado ahí. Sin ellos dos me hubiera costado mucho más el elaborar este trabajo de investigación.

No quiero olvidarme de Manolo Gil por haber diseñado con gran originalidad la portada, y tampoco de Ricardo por haberse prestado a aparecer en ella. Gracias.

Espero no haberme olvidado de nadie, pero por si acaso, lanzo un último GRACIAS.

Índice

Lista de Figuras

Lista de Tablas

Abstract

Sentiment Analysis is the Natural Language Processing task related to the computational treatment of opinion, sentiment and subjectivity in text (Pang & Lee, 2008). The extensive research on Sentiment Analysis is mainly focused on the study of texts written in English. The present thesis is aimed at providing a report concerning the automatic processing of opinions written in Spanish.

An opinion is a linguistic figure formed by different elements, from which the present thesis will only pay attention to the opinion valence. Polarity classification is a problem that can be faced up from different perspectives. The present thesis attempt to cover the largest number of methodologies as a first approach to the polarity classification of Spanish texts. Therefore, classifiers based on supervised learning have been developed; systems that try to take advantage of the similarity of concepts with the intention of extending the sentiment meaning of the words of the texts have been design; and the combination of classifiers of different nature by means the utilization of ensemble methodologies has been studied.

One of the hindrance of Sentiment Analysis in Spanish is the lack of linguistic resources. Therefore, the research underlying to the present thesis also pay attention to the development of new linguistic resources in Spanish, with the aim of providing to the research community of these resources that can be used in their systems, and also to help the development of the research in Sentiment Analysis in Spanish.

Sentiment Analysis is a very domain dependent task, so the present thesis cannot left without studying the development of techniques to adapt the systems to the domain of the texts. Thus, a domain adaptation method

have been developed with the aim of adapting sentiment linguistic resources to a specific domain.

To conclude, the present thesis takes into account the treatment of two kind of texts: long and short. Long texts correspond to long reviews composed by more than two sentences, and short texts are those that are published in Twitter. Thus, the different classification strategies have been applied to long reviews and tweets.

Resumen

El Análisis de Opiniones es la tarea de Procesamiento del Lenguaje Natural que se centra en el tratamiento de opiniones, sentimientos y expresiones subjetivas (Pang & Lee, 2008). El Análisis de Opiniones es una tarea que actualmente aglutina una gran cantidad de publicaciones, pero la mayoría de ellas se limitan al tratamiento de opiniones escritas en inglés. La presente tesis se marca como objetivo aportar a la comunidad investigadora un estudio sobre el procesamiento automático de opiniones en español.

Una opinión es una figura lingüística conformada por distintos elementos, de los cuales la presente tesis sólo le va a prestar atención a la orientación de la propia opinión. La clasificación de la polaridad se puede afrontar aplicando distintas técnicas de clasificación, tratándose en la presente tesis de cubrir el mayor número de ellas. Por ello, se van a describir sistemas de clasificación supervisados; sistemas basados en la extensión del significado de las palabras que se encuentran en los textos; y sistemas que tratan de combinar clasificadores de naturaleza diferente.

Una de las rémoras de la investigación en Análisis de Opiniones en español es la falta de recursos lingüísticos. Por lo tanto, la investigación subyacente a la presente tesis también va a destinar esfuerzos al desarrollo de recursos lingüísticos en español, con el objetivo de proporcionar a la comunidad investigadora nuevos recursos que puedan ser integrados en sus sistemas, así como con el objetivo de promocionar y facilitar la investigación en Análisis de Opiniones en español.

El Análisis de Opiniones es una tarea muy dependiente del dominio, por lo que también se han estudiado técnicas de adaptación de los sistemas de

clasificación de la polaridad al dominio de los textos. Por consiguiente, se presenta un método de adaptación de recursos lingüísticos al dominio de los documentos basado en la frecuencia de las palabras.

Por último, destacar que la presente tesis ha tenido en cuenta dos tipos de documentos a los que se pueden enfrentar los sistemas anteriormente señalados: textos largos y textos cortos. Los textos largos son aquellos que tienen una longitud superior a dos oraciones, mientras que los textos cortos son los que se publican en Twitter.

1

Introducción

1.1. Petición de opinión

El pedir consejo, el pedir ayuda ante una decisión de cualquier índole, es propio de la condiciona humana. El proceso de decisión requiere de un conocimiento amplio de todas las posibilidades existentes, los inconvenientes y los beneficios que pueden conllevar cada una de ellas. Ese amplio conocimiento lleva en muchas ocasiones incluso a intentar conocer el devenir de hechos futuros, los cuales son extremadamente complejos de dilucidar para la mente humana. Dado que es excesivamente complicado, por no decir casi imposible sentirse seguro sobre el desarrollo futuro de los acontecimientos, los seres humanos intentamos ampliar al máximo nuestro conocimiento sobre los distintos elementos relacionados con la decisión a tomar. Una vez dado el primer paso de querer comenzar el proceso de tomar la decisión, el ser humano se afana por conocer todas las características de las distintas opciones que puede tomar.

El proceso de tomar una decisión no es sencillo, y normalmente está constituido por un proceso intrínseco y otro extrínseco. El procedimiento intrínseco o personal está relacionado con todo lo que conlleva el análisis de las distintas opciones: el estudio de lo que conviene más o menos, los beneficios y los perjuicios, y el aprovechamiento en términos de satisfacción personal o material que va a generar cada una de las opciones. La duración y la importancia de este proceso puede ser mayor o menor en función de lo reflexiva que sea la persona en cuestión. La fase externa está motivada por la interna, ya que en el proceso de reflexión surge la necesidad de conocer experiencias similares tras la toma de una decisión parecida. De esta necesidad nace la acción de "pedir opinión". La "petición de opinión" es intentar introducir en el proceso de la toma de la decisión una experiencia semejante con la intención de ayudar a dar el paso final de la decisión. Las opiniones de terceros se suelen utilizar como una manera de intentar conocer lo que ocurrirá tras la toma de la decisión, que debido a la limitación de nuestro conocimiento no podemos vislumbrar. Por lo tanto, esas opiniones suelen ayudar considerablemente a finalizar el proceso de la toma de decisión. Pero la "petición de opinión" no solamente se realiza para la incorporación de la experiencia de un tercero, sino también para poder integrar el conocimiento de esa persona acerca de la decisión que se quiere tomar.

La "petición de opinión" suele realizarse en primer lugar a personas de nuestro círculo de confianza, ya que se confía en la sinceridad de su respuesta, en la veracidad de la información que nos va a aportar, en su criterio, o simplemente se conoce su experiencia. En ocasiones no es suficiente con la opinión del entorno más cercano, sino que se

necesita aumentar aún más la información disponible, o se quiere ampliar el conocimiento de experiencias similares. El proceso de la toma de decisión finaliza cuando las dos etapas, intrínseca y extrínseca, aportan el conocimiento suficiente para la toma de la decisión.

Las opiniones no solamente son ayuda o asistencia en un proceso de toma de decisión. Mediante la opinión las personas expresan su parecer sobre los temas de la más diversa índole: política, religiosa, económica, social, deportiva, comercial, cultural, etc. Si existe una conversación, la expresión de opiniones es fuente de enriquecimiento para los participantes de esa conversación, o incluso de conflicto cuando esas opiniones son muy encontradas. En ocasiones la expresión de la opinión no se realiza en el cuerpo de una conversación, sino que se publica en algún medio. Normalmente ese tipo de publicaciones reciben el nombre de crítica, que ya puede ser política, literaria, cinematográfica o de cualquier otra naturaleza. La crítica, como bien dice nuestro premio Nobel Vargas Llosa en (Vargas Llosa, 2013), ha ayudado al lector a juzgar entre lo bueno y lo malo. La crítica tiene como función el establecimiento de una jerarquía de distintos niveles de calidad, de manera que asiste a las personas en las dos etapas, intrínseca y extrínseca, del proceso de toma de decisión de consumir el objeto de la crítica. La lectura de la crítica ha interesado primeramente a la persona o ente objeto de la misma, pero también ha atraído a su lectura a personas ávidas de información a la hora de tomar una decisión.

1.2. Motivación

Como se puede apreciar en la sección anterior, la expresión, la búsqueda y la atención de la opinión es propia de la condición humana. Desde los inicios del lenguaje se han venido expresando opiniones. La expresión de la opinión ha experimentado una evolución vertiginosa en los últimos años. La opinión en un principio se transmitía por el medio de comunicación más tradicional y antiguo, el "boca a boca". Con la llegada de la escritura, las expresiones subjetivas ya no sólo se comunicaban oralmente sino también por escrito. Llegada la imprenta, surgieron los diarios de prensa, que fueron estupendas plataformas para difundir la opinión de los distintos autores que participaban en su edición. Antes se hablaba de la crítica, pues fue en esos primeros diarios donde la crítica se hizo hueco todos los días, e incluso ocupando lugares destacados los columnistas de opinión política y críticos culturales.

Llegó la Web a finales del siglo XX, y paulatinamente comenzó a revolucionar la sociedad. Internet ha transformado el acceso a la

información, la búsqueda de información, la comunicación, el trabajo, e incluso se podría llegar a decir que hasta las relaciones sociales. En nuestra opinión esa revolución que ha traído Internet distingue dos fases. La primera etapa permitió acceder a información que antes era imposible, dado que a un golpe de *click* se podía conseguir lo que en tiempos anteriores suponía visitar varias bibliotecas y dedicar horas de búsqueda entre las hojas de sus libros. La Web en sus inicios era estática, sólo era un escaparate de información, donde únicamente los administradores de las páginas webs podían editar su contenido. Dicho con otras palabras, Internet ofrecía una comunicación unidireccional. A comienzos del siglo XXI se alumbró un concepto que revolucionó la propia Web, y dio una vuelta de tuerca más al cambio social que estaba impulsando la Red, comenzando así la segunda fase de la revolución de Internet. Ese concepto es Web 2.0, definido por Tim O'Reilly en (O'Really, 2005) como una Web en donde existe una comunicación bidireccional, en la que no solo el administrador de la página web puede editar la información, sino también el usuario que antaño solo se limitaba a leer la información. La Web 2.0 ofrece un paradigma en donde las fronteras entre productores y consumidores de información son muy difusas, y cualquier usuario de Internet puede publicar todo tipo de información. El desarrollo del concepto Web 2.0 fue el que auspició la llegada de los elementos que transformaron la Web, como los blogs, el RSS, las *wikis*, foros, o las redes sociales.

Un blog es una plataforma en las que el administrador o autor publica textos sobre la temática que le plazca en el momento que estime oportuno. Lo más común es que los blogs traten sobre un tema en concreto, sobre el cual su autor publica un texto periódicamente. Las publicaciones de los blogs suelen incluir contenido subjetivo, ya que se trata de una publicación personal y libre, por lo que usualmente el autor expresa su parecer sobre el tema que está desarrollando. Los blogs permiten que los lectores puedan comentar la publicación de su autor, emitiendo de esta manera su opinión favorable o contraria a lo expresado por parte del autor. No son raras las ocasiones en las que se establecen, a través de los comentarios, conversaciones entre los distintos lectores y el autor del blog, de manera que se enriquece aún más el contenido de la publicación.

Los foros se podría decir que son plataformas donde los usuarios conversan sobre un determinado tema que un usuario abre. En las distintas publicaciones de un hilo de conversación de un foro también son abundantes las opiniones, ya que normalmente los usuarios no relatan hechos sino que expresan su parecer sobre el objeto de la conversación. Foros hay de todo tipo, desde los de temática general, a los más específicos, como pueden ser

los relacionados con consultas técnicas.

El punto común de blogs y foros es la comunicación que se establece entre los usuarios, siendo su diferencia el marco en el cual se produce la comunicación. En cambio, las redes sociales dan un paso más, dado que establecen relaciones entre los distintos miembros de la red. En las redes sociales los usuarios se van asociando en función de algún tipo de relación existente en el mundo físico. Actualmente existe una amplia variedad de redes sociales, pudiéndose distinguir entre redes sociales verticales y horizontales. Las redes sociales verticales son las especializadas en una temática concreta, como puede ser música (Last.fm[1]), fotografía (Instagram[2]), vídeo (Youtube[3]), empleo (LinkedIn[4]), deporte (Timpik[5]), investigación (ResearchGate[6]) o literatura (Lecturalia[7]). Una persona siempre se siente reforzada en su afición si hay otras personas que la comparten con él. Ese es el objetivo de las redes sociales verticales, poner en contacto a personas que comparten inquietudes. Al tener el mismo gusto por un tema, las personas se sienten atraídas a compartir su parecer con aquellos que tienen la misma afición. Por ello, las redes sociales verticales están repletas de contenido subjetivo, el cual está constituido principalmente por opiniones. Las redes sociales horizontales no centran la atención en un tema en concreto, o mejor dicho, su punto de mira no es más que poner en contacto virtualmente a personas o entes, y por tanto, cuentan con un mayor número de usuarios. Algunos ejemplos de redes sociales horizontales son: Facebook[8], Google+[9], Tuenti[10], Whatsapp[11], Twitter[12] o Tumblr[13]. Estas redes sociales se caracterizan por el alto nivel de comunicación entre los distintos usuarios. Los usuarios publican cualquier tipo de contenido, desde una experiencia personal, un sentimiento propio, una opinión o simplemente comparten información publicada en otro medio. Cada publicación va acompañada de una sección de comentarios en la cual los usuarios interaccionan, llegando en ocasiones a generarse verdaderas

[1] http://www.lastfm.es/
[2] http://instagram.com/
[3] https://www.youtube.com/
[4] http://www.linkedin.com/
[5] http://www.timpik.com/
[6] http://www.researchgate.net/
[7] http://www.lecturalia.com/
[8] https://www.facebook.com/
[9] https://plus.google.com/
[10] https://www.tuenti.com/
[11] http://www.whatsapp.com/
[12] https://twitter.com
[13] https://www.tumblr.com/

conversaciones donde los usuarios intercambian sus puntos de vista.

Una característica de las redes sociales es su funcionamiento en tiempo real. Cuando un usuario publica un contenido, en ese mismo instante es visto, leído e incluso contestado. Además, debe remarcarse el hecho de que cada publicación está expuesta a una ingente cantidad de personas en todo el mundo. Esta inmediatez en la comunicación es la principal característica de los plataformas de *microblogging*. Los sitios de *microblogging* son redes sociales horizontales que restringen el tamaño del mensaje compartido, el cual puede ser principalmente texto (Twitter) o multimedia (Tumblr), con el fin de que la comunicación sea más inmediata y concisa. El principal representante de las plataformas de *microblogging* es Twitter. La gran masa de usuarios de Twitter hace que fluyan enormes cantidades de información cada segundo. Los usuarios de Twitter tienen toda la libertad de publicar lo que le venga en gana, como puede ser lo que está haciendo en ese momento, lo que está ocurriendo a su alrededor, compartir un contenido publicado en otra página web, o una sesuda opinión concentrada en los 140 caracteres a los que Twitter restringe los mensajes. El contenido subjetivo de la información que se publica en Twitter es elevado, ya que los usuarios suelen compartir su experiencia con productos o servicios que adquieren. Pero los usuarios, no solo comparten sus experiencias, sino que manifiestan libremente su opinión sobre los temas más variados: política, cine, literatura, deporte, música, etc.

La información subjetiva, o dicho de otra manera, la opinión está muy presente en Internet. Ésta omnipresencia se debe a que la manifestación del parecer, como se ha dicho en la sección anterior, es propia de la condición humana, por lo que no es de extrañar que en una de las principales herramientas de comunicación existente en nuestros tiempo esté plagada de opiniones. En la Sección 1.1 se habla de la necesidad de las personas por conocer experiencias similares, es decir las opiniones de sus semejantes para facilitar sus tomas de decisiones. Internet ensancha la probabilidad de encontrar esas experiencias similares, pero la vasta cantidad de opiniones que fluye a través de la Red hace imposible que un humano sea capaz de procesar todas las que pueden influir en su decisión final. Aquí surge la primera justificación para el estudio de la opinión. Las personas necesitan de sistemas, que al igual que los actuales recuperadores de información o buscadores, les permitan acceder a las opiniones útiles para su proceso de decisión.

Ya se ha indicado que la opinión también es un juicio de valor sobre los temas más dispares, que la opinión también es crítica, y que esa opinión tiene dos interesados, el ente objeto de la crítica, y las personas

interesadas en el tema en el que se centra la crítica. Los entes (personas, compañías, instituciones, etc.) sobre los cuales se opinan tienen un alto interés en tener acceso de la manera más rápida, clara y concisa posible a la orientación de las opiniones que se están vertiendo sobre ellas. Aquí surge la segunda justificación para el estudio de la opinión ¿no quisieran los partidos políticos conocer la orientación de la opinión de sus potenciales votantes? ¿No quisiera saber una institución, como puede ser una universidad, saber la opinión de sus estudiantes? ¿No quisiera una empresa enterarse de la opinión de sus clientes? Pues la respuesta es muy clara, sí, sí la quieren conocer, y en el menor tiempo posible.

Como se irá indicando a lo largo de esta memoria, el estudio de la opinión, el desarrollo de sistemas de clasificación automática de la orientación de la opinión se viene realizando desde hace tiempo, pero la lengua que ha acaparado la atención de la comunidad investigadora ha sido la de Shakespeare, el inglés. En el campo de las Tecnologías del Lenguaje Humano (TLH) o del Procesamiento del Lenguaje Natural (PLN) los idiomas que no son el inglés son estudiados por una comunidad infinitamente menor. Hay que resaltar que la segunda lengua del mundo y la tercera en Internet (Fernández Vítores, 2015), el español, cuenta actualmente con una reducida investigación del tratamiento automático de opiniones. Ésta, y no otra, es la tercera justificación para la realización de una tesis titulada Análisis de Opiniones en Español.

1.3. Organización de la memoria

Estando bien justificada la perentoria necesidad de elaborar técnicas automáticas de procesamiento de textos de opinión escritos en español, la presente memoria describe la investigación que se ha llevado a cabo en este campo del Procesamiento del Lenguaje Natural. La memoria se ha estructurado en ocho capítulos que tratan de exponer el estudio que se ha realizado. Con la intención de que el lector tenga una visión inicial de cada uno de ellos antes de comenzar su lectura, se exponen a continuación sus elementos principales:

Capítulo 1: El presente capítulo ha expuesto la necesidad que tenemos las personas de conocer la opinión de nuestros semejantes. Además, ha remarcado que la opinión no solo se materializa en la comunicación oral, sino también en la escrita y por extensión en Internet. Por último, ante la necesidad de las personas de conocer la opinión y la disponibilidad de una enorme cantidad de ellas en Internet, se ha justificado el estudio de técnicas de análisis automático de opiniones.

Capítulo 2: Tratará de definir y poner en contexto la tarea de Análisis de Opiniones, así como realizar una revisión del estado en el que se encuentra la investigación actualmente.

Capítulo 3: Este capítulo se centrará en la introducción de la investigación que se ha llevado a cabo, justificando primeramente el porqué del estudio del español, definiendo los dos tipos de textos que se han tenido en cuenta, y determinando el nivel de análisis de opinión que ha protagonizado la investigación.

Capítulo 4: Desarrollará la experimentación que se ha realizado con métodos de aprendizaje supervisado para la identificación de la orientación de la opinión.

Capítulo 5: También se ha desarrollado un estudio de metodologías basadas en aprendizaje no supervisado, con el fin de conocer su nivel de éxito en la clasificación de la polaridad de una opinión. Este capítulo detallará la experimentación subyacente al análisis de metodologías no supervisadas de clasificación de la opinión.

Capítulo 6: Toda tarea de PLN requiere de recursos lingüísticos que incorporen la información necesaria para facilitar la elaboración de conocimiento. Este capítulo describirá las técnicas que se han llevado a cabo para la generación de recursos para el análisis de opiniones en español.

Capítulo 7: Se circunscribirá al estudio de si la combinación de clasificadores y recursos lingüísticos tiene un aporte positivo a la clasificación automática de opiniones.

Capítulo 8: Es el capítulo donde se recogerán las conclusiones de la investigación que se ha desarrollado. Además, se indicarán cuales serán los pasos siguientes a dar en la investigación del tratamiento automático de la opinión.

2

Análisis de la Opinión

2.1. Introducción

Cuando se habla del tratamiento de un fenómeno lingüístico es muy común pensar que la ciencia encargada de su estudio es la Lingüística. A esa conclusión llegarán muchos lectores que se acerquen a esta tesis titulada "Análisis de Opiniones en Español". Pero no hay que extrañarse si son pocos los que intuyen que ese tratamiento será automático y realizado por un ordenador, dado que el título no da pie a ello. La función de las primeras secciones de este capítulo va a ser esa, situar el título de esta memoria, y más aún, contextualizar el trabajo que se expone en un determinado campo de la ciencia, de manera que sirva de base para su desarrollo ulterior.

Como era de esperar, la Lingüística es una de las ciencias que ha dirigido su mirada al estudio de la opinión, o mejor dicho, del lenguaje subjetivo. El lenguaje subjetivo se define como el lenguaje empleado para expresar estados personales (*private states*) en el contexto de un texto o de una conversación (Wiebe et al., 2004). Según Quirk et al. (1985) los estados personales son los que se reflejan en las opiniones, en las evaluaciones, emociones y en las especulaciones. Quirk et al. (1985) define la opinión como una evidencia de un estado personal, a lo que se le podría añadir, creencia o actitud sobre algo. En el Capítulo 1 se describe el proceso de petición de opinión como la pretensión de insertar el conocimiento, o parafraseando a Quirk, "de inyectar el estado personal de una persona perteneciente a nuestro círculo de confianza en nuestro proceso de toma de decisiones". Amplio ha sido el estudio del lenguaje subjetivo en la Lingüística. Como muestra, se pueden mencionar los trabajos relacionados con la identificación de categorías de oraciones subjetivas en (Doležel, 1973) y (Uspensky, 1973), el estudio de la subjetividad desde un punto de vista pragmático (Kuroda, 1973, 1976), o en el contexto del discurso (Chatman, 1978), el análisis del estilo lingüístico para expresar conciencia (Cohn, 1978), la descripción del lenguaje empleado para la descripción de contextos opacos (Fodor, 1979), y la amplia descripción que Banfield (1982) sobre su teoría de subjetividad y comunicación. Todo repaso relacionado con el estudio de la Lingüística sobre el lenguaje subjetivo no puede dejar de lado los trabajos de Wiebe & Rapaport (1986, 1988, 1991) y Wiebe en solitario (Wiebe, 1990, 1994).

El análisis de la opinión que se presenta en esta obra, aunque relacionado, no es lingüístico, sino computacional, y es una tarea de investigación que se circunscribe en la compleja, ardua y cargada aún de retos por superar, área del Procesamiento del Lenguaje Natural. Pobre quedaría esta obra sobre el análisis automático de opiniones si no se detuviera por un momento en su fuente metodológica, de donde extrae los fundamentos para poder escudriñar el lenguaje con el fin de identificar

la intención de la persona que lo usa. La siguiente sección tiene como fin presentar la matriz del Análisis de Opiniones, el Procesamiento del Lenguaje Natural.

2.2. Procesamiento del Lenguaje Natural

2.2.1. Definición

La comunicación es el instrumento esencial de relación entre elementos con capacidad de relación. La comunicación se puede establecer entre elementos de la misma o disímil naturaleza, pero suele ser más común entre individuos semejantes. La comprobación de la anterior afirmación se encuentra en que la comunicación entre seres humanos es más frecuente, fluida y natural, que la comunicación entre personas y animales. Para que se establezca una comunicación no sólo tienen que estar al menos dos elementos dispuestos a comunicarse y con capacidad de ello, sino que además debe existir un protocolo de comunicación. Cuando la comunicación se produce entre personas, el protocolo de comunicación recibe el nombre de lenguaje o lenguaje natural, el cual se puede manifestar de forma oral o escrita. Comunicación se produce entre seres vivos como los animales, entre entes como pueden ser las administraciones, y entre objetos lógicos como pueden ser los programas informáticos, todos ellos con sus específicos protocolos de comunicación. También existe comunicación entre individuos de las categorías anteriormente mencionadas, como la comunicación existente entre una persona y su mascota, entre una administración y el administrado, o incluso entre una aplicación informática y una persona. Se podría afirmar que existe comunicación siempre y cuando exista relación. La presente sección va a fijar la atención en la comunicación que surge de la relación entre personas y máquinas.

La informática, o la ciencia de la computación, es una ciencia que estudia el procesamiento de información y la automatización de tareas con el fin de facilitar dichas labores a las personas. De la aserción anterior se desprende que las computadoras se tienen que relacionar con personas para poder allanar sus quehaceres, por lo que debe existir un protocolo de comunicación entre ellos para que la relación sea fructífera. Dado que el objetivo es simplificar la vida de las personas, el protocolo de comunicación que se emplee tiene que ser lo más cercano posible al que emplean las personas de manera natural. Las computadoras son máquinas, y hasta la fecha, todavía no están dotadas de la inteligencia suficiente para poder entender y generar lenguaje natural de igual manera que los humanos. Con la intención de que las computadoras lleguen a relacionarse naturalmente

con las personas, surge a mediados del siglo XX una nueva disciplina dentro de las ciencias de la computación, el Procesamiento del Lenguaje Natural (PLN), en inglés *Natural Language Processing* (NLP). El PLN se puede definir como: *Disciplina centrada en el diseño e implementación de aplicaciones informáticas que se comunican con personas mediante uso de lenguaje natural* (Dale et al., 2000). De la definición se extrae fácilmente la conclusión de que la cima del PLN es el desarrollo de sistemas lo suficientemente inteligentes para que la interacción persona-máquina se produzca usando lenguaje natural.

El objeto de estudio del PLN no es otro que el lenguaje. Si se quiere conseguir que las máquinas tengan la capacidad de procesar y generar lenguaje, se tiene que llegar a una comprensión profunda de la lengua con el fin de poder diseñar los formalismos necesarios para que una máquina pueda trabajar con lenguaje natural. La lingüística tiene como meta la caracterización y explicación de las distintas figuras lingüísticas presentes en el uso de la lengua. Parte de esa aprehensión de las figuras lingüísticas tiene que ver con el proceso de adquisición, generación y compresión del lenguaje, otro tanto con la intelección de la relación entre las diversas expresiones lingüísticas y el contexto en el que se producen, y otra parte con la interpretación de las estructuras lingüísticas, gracias a las cuales la lengua es instrumento de comunicación. De esos tres niveles de análisis del lenguaje, el PLN concentra su atención en el último, es decir, en el entendimiento e interpretación de las estructuras lingüísticas. Este estudio, intensificado durante la segunda mitad del siglo XX, ha estado regado por la controversia entre dos paradigmas que formalmente se pueden denominar, lingüística generativa y tecnología de la lengua[1], e informalmente como enfoque basado en reglas lingüísticas y orientación basada en estadística (Manning & Schütze, 1999a).

El paradigma de lingüística generativa dentro del PLN también recibe el nombre de racionalista o *Chomskiana*, debido a que Noam Chomsky es su precursor más prominente. El periodo de esplendor de las teoría racionalista tuvo lugar entre los años 1960 y 1985, en la que la mayoría de los lingüistas y los primeros investigadores en el área del PLN estaban convencidos de que el núcleo central del conocimiento lingüístico de las personas no es adquirido, sino que se encuentra codificado de alguna manera en la mente humana. En otras palabras, la hipótesis sobre la que se asienta la teoría racionalista está en que la esencia del lenguaje no es adquirida por las personas a través de los sentidos, sino que se encuentra esculpida en las concavidades del cerebro humano. Hace recordar esta concepción de la lengua a la manera

[1] También conocido como Ingeniería del Lenguaje o PLN estadístico

en que Platón percibía la realidad, dividida en un mundo real y en un mundo de las ideas, siendo este último no más que una imagen concreta del primero. Fiel reflejo de esta creencia es la afirmación "la pobreza de los estímulos" enunciada por Chomsky en (Chomsky, 1986). Chomsky sostiene que el lenguaje es un mecanismo tan complejo que es harto complicado adquirirlo exclusivamente a través de los sentidos, por lo que no cabe duda que en la mente humana se encuentran definidas un conjunto de reglas lingüísticas que son las que permiten a los humanos poseer capacidad de utilizar la lengua.

La estrategia de afrontar los problemas relacionados con el PLN desde una perspectiva de la tecnología de la lengua formalmente también se la denomina teoría empírica. Los defensores de la teoría empírica no son detractores de la hipótesis generativa en el sentido de que la rechazan de plano, dado que en cierta forma aceptan que en la mente humana se encuentre cincelados elementos básicos de conocimiento lingüístico. En otras palabras, se podría decir que la teoría empírica parte de la premisa de que la mente humana no es una *tabula rasa*. Como mínimo, la teoría empírica acepta que el cerebro de las personas nace con la capacidad de establecer asociaciones, de reconocer patrones y de generalizar ocurrencias de eventos que perciben a través de los sentidos, y por ende, de inferir conocimiento. Partiendo de esas habilidades innatas, los humanos, por medio de la interpretación involuntaria, o mejor dicho, natural de los eventos que continuamente capta, van interiorizando progresivamente la estructura del lenguaje hasta llegar a un punto en que son capaces de entenderlo y generarlo de manera completamente natural. Si mientras leemos estas líneas conseguimos abstraernos un poco, comprobaremos que de manera espontánea brota la idea de que nuestro cerebro está continuamente contando eventos, aplicándoles operaciones aritméticas, estableciendo asociaciones en función del resultado de dichas operaciones, aplicándoles patrones y, finalmente asignándoles una etiqueta que depende del resultado de todos esos cálculos. Esta idea es el hormigón de los cimientos de la ingeniería de la lengua, la cual nos dice que es factible el aprendizaje de la compleja y extensa estructura de la lengua por medio de la definición de un modelo de lenguaje, cuyos parámetros pueden ser inducidos mediante el uso de métodos estadísticos, reconocimiento de patrones, y aprendizaje automático sobre una cantidad considerable de ejemplos de uso de lenguaje.

Una vez expuestas las dos escuelas principales dentro de la ciencia que estudia la lengua, es probable arribar a la conclusión simplista de que los seguidores de las ideas de Chomsky no son más que teóricos de

la lengua, mientras que los empíricos son aquellos que intentan buscar soluciones prácticas a la hora de modelar el lenguaje. Empero, para alcanzar un completo entendimiento de las dos teorías hay que conocer su principal desemejanza, la cual recae en que ambas escuelas intentan describir elementos diferentes de la lengua. La lingüística generativa persigue la descripción de la lengua definida en el cerebro humano, la cual es nombrada como Lenguaje-I (*I-language*). Para llegar a describir el arquetipo de lenguaje residente en el cerebro, la lingüística generativa emplea las derivaciones concretas de dicho modelo de lenguaje, que no son otras que las que se encuentran impresas en textos. A las reproducciones físicas del Lenguaje-I se les conoce como Lenguaje-E (*E-language*). Por contra, los ingenieros del lenguaje dejan de lado el Lenguaje-I, y pretenden modelar el lenguaje mediante el análisis exhaustivo del Lenguaje-E.

Esta memoria no va a revelar el arquetipo de opinión impreso en nuestro cerebro, sino más bien exponer la investigación sobre cómo modelar la opinión a través del estudio profundo de opiniones, es decir, de Lenguaje-E. Por lo tanto, es momento de afirmar que esta memoria relata una investigación circunscrita en el enfoque ingenieril del PLN, el cual también es conocido como Tecnologías del Lenguaje Humano (TLH).

2.2.2. Flujo de trabajo

En el apartado anterior se ha indicado que las TLH tratan de modelar el lenguaje mediante el estudio en profundidad de ejemplos del propio lenguaje. Teniendo como meta la representación abstracta de la lengua, el PLN o las TLH descomponen dicho estudio en fases o etapas de procesamiento. Tradicionalmente, los sistemas de PLN, tomando como referencia la teoría lingüística, dividen el procesamiento de un texto en análisis sintáctico, semántico y pragmático. Siguiendo esta descomposición, los sistemas de PLN se fijan primeramente en estudiar la estructura sintáctica de las oraciones de un documento. Posteriormente centran sus esfuerzos en identificar el significado literal de las oraciones, y como último hito se haya el estudio pragmático de las oraciones, es decir, la identificación del significado de las oraciones dentro del contexto marcado por el documento en estudio. Otras interpretaciones de los niveles de análisis del PLN categorizan los dos primeros niveles como análisis a nivel de oración, mientras al análisis pragmático lo suelen catalogar como análisis del discurso. De otra manera se podría decir que la primera aproximación describe los sistemas PLN de una manera estratificada (sintaxis, semántica, pragmática), mientras que la segunda hace más hincapié en el nivel de profundidad del análisis (oración, discurso) que aplican a los documentos.

La complejidad de la lengua no ayuda a la hora de determinar los distintos análisis que hay que aplicar para obtener una representación del mismo, pero desde un punto de vista técnico y didáctico, la definición estratificada es la más iluminadora, por lo que es la que se va a seguir de aquí en adelante.

La identificación de solo tres capas de análisis no es completamente suficiente para el procesamiento de documentos en un entorno real, debido principalmente a que no es posible lanzarse a la elaboración de un análisis sintáctico de un texto sin haber previamente identificado los elementos mínimos de información, las palabras; los mínimos constituyentes sintácticos, las oraciones; así como la información morfológica de los términos que componen la frase. Por consiguiente, al sistema base de PLN que se ha comenzando a definir en esta sección hay que añadirle dos nuevas capas de análisis: *tokenización* y análisis léxico. En la Figura 2.1 se muestra gráficamente el flujo de trabajo de un sistema de PLN.

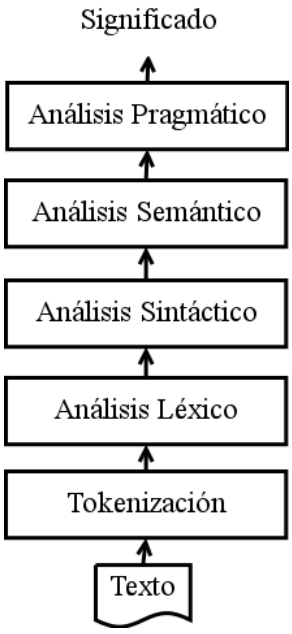


Figura 2.1: Flujo de trabajo en un sistema de PLN.

Sin ánimo de ser muy exhaustivos, y debido principalmente a que a lo largo de esta memoria se mencionarán las distintas fases anteriormente resaltadas, a continuación se va simplemente a indicar la función de cada una de ellas en un sistema de PLN. La primera de las capas de procesamiento es la de *tokenización*, la cual trata de identificar por un lado las unidades que constituyen el mensaje, que también se conocen como *tokens*, y por otro lado las mínimas unidades de significado que son las oraciones. Una vez identificados los elementos que constituyen el mensaje, es hora de obtener información de ellos teniendo en cuenta la función que desempeñan dentro de la oración. De la obtención de esta información se encarga la etapa del análisis léxico. Una vez obtenida la información morfológica, es momento de descubrir la estructura sintáctica de la oración, de las relaciones existentes entre los distintos elementos, y de las dependencias existentes entre ellos. Cuando finaliza su tarea el analizador sintáctico, se tiene una cantidad de información que incluso permite ser escudriñada con el fin de comenzar a extraer la semántica o el significado que el autor del mensaje quiere transmitir. El analizador semántico trata de descubrir el significado de cada frase,

pero el sentido de un mensaje no es el resultado de la suma individual de los sentidos de las frases que lo constituyen. Las oraciones están relacionadas entre sí, y son esas conexiones y referencias entre ellas lo que construyen el discurso, cuyo significado es la cima que todo sistema de PLN busca coronar. Esta última etapa de la ascensión al significado es liderada por el analizador pragmático.

El PLN no sólo trata de entender el lenguaje por medio de su representación y procesamiento, sino que también persigue la generación de lenguaje, es decir, el proyectar una representación determinada de conocimiento en palabras correctamente hilvanadas y relacionadas de manera que sean naturalmente comprendidas por una persona. Si se consulta la bibliografía relacionada con PLN, la diferencia entre publicaciones relacionadas con el análisis del lenguaje y la generación de lenguaje es abultada y a favor del primero. Dos son las principales teorías que intentan explicar la desventaja de la generación con respecto al procesamiento. La primera, y menos consistente, es la que considera que la generación del lenguaje no atrae la atención de los investigadores porque no es un problema con el grado de complejidad suficiente para considerarlo un reto. Parece que los valedores de esta hipótesis no son conscientes de lo difícil que es para algunas personas el poder comunicarse mediante el uso de la lengua de una manera clara y fluida, por lo que se podría decir que la experiencia invalida esta teoría. La segunda teoría, desde su enunciación, presenta más consistencia. Esta segunda hipótesis nos dice que el reducido número de investigaciones científicas relacionadas con la generación del lenguaje se debe a su extremada complejidad. Su aserción se asienta en que el procesamiento del lenguaje se cimenta en la observación de ejemplos de lenguaje, es decir, la entrada de los sistemas de análisis de la lengua es conocida a priori. Los sistemas de generación de lenguaje no cuentan con la fortuna de conocer a priori la entrada, dado que todavía no existe una representación única del conocimiento que permita su traducción a lenguaje natural. Por ende, la construcción de un sistema de generación del lenguaje tiene que comenzar forzosamente por el diseño de una representación del conocimiento que se necesita para la organización correcta de las palabras. Éste y no otros, es el motivo por el cual la investigación relacionada con la generación del lenguaje es mucho más reducida que la ligada al análisis del lenguaje. El Análisis de Opiniones, tiene como último fin la clasificación de las opiniones vertidas por los hablantes, no va detrás de la producción de opiniones, por lo que se encuadra sin ningún género de dudas en el grupo de análisis de la lengua dentro del PLN.

2.2.3. Aplicaciones

Como se ha indicado a lo largo de esta sección, el PLN es la disciplina que se encarga del desarrollo de las metodologías, los métodos y los algoritmos necesarios para representar la estructura del lenguaje con el fin de poder entender su significado y poder producir lenguaje, con el objetivo siempre de conseguir que las personas se puedan comunicar con las máquinas empleando lenguaje natural. Por tanto, el abanico de aplicaciones es tan abierto como el uso de la propia lengua. Éstas son las aplicaciones posibles, pero no las abordables con el estado actual de la técnica. Desde que la necesidad de contar con procedimientos que permitieran la traducción automática de textos durante la Segunda Guerra Mundial, sirviera de reclamo para despertar el interés en el estudio de procedimientos automáticos de procesamiento de la lengua, muchas han sido las aplicaciones que han ido brotando, estando todas ellas siempre impulsadas por una necesidad evidente de la sociedad. Algunos de los problemas a los que el PLN aporta su granito de arena para encontrar una solución son:

Traducción automática La traducción automática sigue siendo todavía un reto dentro del PLN, debido fundamentalmente a que todavía no se ha conseguido que las máquinas puedan convertir, al igual que un humano, un texto escrito en un idioma en otro totalmente distinto. Otro factor que empuja a los profesionales del PLN a seguir buscando la *piedra rosetta* automática es la gran diversidad idiomática existente en el mundo. Internet ha permitido que cualquier persona pueda acercarse a todo tipo de información y además en cualquier idioma. La posibilidad de que se pueda traducir a cualquier idioma un fragmento de texto, posibilita a una persona profana en una lengua comprender documentos escritos en dicho idioma, porque un traductor automático lo ha convertido a una lengua inteligible para dicha persona.

Recuperación de Información De Internet podríamos decir que es un manantial de información, del cual los datos e información no cesan de manar. Al igual que el agua que nace de una fuente natural, que si no se canaliza, trata y proporciona de manera adecuada, no es útil para el consumo humano, esa inmensa cantidad de información que fluye a través de Internet no sería accesible, si no se cuenta con sistemas que permitan recuperar la información pertinente que una persona requiere en un momento determinado. La Recuperación de Información sigue siendo una disciplina con un alto nivel de investigación, porque la necesidad de disponer de la manera más

inmediata posible de información fiable que sacie la demanda de ésta de una persona es cada vez más exigente.

Recuperación de Respuestas En ocasiones la demanda de información consiste en una pregunta muy concreta, y por lo tanto debe tener una respuesta precisa. Verbigracia, si un sistema recibe como consulta "¿Cuál es la superficie de El Escorial?" como respuesta debería devolver los metros cuadrados exactos que tiene de superficie El Escorial. Aunque parezca sencillo y una tarea similar a la Recuperación de Información, no lo es ni mucho menos. Un sistema de Recuperación de Respuestas debe en primera instancia entender la pregunta que se le realiza. La pregunta elegida como ejemplo plantea un problema de desambiguación importante, ya que al cuestionarse sobre la superficie de El Escorial no se determina si se inquiere por la del municipio San Lorenzo de El Escorial o por la del Monasterio. Una vez entendida completamente la pregunta, la siguiente acción es encontrar la información que responde a la consulta, posteriormente extraer la respuesta exacta y, por último, presentar la respuesta al usuario.

Extracción de Información Es la tarea que se encarga de explorar minuciosamente un texto o un mensaje para extraer aquellos elementos de información que son de interés. Por ejemplo, en los textos relacionados con anuncios de eventos de ocio, la identificación de la fecha, hora y lugar en la que tendrá lugar es muy importante, así como es igual de relevante la extracción de los artistas que van a participar. Un gran número de situaciones son las que necesitan de sistemas que permitan extraer entidades en ingentes cantidades de información.

Simplificación de Textos La Simplificación textual se podría definir como la traducción de un texto complejo en un texto mucho más asequible para su entendimiento, en otras palabras, más simple. No se debe confundir la Simplificación de Textos con la tarea de Resumir Textos, ya que un resumen no es igual que una versión simplificada de un documento. Un resumen expone la idea principal de un texto, mientras que un texto simple es el mismo texto complejo pero dispuesto de una manera mucho más apta para su compresión. Un sistema ideal de Simplificación de Textos requiere incluso de la generación de lenguaje, dado que en muchas ocasiones la simplificación precisa parafrasear el texto.

Las posibles aplicaciones del PLN no terminan con esta mínima enumeración, son indudablemente un número mayor, pero el objetivo de

esta memoria no es realizar un tratado sobre las aplicaciones del TLH, sino que se va a centrar en una de esas aplicaciones con profusión, y que está anunciada en el título de esta memoria, el Análisis de Opiniones.

2.3. Análisis de Opiniones

Una vez cumplida la misión propedeútica de la Sección 2.2, y antes de iniciar la exposición de la investigación sobre la que se sustenta esta tesis doctoral, es momento de saber qué es el Análisis de Opiniones (AO), y de conocer la progresión de la investigación sobre el estudio de la opinión.

Todo camino hacia un lugar tiene un punto de partida. El sendero que se va a recorrer en esta sección va a permitir al lector impregnarse gradualmente del conocimiento que se ha ido construyendo en torno al estudio de la opinión, teniendo como meta que el lector tenga una cierta conciencia sobre cual es el estado de la tarea de AO. Pero al igual que un camino termina en un punto, se inicia en otro, y el de este recorrido va a ser la definición de AO:

> *Tratamiento computacional de opiniones, sentimientos y subjetividad en textos.*

La definición de Pang y Lee enunciada en (Pang & Lee, 2008) es la más seguida por la comunidad investigadora en AO. Cambria & Hussain (2012) consideran la definición de Pang y Lee algo general, por lo que ellos proponen una nueva definición que intenta detallar algo más la tarea del AO:

> *Conjunto de técnicas computacionales para la extracción, clasificación, comprensión y evaluación de opiniones expresadas en fuentes publicadas en Internet, comentarios en portales web y en otros contenidos generados por usuarios.*

La definición de Cambria y Hussain es algo más concreta, y sólo menciona como objeto de estudio a las opiniones, dejando de lado a los sentimientos y a la subjetividad. Dejar de lado a los sentimientos puede no ser del todo correcto, porque una opinión favorable puede estar producida por un sentimiento positivo, pero en relación a la subjetividad hay que decir que sí es acertada su exclusión de la definición, dado que, como se verá más adelante, las opiniones no se encuentran solamente en los textos subjetivos, sino también en proposiciones objetivas, las cuales pueden expresar perfectamente una opinión. Además, la definición de Cambria y

Hussain anuncia las tareas que abarca el AO. Hasta la fecha la definición de Pang y Lee es la más extendida, pero no será extraño que con el paso del tiempo se vaya consolidando la propuesta por Cambria y Hussain.

2.3.1. Reseña histórica

Aunque el estudio en profundidad del AO va ligado al inicio del siglo XXI, en el siglo XX se publicaron algunos trabajos que se consideran como precursores del gran número de publicaciones que hoy en día aparecen en las diversas revistas y congresos científicos. Carbonell (1979) propone en su tesis un modelo computacional para representar pensamiento subjetivo de las personas. Carbonell sostiene que los pensamientos que tienen las personas están motivados por los diferentes objetivos que persiguen a través de sus acciones diarias. Por lo tanto, Carbonell se atreve a intentar entender la ideología y los rasgos de personalidad de las personas mediante el estudio de la subjetividad que expresan sus textos. Unos años más tarde en (Wilks & Bien, 1984) también es presentado un estudio sobre la opinión o creencia que tiene un sujeto sobre otro a partir de un diálogo entre ambos. Estos dos estudios sobre las creencias personales no fueron continuados por análisis sobre la opinión, sino que la investigación viró ligeramente la atención hacia la interpretación de metáforas (Hearst, 1992), clasificación de lenguaje afectivo y emocional (Kantrowitz, 2003) e identificación de bloques textuales subjetivos con el punto de vista del autor sobre un determinado agente (Wiebe & Bruce, 1995).

Llegó el siglo XXI y con él, un paulatino incremento en el interés por conocer la actitud de las personas hacia determinados agentes, en definitiva, en procesar automáticamente la opinión de la personas. Pero cabe preguntarse la razón por ese repentino interés en el estudio de la opinión, por esa finalización temprana del periodo de incubación del estudio tranquilo de la opinión. Pang & Lee (2008) mencionan tres posibles factores que ayudaron a sacar de la incubadora el estudio de la opinión:

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.

Pang & Lee (2008) intentan explicar las razones de la casi generación espontánea del celo por el estudio de la opinión desde un punto de vista investigador. Pero cuando la comunidad investigadora, o un simple investigador, se adentra en el análisis de un problema para la consiguiente búsqueda de la solución, se debe, esencialmente, a que su radar de necesidades irresolutas le marca que está surgiendo o emergerá una necesidad a la que se le tendrá que proporcionar una solución. Las necesidades también van precedidas por un interés empresarial, el cual, ante el simple olor de la posibilidad de satisfacer a su clientela y adelantar a la competencia insta a la resolución de la necesidad. Como ya se vio en el Capítulo 1, la necesidad de conocer el parecer de las personas siempre ha existido, y el deseo de conocer el estado de satisfacción de los clientes por parte de las compañías también ha estado presente desde los inicios de los intercambios de servicios entre personas. Por lo tanto, ese auge en el estudio de la opinión no sólo ha estado motivado por la mayor disponibilidad de recursos, que es en lo que se resumen las tres razones aportadas por Pang & Lee (2008), sino también ha estado empujada por el todavía insatisfecho deseo de conocer la opinión de las personas.

Nacido el interés en la búsqueda de la solución del problema del descubrimiento de la orientación de la opinión, aparece la obligación de asignarle un nombre que denomine al proceso de exploración de la solución. Diferentes han sido las maneras de nombrar el estudio de la opinión desde que se comenzara su investigación, entre las que nos encontramos *Opinion Mining*, *Sentiment Analysis*, *Opinion Extraction*, *Sentiment Mining*, *Subjectivitiy Analysis*, *Affect Analysis*, *Emotion Analysis*, *Review Mining*, etc. La multiplicidad de denominaciones iniciales con el paso del tiempo ha ido reorganizándose en torno a dos nombres que han actuado como sumideros equivalentes, y que son *Sentiment Analysis* y *Opinion Mining*. Los términos ingleses *sentiment* y *opinion* fueron usados indistintamente en los primeros trabajos relacionados con la extracción de la orientación de los textos sobre una determinada temática, específicamente sobre documentos en el dominio bursátil (Das & Chen, 2001), y sobre productos comerciales (Tong, 2001).

La primera vez que fue empleada la denominación *Opinion Mining* fue en (Dave et al., 2003), y de acuerdo con los autores, un sistema de *Opinion Mining* 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. La afirmación de Dave et al. (2003) recuerda a un sistema de AO a nivel de aspecto, cuya definición se tratará más adelante.

Mientras tanto, la primera ocasión en la que se empleó el nombre de *Sentiment Analysis* fue en (Nasukawa & Yi, 2003). Como en el párrafo anterior, hay que adelantarse un poco en la estructura de esta memoria para indicar que Nasukawa & Yi (2003) describen un sistema de clasificación de la opinión a nivel de entidad, es decir, el sistema que presentan no se ciñe a calcular la polaridad de la opinión general que se expresa en un documento, sino que se atreve a obtener la orientación de cada opinión vertida sobre cada uno de los sujetos que aparecen en los documentos.

Sentiment Analysis y *Opinion Mining* son las dos denominaciones mayoritariamente empleadas internacionalmente para designar al estudio automático de opiniones, pero qué nombre se utiliza en español. Los investigadores españoles cuando se adentran en una nueva tarea suelen traducir literalmente las denominaciones, y en ocasiones no se detienen a considerar cual sería la traducción más correcta. En el caso del estudio de la opinión, la comunidad investigadora española ha traducido literalmente la designaciones anglosajonas, por lo que es común encontrarse en artículos científicos el nombre de Análisis de Sentimientos y Minería de Opiniones. 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. Esa preferencia se aprecia claramente en el nombre del único *workshop* sobre el estudio de la opinión en español, TASS: Taller de Análisis de Sentimientos en la SEPLN. Pero parece paradójico, que el uso de la palabra española sentimiento en el nombre de la tarea que estudia la opinión no chirríe a la comunidad investigadora, cuando sí extraña a un lego en la cuestión. Mientras que el significado de *sentiment* en inglés sí hace referencia a opinión o punto de vista, el término sentimiento en español sólo señala al estado anímico de una persona, que puede o no estar causado por alguna circunstancia. Por ende, desde nuestro punto de vista, no se considera adecuado el nombre Análisis de Sentimientos para la tarea que tiene como encomienda el estudio de la opinión. El título de esta memoria es toda una declaración de intenciones, dado que se quiere proponer la denominación **Análisis de Opiniones** (AO) como el nombre que debe emplearse en español a la investigación sobre la identificación, recuperación,

extracción, clasificación, visualizado y verificación de opiniones.

2.3.2. La Opinión

Ya está cincelado el nombre de la tarea pero todavía no se ha determinado ni diseccionado su objeto de estudio, el fenómeno lingüístico cuyo significado se quiere inferir para así conocer la actitud de una persona hacia un determinado sujeto. Parece evidente, pero es menester recalcarlo, el AO persigue el examen de las distintas maneras que una persona tiene de manifestar su posición, criterio o juicio sobre un aspecto determinado que le incumbe, en otras palabras, el AO tiene como fin escrutar opiniones. Tantas maneras hay de expresar un juicio de valor, de interpretarse una oración como susceptible de contener una determinad polaridad, y los diversos elementos que componen una opinión, que es preferible dirigir la definición de una opinión a través de un ejemplo.

> "[1] Mi mujer y yo estuvimos en Jaén el fin de semana pasado. [2] Elegimos un coqueto hotel en el centro de la ciudad. [3] La impresión que nos dejó el personal fue buena. [4] Si digo la verdad, la cama era de lo más confortable. [5] Por el contrario mi mujer no pudo dormir en toda la noche por, según ella, la mala calidad de la almohada. [6] El baño no nos gustó del todo a los dos ya que encontramos algunos pelos que no eran nuestros. [7] Lo mejor del hotel, los desayunos, un día más y reviento."

Leyendo el fragmento anterior con la intención de identificar la actitud del autor del comentario hacia el hotel en el que pernoctó durante el fin de semana ¿qué información útil se puede extraer? Primeramente, hay que descubrir si en todas las oraciones se expone una opinión. Es prácticamente evidente que en la oración [1] no se pone de manifiesto ningún juicio de valor, sino que se describe el hecho de que el autor del comentario y su esposa estuvieron en Jaén, mientras que en el resto de frases ([2] [3] [4] [5] [6] [7]) sí se podrían considerar en cierta medida opiniones. Fijándonos únicamente en las oraciones de opinión se aprecia aquellas que valoran al hotel en su conjunto (coqueto hotel [2]), y otras ([3] [4] [5] [6] [7]) en las que se opina sobre sus componentes. Este elemento es importante tenerlo en cuenta, porque dependiendo del nivel de detalle que exija el estudio, se realizarán análisis a nivel de entidad o a nivel de los distintos elementos en los que se conforma dicha entidad. Otro detalle que se aprecia en el ejemplo expuesto es que no siempre es el autor de la publicación la persona que sostiene la valoración, sino que se pueden indicar opiniones de terceras personas

o entes, como ocurre en la oración [5], donde es la mujer del autor del comentario la que se queja de que no ha podido descansar adecuadamente a causa de que la almohada no era de su agrado.

No sería vano insistir en releer de nuevo el ejemplo, dado que hay otro fenómeno lingüístico que podría ayudar a iluminar el concepto de opinión al lector. Se suele estar en el convencimiento de que la construcción subjetiva es el único instrumento para manifestar una opinión, pero ¿no es la frase [6] una situación más que negativa para un hotel? En la oración [6] el autor del comentario deja bien claro el hecho de que hay pelos en el baño, algo que es todo menos positivo para una hospedería. Por lo tanto, no es preciso afirmar que solamente las oraciones subjetivas son las que contienen opinión. La oración [7] deja entrever una cuestión harto importante a la hora de estudiar opiniones, y no es otra que el contexto del mensaje. El término "reventar" no suele tener una connotación positiva, dado que es rara la persona que le plazca "reventar". Pero, de la expresión "he reventado comiendo" suele inferirse que el sujeto en cuestión ha gozado tanto de la comida, que le era imposible detener la ingestión que lo ha llevado a un estado cercano al de "reventar". Por lo tanto, en la oración [7] la palabra "reviento" debe interpretarse positivamente, y no de manera opuesta como sugeriría un análisis literal de la frase.

A partir de los párrafos anteriores en los que se han desgranado los distintos elementos encerrados en un mensaje valorativo, se va a ir moldeando paulatinamente una definición formal del concepto de opinión. La más de las inmediatas conclusiones a la que se llega tras la lectura de la descripción anterior, es que una opinión está constituida por un objeto o sujeto de la opinión (o) y la orientación de la valoración (p). También se ha visto que la persona que opina no es siempre el autor del mensaje, dado que se pueden describir experiencias de otras personas o entes, por ejemplo: "los ciudadanos tienen muy buena opinión de su alcalde, mientras que la administración regional lo tiene entre ceja y ceja". En el ejemplo, el autor del mensaje no opina sobre el alcalde, sino que indica la actitud contrapuesta hacia el regidor municipal de sus administrados y de la administración regional. Por ende, es importante la identificación del elemento que soporta la opinión o autor de la opinión (h) (en inglés *opinion holder* (Kim & Hovy, 2004)). El párrafo anterior se refería a una opinión que se encuadra dentro del ámbito del turismo, y más concretamente en el dominio de hoteles. La valoración positiva o negativa sobre una hospedería puede depender de muchos factores, y uno de ellos, sino el más importante, es el derivado de la gestión del establecimiento. La gestión puede variar con el tiempo, por lo que una experiencia negativa en un determinado momento, puede

tornarse en positiva en otro distinto. Por lo tanto es importante tomar en consideración el momento temporal en el que se produce la opinión (t).

Ya se han citado cuatro elementos de una opinión con los que se puede construir una primera definición formal de lo que es una fragmento de texto valorativo. Se podría decir que una opinión es la cuádrupla *(o, p, h, t)*, donde o es el objeto de la opinión, p es su polaridad, h el autor y t el momento temporal en la que se produce.

La representación formal de opinión como una cuádrupla es concisa, pero deja aún sin cubrir elementos de los cuales es interesante también inferir la opinión que se manifiesta sobre ellos. Si se repara en la oración [5] del ejemplo anterior, no se está valorando negativamente al hotel en su conjunto, sino únicamente se menciona a la almohada de la cama de la habitación en la que pasaron la noche el autor del comentario y su esposa. Según la definición anterior, la oración es como una opinión negativa hacia el hotel en su conjunto, pero nada más lejos de la realidad, el establecimiento puede ser magnifico aunque la almohada no haya sido del gusto del cliente. Eso sí, puede que las almohadas de ese hotel maravilloso sean el garbanzo negro del conjunto de servicios que se ofrecen. Se deduce, que un objeto de opinión se puede dividir en componentes, pudiéndose identificar los distintos elementos de ese objeto. Ese objeto de opinión partitivo se denominará desde este momento entidad. Una entidad e es un producto, servicio, tema, problema, persona, organización, evento o ente. Esa entidad se representa por el par *(T, W)*, donde T es un jerarquía de componentes y sub-componentes, y W es un conjunto de atributos de e. Cada componente tiene su conjunto de atributos propios.

Con ánimo de ayudar a la comprensión de la definición de entidad se podría usar como ejemplo un teléfono móvil. Un teléfono móvil está formado por distintas partes: una pantalla, una batería, un procesador, memoria de almacenamiento, memoria de ejecución, carcasa... A su vez, cada una de esas partes se pueden dividir en más componentes, de manera que progresivamente se va construyendo una estructura jerárquica en la que el nodo principal es la entidad, que en este caso es el teléfono móvil. A su vez cada uno de los nodos de esa estructura tienen asociados un conjunto de atributos, como en el caso de la batería puede ser la autonomía que le aporta al teléfono o su vida útil. Por tanto, el par *(T, W)* que define a una entidad puede llegar a ser verdaderamente complejo. Tal puede ser el nivel de complejidad al que puede llegar la representación de una entidad, que a la hora de implementar esa representación formal se prefiera una versión simplificada. Esa versión simplificada nos dice que la estructura jerárquica de relaciones “parte de” que conforman la entidad con su componentes y

todos los atributos, debe reducirse a una estructura de dos niveles, en el primero debe seguir estando la entidad, y en el segundo deben estar tanto los componentes como todos los atributos de la entidad. A partir de ahora ya se puede hablar de entidad y aspectos, siendo los aspectos el conjunto de los componentes y los atributos de la entidad.

Tras la descomposición del objeto de la opinión, la cuádrupla que se había empleado para la definición formal de la opinión debe ser redefinida. Por lo tanto, el objeto de estudio del AO, la opinión, se define como una quíntupla, $(e_i, a_{ij}, p_{ijkl}, h_h, t_l)$, donde e_i es el nombre de la entidad, a_{ij} es el aspecto de e_i sobre el que se opina, p_{ijkl} es la polaridad de la opinión que se está vertiendo, h_k es el autor del juicio de valor, y t_l es el momento en el que se manifiesta la opinión (Liu, 2012a). El valor que suele tomar p_{ijkl} por regla general es binario, es decir, que oscila entre *positivo* y *negativo*, aunque la polaridad también puede ser representada en una escala de intensidades, verbigracia, los ya clásicos 5 niveles de satisfacción figurados por iconos de estrellas. La definición cuenta con un caso especial, y es la situación en la que la diana de la opinión es la propia entidad. En ese caso, e_i y a_{ij} conjuntamente se refieren al objeto de la opinión.

Se podría caer en la complacencia de que la definición expuesta cubre todas la maneras posibles de manifestar una opinión, pero no es así. Se ha comentado en párrafos anteriores que la dificultad que entraña el estudio de la opinión obliga a simplificar en cierta manera su tratamiento, por lo que la definición resultante de dicha premisa es lógico que no pueda alcanzar a simbolizar las diversas maneras de expresar un juicio de valor. Es importante repasar y tener en cuenta las limitaciones de la definición, por lo que a continuación se enumeran:

1. Si se piensa con cautela, la definición que se ha expuesto de opinión sólo modela aquellas valoraciones que se refieren a un objeto, a una cualidad del mismo, pero complicado es ajustar la quíntupla de opinión a aquellas valoraciones en las que interviene una relación entre dos entidades. Por ejemplo: “el portátil es demasiado grande para mi mesa de escritorio”. La primera interpretación es que se está emitiendo una opinión negativa sobre el tamaño de un ordenador portátil. Pero la oración debe leerse con más cuidado, ya que el autor está indicando que el portátil es demasiado grande para su mesa, es decir, más que exponer que el portátil tiene un tamaño excesivo para ser un portátil, lo que sería una opinión negativa sobre el aspecto “tamaño del portátil”, lo que el autor quiere expresar es una valoración negativa de la relación existente entre el portátil y su escritorio. Dicho ordenador no es apropiado para la mesa del autor de la opinión, lo cual no

quiere decir que su tamaño no sea adecuado. Estas opiniones que dependen de la relación de dos entidades no se pueden representar con la definición anteriormente dada, pero dicha definición lo que quiere es modelar formalmente la opinión independientemente del contexto, en otras palabras, para tratar opiniones como las del ejemplo se necesita entender el contexto, lo cual hace que se tenga que adentrar en otras tareas de las TLH.

2. En el camino que se ha ido recorriendo para llegar a la definición de opinión, se ha relatado como una entidad se subdivide progresivamente en un conjunto considerablemente grande de partes, que a su vez tienen asociadas, en ocasiones, un ramillete importante de atributos. Dicha división confluye en una estructura jerárquica cuya computación puede llegar a ser verdaderamente compleja. Por ello se indica que en el AO se simplifica dicha estructura para sólo tener en cuenta la entidad, los atributos ligados a la entidad, y todos los elementos que componen la entidad. Pero qué hacer ante una opinión sobre un atributo de una parte de una entidad, es decir, ante una opinión sobre la vida útil de la batería de un teléfono móvil. Si se quiere ajustar la opinión a la definición anteriormente dada, la entidad sería el modelo de teléfono móvil en cuestión, mientras que el aspecto sería la vida útil de la batería. Si se quiere afinar más en el análisis y se desea determinar la polaridad de la opinión del aspecto "batería", entonces se tendría que descender en la estructura jerárquica y hacer que la batería ocupe el lugar de la entidad, y "vida útil" sea el aspecto por el cual se opina.

3. Aunque se verá en secciones posteriores, la definición formal de opinión sólo es aplicable a un tipo de opiniones, las conocidas como opiniones regulares. El segundo tipo de opiniones, las conocidas como opiniones comparativas (Jindal & Liu, 2006) necesitan una definición diferente dado que se basan en la contraposición de dos entidades.

Tipos de opiniones

Las infinitas maneras de expresar el parecer sobre un determinado hecho se pueden agrupar siguiendo dos criterios disímiles. Por un lado, se encuentra la diferenciación entre opiniones regulares y comparativas, y por otro, se hallan las opiniones explícitas e implícitas.

Opiniones regulares y comparativas: Las opiniones regulares son aquellas en las que se expresa un punto de vista sobre una entidad o

cualquiera de sus aspectos (Liu, 2006–2011). Dentro de las opiniones regulares se pueden diferenciar las opiniones directas y las indirectas:

Opiniones directas: Son aquellas en las que se manifiesta de una maneara clara y directa una opinión sobre una entidad o uno de sus aspectos, por ejemplo: "La calidad de imagen del monitor es excelente".

Opiniones indirectas: Son aquellas manifestaciones en las que a través de la valoración del efecto que origina la entidad, se expone una evaluación de la entidad. Un ejemplo muy claro ocurre cuando se opina sobre medicamentos. En la oración "el jarabe no consigue quitarme el carraspeo de la garganta" se está exponiendo que el jarabe no está haciendo efecto o que no está realizando adecuadamente su cometido, pero no se está mencionando directamente que el jarabe no es de calidad.

Las opiniones comparativas son aquellas en las que se expone una relación de semejanza o disimilitud entre dos o más entidades y/o la preferencia del autor, basándose para formular la opinión en algún o algunos aspectos compartidos por las entidades en cuestión (Jindal & Liu, 2006). Las opiniones comparativas se escenifican normalmente mediante el uso de las formas comparativas o superlativas de adjetivos y adverbios, siendo la excepción que confirma la regla las expresiones de preferencia.

Opiniones implícitas y explícitas: Una opinión explícita es una opinión regular o comparativa, es decir, una oración en la que se expresa el parecer sobre una entidad o entidades. Verbigracia "las magdalenas están exquisitas"; "La cerveza Cruzcampo elaborada en Jaén es infinitamente mejor que la Mahou". Mientras que las opiniones explícitas son proposiciones subjetivas, las implícitas son enunciados objetivos que implican una opinión regular o comparativa. Por regla general este tipo de oraciones objetivas expresan situaciones deseables o no deseadas, por ejemplo, "la hamaca que compré ha durado dos días"; "un Mercedes tiene más potencia que un Hyundai".

No es extraño que sea mucho más sencillo identificar opiniones explícitas que implícitas, y por tanto que la mayor parte de la investigación actual esté centrada en el trabajo sobre opiniones explícitas. A pesar de ello se podrían destacar dos trabajos en el estudio de la clasificación de opiniones implícitas. Greene & Resnik (2009) aplican un enfoque sintáctico a la identificación de oraciones

objetivas en las que se manifiesta un punto de vista. Los autores afirman que la semántica de un enunciado puede variar en función de la estructuras sintácticas empleadas para su construcción. Por otro lado, se encuentra el trabajo de (Zhang & Liu, 2011) que deja a un lado la sintaxis y se centra en una estrategia combinada de definición de reglas de opinión y de uso de listas de palabras de opinión.

Subjetividad y Emoción

Antes de continuar con la definición del objeto de estudio en AO, y de la propia tarea, es menester despejar una confusión en la que no es complicado tropezar, y no es otra que la distinción entre subjetividad y emoción. Los enunciados de opinión, son proposiciones subjetivas u objetivas que implican una preferencia o un rechazo sobre una determinada entidad, mientras que las frases emotivas son oraciones subjetivas donde se ponen de manifiesto sentimientos y pensamientos.

Las emociones son estudiadas esencialmente por la psicología, filosofía y la sociología. El trabajo sobre la emoción es bastante variado, y va desde el análisis de las respuestas emocionales de reacciones fisiológicas, hasta expresiones faciales, gestos y posturas provocadas por distintos tipos de experiencias que experimenta un individuo. Los investigadores preocupados en el tratamiento de las emociones han intentado, sin mucho éxito hasta la fecha, la categorización de las diversas emociones que puede sentir una persona. A pesar de la falta de acuerdo, algunas clasificaciones están teniendo más predicamento que otras, como es la presentada en (Parrott, 2001). Parrott (2001) identifica 6 emociones primarias: amor, alegría, sorpresa, enfado, tristeza y miedo. Estas emociones a sus vez son divididas en emociones secundarias y terciarias.

Las emociones están íntimamente relacionadas con los sentimientos e incluso con las opiniones. La intensidad de una opinión normalmente está asociada a la magnitud de ciertas emociones, como pueden ser la alegría y el enfado. Las opiniones que se estudian en AO son generalmente evaluaciones. En el contexto de opiniones comerciales, las evaluaciones que realizan los clientes se pueden categorizar en evaluaciones racionales y evaluaciones emocionales (Chaudhuri, 2006).

Evaluaciones racionales: Son evaluaciones basadas en creencias demostrables empíricamente, es decir, fundamentadas en evidencias.

Evaluaciones emocionales: Son las dependientes del estado de ánimo de la persona.

Tras ésta pequeña explicación sobre la investigación ligada a la emoción, se debe hacer hincapié en que no son equivalentes los conceptos de opinión y emoción. Las opiniones racionales (evaluaciones racionales) no expresan ninguna emoción, mientras que las emociones en la mayoría de los casos se deben al estado psicológico de la persona en cuestión. Además, en la expresión de una emoción el objeto de la emoción suele ser el estado anímico del autor de la proposición y no una entidad distinta a él mismo.

2.3.3. Tareas del Análisis de Opiniones

Una vez que se ha definido todo lo concerniente a la opinión, es momento de precisar las tareas involucradas en el AO. Primeramente es menester señalar el objetivo del AO, que no es otro que, dado un documento susceptible de contener opiniones d, identificar cada uno de los elementos de la quíntupla de opinión $(e_i, a_{ij}, p_{ijkl}, h_h, t_l)$ en d.

Que el objetivo del AO sea la de poblar la quíntupla de opinión obliga al AO a dividirse en distintas tareas que se nutren de diversos campos del PLN. Lo primero que hay que afrontar es la identificación de las entidades sobre las que se opinan, en otras palabras, se precisa extraer entidades del documento d. La extracción de entidades es similar al Reconocimiento de Entidades Nombradas (*Named Entity Recognition, NER*) que se realiza en Extracción de Información (Mooney & Bunescu, 2005). En (Hobbs & Riloff, 2010) se puede comprobar que la Extracción de Información es un problema muy interesante y complejo. El lenguaje natural es de todo menos formal, por lo que la referencia a una misma entidad en diferentes proposiciones se puede realizar de diferentes maneras. Por ello, una vez que se han extraído las entidades nombradas del texto, es preciso asociarlas a la entidad a la que realmente hace referencia.

Parece necesario distinguir entre categoría de entidad y expresión de entidad. Cuando se habla de categoría de entidad se está haciendo referencia al nombre de la entidad, mientras que expresión de entidad es la manera de representar a todas las distintas formas que se pueden utilizar para hacer mención a la categoría de entidad. Al proceso de agrupar las distintas expresiones de entidad en una categoría de entidad se le denomina categorización de entidades.

Ya se ha encontrado el primer elemento de la quíntupla, y ahora toca encontrar el aspecto sobre el que se está opinando. Este problema es muy semejante al anterior, dado que un único aspecto puede ser expresado de la más diversas de la maneras. Por ende, hay que distinguir de nuevo entre categoría de aspecto, y expresión de aspecto. Además, hay nuevamente que realizar el emparejamiento entre expresión de aspecto y la categoría

de aspecto que le corresponde. Pero en esta ocasión no es complicado encontrarse con un obstáculo adicional, como es que los aspectos pueden aparecer de manera implícita en los mensajes. Por tanto, hay que distinguir entre expresiones de aspecto explícitas e implícitas.

Expresiones de aspecto explícitas: Normalmente están constituidos por nombres o por sintagmas nominales. Por ejemplo: "la calidad de sonido de unos auriculares Bose es alucinante". Como se puede apreciar con nitidez, se está manifestando una opinión sobre la calidad de sonido, el cual es un aspecto de los auriculares Bose.

Expresiones de aspecto implícitas: Cuando los aspectos están presentes en la oración, pero no están representados por un nombre o por un sintagma nominal. Verbigracia, en la oración, "los auriculares AIWA son muy caros". En este caso se está poniendo de manifiesto que los auriculares son caros, en otras palabras, se está haciendo referencia al aspecto "precio" de los auriculares. Pero el aspecto sobre el que se está opinando no está siendo derivado por un nombre o por un sintagma nominal, sino que está siendo señalado por un adjetivo. La mayoría de las ocasiones se emplean adjetivos y adverbios para hacer referencia de manera implícita a los aspectos de las entidades que son objeto de la opinión.

El tercer componente de la quíntupla es el cálculo de la polaridad de la opinión, es decir, hay que indicar si la valoración que se ha vertido sobre la entidad o sobre un aspecto de la entidad es positiva, negativa, o se acomoda en una escala de intensidades de opinión. El cuarto y el quinto componente de la quíntupla se corresponden con la identificación del autor de la opinión y del momento temporal en el se ha producido respectivamente. Tanto para la identificación del autor de la opinión como para la extracción del tiempo hay que tener en cuenta que, al igual que ocurría con entidades y aspectos, es necesario aplicar un proceso de correspondencia entre expresiones y categorías.

Tomando los elementos que se han ido describiendo en los párrafos anteriores, se puede moldear la estructura formal de una entidad y el esqueleto de modelo de opinión. Formalmente, una entidad e_i se representa por sí misma como un conjunto finito de aspectos $A_i = \{a_{i1}, a_{i2}, ..., a_{in}\}$. La referencia a la entidad e_i puede apreciarse con la detección de un conjunto finito de expresiones de entidad $\{ ee_{i1}, ee_{i2}, ... ee_{in}\}$. Cada aspecto $a_{ij} \in A_i$ de una entidad e_i puede, a su vez, ser expresada por un conjunto finito de expresiones de aspecto $\{ ae_{ij1}, ae_{ij2}, ... ae_{ijm}\}$. Asimismo el modelo de documento de opinión se define como un conjunto de opiniones sobre un

grupo de entidades $\{e_1, e_2, \ldots e_r\}$ y sus aspectos asociados por parte de un conjunto de autores de opinión $\{h_1, h_2, \ldots h_n\}$ en un determinado momento temporal.

Es recomendable finalizar esta sección de definición de la tarea de AO con una síntesis de las tareas que conlleva la extracción de información de opinión de un mensaje:

Tarea 1. Extracción y categorización de entidades: Dado un documento de opinión D, la primera tarea consiste en identificar todas las posibles entidades y agrupar todas las maneras de referenciar a una entidad en torno a su entidad correspondiente. Cada grupo de entidad representa una entidad única e_i.

Tarea 2. Extracción y categorización de aspectos: Tarea similar al anterior, pero en esta ocasión el objetivo de la extracción no es otro que los aspectos, y las diversas formas de expresar dichos aspectos. Una vez terminada la extracción, toca agrupar las distintas expresiones en su aspecto, y asociar cada grupo con su correspondiente entidad. Cada agrupamiento de expresiones de aspecto de una entidad e_i representa un único aspecto a_{ij}.

Tarea 3. Extracción y categorización del autor de la opinión: No salimos de la extracción, porque en esta ocasión hay que identificar el autor de la opinión o como también lo llama Wiebe et al. (2005) fuente de opinión. Al igual que ocurre con las entidades y los aspectos, también hay que aglutinar las variadas maneras de referirse al autor en torno al nombre del mismo.

Tarea 4. Extracción del momento temporal: En el camino que se está recorriendo para completar la quíntupla de opinión, toca ahora detenerse en la identificación del momento temporal en el que tiene lugar la expresión de la opinión.

Tarea 5. Clasificación de la polaridad a nivel de aspecto: Una vez que se conoce la entidad y el aspecto, es momento de clasificar la opinión, es decir, de indicar si el mensaje trasmite una opinión positiva, neutra o negativa.

Tarea 6. Generación de la quíntupla de opinión: La última etapa del camino es tomar todos los elementos identificados y generar la quíntupla de opinión.

El siguiente ejemplo tiene como fin dar aún más luz al proceso que engloba el AO:

> Publicado por: Pedro Martínez. Fecha: 22 de Diciembre, 2014 [1] Durante un mes he estado visitando concesionarios de coches hasta que me he comprado un Hyundai i20. [2] Estuve a punto de comprarme un Ford Fiesta pero probé el Hyundai y me encantó. [3] El Ford tiene un diseño mucho más bonito que el Hyundai. [4] Por contra, el Hyundai tiene un equipamiento mucho más completo que el coche americano. [5] El motor del Hyundai es más potente y silencioso que el del Fiesta. [6] El último detalle que ayudó a mi decisión final fue la gran suavidad de la dirección del i20.

Siguiendo el orden anteriormente indicado, el proceso de estructuración de la información de opinión presente en el ejemplo comienza con la identificación de las entidades. Las dos entidades a las que se hace referencia son Hyundai i20 y Ford Fiesta, dado que son los dos modelos de coche sobre los que se está hablando. Como se puede apreciar se utilizan distintas maneras de referirse a ambos modelos, Hyundai, Ford, coche americano, Fiesta, pero la Tarea 1 los aglutina bajo las entidades Hyundai i20 y Ford Fiesta. La Tarea 2 centra su atención en los aspectos. Se observa con claridad que de manera explícita o implícita se opina sobre el diseño, el equipamiento, la potencia y el ruido que hace el motor. En este caso la identificación del autor no es complicada, dado que el propio autor de la publicación ya se encuentra etiquetado, por lo que la Tarea 3 ya está completada. En el ejemplo, la extracción del tiempo tampoco es un problema dado que no es necesario identificarlo en el texto. La Tarea 5 tendría como resultado que el diseño del Ford Fiesta cuenta con una opinión positiva, que el equipamiento del Hyundai i20 suma en su haber una valoración positiva, al igual que el sonido y la potencia del motor. Para terminar el proceso, la Tarea 6 se corresponde con la generación de las quíntuplas.

(Ford Fiesta, diseño, positivo, Pedro Martínez, 22/12/2014)
(Hyundai i20, equipamiento, positivo, Pedro Martínez, 22/12/2014)
(Hyundai i20, motor_potencia, positivo, Pedro Martínez, 22/12/2014)
(Hyundai i20, motor_ruido, positivo, Pedro Martínez, 22/12/2014)
(Hyundai i20, dirección, positivo, Pedro Martínez, 22/12/2014)

Los sistemas que tengan como meta el análisis de la opinión no están obligados a desarrollar todas las tareas que se han indicado anteriormente.

Dependiendo del nivel de análisis que pretendan llevar a cabo, se realizarán tales tareas o no. La siguiente sección se centrará en esos niveles de análisis, que determinarán las tareas que se realizan o no.

2.3.4. Niveles de análisis

Dado un documento de opinión, la clasificación de la orientación de la opinión u opiniones que encierra dicho documento puede ser más o menos detallista. Existen tres niveles de análisis que se corresponden con el nivel de detalle que se quiera alcanzar. En la literatura relacionada con el AO se pueden hallar tres niveles de detalle:

Nivel de documento: Este estrato de clasificación se refiere a la clasificación de la polaridad de la opinión que expresa en su totalidad un documento de opinión. Se parte de la premisa de que en el documento de opinión únicamente se hace referencia a una entidad, por lo que no es aplicable para aquellos documentos donde se mencionan diversas entidades. Un ejemplo claro de análisis a nivel de documento son (Pang et al., 2002) y (Turney, 2002).

Debe destacarse la tendencia actual de emplear vectores de palabras, en lugar de los clásicos vectores de documentos, como modelo de representación de la información de los documentos cuyo contenido de opinión se pretende clasificar. Uno de los primeros trabajos que introducen los vectores de palabras a la clasificación de la opinión a nivel de documento es (Maas et al., 2011).

Nivel de oración: En este caso se pretende realizar un estudio a nivel de oración, es decir, identificar las oraciones que constituyen un texto e intentar inferir la orientación de la opinión que encierran. Dos ejemplos de trabajos que tratan de extraer la opinión de cada oración son (Wilson et al., 2005b) y (Meena & Prabhakar, 2007). Algunos autores, como es el caso de Wilson et al. (2004), han querido profundizar hasta el nivel de los sintagmas y construir la polaridad de una oración como la combinación de las polaridades de los sintagmas que la componen.

Al igual que en el caso de clasificación a nivel de documento, al estudio de la opinión a nivel de oración también ha llegado el modelo de vectores de palabras, siendo un reciente ejemplo el trabajo (Zhang & He, 2015).

Nivel de entidad y aspecto: Éste es el nivel de análisis que entiende que una opinión trata sobre una entidad, y que quiere afinar más

sobre un aspecto de dicha entidad. El análisis a nivel de entidad y aspecto es el que se corresponde sin ningún genero de dudas con la quíntupla de opinión definida anteriormente. En la bibliografía se puede encontrar este nivel de análisis referido también como análisis a nivel de característica (Hu & Liu, 2004).

Este nivel de análisis está atrayendo cada vez más el interés de la comunidad investigadora en AO, y una muestra de de ello son los recientes trabajos (Carter & Inkpen, 2015; Jiménez-Zafra et al., 2015), las dos ediciones de la tarea *Aspect Based Sentiment Analysis (ABSA)* (Pontiki et al., 2014, 2015) y también la tarea clasificación de la polaridad a nivel de aspecto del principal taller de AO en español, TASS, (Villena Román et al., 2015b,a).

2.4. Recursos lingüísticos para el AO

El AO, al igual que otras tareas de PLN, requiere del uso de recursos lingüísticos para introducir información en los sistemas de clasificación de la opinión. La mayoría de los recursos están desarrollados para su aplicación sobre textos en inglés, aunque cada vez más la comunidad investigadora va contando con recursos en un idioma diferente al inglés. A continuación se van a describir algunos recursos que se pueden emplear para la investigación en AO.

2.4.1. Corpus

En muchas publicaciones se puede leer como los investigadores construyen colecciones de documentos de opinión con el único fin de llevar a cabo sus experimentos. En ocasiones puede estar justificado por las determinadas características que puede tener el sistema que hayan desarrollado, pero en otras puede estar motivado por el simple desconocimiento de la existencia de corpus de opiniones. No aprovechar los corpus que están a disposición de la comunidad investigadora imposibilita la comparación del rendimiento del sistema con los existentes en el estado del arte, lo cual no es positivo para el desarrollo de la investigación, dado que dificulta medir comparativamente la calidad de un nuevo sistema. A continuación se van a destacar algunos *corpora* de opiniones que están disponibles para la investigación en AO:

[2] `http://www.cs.cornell.edu/people/pabo/movie-review-data/`

Cornell movie-review datasets[2]**:** Corpus conformado por opiniones con diferentes tipos de etiquetas de opinión. Dichas opiniones se encuentran agrupadas en tres subcolecciones:

1. *Sentiment polarity datasets*: corpus etiquetado a nivel de documento y a nivel de oración. Los datos del corpus son:
   - A nivel de documento: Conjunto de opiniones etiquetadas como positivas y negativas. En la versión 2.0 el corpus está formado por 1.000 opiniones positivas y 1.000 negativas. La primera versión fue utilizada por primera vez en (Pang & Lee, 2004).
   - A nivel de oración: Conjunto de opiniones cuyas oraciones están etiquetadas como positivas y negativas. En total cuenta con 5331 oraciones positivas y 5331 oraciones negativas. Este conjunto de oraciones fue presentado en (Pang & Lee, 2005).
2. *Sentiment scale datasets*: Colección de opiniones que no se encuentran etiquetadas como positivas o negativas, sino en una escala de polaridad.
3. *Subjectivity datasets*: Conjunto de documentos en el que sus oraciones se encuentran etiquetadas como subjetivas y objetivas. En total la colección cuenta con 5000 oraciones subjetivas y 5000 oraciones objetivas.

Éste fue uno de los primeros corpus que se puso a disposición de la comunidad investigadora en AO. La relevancia de este corpus se ve reflejada en la larga lista de publicaciones que lo utilizan[3].

Economining[4]**:** Corpus utilizado en el trabajo (Ghose et al., 2007), el cual está centrado en el análisis de la interacción entre las opiniones, la subjetividad y los indicadores económicos. El corpus está constituido por tres conjuntos de datos que versan sobre:

- Transacciones económicas y precios.
- Comentarios de usuarios de Amazon.com sobre diversos productos.
- Grado de polaridad de oraciones extraídas de comentarios de Amazon.com.

[3] `http://www.cs.cornell.edu/people/pabo/movie-review-data/otherexperiments.html`

[4] `http://economining.stern.nyu.edu/datasets.html`

Corpus MPQA[5]**:** Corpus formado por 535 artículos periodísticos (noticias), en el que cada una de sus oraciones tienen asociadas etiquetas de opinión y otros estados personales (creencias, emociones, sentimientos, especulaciones...). En total, el corpus cuenta con 10 000 oraciones etiquetadas. El corpus fue presentado por primera vez en (Wiebe et al., 2005), la segunda versión, en la que se añaden etiquetas cercanas al AO a nivel de aspecto, se presenta en (Wilson, 2008) y la tercera versión, en la que se añaden etiquetas de opinión a nivel de entidad y eventos, se describe en (Deng & Wiebe, 2015).

Costumer Review Dataset[6]**:** Colección de opiniones de 5 productos electrónicos distintos. Las opiniones fueron obtenidas de Amazon.com y C|Net.com. La primera utilización del corpus se encuentra descrita en (Hu & Liu, 2004).

SFU Review Corpus[7]**:** Corpus de 400 opiniones en inglés sobre productos comerciales. Las opiniones están organizadas en 8 dominios diferentes: libros, coches, ordenadores, lavadoras, hoteles, cine, música y teléfonos. Cada dominio cuenta con 50 opiniones, de las cuales 25 son positivas y 25 son negativas. El corpus también cuenta con su versión en español, convirtiéndose así el *SFU Review Corpus* en un corpus comparable en inglés y en español. La versión inglesa se publicó por primera vez en (Taboada & Grieve, 2004) y la española en (Brooke et al., 2009). Este corpus se ha utilizado en la experimentación relacionada con técnicas de adaptación al dominio (ver Sección 6.4).

NTCIR multilingual corpus[8]**:** Corpus multilingüe de artículos periodísticos escritos en inglés, japonés y chino que se utilizó en la tarea *Multilingual Opinion Analysis* del congreso NTCIR[9].

Amazon Product Review Data:[10] Una tarea muy interesante y relacionada con el AO es la clasificación automática de opiniones falsas, en inglés *opinion span detection*. El corpus *Amazon Product Review Data* está compuesto por 5,8 millones de opiniones publicadas en Amazon. Dichas opiniones, además de contar con una etiqueta de veracidad (opinión verdadera/opinión falsa), cuenta también con información

[5] `http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/`
[6] `http://www.cs.uic.edu/~liub/FBS/CustomerReviewData.zip`
[7] `https://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html`
[8] `http://research.nii.ac.jp/ntcir/permission/ntcir-7/perm-en-MOAT.html`
[9] `http://research.nii.ac.jp/ntcir/index-en.html`
[10] `http://liu.cs.uic.edu/download/data/`

relacionada con la propia opinión y el producto sobre la que trata, así como con su valor de polaridad, por lo que también se puede emplear la colección para un estudio de la clasificación de la polaridad. El corpus fue presentado en (Jindal & Liu, 2008).

Emotiblog: Corpus multilingüe inspirado en el MPQA para la detección de emociones, cuyo uso está orientado a la clasificación y extracción de opiniones. El corpus está formado por textos de *blogs* sobre tres temas muy concretos: el protocolo de Kioto, las elecciones de 2008 a la presidencia de Estados Unidos de América, y las elecciones de Zimbabue. Está formado por 30 000 palabras por idioma y tema. La descripción completa del corpus se puede leer en (Boldrini et al., 2010).

Corpus de Opiniones en Frances[11]**:** Corpus presentado en (Bestgen et al., 2004), el cual está formado por 702 oraciones recogidas de periódicos franceses y belgas etiquetadas por un grupo de 10 expertos como positivas, neutras o negativas.

English Sentiment Quotes[12]**:** Corpus constituido por 1590 citas periodísticas en inglés. Cada una de las citas son una opinión sobre una determinada entidad, estando cada una de ellas etiquetadas por cuatro expertos. El corpus fue presentado en (Balahur et al., 2010), y se describe más detalladamente en la Sección 5.4.1.

Stanford Twitter Corpus[13]**:** Primer corpus de *tweets* en inglés para la investigación en AO. El corpus se divide en dos subconjuntos: entrenamiento y evaluación. El conjunto de entrenamiento es un corpus completamente balanceado compuesto por 1,6 millones de *tweets*, de los cuales 800 000 son positivos y 800 000 son negativos. La etiqueta de opinión de los *tweets* del conjunto de entrenamiento se encuentra determinada en función de los emoticonos que aparecen en los *tweets*. Por contra, el conjunto de evaluación está conformado por 182 *tweets* positivos y 177 *negativos*, los cuales sí están etiquetados manualmente. El corpus fue presentado en (Go et al., 2009). Una descripción más detallada del corpus se encuentra en la Sección 4.4.

[11] `https://sites.google.com/site/byresearchoa/home/resources-for-opinion-mining-content-analysis-and-psycholinguistics`

[12] `http://islrn.org/resources/574-735-957-886-6/`

[13] `http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip`

[14] `https://www.cs.york.ac.uk/semeval-2013/task2/index.php\%3Fid=data.html`

Corpus de la tarea de AO en Twitter del taller SemEval[14]: SemEval[15] es un taller de evaluación de sistemas relacionados con el análisis semántico. En el año 2013 SemEval incluyó entre sus evaluaciones una específica relacionada con el AO sobre *tweets* en inglés, la cual se ha reeditado en 2014 y en 2015. En las ediciones de 2013 y 2014 el grupo SINAI de la Universidad de Jaén participó en la tarea de AO en Twitter (Martínez-Cámara et al., 2013, 2014a).

La tarea de AO en Twitter ha dado como resultado un conjunto de *tweets* etiquetados con tres niveles de polaridad: positivo, negativo y neutro. En concreto, al corpus de *tweets* al que se está haciendo referencia es al conformado por el conjunto de entrenamiento y desarrollo. El conjunto de evaluación no se considera porque la orgnización del taller lo modifica cada año. Los *tweets* que conforman el corpus fueron descargados entre enero de 2012 y enero de 2013. De los *tweets* descargados se seleccionaron aquellos que contuvieran al menos una palabra con al menos un sentido con un grado de polaridad superior a 0,3 según SentiWordNet. El resultado fue un conjunto de 11 392 *tweets*. Una descripción más amplia del corpus se encuentra en (Nakov et al., 2013).

Compus de RepLab[16]: La reputación es un concepto relacionado con la opinión, dado que la reputación de una entidad suele estar asociada al estado de la opinión existente sobre ella. El taller RepLab tiene como objetivo la evaluación de sistemas orientados a la determinación automática de la reputación de una entidad. El corpus asociado al taller es una colección de 142 000 *tweets* en inglés y español, que se encuentran etiquetados en tres niveles de polaridad: positivo, neutro y negativo.

Los *corpora* anteriores se corresponden con conjuntos de opiniones no escritas en español, que pueden ser descargadas para el estudio de la opinión. El español cuenta con un menor número de corpus de opinión, pero poco a poco va aumentando el repertorio disponible para la comunidad investigadora. A continuación se van a destacar algunos de esos corpus.

***Spanish Movie Reviews*[17]:** Aunque se describirá con mayor profusión en la Sección 4.3, no puede faltar en un lista de *corpora* de opiniones en español. El corpus *Spanish Movie Reviews* es el primer corpus de

[15] `http://aclweb.org/aclwiki/index.php?title=SemEval_Portal`
[16] `http://nlp.uned.es/replab2013/`
[17] `http://www.lsi.us.es/~fermin/corpusCine.zip`

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.

Corpus General de TASS[18]**:** TASS es el primer taller enfocado a la promoción de la investigación de técnicas de AO en español aplicadas a textos publicados en Twitter. El taller se viene organizando interrumpidamente desde el año 2012 (Villena-Román et al., 2013; Villena-Román et al., 2014; Villena Román et al., 2015b,a). Uno de los frutos más valiosos del taller, además del avance de los sistemas que cada año presentan los investigadores, radica en los recursos lingüísticos generados para la investigación en AO. 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, y en la Tabla 2.1 se muestran los datos que lo caracterizan al detalle.

Tweets	68 017
Tweets (entrenamiento)	60 798 (89 %)
Tweets (evaluación)	7219 (11 %)
Temáticas	10
Idiomas	1 (español)
Usuarios	154
Niveles de polaridad	6 y 4
Inicio descarga (entrenamiento)	02/12/2011 00:47:55
Fin descarga (entrenamiento)	10/04/2012 23:40:36
Inicio descarga (evaluación)	02/12/2011 00:03:32
Fin descarga (evaluación)	10/04/2012 23:47:55

Tabla 2.1: Descripción del Corpus General de TASS.

El Corpus General del TASS 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).

[18] `http://www.daedalus.es/TASS2015/private/general-tweets-train-tagged.xml`

El segundo corpus a destacar del TASS se publicó para la edición de 2013[19]. Dicho corpus se circunscribe en el dominio político, y alberga *tweets* descargados durante la campaña electoral de las elecciones a las Cortes Generales de España que tuvieron lugar en noviembre de 2011. El corpus está formado por 2500 *tweets* relacionados con los cuatros principales partidos políticos del momento: Partido Popular, Partido Socialista Obrero Español, Izquierda Unida y Unión Progreso y Democracia. Cada *tweet* del corpus cuenta con una etiqueta de polaridad general del *tweet*, y con una etiqueta de polaridad que indicaba la orientación de la opinión con respecto a cualquiera de los cuatro partidos políticos indicados que se referencian en el *tweet*. Los niveles de polaridad que se consideran son: P (positivo), NEU (neutro), N (negativo) y NONE (sin polaridad). Se debe destacar que el etiquetado de este corpus se realizó manualmente.

El tercer corpus a destacar se preparó para la edición del año 2014, y en concreto para dos tareas: detección de aspectos y clasificación de la opinión con respecto a los aspectos identificados. En el año 2014 se produjo en España la irrupción del seguimiento masivo de programas de televisión a través de las redes sociales, en especial a través de Twitter. Ese fenómeno se le conoce en inglés como *Social TV* y en español como Televisión Social. La denominación inglesa del comportamiento social se tomó prestada para nombrar al corpus, que sin más circunloquios es corpus Social-TV[20]. El dominio del corpus es el futbolístico, dado que está constituido por *tweets* publicados durante la final de la Copa del Rey de fútbol de 2014, la cual enfrentó a los equipos con mayor rivalidad de España: Fútbol Club Barcelona y Real Madrid Club de Fútbol. El número de *tweets* del corpus asciende a 2773, y cada uno de ellos están etiquetados a nivel de *tweet* y de aspecto.

El cuarto corpus ha sido publicado para la edición de 2015 y vuelve a centrarse en el dominio político. Éste nuevo corpus, llamado STOMPOL (*corpus of Spanish Tweets for Opinion Mining at aspect level about POLitics*), está formado por 1284 *tweets* que fueron descargados durante los días 23 y 24 de abril de 2015. La intención que subyace al desarrollo del corpus se corresponde con contar con una colección de *tweets* con opiniones sobre distintas políticas que desarrollan los partidos políticos. Las políticas sobre las que versan los *tweets* de STOMPOL son: Economía, Sanidad y Educación.

[19] `http://www.daedalus.es/TASS2013/corpus.php`

[20] `http://www.daedalus.es/TASS2014/tass2014.php#corpus`

STOMPOL también recoge *tweets* que hablan sobre los propios partidos políticos, los cuales en la mayoría de los casos se refieren a los casos de corrupción de cada uno de ellos. Todos aquellos *tweets* que no hablan de las políticas indicadas ni del propio partido se han catalogado como "Otros". Los 1284 *tweets* de STOMPOL fueron etiquetados manualmente por varios expertos, a nivel de *tweet* y a nivel de cada uno de los aspectos que aparecen referenciados, los cuales se corresponden con las políticas señaladas y con los propios partidos políticos.

Por último, se van a destacar los recursos en el ámbito del AO que ha desarrollado el grupo SINAI de la Universidad de Jaén, dado que es el grupo al que pertenece el autor y directores de la presente tesis.

OCA[21]**:** 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. El corpus fue presentado en (Rushdi-Saleh et al., 2011b) y se emplea en la experimentación descrita en la Sección 7.3.1.

EVOCA[22]**:** Este corpus es la traducción automática del corpus OCA a inglés, por lo que está conformado por el mismo número de opiniones que OCA. El corpus fue presentado en (Rushdi-Saleh et al., 2011a) y, al igual que OCA, se emplea en la experimentación descrita en la Sección 7.3.1.

COAH[23]**:** 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. El corpus fue presentado en (Molina-González et al., 2014) y en la Sección 6.2.2 se describe con mayor profusión.

Corpus MCE[24]**:** 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. El corpus fue utilizado en la experimentación que se describe en la Sección 7.3.2.

[21] `http://sinai.ujaen.es/oca-corpus/`
[22] `http://sinai.ujaen.es/evoca-corpus/`
[23] `http://sinai.ujaen.es/coah/`
[24] `http://sinai.ujaen.es/mce-corpus/`

COST[25]**:** El corpus COST 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*. Una descripción más amplia del corpus se encuentra en la Sección 4.4.

2.4.2. Bases de conocimiento de opinión

Los corpus etiquetados son un recurso de gran valor para la investigación de la opinión porque posibilitan la construcción de sistemas de aprendizaje supervisados, así como la evaluación de sistemas independientemente del tipo de aprendizaje que desarrollen. Pero, otro recurso fundamental en el AO son las bases de conocimiento de opinión, que pueden ir desde las simples listas de palabras de opinión, a bases de conocimiento léxicas y conceptuales con información de opinión. Al igual que se ha hecho con los corpus, se va a realizar un repaso de las principales bases de conocimiento, destacando al final las desarrolladas por el grupo SINAI.

General Inquirer[26]**:** Se trata de una lista de funciones morfológicas de un conjunto de palabras[27] con información sintáctica, semántica y pragmática. Asimismo las palabras tienen asociadas su valor de polaridad, que en el caso de este lexicón puede ser positivo o negativo. General Inquierer cuenta con 1915 palabras positiva y 2291 palabras negativas. General Inquierer fue presentado en (Stone et al., 1966).

SentiWordNet[28]**:** Base de conocimiento de opinión construida a partir de WordNet, que asocia a cada *synset* de WordNet tres valores de polaridad: Positivo, Objetivo y Negativo. En realidad esos tres valores de polaridad son la probabilidad de que el sentido correspondiente de WordNet sea Positivo, Negativo u Objetivo. Al ser valores de probabilidad es evidente que su suma tiene que ser igual a 1. El recursos fue presentado en (Esuli & Sebastiani, 2006). En la Sección 6.2.1 se describe con algo de más detalle SentiWordNet y se cataloga como léxico generado a partir de un diccionario.

BLOL[29]**:** Se trata del léxico de opinión desarrollado por Bing Liu, el cual aglutina 4783 palabras negativas y 2006 palabras positivas. Al

[25] `http://sinai.ujaen.es/cost-2/`
[26] `http://www.wjh.harvard.edu/~inquirer`
[27] Cada entrada se corresponde con una palabra y su función morfológica (PoS).
[28] `http://sentiwordnet.isti.cnr.it/`
[29] `http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar`

igual que SentiWordNet, BLOL se describe con mayor profusión en la Sección 6.2.1.

MPQA Subjectivity Lexicon[30]**:** Se trata de una lista de palabras y expresiones indicadoras de subjetividad, o mejor dicho, de expresiones que pueden representar un estado personal (*private state*). Cada una de las expresiones está acompañada por categoría morfológica; por su nivel de intensidad de subjetividad (*strong subjetive* o *weak subjective*); y por un valor de polaridad, que puede ser positivo, negativo, neutro y ambos. Actualmente el léxico cuenta con 8222 entradas y forma parte de la aplicación de clasificación de opiniones OpinionFinder (Wilson et al., 2005a). La lista fue utilizada por primera vez en (Wilson et al., 2005b).

Q-WordNet[31]**:** Es un recurso similar a SentiWordNet, en el sentido de que se ha construido utilizando como referencia a WordNet. Los autores de SentiWordNet asignan a cada *synset* de WordNet un valor de probabilidad de pertenencia a tres niveles de polaridad, mientras que los autores de Q-WordNet consideran la polaridad como una cualidad, de manera que simplemente asocian a cada *synset* el valor de Positivo o Negativo. Al no considerar la posibilidad que un sentido pueda ser objetivo, la versión más reciente de Q-WordNet está formada por 7402 sentidos positivos y 8108 sentidos negativos. El recurso fue presentado en (Agerri & García-Serrano, 2010).

WordNet-Affect[32]**:** Se trata de una extensión de WordNet-Domains[33] (Magnini & Cavaglià, 2000), el cual es una ampliación de WordNet en el sentido de que asigna a cada *synset* el dominio semántico al que pertenece. De la misma manera que WordNet-Domain añade etiquetas de dominio, WordNet-Affect añade etiquetas relacionadas con estados emocionales, como pueden ser: miedo, enfado, sorpresa, frío, confusión, etc. El recurso, además, tiene los distintos estados emocionales que considera catalogados como positivos, negativos, neutros o ambiguos. El recurso fue presentado por primera vez en (Strapparava & Valitutti, 2004).

SenticNet[34]**:** Este recurso pretende dar un paso más en el ámbito del AO. Los recursos anteriores son un conglomerado de palabras, o a lo sumo

[30] `http://mpqa.cs.pitt.edu/lexicons/subj_lexicon`
[31] `https://docs.google.com/open?id=0B9KQ2oHqSfwmSDRiMWZwNmJhMzA`
[32] `http://wndomains.fbk.eu/wnaffect.html`
[33] `http://wndomains.fbk.eu/`
[34] `http://sentic.net/`

un listado de los sentidos de WordNet con su valor de polaridad. SenticNet es un recurso que está más cerca del nivel semántico que del nivel léxico, dado que no está formado por palabras o sentidos, sino por conceptos. Los conceptos de SentiNet están formados por varias palabras, que por sí solas no expresan directamente un estado de opinión, pero conjuntamente sí tienen un valor de polaridad. La última versión de SenticNet, la 3.0, está formada por 30 000 conceptos con su valor de polaridad identificado. SenticNet fue presentado en (Cambria et al., 2010).

Es momento ahora de destacar las bases de conocimiento de opinión orientadas al tratamiento del español.

Léxico de Pérez Rosas[35]**:** Antes de describir someramente la lista de palabras de opinión, debe indicarse que la denominación "Léxico de Pérez Rosas" se ha utilizado para diferenciar a la lista de palabras en esta memoria, ya que sus autores no le han asignado ningún nombre a su léxico. Dicho lo cual, el léxico de Pérez Rosas es una lista de palabras de opinión en dos niveles que los autores denominan como: *strength lexicon* (lexicón preciso) y *medium strength lexicon* (lexicón menos preciso). La diferencia entre ambos, además del número de palabras que incluyen (*strength*: 1347 y *medium*: 2496), estriba en los términos semilla que se han empleado para su generación. El método consiste en tomar palabras de opinión en inglés, identificar su *synset* de WordNet y obtener en una versión en español de WordNet la traducción de la palabra en español. Para el nivel preciso del léxico, los autores toman como semilla la lista de palabras de opinión en inglés MPQA (MPQA *subjective lexicon*). De dicha lista, solamente consideran aquellas palabras marcadas como muy positivas (*strong positive*) o muy negativas (*strong negative*). Tras la selección, acuden a SentiWordNet, quedándose con el *synset* que tiene un valor más alto de la polaridad original asociada al término. Una vez obtenido el *synset* solamente tienen que acudir a una versión en español de WordNet para obtener la palabra en español.

El segundo nivel del léxico se construye a partir de SentiWordNet. Los autores se quedan con aquellos *synsets* que en algunos de sus dos valores de polaridad cuentan con un valor superior a 0,5. Una vez que tienen seleccionados los sentidos llevan a cabo el mismo proceso que

[35] `http://web.eecs.umich.edu/~mihalcea/downloads/SpanishSentimentLexicons.tar.gz`

para el primer nivel de la lista de palabras. El léxico fue presentado por primera vez en (Pérez-Rosas et al., 2012).

ML-SentiCon[36]**:** 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. Los autores aprovechan las versiones de WordNet de las distintas lenguas de España recogidas en *Multilingual Central Repository* (ver Sección 5.3.1) para generar el definitivo listado estratificado de lemas. Aunque, cada lema cuenta con una puntuación de polaridad, en la práctica ML-SentiCon se puede considerar como una lista de lemas positivos y negativos. El recurso fue presentado en (Cruz et al., 2014).

Spanish Emotion Lexicon[37]**:** 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. La lista fue presentada en (Sidorov et al., 2013).

ElhPolar[38]**:** Lista de palabras de opinión construida a partir de varias fuentes de datos. Primeramente los autores tradujeron al español el *MPQA Subjectivity Lexicon*. Seguidamente, mediante la aplicación de un método basado en el 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. Por último, los autores complementan el lexicón con expresiones coloquiales. La lista fue presentada en (Saralegi & San Vicente, 2013).

Al igual que se han destacado por separado los corpus desarrollados por el grupo SINAI, se va a realizar lo mismo con las listas de palabras generadas.

iSOL[39]**:** Lista de palabras de opinión en español desarrollada durante el transcurso de la investigación que se expone en esta memoria. 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

[36] `http://www.lsi.us.es/~fermin/ML-SentiCon.zip`
[37] `http://www.cic.ipn.mx/~sidorov/SEL.zip`
[38] `http://komunitatea.elhuyar.org/ig/files/2013/10/ElhPolar_esV1.lex`
[39] `http://sinai.ujaen.es/isol/`

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. En la Sección 6.2.1 se puede leer una descripción más amplia de iSOL. La lista fue presentada por primera vez en (Molina-González et al., 2013).

eSOLDomainGlobal[40]**:** La experimentaciones relacionadas con la adaptación al dominio de iSOL (ver Sección 6.4) 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. Las versiones de iSOL adaptadas a cada uno de estos 8 dominios fueron presentados por primera vez en (Molina-González et al., 2014).

[40] `http://sinai.ujaen.es/esoldomainglobal/`

3

Análisis de la Opinión en Español

3.1. La importancia del español

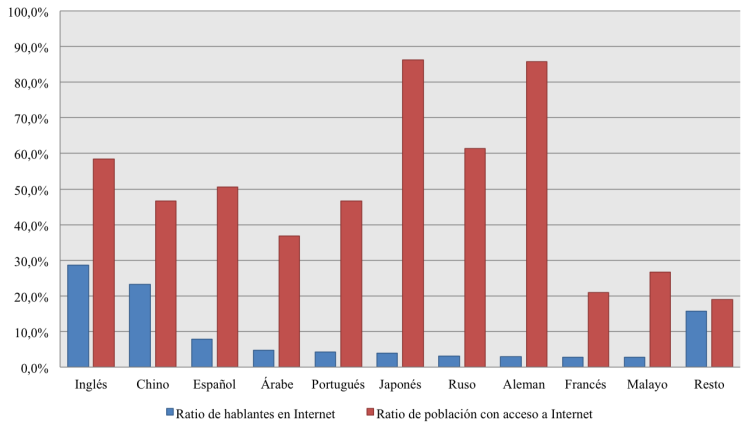
El Capítulo 2 ha definido la tarea de AO, así como ha expuesto los principales recursos lingüísticos existentes para la investigación en AO. De la lectura del Capítulo 2 se puede extraer la conclusión de que la investigación existente está concentrada mayoritariamente en la lengua inglesa, siendo escasos los trabajos cuyo idioma de estudio es distinto al inglés. La fijación por el inglés se debe a que es el idioma más hablado del mundo, y porque la comunidad investigadora inglesa es superior a cualquier otra. Pero a pesar de estas dos razones de peso, no huelga plantearse la pregunta de por qué no prestar atención a otras lenguas.

Según el Instituto Cervantes (Fernández Vítores, 2015) el español es la segunda lengua materna del mundo por número de hablantes, tras el chino mandarín, y también la segunda lengua mundial en cómputo global de hablantes[1]. En 2015 el porcentaje de población mundial hispanohablante alcanza la cifra del 6,7 %, contando con la estimación de alcanzar el 7,5 % en 2030, y en dos o tres generaciones con el 10 %. Asimismo, debe destacarse que se prevé que en 2050 Estados Unidos sea el primer país hispanohablante del mundo. Ésto es en cuanto a datos demográficos naturales, pero en esa proyección del mundo denominada Internet, el español también se encuentra bien representado. El español, como informa el Instituto Cervantes y como se muestra en la figura 3.1, es la tercera lengua más usada en la Red, aunque si se analiza el número de visitas a artículos de Wikipedia por idioma, se puede comprobar que el español se sitúa en segunda posición tras el inglés. Si se quiere aún más destacar el uso del español en Internet, hay que centrar la mirada en su penetración en las redes sociales. El español es la segunda lengua más usada en Twitter y en Facebook. Como último dato curioso, si se ajusta el nivel de detalle hasta el punto de únicamente considerar la red social Twitter y dos ciudades anglófonas como Londres y Nueva York, el español vuelve a ser el segundo idioma más usado por los usuarios[2].

De la Figura 3.1 se puede extraer otra lectura interesante, y es que si se tiene en cuenta que solo el 50,6 % de los hispanohablantes tienen acceso a Internet, entonces queda todavía un amplio margen para incrementar el potencial del español en la Red. Pero si el punto de mira se fija en el ratio de usuarios de Internet, el 7,9 % del total habla la lengua de Cervantes. A pesar de no ser un porcentaje alto, hace que el español salga del pelotón de cola formado por todos los idiomas a excepción del inglés y

[1] Cómputo global de hablantes = dominio nativo + competencia limitada + estudiantes de español.

[2] En (Fernández Vítores, 2015) se puede encontrar un completo estudio sobre la presencia del español en Internet.



Fuente: *Internet World Stats*[3]*, información del 31 de diciembre de 2013, consultada el 5 de julio de 2015.*

Figura 3.1: Millones de usuarios de Internet por idioma.

del chino mandarín, y también es un indicativo de que tiene todavía una alta capacidad de crecimiento. Ese potencial también se ejemplifica en el porcentaje de crecimiento del español en Internet en el periodo 2000-2013 que ha sido de un 1123,3 %, mientras que el del inglés fue de un 468,8 %.

Parece que queda justificado el cuestionarse investigar metodologías para la inferencia de la orientación de la opinión en otros idiomas que no sea el de Shakespeare, y más si se trata de un idioma con el potencial del español, lengua materna no solo en España sino de gran parte de América del Sur, América Central y México. Este hecho fue la chispa que inició un extenso estudio sobre cómo extraer la polaridad de opiniones escritas en español. Este extenso análisis es el que se irá paulatinamente desgranando a lo largo de los siguiente capítulos, y que va a ir desde el tratamiento de textos largos al análisis de *tweets*, o que se va a pasear por la clasificación supervisada al aprendizaje sin supervisión, y no dejará de lado la generación de recursos lingüísticos para el AO en español.

3.2. Textos largos y textos cortos

El impulso que el concepto Web 2.0 proporcionó a Internet provocó que progresivamente se fuera convirtiendo en un repositorio lingüístico con una alta capacidad de almacenamiento de textos. Por ende, la mayor

[3] `http://www.internetworldstats.com/stats7.htm`

parte de los documentos que se han empleado para realizar el estudio provienen de Internet. En ocasiones, la fuente ha sido un portal de opiniones cinematográficas, en otros, un sitio web centrado en el negocio hotelero, y en otros, la plataforma de *microblogging* Twitter.

Dependiendo del origen de los textos, éstos tienen unas determinadas características, siendo la más palpable la longitud. Normalmente los documentos que se han generado a partir de textos publicados en sitios web especializados en opiniones suelen tener una longitud que supera en gran medida una oración. Por contra, los documentos que se generan en las plataformas de *microblogging*, y más concretamente en Twitter, no suelen superar la longitud de una o dos oraciones. Atendiendo a esta diferencia, se ha considerado oportuno establecer la diferencia entre textos largos, es decir, aquellos que suelen tener una longitud de varias oraciones, y textos cortos que son los que su extensión no supera las dos oraciones.

Además de su extensión, los textos largos se caracterizan porque pueden emplear varias oraciones e incluso párrafos para transmitir un mensaje. Dependiendo de la extensión, en un texto largo pueden llegar a subyacer varios mensajes, constituyendo de esta manera un reto mayor para el PLN. Por último decir, que un texto largo suele estar circunscrito a un domino o temática concreta, y en su conjunto constituye una unidad de discurso.

Los textos cortos son aquellos que intentan transmitir un mensaje en una o en muy pocas oraciones. Se emplea de manera correcta el verbo intentar, porque en ocasiones algunos textos cortos no transmiten una idea coherente o simplemente son ininteligibles. Aunque mayoritariamente el PLN ha trabajado sobre textos largos, también se debe destacar que se ha desarrollado investigación sobre textos cortos, como puede ser la clasificación de *spam* (Sahami et al., 1998) o la recuperación de información multimodal, que entre una de sus estrategias se encuentra el aprovechamiento del título que puede acompañar a una imagen (Díaz Galiano, 2011).

La popularización de las plataformas de *microblogging* ha provocado que aumente la necesidad de tratar textos cortos, dado que su disponibilidad es mucho mayor, y constituyen actualmente una unidad de comunicación relevante. Al tratamiento de esta categoría de textos se le podría aplicar el conocimiento y experiencia de las TLH en el procesamiento a nivel de oración. Pero los textos provenientes de las plataformas de *microblogging* tienen una serie de peculiaridades, que obligan a que se le aplique un tratamiento que no sea igual al que se puede realizar a una oración de un texto largo. La plataforma de referencia de *microblogging* actualmente, y la que ha sido origen de los documentos de textos cortos empleados en

el estudio que aquí se presenta, es Twitter. La principal singularidad de las publicaciones de Twitter es su longitud, la cual no pueden exceder los 140 caracteres. En 140 caracteres pocas oraciones formadas por un sujeto y un predicado se pueden condensar, por lo que sin ningún género de dudas estos textos encajan perfectamente en el seno del concepto de textos cortos. Además de la particularidad de la longitud del mensaje, conocida por todo usuario de Twitter, los textos de esta red social también se caracterizan por:

1. Los mensajes de Twitter suelen seguir un estilo informal, con un uso, en ocasiones, excesivo de abreviaturas, modismos y jergas.

2. Una gran mayoría de los usuarios de Twitter no tienen como costumbre el cuidado de la gramática y la ortografía a la hora de escribir sus mensajes. Este hecho dificulta considerablemente el procesamiento automático de estos mensajes.

3. La restricción de la longitud de las publicaciones, hace que los usuarios agudicen el ingenio y se refieran a un mismo concepto de distintas maneras. Esto es lo que se ha venido a llamar dispersión de datos (*data sparsity*).

4. En mensajes tan sumamente cortos es harto complicado que exista un contexto, lo cual dificulta tareas del PLN como la desambiguación.

3.3. Análisis de Opiniones a nivel de documento

En la Sección 2.3.3 se detallaron las tareas que componen el AO con la mira siempre puesta en la generación de la quíntupla de opinión. Muchas de esas tareas son más propias de otras especialidades del PLN, en concreto de la Extracción de Información, que propiamente dicho del AO. En los capítulos subsiguientes no se van a explicar metodologías para la generación de la quíntupla de opinión, sino que se van a ir desgranando diversos estudios emprendidos para el cálculo de la orientación de la opinión en textos escritos en español. De una manera más clara, se podría decir que los siguientes capítulos se centran en describir las metodologías empleadas para saber si una opinión es positiva, neutra o negativa. El lector más perspicaz ya habrá intuido que se van a ir detallando distintas técnicas de clasificación, porque tras un proceso de abstracción, el AO no es más que una tarea de clasificación con unas determinadas peculiaridades, que la convierten en una tarea distinta a una simple clasificación de textos.

Una clasificación es un proceso de aprendizaje de la clase que se tiene que asignar a un ejemplo dependiendo de las características del mismo. Ese aprendizaje puede ser con supervisión o sin ella, es decir, con el requerimiento de disponer de ejemplos con su categoría o clase ya asignada, o sin la necesidad de ellos. Estos dos tipos de aprendizaje son los que van a articular los siguientes capítulos. Se va a diferenciar entre los métodos que se han empleado para la clasificación supervisada de la polaridad y los no supervisados. Pero los métodos de clasificación tienen la ventaja de que pueden combinarse con el fin de aprovechar las ventajas de cada uno, y así mejorar el resultado de clasificación individual. Esta posibilidad también ha sido estudiada, y se expondrán las mejoras que se han obtenido para la clasificación de la opinión en español.

Se podría incluso llegar a manifestar que la generación de recursos para AO es una necesidad casi vital. En este estudio sobre el AO en español no se ha pasado por alto esta cuestión, y se van a detallar los trabajos que se han realizado con la meta puesta en la generación y validación de recursos lingüísticos para el AO en español. El AO es una tarea que está fuertemente ligada al dominio o temática sobre la que versa un documento, por lo que la adaptación de los recursos lingüísticos a la temática de los documentos que se están tratando de clasificar es un factor clave para mejorar la clasificación. Debido a su relevancia, la adaptación al dominio de recursos lingüísticos destinados al AO se abordará más adelante.

En la Sección 2.3.4 del Capítulo 2 se indica que en el AO existen tres niveles de análisis, nivel de documento, de oración y de aspecto. Los siguientes capítulos se van a centrar en el cálculo de la polaridad a nivel de documento, es decir, se van a detallar métodos que asigna a cada documento la orientación semántica global que expresan. Un sistema de clasificación de opiniones a nivel de documento debería contar con un módulo que distinguiera entre documentos susceptibles de manifestar una opinión[4] o no. La mayor parte de los trabajos que describen sistemas de clasificación de opiniones no realizan este paso previo, pero aquellos que sí lo intentan, llevan a cabo el análisis a nivel de oración. De los primeros trabajos que intentan determinar si una oración es subjetiva siguiendo un enfoque supervisado se encuentra (Wiebe et al., 1999). En dicho trabajo se genera un corpus de documentos en el que se distinguen las oraciones subjetivas de las objetivas. En la Sección 2.3.2 se indicó que las oraciones subjetivas de interés en el AO son las que implican una evaluación. Este detalle fue tenido en cuenta por Benamara et al. (2011), dado que tratan de dar un paso más allá de la simple clasificación binaria de subjetividad, tratando de descubrir

[4]Debe recordarse que incluso un texto objetivo puede contener una opinión.

oraciones subjetivas evaluativas, subjetivas no evaluativas, objetivas con opinión, objetivas sin opinión. Más recientemente nos podemos encontrar con el trabajo de Chenlo & Losada (2014), en el que se combina un sistema de clasificación de la subjetividad y de la polaridad. A pesar de estos ejemplos, la mayoría de los trabajos relacionados con la clasificación de la polaridad a nivel de documento parten de la premisa de que los documentos contienen opiniones, de manera que se saltan el paso previo de la clasificación de la subjetividad. Los métodos que se describirán en los siguientes capítulos se sustentan sobre esa misma premisa, de manera que sólo se preocuparán por la clasificación de la polaridad de los documentos.

3.4. Evaluación

Debido a que los sistemas de clasificación que se van a presentar en los siguientes capítulos requieren de la validación de su bondad, a continuación se van a describir los métodos de evaluación que se han utilizado.

Los sistemas desarrollados que siguen una estrategia de aprendizaje supervisado se han evaluado a través de un esquema de validación cruzada. Someramente, la validación cruzada es un método de evaluación que trata de reducir la relación de dependencia entre los resultados del experimento y los datos empleados para entrenar el algoritmo. Este enfoque divide el conjunto de datos de entrenamiento en k subconjuntos, utilizando *k-1* particiones para construir el modelo, y una sola para evaluar el modelo resultante. Ese proceso es repetido k veces, y en cada iteración el subconjunto de evaluación es sustituido por alguna de las otras fracciones del corpus original. Normalmente se emplea un valor de k igual a tres, cinco o diez para evaluar clasificadores. En el caso que nos atañe en este momento, se eligió diez como valor de k.

Independientemente del tipo de aprendizaje que desarrolle un clasificador, el resultado que proporciona debe ser medido por alguna función que permita cuantificar la bondad del clasificador, o de la configuración específica del método de clasificación. Cuando se clasifica un ejemplo, el sistema puede acertar o errar, por lo que se podría pensar que del resultado del clasificador sólo se pueden identificar ejemplos bien clasificados y mal clasificados. Pero, en realidad, también se identifican aquellos ejemplos mal clasificados que podrían haberse inferido bien. Para que se pueda entender mejor, en un sistema de recuperación de información, cuando se le realiza una consulta al sistema, éste devuelve una serie de documentos relevantes en relación a la consulta. Los documentos devueltos son los clasificados como relevantes, y los que no se han devuelto, el sistema les ha asignado la

etiqueta de no relevantes. Dentro del conjunto considerado como relevantes, habrá un subconjunto que en realidad sí son relevantes, pero otro en los que ha errado el sistema al tomarlos como relacionados con la consulta. Igual ocurre en el subconjunto considerado como no relevantes, es decir, habrá una serie de documentos que realmente no están relacionados con la consulta y otros que sí, con los cuales el sistema se ha equivocado. De una manera más formal, los cuatro estados en los que puede encontrarse un ejemplo clasificado son:

1. **TP**[5] **(*True Positives* / Verdaderos positivos)**: Número de documentos clasificados correctamente, o en términos de recuperación de información, número de documentos recuperados correctamente.

2. **FP (*False Positives* / Falsos Positivos)**: Número de documentos que no han sido clasificados correctamente, o en términos de recuperación de información, conjunto de documentos que han sido recuperados incorrectamente.

3. **TN (*True Negatives* / Verdaderos Negativos)**: Número de documentos clasificados correctamente, o en términos de recuperación de información, conjunto de documentos que no son relevantes con respecto a la consulta, y el sistema no los ha considerado como relevantes.

4. **FN (*False Negatives* / Falsos Negativos)**: Número de documentos clasificados erróneamente, y en términos de recuperación de información, conjunto de documentos tomados como no relevantes cuando sí lo son.

Catalogando los resultados de la clasificación en las cuatro categorías que se acaban de definir, se puede medir el grado de bondad de la clasificación realizada. Las medidas de evaluación más usadas en el contexto de la clasificación de textos son:

Precisión: Mide el ratio de documentos correctamente clasificados por clase por el total de documentos clasificados, siendo su ecuación:

$$\text{Precisión} = \frac{TP}{TP+FP} \tag{3.1}$$

[5]La abreviatura en español debería ser VP (Verdaderos Positivos) pero el uso extendido de la abreviatura inglesa obliga a su utilización.

Cobertura (*Recall*): Mide la capacidad del sistema para clasificar correctamente todos los posibles documentos de una clase, siendo su ecuación:

$$Recall = \frac{TP}{TP + FN} \tag{3.2}$$

F-medida: Precisión y *recall* son dos medidas que suelen tener una relación inversamente proporcional, es decir, que una alta precisión es probable que implique un bajo *recall* y viceversa. La consideración que se le desee otorgar a cada medida va a depender de la aplicación en cuestión, pero usualmente se prefiere tener en cuenta por igual a las dos métricas. F-medida (Rijsbergen, 1979) es una métrica que intenta combinar la precisión y el *recall*. F-medida dispone de un parámetro β que permite modular la importancia de la precisión. La ecuación de F-medida es:

$$\mathrm{F}\beta = \frac{(\beta^2 + 1)\, precision\, recall}{(\beta^2\, precision) + recall} \tag{3.3}$$

Normalmente se suele asignar la misma importancia a la precisión que a la cobertura, de manera que β suele ser igual a 1, haciendo que a la F-medida o Fβ se suela referenciar como F1.

Exactitud (*Accuracy*): Esta medida tiene en cuenta todas las clases a la hora de evaluar el sistema. Su formula es:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

En una clasificación multiclase, como es la clasificación de la polaridad, se debe calcular la Precisión, la Cobertura y el F1 para cada clase, es decir, que tiene que ser obtenida la Precisión para la clase Positivo y para la clase Negativo, y de igual manera debe hacerse con el *Recall* y con el F1. Para obtener una evaluación completa del sistema se deben combinar los resultados de la evaluación de cada una de las clases. Para este fin están definidas las medidas *macroavering* y *microavering* (Lewis, 1992), las cuales se definen a continuación:

***Macroavering*:** De una manera directa se puede decir que el *macroavering* es la media aritmética de la precisión, del *recall* y del F1, siendo el

cociente el número de clases con los que cuente el problema.

$$\text{Macro-Precisión} = \frac{\sum_{i=1}^{|C|} Precisión_i}{|C|} \tag{3.5}$$

$$\text{Macro-}Recall = \frac{\sum_{i=1}^{|C|} Recall_i}{|C|} \tag{3.6}$$

***Microavering*:** Si en el *macroavering* a las clases se las considera por igual, en el *microavering* son los documentos a los que se le da la misma importancia. Para calcular el *microavering* es necesario sumar los documentos TP, TN, FP, FN de todas las clases por separado, y posteriormente aplicar la formulas de precisión y *recall*. Formalmente:

$$\text{Micro-Precisión} = \frac{\sum_{i=1}^{|C|} TP}{\sum_{i=1}^{|C|} TP + \sum_{i=1}^{|C|} FP} \tag{3.7}$$

$$\text{Micro-}Recall = \frac{\sum_{i=1}^{|C|} TP}{\sum_{i=1}^{|C|} TP + \sum_{i=1}^{|C|} FN} \tag{3.8}$$

4

Aprendizaje Supervisado

4.1. Introducción

El aprendizaje supervisado es la técnica de aprendizaje automático que requiere de un conjunto de datos etiquetados para la generación de un modelo con la capacidad de clasificar nuevos documentos. De una manera más técnica, el aprendizaje automático precisa de un conjunto de ejemplos de los que se conozca la clase a la que pertenecen. Esos ejemplos deben representarse por medio de un conjunto de características, que se emplean para la construcción de un modelo estadístico, el cual modela la distribución que siguen los datos, de manera que cuando se toma un ejemplo sin clase, pero representado de igual forma que el conjunto de entrenamiento, el algoritmo de clasificación es capaz de determinar a qué distribución de datos más se ajusta, y en función de ello le asigna la clase de la distribución a la que más se asemeja.

La capacidad de clasificación del modelo estadístico depende principalmente del algoritmo de aprendizaje automático, y de las características que se hayan determinado para la representación del conjunto de entrenamiento. Por ende, la elección del algoritmo y la correcta determinación de las características que representan los ejemplos es más que fundamental para obtener un buen resultado en la clasificación. Los algoritmos de clasificación son mucho más maleables que las características, y se suelen adaptar con más facilidad a cada problema. Además, se dispone de una gran variedad de métodos de inferencia, pudiéndose seleccionar en cada momento el que mejor se adapta al problema a resolver. Por contra, la adecuada definición de las características es vital, dado que son ellas las que utiliza el algoritmo para determinar la distribución estadística que representa a cada clase. Dicho de una manera más sencilla, dependiendo de la cualidad que se quiera clasificar, se extraerán de los documentos unas características u otras. No es una tarea exenta de complejidad la selección de las características idóneas, e incluso se podría llegar a decir que es un arte en sí mismo.

Aunque muchas tareas propias del PLN se pueden abstraer como un problema de clasificación, dos son las que destacan en esta analogía: la Categorización de Textos y la Recuperación de Información. La Categorización de Textos es la tarea que trata de resolver la problemática de determinar la clase, categoría o temática de un texto entre un conjunto de dos o más de esas categorías. La Recuperación de Información aunque no es propiamente dicho un problema de categorización, sí se puede considerar como un caso especial. La Recuperación de Información clasifica los documentos en relevantes o no relevantes para una consulta de un usuario. Pero lo que hace a la Recuperación de Información un problema algo distinto a la Categorización de Textos, es la manera de presentar los

resultados. Mientras que en la Categorización de Textos el sistema debe devolver única y exclusivamente las clases que asigna a cada documento, un sistema de Recuperación de Información tiene que generar una lista de documentos ordenada según su relevancia con respecto a la necesidad de información manifestada por el usuario a través de su consulta. Ambos problemas necesitan representar los documentos como un conjunto de términos ponderados por una determinada medida de importancia, que en cierta manera proyecte la información semántica de dicho término en el documento. Por tanto, se requiere convertir los documentos a tratar en una bolsa de términos con un valor de relevancia asociado. A esta conversión es a lo que se llama proceso de indexación.

La indexación está conformada por cuatro etapas consecutivas (Savoy & Gaussier, 2010): identificación de la estructura del documento y *tokenización*, reducción de características (opcional), normalización morfológica (opcional) y asignación de valores de importancia. El resultado del proceso es la indexación de los documentos, o dicho de una manera mucho más gráfica, la proyección de los documentos en vectores con tantas dimensiones como *tokens* se hubieran identificado en el documento, conformando el conjunto de todos ellos un espacio vectorial. Este procedimiento de generación del espacio vectorial, unido al posterior tratamiento de los vectores, es denominado modelo de espacio vectorial (Salton et al., 1975), que aunque primeramente se definió para Recuperación de Información, su simplicidad conceptual y la ligazón que posibilita entre la idea de cercanía espacial y semántica han posibilitado extender su uso para la mayoría de las tareas de clasificación textual.

Una vez que se han determinado los *tokens* o las unidades mínimas de información en las que se descompone un documento, la siguiente fase de la indexación es la reducción de características léxicas. No es objetivo de esta memoria profundizar sobre cómo purgar aquellos *tokens* que no serán útiles en el posterior proceso de clasificación, pero sí es preciso al menos comentar uno de los métodos más usados en clasificación textual. De nuevo, se trata de un método tomado prestado de la Recuperación de Información, que sin más rodeos es la eliminación de *stopwords*[1]. El proceso de indexación en Recuperación de Información tiene como fin representar los documentos por los *tokens* que proporcionen un mayor nivel de distinción entre los diversos documentos que constituyen la colección de documentos en la que se está realizando la búsqueda. Todos los idiomas disponen de una serie de vocablos, que aunque son necesarios para la construcción de

[1] Aunque la traducción natural en español es "palabras vacías", se ha preferido emplear la denominación inglesa por ser la más usada en la bibliografía.

mensajes con sentido, no son válidos para diferenciar unos mensajes de otros, ya que al ser parte esencial del esqueleto de cualquier comunicación no cuenta con capacidad diferenciadora. Ejemplos en español de este tipo de términos son los determinantes, las conjunciones, las preposiciones, los pronombres o el verbo haber. Dado que esta clase de *tokens* no van a ser útiles para la recuperación de documentos, se eliminan, consiguiendo de esta manera reducir la dimensión del espacio vectorial en el que se han proyectado los documentos. No es extraño encontrarse el término *stopper* como denominación del proceso de borrado de las *stopwords*.

La normalización morfológica es otro procedimiento que permite decrecer las dimensiones del espacio vectorial del documento proyectado en el espacio vectorial con. Dos procesos clásicos de normalización morfológica son la lematización y el *stemming*. Ambos métodos buscan la poda de las inflexiones y las derivaciones de las palabras. Para alcanzar tal fin, la lematización, tras la aplicación de un análisis morfológico y mediante el uso de un diccionario, trata de reducir cada palabra a su lema. El *stemming* es un proceso heurístico que recorta la terminación de las palabras con la intención de eliminar la inflexiones y los afijos derivativos. Si se comparan ambos procesos, el *stemming* es mucho más agresivo que la lematización, dado que en ocasiones la palabra transformada no es una unidad con significado como sí ocurre con la lematización. Verbigracia, la palabra "casas" puede tener como lema "casa" o "casar" dependiendo del contexto que se esté considerando, mientras que su *stem*[2] es "cas".

Una vez que se cuenta con *tokens* válidos para el ulterior procesamiento, ya es momento de asignarle a cada *token* un valor que represente la información que aporta al proceso. Es bastante extensa la variedad de medidas que se pueden usar para ponderar la relevancia de un *token*, pero como se ha hecho anteriormente, solo se van a citar las que con mayor profusión se emplean y que también vieron la luz en el seno de la investigación en Recuperación de Información.

1. Ponderación binaria: El valor de relevancia es 1 si el término está presente en el documento, 0 en caso contrario.

2. Frecuencia absoluta: Una idea intuitiva es que la importancia de una palabra esté medida por su número de apariciones en el documento, de manera que cuanto mayor sea la frecuencia mayor será la relevancia del término.

[2]Para el ejemplo se ha usado el *stemmer* del proyecto Snowball `http://snowball.tartarus.org/`

3. Frecuencia relativa: Dado que lo común es trabajar con colecciones de documentos de diferente longitud, es muy recomendable normalizar el valor de la frecuencia, dividiendo la frecuencia absoluta por el número de términos del documento, obteniéndose así el valor de frecuencia relativa.

4. Esquema TF-IDF: En el caso que nos atañe, la Recuperación de Información, el uso de la frecuencia de un término en cada documento para medir su relevancia conlleva un riesgo que debe ser valorado. Tanto si la frecuencia es absoluta o relativa, los términos que son muy frecuentes en todos los documentos del corpus aminoran la capacidad de discriminación del sistema de recuperación de información, porque dificultan el descubrimiento del grupo de documentos que son relevantes a una consulta. Por lo tanto, se precisa de una medida que asigne una mayor importancia a términos muy frecuentes en pocos documentos de la colección, dado que ellos serán los que identifiquen a esos grupos de documentos. Ésa es la idea que perseguía Sparck Jones (1972) al definir la frecuencia inversa por documento, en inglés *inverse document frequency* (idf). El *idf* asigna una mayor relevancia a aquellos términos que aparecen en pocos documentos, y una menor importancia a los que están presentes en la mayoría. Si nos damos cuenta, el *idf* proporciona una medida de importancia en relación a todo el corpus, por lo que si se combina con la frecuencia de cada término en cada documento, se tendría una representación clara de la importancia de cada término en relación a cada documento y al corpus completo. Ésta es la idea que subyace en la medida TF-IDF (Salton & Yang, 1973), que no sólo es ampliamente utilizada en Recuperación de Información, sino en cualquier tarea de clasificación de textos. La fórmula que calcula el valor de TF-IDF es la siguiente:

$$\text{TF-IDF} = tf_d * \log \frac{N}{n_d} \tag{4.1}$$

donde tf_d es la frecuencia del término en el documento, N es el número total de documentos y n_d el número de documentos donde aparece el término.

Se indicaba al inicio de la sección que muchas tareas del PLN se pueden abstraer como un proceso de clasificación. El AO, y más concretamente la determinación de la orientación semántica de un documento, es eso, un proceso de clasificación, en el que a dicho documento se le asigna la clase, positivo, neutro, negativo, o un valor perteneciente a una escala de

intensidad de valores de polaridad que se determine. Por lo tanto, cabría cuestionarse si se podría aplicar un esquema similar a la clasificación de la opinión a nivel de documento. Una respuesta inmediata diría que no es posible, dado que una opinión, como podría ser "El restaurante no me pareció muy bueno", es complicado representarla mediante características basadas en la frecuencia de los términos. Una respuesta algo más reflexiva, por lo menos tendría la prudencia de considerar la capacidad de los métodos de aprendizaje automático para determinar la distribución de las distintas clases de un conjunto de datos. Esa prudencia es la que le valió a Pang et al. (2002) para realizar una experimentación que le permitió validar si un esquema basado en Categorización de Textos puede ser suficiente para la adecuada clasificación de la polaridad en inglés.

De la misma manera que se ha mencionado anteriormente, Pang et al. (2002) comenzaron con la determinación de las características que representarían a los documentos de opinión. La opción más inmediata es la de considerar a cada término o *token* como característica, es decir, se comprobó la capacidad de representación de un esquema basado en *unigramas*. La opinión es un fenómeno lingüístico que es altamente dependiente del contexto, de manera que las palabras se encuentran afectadas por las que están en su vecindad, verbigracia "No me gustan las magdalenas", en inglés "*I don't like muffins*". En el ejemplo anterior el verbo "gustar" (*like*) se encuentra afectado por la partícula negativa "no" (*don't*). Por este motivo los autores también plantean la evaluación de un conjunto de características construido siguiendo un esquema basado en *bigramas*. Los autores no sólo consideran la valoración de la idoneidad de usar como características los términos, sino también información morfológica, e información relacionada con la posición de los términos en el texto. Pang et al. (2002) evalúan la ponderación de las características siguiendo un enfoque basado en frecuencia y en presencia (binario). Para que el estudio sea completo, se comprueba la validez de tres algoritmos de aprendizaje automático: Naïve Bayes (Lewis, 1998; McCallum & Nigam, 1998), Máxima Entropía (Berger et al., 1996) y SVM (Vapnik, 1982, 1995; Cortes & Vapnik, 1995). Continuando con la analogía de la categorización de textos, lo único a lo que dan de lado los autores es a la eliminación de aquellos términos sin capacidad de discriminación (*stopwords*), y a la aplicación de un proceso de normalización morfológica (*stemming*).

De todas las variables evaluadas, la configuración que proporcionó un mejor resultado de clasificación fue la que combinó una representación de los documentos como un conjunto de *unigramas*, cuya relevancia estaba ponderada por un valor binario de presencia, y SVM con *kernel* lineal

como algoritmo de clasificación. A pesar del buen resultado (82,9 % de *accuracy*), los autores concluyen que un modelo basado en bolsa de palabras (*unigramas*) no es suficiente para representar toda la información de opinión que contiene una opinión, y que se debe seguir investigando hasta encontrar la manera de introducir características que aporten información sobre el contexto general o discurso de la opinión.

Si (Pang et al., 2002) nos ha indicado que en un esquema basado en clasificación de texto la mejor configuración consiste en usar *unigramas* como dimensión del vector representante del documento, que la ponderación de los *unigramas* debe ser un valor binario que indique la presencia o ausencia del *token*, y que el algoritmo de aprendizaje automático que parece más idóneo es SVM, Dave et al. (2003) no propugnan lo mismo. En (Dave et al., 2003) se afirma que el empleo de *n-gramas* es mucho más efectivo para la captura del contexto de un mensaje, y por tanto más útil para la subsiguiente clasificación de la polaridad. El método de medición de la cantidad de información que aporta cada término también se erige como una desemejanza entre estos dos primitivos trabajos, dado que mientras Pang et al. (2002) utilizan una medida basada en la presencia, Dave et al. (2003) sostienen que es preferible utilizar la frecuencia relativa de los *n-gramas*. Las diferencias también llegan al método de inferencia, ya que, como se ha visto anteriormente, en la experimentación desarrollada por Pang et al. (2002), SVM es el algoritmo que mejores resultados ofrece, mientras que Dave et al. (2003) muestran como una versión propia de Naïve Bayes da como fruto una clasificación con un poco más de calidad que SVM.

4.2. Algoritmos

En las secciones 4.3 y 4.4 se va a evaluar la efectividad de diversos algoritmos de aprendizaje automático para la clasificación de la polaridad sobre textos largos y cortos escritos en español. De entre la multitud de métodos de aprendizaje automático, se han seleccionado cinco de los más utilizados en Categorización de Textos, y que se van a intentar explicar a continuación.

4.2.1. Máquina de soporte de vectores

La máquina de soporte de vectores, en ingles *Support Vector Machines* (SVM), es un algoritmo de aprendizaje automático lineal[3] definido en

[3]Es preciso indicar que aunque en su definición original estaba constituido por un núcleo lineal, posteriormente se fueron definiendo núcleos no lineales.

(Vapnik, 1982). Someramente, el algoritmo se basa en el principio de Minimización del Riesgo Estructural (*Structural Risk Minimization*) de la teoría del aprendizaje computacional (Vapnik, 1995), la cual se fundamenta en la búsqueda de un *hiperplano* que maximice la separación entre ejemplos pertenecientes a dos categorías diferentes. Desde que Joachims (1998) demostrara su superioridad en el ámbito de la clasificación de textos, SVM ha sido ampliamente empleado en cualquier tarea que requiriera la catalogación automática de textos. En AO SVM ha sido también muy usado y algunos ejemplos son (Pang & Lee, 2004; Gamon, 2004; Abbasi et al., 2008; Rushdi Saleh et al., 2011).

4.2.2. Naïve Bayes

Naïve Bayes (Duda & Hart, 1973; Langley et al., 1992) es un método de aprendizaje automático fundamentado en el Teorema de Bayes (Bayes & Price, 1763). Dicho Teorema nos dice que, dado un conjunto de sucesos mutuamente excluyentes y exhaustivos *(A_1,..., A_n)* cuya probabilidad es distinta de cero, y un suceso cualquiera del que se conocen las probabilidades condicionadas *$P(B/A_i)$*, la probabilidad *$P(A_i/B)$* viene dada por la expresión:

$$P(A_i/B) = \frac{P(B/A_i)P(A_i)}{P(B)} \tag{4.2}$$

Reformular el Teorema para la clasificación de un conjunto de ejemplos no es una tarea ardua, dado que un proceso de clasificación sería equivalente al cálculo de la probabilidad de que un documento de ejemplo sea de una determinada clase, o dicho de otra manera, la probabilidad condicionada de que la clase c_i sobre un documento *d*. En el caso de la clasificación, los sucesos mutuamente excluyentes y exhaustivos con probabilidad distinta de cero son los diversos valores que puede tomar una clase *($C=[c_1...c_n]$)*, que en un sistema binario de AO se correspondería con positivo y negativo. El suceso cualquiera de la definición se asociaría con el documento a clasificar *d*. Como ya se ha indicado anteriormente, el documento a clasificar se representa como un vector de características *($d=[A_1=a_1,...,\ A_{|A|}=a_{|A|}]$)*, por lo que la fórmula del Teorema de Bayes adaptada a un problema de clasificación quedaría redefinida de la siguiente manera[4]:

$$P(C = c_j/A_1 = a_1, ..., A_{|A|}) = \frac{P(A_1 = a_1, ..., A_{|A|})/C = c_j}{P(A_1 = a_1, ..., A_{|A|})} \tag{4.3}$$

[4]En (Liu, 2007) se puede consultar una definición completa y didáctica de la aplicación del Teorema de Bayes a problemas de clasificación.

Se debe destacar que para poder aplicar el Teorema de Bayes en un problema de clasificación, las características de los documentos tienen que ser estadísticamente independientes entre sí. Si el problema de clasificación se circunscribe a textos, dicha asunción se multiplica por dos: las palabras que conforman un documento son independientes del contexto, es decir, que ninguna palabra se ve afectada por sus vecinas; y que la probabilidad de una palabra es independiente de la posición que ocupa en el texto, o dicho de otra manera, que la probabilidad de ver la palabra "actor" en la primera posición de un documento es la misma que la de ver a dicha palabra en cualquier otro lugar del documento.

En el ámbito del AO, Naïve Bayes es un algoritmo que ha sido ampliamente utilizado, siendo algunos ejemplos los siguientes trabajos (Xia et al., 2011; Kang et al., 2012; Zhang et al., 2013; Fersini et al., 2014).

4.2.3. Regresión Bayesiana Binaria

La regresión Bayesiana Binaria, en inglés *Bayesian Binary Regresion* (BBR), es un método propuesto por Genkin et al. (2007) con una fuerte fundamentación matemática, que aunque sea de manera resumida se va a intentar explicar para la compresión del método empleado.

No son escasas las ocasiones en las que nos enfrentamos a problemas en los que a partir de un conjunto de datos observados se quiere obtener un siguiente dato o los subsiguientes valores. Las matemáticas denominan a las técnicas orientadas a la resolución de esta problemática interpolación. La interpolación pretende a partir de una conjunto de observaciones generar una función que no sólo verifique esos datos, sino que también permita obtener nuevos valores. Dicho de una manera más simple, la interpolación engloba a un conjunto de métodos de análisis numérico orientados a la obtención de una función continua en los puntos empleados en su construcción, y por ende posibilita la generación de más puntos. Por lo tanto, la interpolación permite el modelado del comportamiento de una variable.

Los problemas a los que se enfrenta la inteligencia artificial, el PLN y más concretamente el AO, no buscan la modelización de una variable individual, sino que tratan de explicar el comportamiento de una variable a partir del conocimiento de otras, es decir, que por medio de un conjunto de variables explicativas se pretende predecir el comportamiento de una variable de salida. Si anteriormente eran las matemáticas las que proponían la solución, es ahora la estadística la que postula como alternativa los modelos de regresión. La regresión, como método de ajuste, tiene como fin encontrar una función que determine una curva que se ajuste lo

máximo posible a las variables explicativas, permitiendo de esa manera determinar el valor de la variable de salida. Por definición los modelos de regresión únicamente son aplicables a los problemas en donde las variables explicativas y la de salida son cuantitativas, es decir, en aquellos problemas en los que se busca predecir una cantidad, un número. El método que se emplee para la generación de la función de regresión va a determinar el apellido de la regresión, pudiéndose distinguir entre regresión lineal y regresión no lineal.

Las tareas de clasificación en inteligencia artificial, en PLN y en particular en AO no tienen como salida un valor numérico. Por lo tanto, una conclusión inmediata es que no se puede aplicar un modelo de regresión a un problema de clasificación. Pero si al modelo de regresión se le obliga a que el valor numérico que tiene que devolver es la probabilidad de que los datos explicativos pertenezcan o no a una clase, entonces sí se podría aplicar un modelo de regresión a una tarea de clasificación. Esto mismo es lo que hacen los modelos lineales generalizados de regresión. A este grupo de funciones de regresión pertenece la regresión logística, base del algoritmo BBR. En el caso de la regresión logística, la salida de la función es una probabilidad de pertenencia a cada uno de los posibles valores de la variable a predecir, en otras palabras, la probabilidad de pertenencia a cada una de las clases en las que se quiera clasificar a los datos. La función de regresión que se emplee en un modelo lineal generalizado, como es la regresión logísitca, tiene que transformar el valor de la predicción lineal al intervalo probabilístico [0, 1]. En el caso de la regresión logística la función que realiza tal transformación es:

$$p_i = \frac{1}{1 + e^{-\beta' x_i}} \tag{4.4}$$

que también se puede escribir como:

$$log(\frac{p_i}{1 - p_i}) = \beta' x_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_n x_{in} \tag{4.5}$$

A la Ecuación 4.5 también se le conoce como *logit*. Los distintos valores x_i son los valores de las variables explicativas, que en un problema de clasificación se corresponden con los valores que toman las características. Por otra parte, los β_i son los coeficientes que hay que determinar para llegar al valor de probabilidad que se pretende ajustar. El problema radica entonces en el descubrimiento de los valores de β_i. Siguiendo el enfoque clásico de la regresión logística, los valores de los coeficientes β_i se pueden obtener mediante la aplicación de la técnica conocida como "mínimos cuadrados ponderados".

Los modelos de regresión tienen como objetivo ajustarse el máximo posible a las variables explicativas. En problemas de clasificación de texto no conviene que el modelo de regresión se adapte excesivamente a los datos que se emplean de entrenamiento, dado que si los documentos a clasificar difieren demasiado de los de entrenamiento, el modelo de regresión no va a saber determinar la probabilidades de pertenencia a las distintas clases. Para evitar esto se propone aplicar en el modelo de regresión logística un enfoque *bayesiano*. El enfoque *bayesiano* tiene como objetivo insertar más información en el modelo de aprendizaje, mediante el cálculo de una distribución de probabilidad sobre los coeficientes β. Esa distribución de probabilidad tiene que ser definida de manera que se estime que el valor de los coeficientes esté próximo a cero. Genkin et al. (2007) permiten en su algoritmo la elección de tres distribuciones de probabilidad distinta para la determinación de la probabilidad, teniendo en común sus parámetros media (μ) = *cero*, y varianza (τ) >*0*. Esas tres distribuciones de probabilidad son Gaussiana, Laplaciana y Lasso (Tibshirani, 1996). La distribución de probabilidad Gaussiana y Lapliaciana requieren de un valor inicial de varianza τ.

La consideración del uso de BBR para la clasificación de la polaridad viene justificado por los buenos resultados ofrecidos en investigaciones sobre clasificación automática de textos (Genkin et al., 2007; Martín Valdivia et al., 2005; Laws & Schätze, 2008; Koppel et al., 2009; Lee et al., 2011; Taddy, 2013).

4.2.4. K Vecinos más Cercanos

Es momento ahora de utilizar también en el estudio un algoritmo de aprendizaje perezoso o retardado. Dentro de los métodos de aprendizaje automático supervisado se distinguen los métodos no retardados (*eager*) y los perezosos o retardados (*lazy*). La diferencia entre ambos estriba en que los no retardados crean un modelo de generalización a partir de los datos de entrenamiento antes de que se reciba un nuevo ejemplo a clasificar. Esto es lo que precisamente hacen los algoritmos anteriormente descritos. Por contra, los métodos retardados esperan hasta tener un ejemplo a clasificar para utilizar los datos de entrenamiento.

Los métodos retardados siguen la filosofía de que la clase a asignar a un ejemplo debe ser la misma que la del subconjunto de ejemplos más similares que se encuentran presentes en los ejemplos de entrenamiento. Dentro de esta categoría de clasificadores se encuentra el conocido como K Vecinos más cercanos, en inglés *k-Nearest Neighbor* (Cover & Hart, 1967). El *modus operandi* de este algoritmo es muy sencillo: dado un ejemplo a clasificar,

que en nuestro caso sería un documento de texto, se calcula la similitud de ese documento con todos los documentos contenidos en el conjunto de entrenamiento, se seleccionan los k con máxima similitud y se asigna como clase la mayoritaria o más frecuente entre los vecinos de máxima similitud. De la descripción es más que sencillo extraer como conclusión, que el rendimiento del algoritmo está determinado por la correcta selección de una medida de similitud y por la óptima elección de k, es decir, del número de vecinos de máxima similitud a tener en consideración. En (Cover & Hart, 1967) se demuestra que KNN tiene como propiedad, que cuando el número de ejemplos de entrenamiento tiende a infinito, el ratio de error de KNN es como máximo el doble del error bayesiano, es decir, el doble del error mínimo que se puede cometer en un problema de clasificación.

El empleo de KNN viene justificado por los buenos resultados alcanzados en clasificación de textos (Yang, 1994, 1995; Guo et al., 2006; Xiao-fei et al., 2009). También, KNN, ha sido utilizado en estudios comparativos de algoritmos en el ámbito del AO (Tan & Zhang, 2008).

4.2.5. Árboles de decisión

Los árboles de decisión son aquellos algoritmos que fraccionan el conjunto de entrenamiento en sectores, y mediante el escrutinio secuencial de cada característica del ejemplo a clasificar determina el sector, y por ende la clase que se le tiene que asignar a dicho ejemplo. Ese proceso secuencial de inspección de cada atributo se puede representar gráficamente como un árbol, y se puede leer como un conjunto de reglas de clasificación, de manera que es un método de inferencia de cómoda interpretación para una persona. Los algoritmos basados en árboles de decisión son realmente semejantes en su estructura, diferenciándose únicamente en cuatro aspectos: el método de selección del atributo representante del nodo, la estrategia de poda, la capacidad de procesar datos continuos y el tratamiento de datos perdidos.

Para la elaboración de la experimentación, se seleccionó un algoritmo basado en árboles de decisión con un amplio predicamento en el seno de la comunidad investigadora en inteligencia artificial. El algoritmo elegido fue el sucesor de IDE3 (Quinlan, 1986), es decir, C4.5 (Quinlan, 1993). C4.5 se caracteriza por emplear el ratio de ganancia de información (*gainRatio*) como medida de selección de los atributos que mejor fraccionan el conjunto datos; como estrategia de poda aplica un método de simplificación del antecedente de las reglas consiguientes al árbol formado; al contrario que su antecesor ID3, C4.5 tiene la suficiencia para operar con datos continuos mediante la aplicación a dichas variables de un proceso de *discretización*; y para la resolución de los datos perdidos el algoritmo tiene en cuenta todos

los valores del atributo cuya cifra no se haya registrado.

Pocos son los trabajos en los que se concluye la superioridad de C4.5 para la clasificación de textos. En (Gabrilovich & Markovitch, 2004) se afirma que tareas de clasificación en las que se encuentre involucradas un excesivo número de características, C4.5 es más competitivo que SVM, y para que éste último pueda alcanzar al algoritmo de Quinlan es preciso el desempeño de una profunda selección de características. En cierta medida esta aseveración está relacionada con la formulada por Manning & Schütze (1999b), que consiste en que los árboles de decisión no son muy apropiados para la clasificación de textos, dado que para una correcta categorización sería preciso una gran cantidad de ejemplos de entrenamiento. Por ende, la justificación de su uso estriba exclusivamente en que los árboles de decisión son un grupo importante de algoritmos de inteligencia artificial.

4.3. Experimentación sobre textos largos

Los primeros pasos en la clasificación de la polaridad en inglés consistieron en evaluar la efectividad de esquemas exitosos en Categorización de Textos. Los resultados evidenciaron por un lado que se obtenían resultados aceptables, pero, por otro, que para una correcta identificación del sentido de la opinión todavía quedaba un largo camino por recorrer. El español se encontraba huérfano de una experimentación similar, en la que se pusiera a examen si una aproximación semejante a la que se emplea en Categorización de Textos, podría ser al menos aceptable para el descubrimiento de la inclinación de la opinión que manifiesta un autor.

En (Martínez Cámara et al., 2011b) se pretendió poner coto a esta ausencia de investigación, y se inició un estudio que tuvo como fin la evaluación de la efectividad de diversos algoritmos de aprendizaje automático para la determinación de la polaridad en español. Al igual que (Pang et al., 2002), la temática o el dominio del cine fue el seleccionado para esta primera evaluación. Las opiniones o críticas cinematográficas abundan en los estudios de AO en la mayoría de los idiomas, pudiendo ser el motivo principal, su enorme disponibilidad en la Red.

El corpus seleccionado para la experimentación es *Spanish Movie Review* corpus[5] (SMR), el cual fue presentado en (Cruz et al., 2008). Antes de proseguir es preciso advertir al lector que SMR es el mismo corpus que en muchos artículos encuadrados en el AO en español se nombra como corpus MuchoCine. El motivo de la diferencia en la denominación se debe a que el corpus no es bautizado en el artículo en el que se presenta, por lo que los

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

trabajos relacionados con dicho conjunto de opiniones toman como nombre el de la fuente empleada para su compilación, que no es otra que la web de opiniones cinematográficas MuchoCine[6]. Pero, si se consulta la web del autor, el corpus ya aparece referenciado con un nombre propio, *Spanish Movie Reviews*.

El corpus SMR está compuesto por un total de 3878 críticas de cine que fueron recolectadas, como se ha indicado anteriormente, de la web de opiniones MuchoCine. Las opiniones de SMR son una inmejorable representación de los textos que pueden hallarse en Internet, y que en realidad son el tipo de mensaje potencial a analizar por sistemas reales. Esto significa que los textos han sido redactados por usuarios de Internet, que lo más probable no sean expertos críticos de cine o virtuosos en el arte de la redacción, por lo que no será extraño encontrarse con errores ortográficos y gramaticales, aumentando en cierta manera la dificultad de la tarea. Las opiniones están catalogadas según una escala de opinión de cinco niveles, comenzando por el valor 1 (muy negativo) hasta llegar a 5 (muy positivo), mostrándose en la Tabla 4.1 el número de opiniones que hay por clase. El estudio que se desarrolló en (Martínez Cámara et al., 2011b) consistió en el desarrollo de un clasificador binario de opiniones, de manera que las cinco clases de opinión del corpus fueron reducidas a dos: Positivo y Negativo. Para ello, las opiniones en el nivel 1 y 2 de polaridad fueron consideradas como Negativas, y las situadas en los niveles 4 y 5 como positivas. Al no estudiar la identificación de opiniones Neutras, las opiniones con un nivel de intensidad de opinión 3 no fueron tenidas en cuenta. En la Tabla 4.2 se recoge el número de opiniones que finalmente se emplearon en el estudio.

Nivel de opinión	**Número de opiniones**
1	351
2	923
3	1253
4	890
5	461
Total	3878

Tabla 4.1: Número de opiniones por nivel de opinión.

Los algoritmos que se seleccionaron en (Martínez Cámara et al., 2011b) han sido descritos en la Sección 4.2, y una vez conocidos, es momento de adentrarse en la propia experimentación. El corpus SMR tiene una

[6] `http://www.muchocine.net/`

Clase	Número de opiniones
Positivo	1351
Negativo	1274
Total	2625

Tabla 4.2: Número de opiniones por cada clase considerada en el estudio.

peculiaridad bastante interesante, y es que las opiniones constan de dos secciones: resumen (*summary*) y cuerpo (*body*). La sección resumen no es más que una sucinta, clara y directa expresión del parecer del autor del texto sobre la película objeto de crítica, mientras que el cuerpo es la opinión completamente desarrollada. Cruz et al. (2008) dejan bien claro que el resumen y el cuerpo son totalmente distintos, es decir, que el resumen no es un extracto del cuerpo de la opinión. Este hecho, llevó a plantearse el estudio de la capacidad de clasificación que tendrían las dos secciones por separado, dado que se consideró que si la síntesis de la propia opinión proporciona datos suficientes para la clasificación de la polaridad, se podría simplificar en gran medida la tarea de AO. Por este motivo, se llevaron a cabo una experimentación sobre tres conjuntos de datos. Un primer conjunto constituido exclusivamente por la sección resumen (R), una segunda colección conformado solamente por la sección cuerpo (C), y un tercer corpus (R_C) constituido por la unión de las dos secciones.

Cada una de las opiniones de SMR se representaron como un vector de *unigramas*, cuya relevancia se ponderó con TF-IDF. En esta primera aproximación, se estimó que era conveniente reducir el conjunto de características, por lo que se eliminaron las palabras no útiles, es decir, las *stopwords*, y de normalización morfológica. Para la elección de las palabras inútiles se seleccionó la lista de *stopwords* que ofrece el proyecto Snowball[7]. La normalización morfológica consistió en proyectar cada subcorpus a su equivalente conformado exclusivamente por *stems*.

Algunos de los métodos de aprendizaje automático considerados en (Martínez Cámara et al., 2011b), requieren que se fije el valor de una serie de parámetros para que puedan emprender la clasificación. Para la aplicación de SVM se prefirió la implementación proporcionada por el proyecto LibSVM[8] (Chang & Lin, 2011). En concreto se ha optado por el uso del tipo C-SVC, es decir, la versión de SVM que emplea un parámetro C para determinar el grado de penalización de los errores de clasificación.

[7] `http://snowball.tartarus.org/algorithms/spanish/stop.txt`
[8] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

Para este parámetro el valor asignado es 0, es decir, que los errores en la construcción del modelo no se penalizan. SVM para la formación del hiperplano separador de las clases, positiva y negativa, requiere de una función *kernel* que calcule los márgenes máximos de separación de las dos clases, seleccionándose para la evaluación un *kernel* lineal. Se estimó la evaluación del rendimiento de una función lineal.

Naïve Bayes es un algoritmo de clasificación destinado principalmente para problemas que tratan con datos discretos. A pesar de ello existen mecanismos que dan la posibilidad de emplear el algoritmo basado en el Teorema de Bayes sobre datos continuos. La clasificación de texto es un claro ejemplo donde se trabajan con datos continuos, dado que los términos que componen un documento suelen representarse con valores que indican su relevancia en el documento y en el conjunto del corpus. Para dar la posibilidad a Naïve Bayes de trabajar con datos continuos, se ha estudiado la discretización mediante la consideración de que los valores de los atributos siguen una distribución normal. El problema de esta consideración es que se trata de un método paramétrico, es decir, que se deben fijar una serie de parámetros, en este caso la media y la varianza, para poder aplicar la discretización. La necesidad de la determinación de los parámetros rebaja el nivel de flexibilidad del método, por lo que se ha estudiado también el uso de técnicas no paramétricas para la discretización de los valores de los atributos. La elección ha estado dirigida por la de maximizar la flexibilidad del modelo, por lo que se estimó que sería conveniente emplear una implementación de Naïve Bayes fundamentada en un discretización basada en una función de densidad (*Kernel Density Estimation* (John & Langley, 1995)).

KNN es un algoritmo cuyo comportamiento, como ya se ha indicado, está determinado por dos parámetros, la función de similitud y el número de vecinos de máxima semejanza. Como primera aproximación se estimó la evaluación de la distancia Euclídea como función de semejanza. Para la correcta elección del valor de K se llevó a cabo un estudio sobre el valor que podría ser más idóneo, el cual consistió en evaluar el algoritmo con distintos valores de K. Se realizaron 90 ejecuciones, probándose los valores de K comprendidos en una horquilla de 1 a 90 (ver Figura 4.1). El mejor resultado de macro-precisión fue de 78,81% con *K=72*. La detención de la evaluación en *K=90* es debido a que se comprobó que tras obtener el mejor resultado con *K=72* el algoritmo no mejoraba su rendimiento con cifras superiores de K. En cuanto a C4.5 se debe manifestar que ya es una configuración en sí misma de un árbol de decisión, por lo que no se tuvo que determinar ningún parámetro específico.

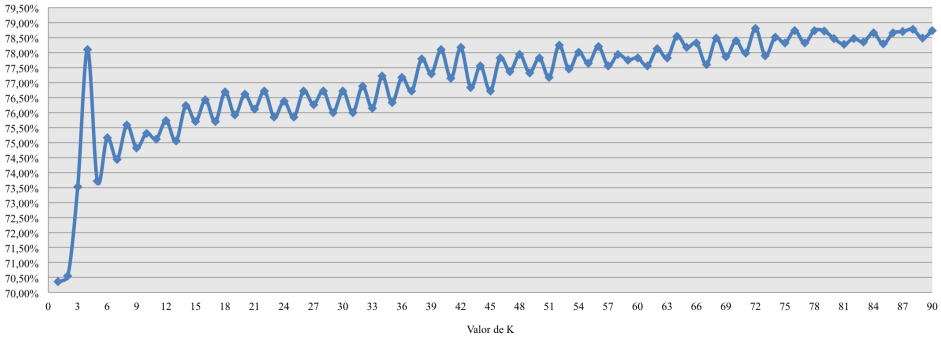


Figura 4.1: Valor de Precisión obtenido por KNN con distintos valores de K.

Se va a tomar la Figura 4.2 como ayuda para la interpretación de los resultados que se han obtenido (ver Tabla 4.3). En la descripción de los algoritmos basados en árboles, se ha mencionado que la propia naturaleza de estos algoritmos hace que sea complicado que puedan tener éxito en tareas de clasificación de textos, como ésta de clasificación de la polaridad, debido principalmente a que el exhaustivo fraccionamiento del conjunto de entrenamiento limita sobremanera su capacidad de generalización. En la gráfica se puede comprobar claramente que la diferencia con sus compañeros es abismal, y sin dilatar más la conclusión, se puede afirmar que, al menos, el algoritmo basado en árboles de decisión C4.5 no es adecuado para la clasificación de la polaridad en español.

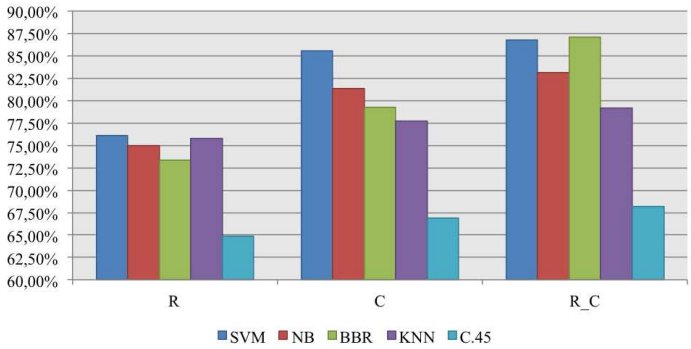


Figura 4.2: Comparativa de los resultados obtenidos por los cinco algoritmos.

Algoritmo	**Corpus**	**Precisión**	***Recall***	**F1**	***Accuracy***
SVM	R	76,23 %	76,07 %	76,14 %	76,15 %
	C	85,67 %	85,49 %	85,57 %	85,56 %
	R_C	85,67 %	85,49 %	85,57 %	85,56 %
NB	R	75,37 %	74,63 %	74,99 %	74,86 %
	C	81,35 %	81,32 %	81,33 %	81,33 %
	R_C	83,16 %	83,14 %	83,14 %	83,16 %
BBR	R	73,51 %	73,31 %	73,40 %	73,11 %
	C	81,42 %	77,14 %	79,22 %	76,61 %
	R_C	87,21 %	87,01 %	87,10 %	87,08 %
KNN	R	76,00 %	75,53 %	75,76 %	75,7 %
	C	78,62 %	76,92 %	77,76 %	77,25 %
	R_C	80,03 %	78,42 %	79,21 %	78,74 %
C4.5	R	65,00 %	64,73 %	64,87 %	64,87 %
	C	66,80 %	66,88 %	68,88 %	66,86 %
	R_C	68,15 %	68,20 %	68,17 %	68,13 %

Tabla 4.3: Clasificación de la polaridad con cinco algoritmos de aprendizaje automático supervisado.

La comparación de los otros cuatro algoritmos depende del subcorpus. Por ejemplo, si se centra el foco de atención en el conjunto de datos más pequeño (R) primeramente habría que decir que los resultados no son buenos, pero tampoco podrían catalogarse como malos. Se debe tener en cuenta que, en comparación, los resúmenes contienen considerablemente menos texto que la opinión en sí (C). Este comportamiento permite concluir que no se deben desdeñar los textos cortos, porque la limitación de longitud obliga a los autores a ser claros y concisos, de manera que la orientación de su pensamiento es más probable que no esté difuminada por circunloquios. SVM es el algoritmo que mejor se comporta sobre el subcorpus R. El simple algoritmo KNN no se queda muy rezagado, aunque como se ha visto anteriormente no destaca en las tareas de clasificación textual. La razón del buen comportamiento de KNN puede residir en el hecho de que en el subcorpus R, los textos son cortos, con una diversidad léxica reducida, lo cual favorece enormemente a un método basado simplemente en similitud léxica.

En el subcorpus C, el cual sólo contiene el cuerpo de la opinión, SVM destaca como el algoritmo que devuelve mejores resultados, y todo hay

que decirlo, como cabía esperar, dado que SVM es de los algoritmos más exitosos en general en la tarea de clasificación de textos, y en particular en AO. El comportamiento de Naïve Bayes y BBR siguen en la misma línea que en el subcorpus R, es decir, devolviendo unos resultados moderados. El comportamiento en este caso de KNN reafirma la aseveración de que KNN en un conjunto de reducida diversidad léxica ofrece un mejor rendimiento que en otro con una mayor variedad. Los resultados aquí alcanzados por el método KNN son similares a los obtenidos en (Tan & Zhang, 2008), en donde también queda por detrás de SVM y Naïve Bayes. Parecía que la regresión debía ser rechazada de plano para la clasificación de la polaridad en español, pero la experimentación sobre el subcorpus R_C sorprende por el buen resultado devuelto por BBR, dado que incluso supera, aunque por muy poco, a SVM. Parece que en este caso la función logística alimentada con una distribución gaussiana consigue adaptarse a la distribución de los datos de una manera adecuada.

Si sólo se tienen en cuenta los resultados, se puede concluir que para conjuntos de datos de tamaño considerable y amplia diversidad léxica, como es el caso del corpus R_C, la regresión logística bayesiana (BBR) y SVM son los mejores candidatos para clasificar opiniones en español. Pero, si como se manifiesta en (Martínez Cámara et al., 2011b) se tiene en cuenta el costo computacional de los algoritmos, en la elección no caben dudas, el algoritmo a emplear es SVM. BBR, al ser un algoritmo de regresión, necesita estimar un parámetro para cada característica de los datos. Los problemas de clasificación de textos o de clasificación de polaridad son problemas con una gran cantidad de características, por lo que el número de parámetros a estimar es muy elevado. Esto hace que BBR consuma una gran cantidad de recursos, de manera que es preferible perder un pequeño grado de exactitud en la clasificación, pero ganarlo en rendimiento computacional.

Una vez analizados los resultados, ya se pueden emitir las conclusiones de este primer estudio de clasificación de la polaridad en español. Primeramente hay que destacar dos diferencias con la clasificación de la polaridad en inglés, siendo la primera la consideración de características. En inglés hay estudios como el de (Dave et al., 2003) que afirman que los *bigramas* y los *trigramas* representan de mejor manera la información de opinión que los *unigramas*. En este estudio se coincide con (Pang et al., 2002) en la determinación de que los *unigramas* proporcionan una mayor capacidad de representación. La segunda diferencia estriba en la ponderación de las características. Desde (Pang et al., 2002) se afirma que para representar la información de opinión en inglés conviene más el uso de valore binarios, es decir, indicar si aparece o no la palabra. Por contra,

para el español, como se ha podido comprobar, TF-IDF ofrece muy buenos resultados. Aunque esta afirmación es conveniente dejarla en cuarentena hasta que no sea demostrada.

Ya se ha mencionado en varias ocasiones el grado de dependencia de los algoritmos de clasificación del conjunto de características que representan los datos. Esa dependencia se refleja en dos factores, por un lado, cuanto mejor describan las características los datos, más sencillo será para el clasificador determinar la clase a la que pertenecen los textos a clasificar; y por otro, en los recursos computacionales que emplea el método de clasificación. Cuantas más características se utilicen para representar una opinión mayor será el vector que lo representa, y por ende, de mayor número de dimensiones será el espacio en el que se proyectan las opiniones. Por lo tanto, cuanto mayor sea el espacio en el que se proyectan las opiniones, mayor será la cantidad de recursos computacionales requeridos para su procesamiento.

Hasta el momento se ha descrito una primera aproximación en la que se estudia la capacidad de clasificación de cinco algoritmos de aprendizaje automático sin prestar demasiada atención a las características usadas. Para completar el estudio de la resolución del problema de la clasificación de la polaridad siguiendo una estrategia propia de la clasificación de textos y recuperación de información, se requiere realizar un estudio sobre la importancia en clasificación de la polaridad en español de la eliminación de las palabras con una reducida capacidad de discriminación, de la normalización morfológica, que en este caso se reduce a la aplicación de un *stemmer*, y de la medida de relevancia para la ponderación de las características.

En (Martínez Cámara et al., 2011a) se realiza tal estudio, y a la descripción del mismo se van a dedicar los siguientes párrafos, para seguir dando luz a la investigación relacionada con la clasificación de la polaridad en español. El trabajo se realiza sobre el mismo corpus que en (Martínez Cámara et al., 2011b), es decir, SMR, y de nuevo se tienen en consideración los tres subcorpus anteriormente detallados (R, C, R_C). Como ya se ha indicado, en Recuperación de Información es común la práctica de eliminar los términos con escasa capacidad discriminativa entre documentos, con el objetivo claro de reducir la dimensión del espacio de características, consiguiendo de esta manera una aminoración en el uso de recursos computacionales. La eliminación de *stopwords* es llevada a cabo por la inmensa mayoría de los sistemas de Recuperación de Información, por no decir todos, porque se ha demostrado empíricamente que los sistemas mejoran su rendimiento cuando se desechan las *stopwords*, mientras que

en AO y más concretamente en AO en español no se ha evaluado si es conveniente la eliminación de palabras sin capacidad de discriminación. Para el estudio se ha usado idéntica lista de palabras *stopwords* que en (Martínez Cámara et al., 2011b).

La normalización morfológica es un proceso que trata de alcanzar la misma meta que la no consideración de las *stopwords*, es decir, la reducción de características. En Recuperación de Información y en Clasificación de Textos la técnica de normalización morfológica más usada es la de reducir las palabras a su *stem*, que consiste fundamentalmente en la poda de los prefijos y sufijos de las palabras. El *stemming* es una técnica que posibilita la reducción de la familia semántica de un término en una unidad léxica representante de la familia semántica del término. Este proceso permite la reducción considerable del espacio de características permitiendo de este modo la mejora del rendimiento de los sistemas. Al igual que con la eliminación de *stopwords*, se ha comprobado empíricamente del buen hacer en Recuperación de Información y en Clasificación de Textos del *stemming*, pero no en AO. Para la elaboración de los experimentos se ha empleado el mismo algoritmo de normalización morfológica que en (Martínez Cámara et al., 2011b).

En la literatura relacionada con la Recuperación de Información se describen una gran diversidad de medidas para medir la relevancia de los términos (Manning et al., 2008). En cambio, la investigación en AO ha tomado prestadas muy pocas de esas medidas para determinar la importancia de los términos, siendo la más explotada la métrica binaria, basada en si el término aparece en el documento o no se encuentra. Paltoglou & Thelwall (2010) realizaron un estudio sobre la idoneidad de medidas de evaluación propias de recuperación de información para clasificación de la polaridad en inglés. Los autores, al contrario que en (Pang et al., 2002), concluyen que las medidas que tienen en cuenta la frecuencia de aparición de los términos proporcionan mejores resultados. El estudio descrito en (Martínez Cámara et al., 2011a) para español no es tan ambicioso como el de (Paltoglou & Thelwall, 2010), pero cubre cuatro extendidas medidas de ponderación de la importancia de los términos. Dichas medidas son:

1. TF-IDF: Medida propuesta en (Salton & Yang, 1973), y que es muy utilizada en el ámbito de la Recuperación de Información. En el trabajo (Martínez Cámara et al., 2011b) se ha comprobado su buen hacer en tareas de clasificación de la polaridad en español. TF-IDF es una medida que premia a términos frecuentes en subconjuntos de la colección de búsqueda, pero que no lo son tanto si se tiene en cuenta

el corpus completo. Dicho de otra manera, TF-IDF otorga una mayor importancia a los términos discriminantes.

2. Frecuencia relativa, en inglés *Term Frequency* (TF): Frecuencia relativa de los términos con respecto al corpus en su conjunto.

3. Frecuencia absoluta, en inglés *Term Occurrences* (TO): Frecuencia absoluta de los términos.

4. Presencia, en inglés *Binary Term Occurrences* (BTO): Si el término se encuentra en el documento su valor en el vector de dimensiones es 1, y si no está es 0.

Tomando como referencia la experiencia adquirida en (Martínez Cámara et al., 2011b), los algoritmos empleados para llevar a cabo la experimentación han sido SVM y Naïve Bayes. Se eligió SVM porque, como se ha podido comprobar antes, es el algoritmo que ofrece un mejor rendimiento. La evaluación se ha llevado a cabo con Naïve Bayes por varias razones: por ofrecer consistencia en los resultados que devuelve; por ser más eficiente que BBR; y por ser muy usado por la comunidad investigadora en AO. Los resultados que se han obtenido se pueden consultar en las Tablas 4.4 y 4.5.

			R			**C**			**R_C**		
	Stop	Stem	Prec.	*Recall*	F1	Prec.	*Recall*	F1	Prec.	*Recall*	F1
TF-IDF	✓	✓	76,23 %	76,07 %	76,15 %	85,67 %	85,49 %	85,58 %	86,84 %	86,67 %	86,75 %
	✓		75,75 %	75,51 %	75,63 %	86,40 %	86,34 %	86,37 %	87,66 %	87,60 %	87,63 %
		✓	75,95 %	75,80 %	75,87 %	85,77 %	85,56 %	85,66 %	86,80 %	86,64 %	86,72 %
			75,57 %	75,37 %	75,47 %	86,18 %	86,10 %	86,14 %	87,73 %	87,69 %	87,71 %
TF	✓	✓	74,74 %	74,51 %	74,62 %	78,47 %	78,20 %	78,33 %	79,81 %	79,44 %	79,62 %
	✓		72,12 %	71,92 %	72,02 %	76,12 %	75,65 %	75,88 %	77,48 %	77,08 %	77,28 %
		✓	74,74 %	74,56 %	74,65 %	78,16 %	77,91 %	78,03 %	79,74 %	79,42 %	79,58 %
			72,29 %	72,07 %	72,18 %	75,83 %	75,34 %	75,58 %	77,06 %	76,65 %	76,85 %
TO	✓	✓	73,82 %	73,51 %	73,66 %	74,66 %	73,13 %	73,89 %	77,66 %	76,89 %	77,27 %
	✓		71,84 %	71,49 %	71,66 %	72,59 %	70,96 %	71,77 %	74,05 %	73,00 %	73,52 %
		✓	74,08 %	73,81 %	73,94 %	74,64 %	73,03 %	73,83 %	77,86 %	77,09 %	77,47 %
			71,87 %	71,49 %	71,68 %	72,45 %	70,77 %	71,60 %	74,25 %	73,12 %	73,68 %
BTO	✓	✓	75,24 %	75,07 %	75,15 %	83,61 %	83,63 %	83,62 %	84,23 %	84,20 %	84,21 %
	✓		73,53 %	73,24 %	73,38 %	83,69 %	83,62 %	83,65 %	83,94 %	83,91 %	83,92 %
		✓	74,63 %	74,45 %	74,54 %	84,10 %	84,09 %	84,09 %	84,13 %	84,12 %	84,12 %
			73,96 %	73,70 %	73,83 %	83,84 %	83,78 %	83,81 %	84,11 %	84,16 %	84,13 %

Tabla 4.4: Evaluación de la normalización morfológica y de las medidas de ponderación con SVM.

			R			**C**			**R_C**		
	Stop	Stem	Prec.	*Recall*	F1	Prec.	*Recall*	F1	Prec.	*Recall*	F1
TF-IDF	✓	✓	75,37 %	74,63 %	75,00 %	81,35 %	81,32 %	81,33 %	83,16 %	83,14 %	83,15 %
	✓		75,23 %	74,86 %	75,04 %	82,10 %	82,01 %	82,05 %	84,08 %	84,01 %	84,04 %
		✓	75,87 %	75,22 %	75,54 %	81,40 %	81,35 %	81,37 %	83,47 %	83,44 %	83,45 %
			74,83 %	74,43 %	74,63 %	81,91 %	81,82 %	81,86 %	83,62 %	83,55 %	83,58 %
TF	✓	✓	65,45 %	65,25 %	65,35 %	67,18 %	66,89 %	67,03 %	68,49 %	68,19 %	68,34 %
	✓		62,48 %	62,40 %	62,44 %	64,97 %	64,63 %	64,80 %	66,54 %	66,24 %	66,39 %
		✓	65,77 %	65,62 %	65,69 %	66,72 %	66,47 %	66,59 %	67,53 %	67,24 %	67,38 %
			63,32 %	63,24 %	63,28 %	64,37 %	64,10 %	64,23 %	65,57 %	65,29 %	65,43 %
TO	✓	✓	74,80 %	74,76 %	74,78 %	68,61 %	66,71 %	67,65 %	69,21 %	68,65 %	68,93 %
	✓		73,59 %	73,56 %	73,57 %	67,62 %	65,49 %	66,54 %	72,12 %	71,72 %	71,92 %
		✓	74,62 %	74,60 %	74,61 %	68,56 %	66,66 %	67,60 %	69,25 %	68,69 %	68,97 %
			73,46 %	73,41 %	73,43 %	67,72 %	65,61 %	66,65 %	72,32 %	71,94 %	72,13 %
BTO	✓	✓	74,61 %	74,58 %	74,59 %	75,18 %	74,91 %	75,04 %	75,32 %	78,58 %	76,92 %
	✓		73,59 %	73,54 %	73,56 %	76,29 %	75,74 %	76,01 %	76,40 %	75,88 %	76,14 %
		✓	74,23 %	74,20 %	74,21 %	75,37 %	75,10 %	75,23 %	75,53 %	75,28 %	75,40 %
			73,44 %	73,38 %	73,41 %	76,21 %	75,63 %	75,92 %	76,58 %	76,07 %	76,32 %

Tabla 4.5: Evaluación de la normalización morfológica y de las medidas de ponderación con NB.

Las dos tablas de datos (Tabla 4.4 y Tabla 4.5) van a ser desgranadas con varios gráficos que van a ayudar a entender los resultados obtenidos. La Figura 4.3 muestra los valores de F1 de los dos algoritmos por cada una de las medidas de ponderación, sin tener en cuenta la eliminación de *stopwords* y la aplicación de *stemming*. De dicha gráfica se pueden extraer dos conclusiones determinantes, siendo la primera más clara y esperable, la superioridad de SVM en los tres subcorpus y en casi todas las medidas de valoración de relevancia de los *unigramas*. Se precisa casi todas, porque si se detiene la mirada en el gráfico se podrá ver que para la medida TO Naïve Bayes proporciona un resultado ligeramente superior al de SVM, pero si se atiende a las tablas de resultados se puede comprobar que la diferencia es ínfima. Por lo tanto, se puede continuar afirmando la superioridad de SVM en la tarea de clasificación de la polaridad en español. La segunda conclusión también es bastante evidente, e incluso marca una diferencia con respecto al inglés, y no es otra que TF-IDF proporciona unos mejores resultados que BTO. Por lo tanto, queda demostrado que para clasificación de la polaridad en español, al contrario de lo que asevera (Pang et al., 2002) para el inglés y en la línea de (Paltoglou & Thelwall, 2010), una métrica a nivel de corpus basada en la frecuencia de los términos proporciona unos mejores

resultados, que una medida fundamentada únicamente en la presencia o no del término.

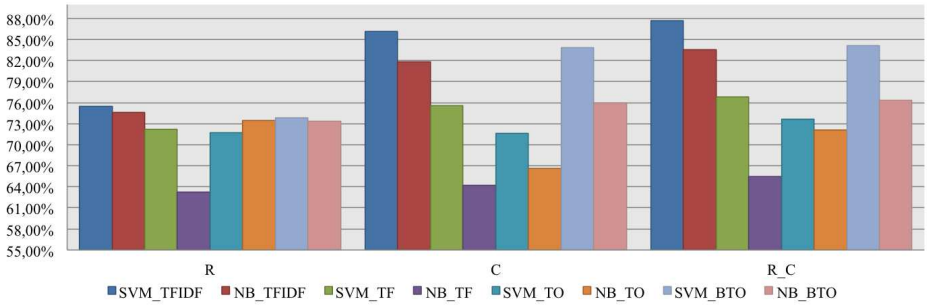


Figura 4.3: F1 sin tener en cuenta la aplicación de *stopper* y *stemming*.

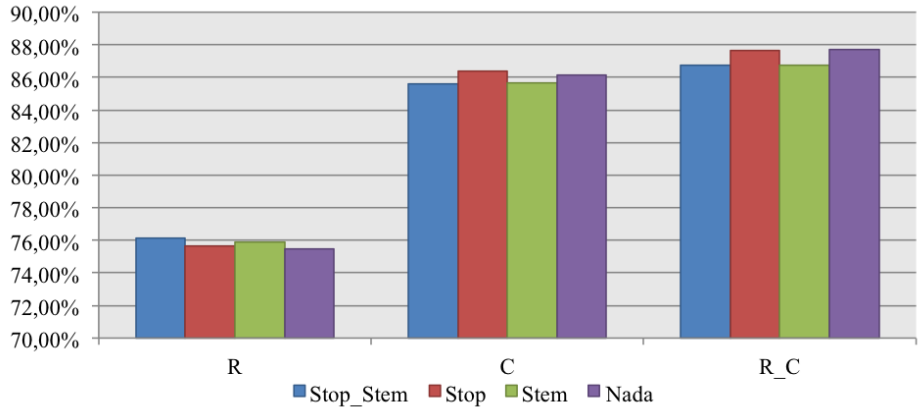


Figura 4.4: F1 obtenido por SVM teniendo en cuenta la aplicación de *stopper* y *stemming*.

Sabiendo de la superioridad de SVM y de TF-IDF, la Figura 4.4 solo representa los resultados obtenidos por SVM ponderando los *unigramas* con TF-IDF y teniendo en cuenta las distintas configuraciones posibles de combinación de uso de *stopper* y *stemming*. En la figura se observan dos patrones de comportamiento, el ofrecido por el subcorpus conformado únicamente por los resúmenes (R) y el proporcionado por los otros dos, de mucha más cantidad de información. Si sólo se fija la atención en el corpus R es evidente que cuanto mayor sea el nivel de normalización, mejor

es el resultado proporcionado por los algoritmos. Dicho de otra manera, cuando las opiniones son más directas o, como se ha indicado anteriormente, existe una menor diversidad léxica, la agrupación de los *unigramas* en sus representantes normalizados ayuda al proceso de clasificación. Por lo tanto para el subcorpus R, o para un corpus con una reducida diversidad léxica, interesa la eliminación de palabras sin poder de discriminación y la aplicación de un proceso de *stemming*. En cambio, en los subcorpus más extensos y con una mayor diversidad léxica el comportamiento es distinto. Hay un hecho claro que sucede tanto en C como en R_C, y es que cuando se aplica a los textos un *stemmer* la exactitud de los resultados decae. La aplicación de *stopper* sí es beneficiosa, sobre todo para el subcorpus C en el que dicha configuración es la que mejor resultado alcanza. En cambio, en el surcorpus R_C el rendimiento más alto se obtiene cuando no se aplica ningún método de reducción de características. A pesar de esto la diferencia con respecto a eliminar las denominadas *stopwords* es de un 0,09 %, es decir, la diferencia es prácticamente nula. Esa diferencia nimia se recupera con el mejor aprovechamiento de los recursos computacionales, debido a que la eliminación de las *stopwords* reduce considerablemente la dimensión del espacio en el que se representan las opiniones.

Tras repasar las dos experimentaciones anteriores (Martínez Cámara et al., 2011b) y (Martínez Cámara et al., 2011a) se puede concluir que la mejor configuración de un sistema para clasificación de la polaridad en español, con un enfoque similar al que se sigue en Recuperación de Información, es la que divide las opiniones en *unigramas*, elimina las palabras carentes de poder discriminatorio (*stopwords*), y utiliza como algoritmo de aprendizaje automático SVM.

4.4. Experimentación sobre textos cortos

A tenor de las desemejantes propiedades de los textos largos y cortos (ver Sección 3.2), cabe plantearse la cuestión de si los mismos sistemas configurados para la clasificación de la polaridad sobre textos largos van a rendir al mismo nivel sobre textos cortos. Esa pregunta sólo es factible resolverla mediante la aplicación de un marco experimental similar al suministrado a los textos largos. Al igual que en el caso anterior, la mayoría de los trabajos centrados en el trabajo sobre textos cortos están orientados al tratamiento del inglés. La red social Twitter y sus denominados mensajes, *tweets*, se han convertido en la principal fuente para las investigaciones centradas en la clasificación de la polaridad en textos cortos. Las experimentaciones en AO relacionadas con Twitter han

tenido una evolución similar a las desarrolladas sobre textos largos, es decir, comenzaron con la generación de recursos, y le siguieron el análisis de distintos métodos para la identificación de la orientación semántica que los usuarios pretenden impregnar en sus mensajes.

Go et al. (2009) fueron pioneros en el estudio de la polaridad en Twitter. Debido a que en aquella época en la que comenzaba Twitter a popularizarse no se disponía de corpus de *tweets* prestos para su uso, los autores tuvieron que compilar uno. Éstos tenían en mente el desarrollo de un método basado en aprendizaje supervisado, por lo que se requería de la elaboración de un conjunto de datos de entrenamiento.

Los conjuntos de entrenamiento en aprendizaje automático deben cumplir la premisa de ser representativos. La condición de representatividad en PLN se traduce en que los conjuntos de datos deben estar conformados por grandes colecciones de textos. El lector ya conoce que los textos de Twitter están limitados a 140 caracteres, es decir, por definición los documentos que se generan a partir de Twitter son pequeños, circunstancia que obliga a que un corpus de *tweets* esté formado por una cantidad ingente de documentos. El aprendizaje supervisado, además, necesita que se conozca la clase a la que pertenecen los documentos que se emplean para la generación del modelo de clasificación. En los experimentos relacionados con textos largos, los textos, o mejor dichos las opiniones, son extraídas de *webs* de opiniones donde la orientación de la opinión ya la ha determinado el propio autor de la opinión. Por lo que cabe preguntarse, cómo generar un corpus de *tweets* con etiquetas de opinión sin consumir una elevada cantidad de recursos humanos y económicos. Go et al. (2009) obtuvieron la respuesta en el trabajo de Read (2005).

En (Read, 2005) se afirma que cuando un usuario emplea un emoticono en un texto, lo que realmente está haciendo es marcarlo con su estado emocional o representando gráficamente su posición verbalizada en el texto. Esta premisa nos dirige a la afirmación de que es factible considerar los emoticionos de un texto como marcadores de la clase a la que pertenece. Read (2005) pretende diseñar una metodología de trabajo que reduzca las dependencias clásicas de la clasificación de la polaridad, es decir, de la rígida vinculación existente al dominio sobre el que versen los documentos, y al periodo temporal al que se circunscriben los mismos. Para ello construye un corpus de opiniones cuyas etiquetas de clase se han obtenido a partir de los emoticonos que están presentes en los textos. Con la colección de opiniones construida a partir de emoticonos, los autores consiguen aminorar la subordinación al dominio y al tiempo, demostrando así la utilidad de los emoticonos para la compilación de corpus de opiniones. Esta demostración

es la base sobre la que se apoyan Go et al. (2009) para la generación del primer corpus de opiniones o estados anímicos a partir de publicaciones de Twitter. Read (2005) no bautiza a las etiquetas de clase generadas a partir de emoticonos, de manera que para denominarlas de una manera más técnica Go et al. (2009) le asignan el nombre de *noisy labels*, cuya traducción al español podría ser "etiquetas impuras". El resultado es un corpus compuesto por *tweets* positivos, o mejor dicho con emoticonos positivos (":-)"), y *tweets* con emoticonos negativos (":-(").

Go et al. (2009) plantean una evaluación consistente en comprobar cuales son las características idóneas y los algoritmos de aprendizaje automático adecuados para la clasificación de la polaridad en *tweets*. Las características que examinan son *unigramas*, *bigramas*, *unigramas* más *bigramas* y *unigramas* con sus respectivas categorías morfológicas. Al igual que en (Pang et al., 2002) las características están ponderadas en función de la presencia de las mismas. Los algoritmos de aprendizaje automático incluidos en la experimentación son los mismoss que fueron valorados en (Pang et al., 2002), SVM, Máxima Entropía y Naïve Bayes. A tenor de los resultados, los autores concluyen que para la clasificación de la polaridad en Twitter lo más conveniente es representar los *tweets* a partir de una combinación de *unigramas* y *bigramas*, con una ponderación binaria de presencia, y empleando como algoritmos de aprendizaje automático Máxima Entropía.

La experimentación sobre textos cortos en español se inició a la par que se atisbaba una pronta popularización del uso de Twitter en España. La primera declaración de intenciones de elaborar un estudio encaminado a la clasificación de la polaridad de *tweets* en español se encuentra en (Martínez Cámara et al., 2011c). Pero para alcanzar tal objetivo, se requería de un colección de *tweets* en español en la que cada *tweet* tuviera asociado una etiqueta de polaridad. Por aquellas fechas, en las que todavía Twitter no desautorizaba la distribución de colecciones de *tweets*, no estaba disponible para la comunidad investigadora ningún corpus de *tweets* en español. Por consiguiente, si se quería iniciar el estudio sobre cómo clasificar la opinión de *tweets* escritos en español siguiendo un enfoque basado en aprendizaje automático supervisado, se debía empezar por la compilación de un corpus.

A la hora de generar un corpus hay que cuestionarse sobre cuál va a ser la fuente, cómo se van a obtener los datos, cuáles van a ser las etiquetas clase y cómo se va a almacenar el corpus. La primera pregunta tiene como respuesta directa, la red social Twitter. Para la segunda se requiere de una explicación algo más extensa. Twitter permite la descarga de las publicaciones de sus usuarios, y además la facilita ofreciendo una interfaz de

programación de aplicaciones[9], en inglés *Application Programming Interface* (API). Actualmente Twitter ofrece dos tipos de APIs[10], la conocida como REST API, y la STREAMING API. La REST API, es la API que posibilita el acceso a los perfiles de los usuarios, a realizar acciones sobre los perfiles como es la publicación de nuevos *tweets*, a obtener los *tweets* de los usuarios y a buscar *tweets* a partir de una consulta. Mientras tanto, la STREAMING API ofrece la capacidad de acceder en tiempo real al 1 % total de todos los *tweets* que se están publicando en Twitter. Otra diferencia, y técnicamente es la esencial, radica en que la REST API requiere de llamadas continuas al servicio de Twitter para obtener información, mientras que la STREAMING tiene una aquitectura PUSH, es decir, con la primera petición a la API se obtiene un manipulador o escuchante (*listener*), a través del cual sin ninguna petición adicional se van obteniendo paulatinamente los *tweets* que se van publicando asociados a una consulta. La API seleccionada fue la REST, porque la intención era recolectar *tweets* durante un periodo de tiempo, sin importar que se estuvieran publicando en el mismo instante en que se iniciara la consulta.

Sabiendo cómo se van a conseguir los *tweets* queda por determinar las clases. El objetivo era la construcción de una colección de *tweets* escritos en español para el entrenamiento de sistemas de clasificación de la polaridad de opinión, de manera que se consideró oportuno, para una primera aproximación, la generación de un corpus con etiquetas binarias de polaridad (POSITIVO, NEGATIVO). Tomando como referencia el trabajo de (Go et al., 2009), y partiendo de sus mismas premisas, se estimó oportuno seguir un enfoque de etiquetas impuras. Al igual que (Go et al., 2009), se escogieron como etiquetas impuras los emoticonos que los usuarios emplean en los *tweets* para concretar la orientación de su opinión. Con la intención de encontrar el máximo número de *tweets* con emoticonos susceptibles de contener opinión, se determinaron un conjunto de emoticonos que transmiten una idea positiva, y otros que manifiestan un mensaje negativo. Teniendo en cuenta los emoticonos expuestos en la Tabla 4.6 se lanzaron las correspondientes consultas a la plataforma Twitter, de manera que se fueron recopilando progresivamente un conjunto de *tweets* positivos y negativos.

Antes de determinar la representación del corpus almacenado y como se va a conservar, se requiere descartar del corpus todos aquellos elementos, en este caso *tweets*, que no van a ser beneficiosos para la tarea para la cual se está preparando la colección de datos. A continuación se van a listar las

[9] `https://dev.twitter.com/overview/api`

[10] En el año que se comenzó la generación del corpus, 2011, Twitter proporcionaba tres APIs, cuyos nombres eran REST, SEARCH, STREAMING.

	Emoticons
Emoticonos positivos :) (*tweets positivos*)	:) :) :-) ;) ;-) ^_^ :D :d =D C: XD xD Xd (x (= ^^ ^o^ 'u' n_n *_* *O* *o* *-*
Emoticonos negativos :((*tweets negativos*)	:-(:(:((: (D: Dx 'n' :\ /:):-/ :' ='[/T_T TOT ;_;

Tabla 4.6: Emoticonos considerados en la generación del corpus.

tareas emprendidas para la limpieza del corpus:

1. *Retweets*: En el momento que se elaboró el corpus, en Twitter existían dos métodos diferentes para realizar un *retweet*. El primero consistía en copiar el contenido del *tweet* que se quería difundir de nuevo precedido de una declaración de la autoría del mismo. Esa declaración de autoría estaba compuesta por tres elementos: las siglas RT de *retweet*, el nombre de usuario del autor precedido por el símbolo @, y por dos puntos. Un proceso de abstracción permite comprender que la presencia en un corpus de *retweets* implica la repetición de información sin aportar nada nuevo. Esa reiteración de información para experimentos basados en aprendizaje automático constituyen un riesgo, dado que la repetición del mismo fragmento de texto puede conducir a los algoritmos a otorgar demasiada importancia a unos términos, no por que la tengan, sino por el simple hecho circunstancial de que el fragmento de texto en la fuente de datos ha sido replicado varias veces[11]. Por este motivo se eliminaron del corpus en construcción todos los *retweets* que se descargaron.

2. Menciones: Las menciones, es decir, las referencias a otros usuarios en un *tweet* tampoco aportan información relevante al proceso de clasificación de la polaridad, de manera que se estimó oportuno el borrado de la mención en sí (@nombre_de_usuario) del corpus en construcción.

3. Enlaces: Es muy común que los usuarios enlacen contenido en sus *tweets*. Para una tarea de clasificación de la opinión los enlaces en sí mismos sólo serían relevantes si se añadiera al corpus el contenido de la

[11] Hay que recordar que se está compilando un corpus destinado a tareas de clasificación de la polaridad a nivel de documento, considerando únicamente el contenido textual de cada *tweet* individualmente. Si el objetivo hubiera sido la generación de un corpus para la tarea de Análisis de Reputación, sí se hubiera incluido información relativa a los *retweets* que ha ido cosechando un determinado *tweet*.

página que se está enlazando. Dado que se pretendía la generación de un corpus únicamente constituido por *tweets*, se decidió la eliminación de los mismos de los *tweets*.

4. *Hashtags*[12]: Estas etiquetas de Twitter identifican de manera única a un tema en Twitter. Dado que el objetivo del corpus es simplemente proveer a la comunidad investigadora de una colección de *tweets* sobre la que estudiar técnicas específicas de clasificación de la opinión, se consideró que no aportaban información para la tarea de clasificación de la polaridad, y por ende se eliminaron de todos los *tweets* en los que aparecían.

5. Eliminación de retornos de carro: Muchos *tweets* al ser descargados aparecían escritos en varias líneas, lo cual entorpecía su procesamiento. Para facilitar el trabajo de la comunidad investigadora, se estimó suprimir los caracteres de retornos de carro, de manera que los *tweets* aparecieran escritos en una sola línea.

6. Emoticonos opuestos: No es raro toparse con *tweets* en los que aparecen emoticonos con orientaciones semánticas opuestas, es decir, *tweets* donde se encuentra presentes emoticonos del grupo representado por ":)", y emoticonos simbolizados por ":(". La interpretación de este tipo de *tweets* no es directa, dicho de otra manera, no se puede afirmar que son claramente positivos o negativos, por lo que lo más prudente, teniendo en cuenta el objetivo del corpus, es la eliminación de este tipo de *tweets*. Un ejemplo de esta clase de *tweets* sería el que se muestra en la Figura 4.5.

```
@fragilejunkie Yo también te extraño Marian! :(
decime yaa un día de la semana que viene cuando
quieras que nos juntamos a la tarde! :)
```

Figura 4.5: *Tweet* de ejemplo de emoticonos opuestos.

7. Emoticonos carentes de orientación semántica evidente: La polaridad de una opinión, como ya se ha precisado varias veces, puede variar en su caso más simple entre positivo o negativo, o oscilar en una escala más amplia de valores de intensidad de opinión. Como se

[12]Se emplea el término inglés porque se ha convertido de facto en un nombre propio que denomina a las etiquetas que distingue a un tema en Twitter.

Emoticonos sin orientación semántica evidente	:-P :P :PP \(

Tabla 4.7: Emoticonos identificados sin orientación semántica evidente.

ha indicado antes, el corpus únicamente va a albergar documentos positivos y negativos. Por este motivo se ha extremado el cuidado para no dejar *tweets* cuya intensidad de polaridad sea tan leve que no pudiera precisarse si es positivo o negativo, teniéndose en dicho caso que asignársele la clase neutro. Durante el proceso de generación del corpus se cayó en la cuenta de que existen un grupo de emoticonos que los usuarios españoles no los emplean para describir gráficamente la opinión que acaban de publicar, sino como un elemento más bien decorativo. A este tipo de emoticonos se le ha asignado el nombre de emoticonos carentes de una orientación semántica evidente. Entre la gran diversidad de emoticonos que existen, se han identificado los que se recogen en la Tabla 4.7 como emoticonos carentes de orientación semántica evidente.

8. Letras repetidas: El estilo de redacción en Twitter es informal, por lo que es muy común el uso de la repetición de caracteres como recurso para la expresión de intensidad. Se puede afirmar que no existe ninguna norma que determine la cantidad de caracteres que se corresponde con un nivel de intensidad de opinión, de manera que se podrían solamente distinguir entre términos con y sin intensidad. A los términos que manifiestan intensidad se le debe aplicar un procesamiento para unificarlos, ya que cada usuario estima que la intensidad del mensaje que quiere transmitir se corresponde con un número determinado de reiteración de caracteres. Por lo tanto, en la línea de purificar al máximo el corpus, se aplicó un proceso de homogeneización de los términos con caracteres repetidos, de modo que independientemente del número de letras repetidas, éstas se reducían a solo dos repeticiones. Seguramente con un ejemplo (Figura 4.6) se pueda comprender mejor el proceso.

```
cosaaaa hermosaaaa te quiero mucho :)
```

Figura 4.6: *Tweet* con algunos términos con letras repetidas.

Como se puede comprobar dos de los términos cuentan con letras

repetidas. El proceso que se ha aplicado es transformar la versión del término con x repeticiones de una letra en otra versión con solo 2 reiteraciones. El resultado se puede ver en la Figura 4.7.

```
cosaa hermosaa te quiero mucho :)
```

Figura 4.7: *Tweet* transformado sin términos con letras repetidas

9. Carcajadas: Es complicado encontrar una manera común entre los usuario de la red social Twitter de expresar una carcajada, y además ocurre como con los términos con letras repetidas, que el número de reiteraciones de la onomatopeya de risa depende de la gracia que le cause el mensaje al autor del mismo. Respetando las distintas interpretaciones que pueden tener las diversas formas de representar una carcajada, risa o sonrisa, se han homogeneizado dichas expresiones intentando que la onomatopeya de risa únicamente tenga dos repeticiones. En la Tabla 4.8 se puede apreciar el proceso de transformación que se llevó a cabo.

Carcajada	**Transformación**
jajajajaja...	jaja
jejejejeje...	jeje
jijijijiji...	jiji
jujujujuju...	juju
Lol	jaja
Juasjuasjuas...	juas
Muajajajaja...	Buaja
Buajajajaja...	Buaja

Tabla 4.8: Transformación de expresiones de carcajada.

10. Por último se debe indicar que, dado que los emoticonos se han empleado para la determinación de las clases, se han eliminado toda presencia de emoticonos en los *tweets*.

Queda aún por indicar cómo se almacenó el corpus. Se determinó que para facilitar la distribución del conjunto de *tweets* se deberían almacenar en un fichero XML.

El resultado de todo el proceso fue un corpus balanceado de 34634 *tweets* de los cuales 17317 son positivos y 17317 son negativos, los cuales fueron publicados entre el 3 de marzo de 2011 y el 4 de marzo de 2011. El corpus se bautizó con el nombre anglosajón *Corpus of Spanish Tweets* (COST)[13], que en español quiere decir Corpus de *tweets* en español. Para que el lector se pueda hacer una idea del contenido de COST, el extracto Código 4.1 muestra un pequeño ejemplo del mismo. Para facilitar la compresión de todas las etapas que constituyen el proceso de generación de COST, se recomienda al lector consultar la Figura 4.8.

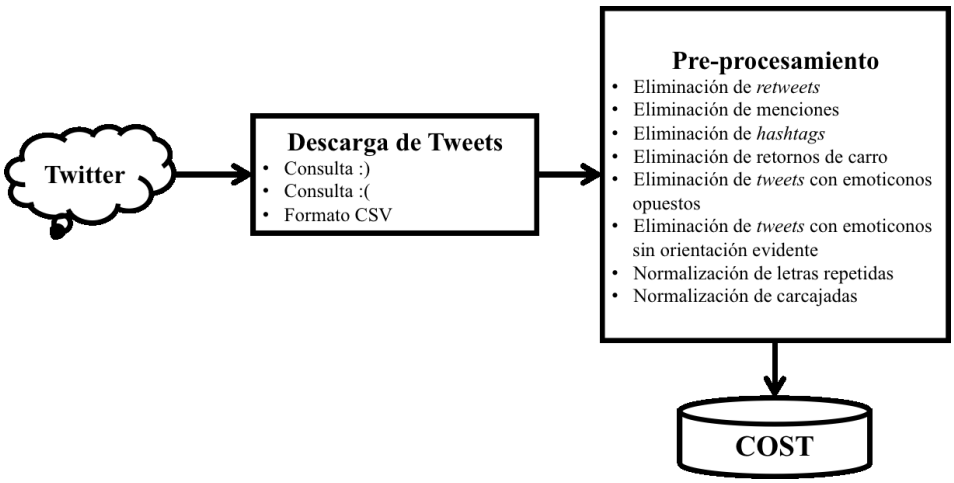


Figura 4.8: Proceso de generación del corpus COST.

Ya se dispone de un conjunto de datos sobre el cual realizar una experimentación, que permita llevar a cabo una primera aproximación al estudio de la clasificación de la polaridad sobre *tweets* en español. Se van a detallar la experimentación publicada en (Martínez-Cámara et al., 2015). En dicho trabajo, la experimentación que se realizó es similar a la realizada en (Martínez Cámara et al., 2011b) y (Martínez Cámara et al., 2011a). La intención no es otra que comprobar si se pueden extraer unas conclusiones similares a la obtenidas con opiniones largas. Por lo tanto, la evaluación ha consistido en determinar la validez de las mismas cuatro medidas de representación de la relevancia de los términos que se han

[13]La información de cómo se puede descargar el corpus se encuentra en `http://sinai.ujaen.es/cost-2/`

```
<tweet>
        <id>43297892068372480</id>
        <date>Thu Mar 03 14:13:16 CET 2011</date>
        <lang>es</lang>
        <geo>null</geo>
        <text>
                a comer y volviendo poco a poco a
                   currar de forma normal ya me
                   empiezo a encontrar mucho mejor y
                   me estoy emocionando
        </text>
        <polarity>1</polarity>
</tweet>
<tweet>
        <id>43335603206635520</id>
        <date>Thu Mar 03 16:43:07 CET 2011</date>
        <lang>es</lang>
        <geo>null</geo>
        <text>
                la comidaa de mi casaa huelee buueenaa
                    mm
        </text>
        <polarity>1</polarity>
</tweet>
```

Código 4.1: Extracto del corpus COST.

visto anteriormente, es decir, TF-IDF, TF, TO y BTO. También se ha validado el aporte que realiza la eliminación de las palabras carentes de potencial identificativo de clases, *stopwords*, y la aplicación de un proceso de *stemming*.

En (Martínez Cámara et al., 2011b) se evaluaban cinco algoritmos de aprendizaje automático, de los cuales solo tres destacaban, o mejor dicho, solo tres obtenían resultados aceptables. Esos tres algoritmos eran, SVN, Naïve Bayes y BBR. El algoritmo BBR obtuvo muy buenos resultados, sobretodo en la experimentación realizada con el corpus completo, lo cual nos indicaba que BBR requería de grandes volúmenes de información para poder construir correctamente la función de regresión que se ajusta a la distribución de los datos. El problema de BBR radicaba en su elevado consumo de recursos computacionales, fundamentado principalmente en las

diversas iteraciones que realiza el algoritmo para la determinación de los numerosos coeficientes de regresión. Ésto llevaba a justificar, a pesar de los buenos resultados, el descarte, al menos, de ese algoritmo de regresión para la clasificación de la polaridad en español. A pesar de este hecho, no se quería dejar huérfana una experimentación de clasificación de la polaridad sobre *tweets* de un algoritmo de regresión, de manera que se intentó encontrar otro método de regresión que requiriera de menos recursos computacionales. Se estimó que sería un buen candidato el algoritmo definido en (Keerthi et al., 2005). Este algoritmo de regresión logística está basado en el algoritmo de Optimización Secuencial Mínimo o SMO por sus siglas en inglés (*Sequential Minimal Optimization*) (Platt, 1998; Keerthi et al., 2001). El algoritmo de Keerthi et al. (2005), a diferencia de BBR, minimiza considerablemente el número de iteraciones a llevar a cabo para el cálculo de los coeficiente de regresión, dado que, como indican sus autores, el algoritmo no realiza ninguna operación sobre la matriz núcleo, y por tanto recomiendan su uso para problemas con un elevado número de datos y características, como es el caso que nos ocupa en este momento. En las tablas de resultados (Tablas 4.9 y 4.10) el algoritmo de Keerthi et al. (2005) se referencia como RL.

Conociendo lo que se va evaluar, el siguiente paso es mostrar los resultados. La exposición va a seguir el mismo orden que en el caso de textos largos, es decir, primeramente se va a evaluar la efectividad de los algoritmos con distintas métricas de relevancia de las características, y posteriormente se mostrará el aporte al proceso de los métodos de eliminación de términos y normalización morfológica. En la Tabla 4.9 se pueden leer los resultados alcanzados por los tres métodos de clasificación.

La Figura 4.9 va a asistir al proceso de análisis de los resultados presentados en la Tabla 4.9. Al mirar la gráfica se obtiene una conclusión que no requiere de mucha reflexión, y no es otra que el algoritmo más adecuado para la clasificación de la polaridad en *tweets* en español es SVM. En lo que respecta a las medidas de mesura de la importancia de los *unigramas* se debe afinar algo más. Como se puede comprobar todas están entre un 73 % y un 74 % de F1, siendo las diferencias entre ellos muy reducidas. Por contra, de las cuatro medidas, la que valora la frecuencia relativa de los *unigramas* con respecto al tamaño del corpus es la que destaca en cierta manera sobre las otras tres. Si en los textos largos el uso de una medida que valora, al igual que en Recuperación de Información, la cualidad de los términos de concentrarse en subgrupos de documentos correspondientes a las clases, es decir TF-IDF, en la clasificación sobre textos cortos (*tweets*), parece que los clasificadores valoran más los términos

Algoritmo	**Relevancia**	**Precisión**	***Recall***	**F1**
SVM	TF-IDF	73,42%	73,24%	73,33%
	TF	73,96%	73,98%	73,92%
	TO	73,56%	73,21%	73,38%
	BTO	73,68%	73,20%	73,44%
NB	TF-IDF	66,01%	65,44%	65,72%
	TF	64,22%	62,99%	63,60%
	TO	62,70%	60,72%	61,69%
	BTO	62,92%	60,93%	61,91%
RL	TF-IDF	66,28%	63,65%	64,94%
	TF	66,88%	63,52%	65,16%
	TO	65,14%	59,16%	62,01%
	BTO	65,40%	59,62%	62,38%

Tabla 4.9: Clasificación de la polaridad con diferentes algoritmos y medidas de relevancia sobre COST.

que tienen una mayor frecuencia relativa en relación a todo el corpus. Se emplea bien el verbo "parece" cuando se indica que TF es la medida de importancia que mejor valora la relevancia de los *unigramas*, porque todavía no se ha evaluado la influencia en el proceso de clasificación de la eliminación de *stopwords*, y de la aplicación de *stemmer*. La Tabla 4.10 muestra los resultados alcanzados en esta evaluación.

De nuevo se recurre a una gráfica (Figura 4.10) para facilitar el entendimiento de los resultados obtenidos, y asistir a la comprensión del respectivo análisis. Ya se había comprobado en la anterior experimentación que TF es claramente superior a las otras medidas de graduación de la importancia, y en la Figura 4.10 se puede observar de nuevo. En la gráfica también se puede observar un comportamiento muy claro, como es que la eliminación de las *stopwords* perjudica seriamente el proceso de clasificación. Se puede ver que siempre que se usa *stopper* (Stop_Stem, Stop) se obtienen peores resultados que en el caso base (Nada). Por contra, el uso de *stemmer* se traduce en todos los casos en una mejora significativa con respecto al caso base. Este comportamiento pone aún más en evidencia la rémora que supone el empleo de *stopper*, dado que la eliminación de las palabras sin capacidad de diferenciación limita la capacidad del *stemmer* de representar mejor los *tweets*. Una lectura más simple de los resultados nos indica que la limitada longitud que presentan los *tweets* obliga a que si se aplica un proceso de reducción de características éste sea muy

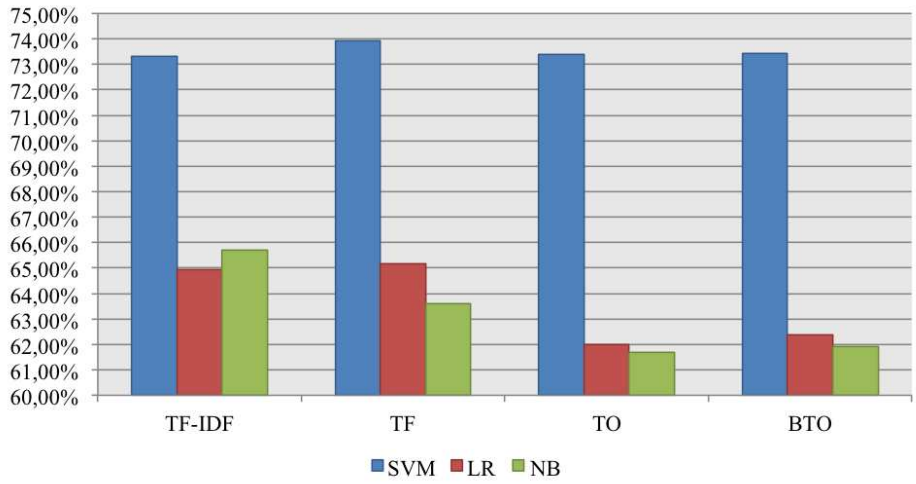


Figura 4.9: Comparación de resultados obtenidos por tres algoritmos con cuatro medidas de relevancia

cuidadoso, dado que es muy probable que se cercene considerablemente la de por sí reducida información que contiene un *tweet*. Aunque técnicamente un *stopper* no es muy ortodoxo denominarlo como proceso de reducción de características, las verdad es que su uso implica la eliminación de un gran número de ellas, al borrar todas aquellos términos que se encuentran presentes en una lista de *stop*. A la luz de los resultados, la simpleza que implica un *stopper* perjudica considerablemente a la clasificación de la polaridad sobre *tweets*. Por contra, esos mismos resultados, desvelan que un proceso de normalización morfológica, por simple que sea, ayuda ostensiblemente al proceso de clasificación. A tenor de los resultados, la configuración más adecuada para emprender una clasificación de la polaridad con aprendizaje supervisado es la que toma como características *unigramas*, aplica un proceso de *stemmer*, mide la importancia de las características con la métrica TF, y emplea el algoritmo SVM.

4.4.1. Participación en campañas de evaluación

Leyendo los párrafos anteriores, tanto los relacionados con la experimentación sobre textos largos como de textos cortos, se podría caer en la equivocación de pensar que cuando se elige como método de clasificación un algoritmo de aprendizaje automático sólo se pueden emplear como características los propios términos del documento, pero nada más lejos de la realidad. Lo que requieren los algoritmos de aprendizaje automático es

Relevancia	Stop	Stem	Precisión	*Recall*	F1
TF-IDF	✓	✓	71,88 %	71,86 %	71,87 %
	✓		71,69 %	71,68 %	71,68 %
		✓	73,56 %	73,43 %	73,49 %
			73,42 %	73,24 %	73,33 %
TF	✓	✓	72,61 %	72,60 %	72,60 %
	✓		72,25 %	72,24 %	72,24 %
		✓	74,29 %	74,27 %	74,28 %
			73,96 %	73,88 %	73,92 %
TO	✓	✓	72,73 %	72,73 %	72,73 %
	✓		72,44 %	72,43 %	72,43 %
		✓	74,20 %	73,94 %	74,07 %
			73,56 %	73,21 %	73,38 %
BTO	✓	✓	72,89 %	72,89 %	72,89 %
	✓		72,55 %	72,55 %	72,55 %
		✓	74,16 %	73,94 %	73,99 %
			73,68 %	73,20 %	73,44 %

Tabla 4.10: Evaluación con SVM de diferentes medidas de ponderación de la importancia de *unigramas* sobre COST.

que se represente a los objetos a clasificar, en este caso documentos de opinión, de la manera más fidedigna posible. Esas características pueden ser los mismos términos, o pueden ser características derivadas de operaciones aritméticas relacionadas con propiedades y relaciones de los propios términos del documento.

La inclusión de características adicionales a los propios términos del documento es lo que se intentó en la participación en el taller TASS , del cual ya se ha hablado en el Capítulo 2. El taller proponía una clasificación algo más complicada de lo que se había intentado hasta ese momento, como atestiguan los párrafos anteriores. La primera edición del TASS, en la tarea concerniente a la clasificación de la polaridad, propone una clasificación en seis niveles de opinión, y otra clasificación en cuatro niveles de intensidad de opinión. Martínez Cámara et al. (2013b) explican las decisiones que se tomaron para aplicar la experiencia adquirida en clasificación de la polaridad binaria sobre textos cortos, con el fin de construir un sistema que tuviera la capacidad de clasificar *tweets* en seis y cuatro clases. Recordando, lo indicado en el Capítulo 2, los seis niveles de intensidad son: muy positivo

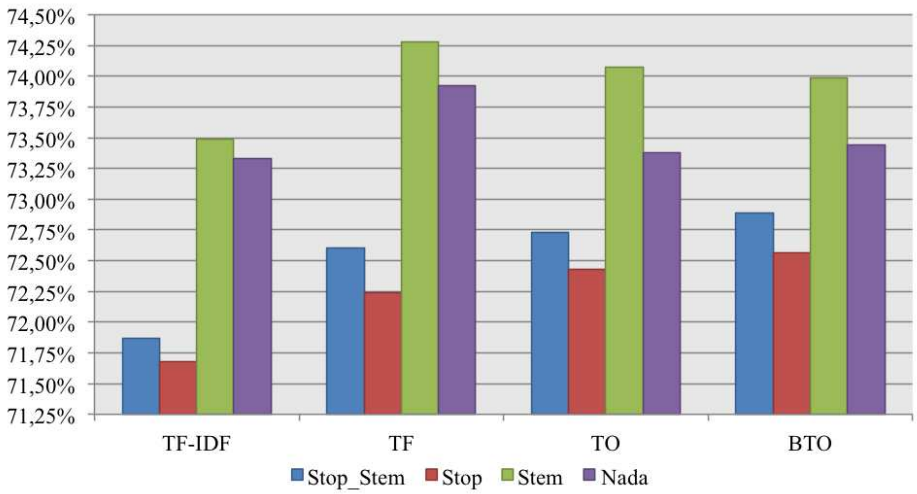


Figura 4.10: Comparación de resultados en COST empleando SVM, diferentes medidas de ponderación, y diversas combinaciones de *stopper* y *stemmer*.

(P+), positivo (P), neutro (NEU), negativo (N), muy negativo (N+), texto carente de opinión (NONE). La subtarea que exige la clasificación en cuatro estratos de intensidad unifica en dos clases los dos niveles de intensidad de positivo y negativo, quedándose, por tanto, las siguientes clases: positivo (P), neutro (NEU), negativo (N), texto carente de opinión (NONE).

Para intentar modular los distintos estratos de intensidad de opinión se han intentado definir una serie de características que aporten algo más de información a un algoritmo de aprendizaje automático. El algoritmo empleado atendiendo a la experiencia adquirida fue SVM. Sin más circunloquios, las características que se analizaron fueron:

- *Unigramas*: Se toman los términos que aparecen en los *tweets* como características. Como métrica de ponderación de la relevancia se utiliza TF debido a su buen rendimiento mostrado en experimentaciones anteriores (ver Figura 4.10). Como *unigramas* se pueden considerar las menciones y las URLs que puedan aparecer en el *tweet*. Tratar las menciones y URLs como términos independientes no constituye un suministro importante de información al proceso de clasificación. Se diseñaron evaluaciones del sistema en los que se unificaban todas las menciones que aparecían en los *tweets* como MENTION, y las direcciones de recursos en Internet como URL. Con esta unificación se pretendía evaluar si la presencia de estos dos tipos de elementos proporcionaban información de calidad al clasificador.

- Emoticonos: Se añade como característica el número de emoticonos positivos o negativos presentes en el *tweet*. Para ello se emplea el conjunto de emoticonos listados en la Tabla 4.6.

- Palabras positivas y negativas: Como características se han añadido el número de palabras positivas y negativas que hay en los *tweets*. El lector avezado ya estará intentando adelantarse a la lectura pensando en los posibles listas de palabras de opinión en español existentes en la actualidad, pero en la época en la que se diseñó el sistema que participaría en el TASS, ninguna lista de palabras indicadoras de opinión en español estaba disponible. Éste hecho motivó la traducción automática de un léxico de opinión en inglés muy usado por la comunidad investigadora en AO. Se tomó el conjunto de palabras de opinión de Bing Liu (Hu & Liu, 2004), se tradujo al español y se empleó para contar el número de palabras positivas y negativas.

- Intensidad: La organización del taller requería una clasificación en varios niveles de intensidad. Si se analiza el lenguaje empleado por los usuarios de Twitter, es fácil llegar a la conclusión de que muchos usuarios emplean determinados recursos léxicos para pincelar notas de intensidad en el *tweet*. Tomando el requisito de la tarea, y las propias características del lenguaje empleado en Twitter, parece recomendable la inclusión de características que representen la intensidad con la que se expresan los usuario de la red social.

  Para la representación de la intensidad se han considerado dos recursos lingüísticos. En primer lugar se ha atendido a la reiteración de letras en los vocablos. Cuando un usuario escribe repetidamente un carácter de un término es evidente que lo que quiere es que el significado de esa palabra sea transmitido con una mayor intensidad. Por ende, todas aquellas palabras indicadoras de opinión que aparecen con letras repetidas se considera como si fuera dos palabras, de manera que si es positiva, el contador de palabras positivas se incrementa en dos en lugar de en uno, mientras que si es negativa, en el contador de palabras negativas se aplica la misma operación.

  El segundo recurso lingüístico que se emplea para transmitir intensidad es idéntico al usado en la lengua escrita convencional, que no es otro que el signo de admiración. Por consiguiente, toda aquella palabra de opinión que esté acompañada por un signo de admiración se la considerará doble, de manera que sumará dos en su correspondiente contador de polaridad.

- Negación: La negación es un fenómeno lingüístico extremadamente complejo cuyo tratamiento en AO es de suma importancia. El procesamiento de la negación se realizó siguiendo un enfoque simple basado en la toma en consideración de una ventana de palabras, es decir, que si en el ámbito de la palabra indicadora de opinión, constituido por tres palabras que la preceden (ventana de longitud tres) aparece una partícula negativa, entonces se invierte la orientación semántica del término, y se suma en el contador de polaridad opuesto. En otras palabras, que si un vocablo positivo está precedido por una partícula negativa, entonces la palabra se interpreta como negativa, y si es un término negativo el afectado por una partícula de negación, entonces se toma como positiva.

Las posibles combinaciones que pueden dar lugar la consideración de cada una de las características listadas anteriormente originaron cuarenta y dos experimentos diferentes, los cuales se van a intentar explicar para mostrar la envergadura del trabajo previo realizado a la presentación del sistema definitivo en la competición organizada por el taller.

EXP1: El conjunto de características estuvo constituido exclusivamente por *unigramas*, eliminando menciones y URLs. En esta experimentación no se normaliza morfológicamente los *unigramas* mediante el uso de un *stemmer*.

EXP2: Igual que el caso anterior, pero los *unigramas* son normalizados por un *stemmer*.

EXP3: Igual que EXP2, pero en éste caso se normalizan las menciones, se eliminan las URLs, y no se aplica normalización morfológica.

EXP4: Igual que EXP3, pero en éste caso sí se aplica *stemmer*.

EXP5: Igual que EXP3, pero en éste caso se normalizan además las URLs.

EXP6: Igual que EXP5, pero aplicando *stemmer*.

EXP7: Se añaden como características el número de emoticonos positivos y negativos. También se incluyen como características el numero total de palabras positivas y negativas, así como el número de palabras positivas y negativas que cuentan con letras repetidas. En este experimento no se normaliza morfológicamente los *unigramas*

EXP8: Igual que el EXP7, pero aplicando un proceso de *stemming* a los *unigramas*.

EXP9: Igual que el EXP7, pero sin incluir las características léxicas, es decir, en este caso no se emplean los *unigramas* como características, de manera que solamente el número de emoticonos positivos y negativos, el número de palabras positivas y negativas, y el número de palabras positivas y negativas con letras repetidas representan los *tweets*.

EXP10: Igual que el EXP7, pero los contadores de términos con caracteres repetidos de palabras positivas y negativas, así como de emoticonos positivos y negativos en lugar de ser absolutos son relativos al número total de palabras del *tweet*.

EXP11: Igual que EXP10, pero aplicando *stemming*.

EXP12: Igual que EXP10, pero sin características léxicas, es decir sin considerar los *unigramas*.

EXP13: En esta configuración se tienen en cuenta los *unigramas* ponderados por su frecuencia relativa (TF), las palabras indicadoras de opinión no se consideran una característica distinta de las las palabras indicadoras de opinión con caracteres repetidos, y además, se añaden al conjunto de características, como desde EXP7, el número de emoticonos positivos y negativos.

EXP14: Lo mismo que EXP13, pero en esta ocasión a los *unigramas* se le aplica un *stemmer*.

EXP15: Lo mismo que EXP13, pero sin considerar los *unigramas*.

EXP16: Lo mismo que EXP13, pero en lugar de las características estar expresadas en término absolutos, están medidas en términos relativos.

EXP17: Igual que EXP14, pero en lugar de las características estar expresadas en término absolutos, están medidas en términos relativos.

EXP18: Lo mismo que EXP15, pero en lugar de las características estar expresadas en término absolutos, están medidas en términos relativos.

EXP19: Igual que EXP13, pero en este caso ya sí se consideran con un valor doble los términos indicadores de opinión con letras repetidas.

EXP20: Como EXP19, pero en esta ocasión aplicando normalización morfológica (*stemmer*).

EXP21: Igual que EXP20, pero sin tener en cuenta las característica que se generan a partir de los *unigramas*.

EXP22: Igual que EXP19, pero en este caso las características numéricas se miden en términos relativos.

EXP23: Igual que EXP22, pero en esta ocasión aplicando *stemmer*.

EXP24: Igual que EXP22, pero sin tener en cuenta las características que se generan a partir de los *unigramas*.

EXP25: Igual que EXP7, pero teniendo en cuenta la inversión de la polaridad provocada por la presencia de partículas negativas cerca de los términos indicadores de opinión.

EXP26: Igual que EXP25, pero en esta ocasión aplicando un proceso de *stemming* a los *unigramas*.

EXP27: Lo mismo que EXP25, pero sin características léxicas, es decir, sin las características que se generan a partir de los *unigrmas*.

EXP28: Igual que EXP13, pero llevando a cabo un proceso de tratamiento de la negación.

EXP29: Lo mismo que EXP28, pero en esta ocasión utilizando un *stemmer* sobre los *unigramas*.

EXP30: Igual que EXP28, pero sin considerar las características léxicas.

EXP31: Exactamente igual que EXP19, pero realizando el proceso de tratamiento de negación diseñado.

EXP32: Igual que EXP31, pero aplicando *stemming*.

EXP33: Lo mismo que EXP31, pero sin tener en cuenta las características léxicas.

EXP34: Igual que EXP25, pero en esta ocasión considerando los signos de exclamación para el cálculo de la intensidad.

EXP35: Lo mismo que EXP34, pero en esta configuración del sistema se aplica *stemming* a los *unigramas*.

EXP36: Exactamente igual que EXP34, pero sin tener en cuenta las características que se generan a partir de los *unigramas*.

EXP37: Igual que EXP28, pero considerando los signos de exclamación.

EXP38: Igual que EXP37, pero aplicando un proceso de *stemming*.

EXP39: Igual que EXP37, pero sin tener en cuenta las características léxicas (*unigramas*).

EXP40: Lo mismo que EXP31, pero considerando los signos de exclamación.

EXP41: Igual que EXP40, pero además aplicando un proceso de *stemming*.

EXP42: Igual que EXP40, pero sin tener en cuenta las características léxicas.

Todos estos experimentos se evaluaron para seleccionar los que podrían obtener un mejor rendimiento en el taller. A priori se tenía el convencimiento de que las configuraciones en las que se tenía en cuenta la negación, e incluso aquellas en las que se considerara las exclamaciones con la intención de incluir información que representara la intensidad, devolverían mejores resultados y, por ende, serían las candidatas para ser enviadas a la organización del TASS. La Figura 4.11 muestra una gráfica con los valores de F1 obtenidos por los experimentos más relevantes ordenados de mayor a menor.

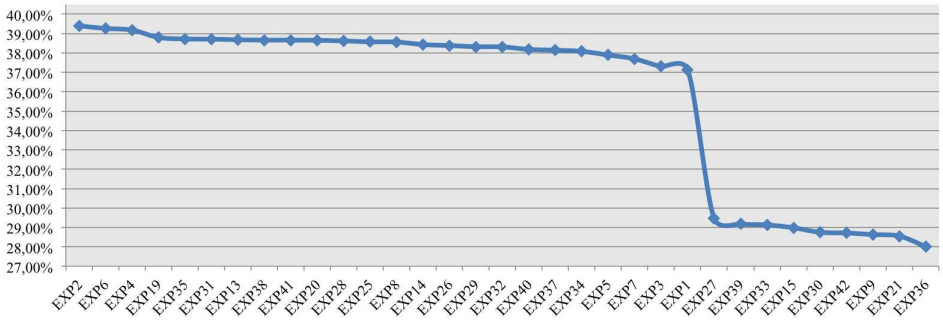


Figura 4.11: Valores de F1 obtenidos por los experimentos más relevantes.

La Figura 4.11 visualiza que las cuatro primeras configuraciones no incluyen ni tratamiento de la negación, ni consideración del signo de admiración. El quinto y el sexto resultado, que se corresponden con la configuración EXP35 y EXP31 respectivamente, son configuraciones que

sí incluyen el procesamiento de esos dos fenómenos lingüísticos, pero esos sistemas no fueron seleccionados para su participación en el TASS, porque la organización limitaba a cuatro el número de sistemas que cada equipo podía enviar. Otra conclusión, que a simple vista se puede extraer, es que todas aquellas configuraciones que no incluyen características léxicas, es decir, aquellas que se obtienen a partir de los *unigramas*, obtienen peores resultados que las configuraciones que sí las consideran.

Centrando el análisis en los cuatro sistemas que ofrecieron mejores resultados, se puede comprobar que EXP2, el cual se corresponde con una configuración idéntica al sistema descrito en (Martínez-Cámara et al., 2015), es decir, en representar los *tweets* como vectores de *unigramas* ponderados por su valor de frecuencia relativa a los que previamente se le ha aplicado un proceso de *stemming*. La siguiente configuración mejor posicionada es aquella en la que de nuevo se aplica *stemming*, por lo que se sigue reafirmando las conclusiones de Martínez-Cámara et al. (2015), pero además se homogeneizan las menciones y las URLS. Este hecho manifiesta que algo aparentemente estéril para la indicación de opinión como es la presencia de menciones y direcciones web, parece que ayuda a SVM a separar los *tweets* en diferentes escalas de intensidad de opinión. El experimento subsiguiente a EXP6 es EXP4, que es exactamente igual pero en el que las URLs no son normalizadas sino eliminadas. Parece que el eliminar las direcciones web no le sienta muy bien al proceso de clasificación, de manera que es un nuevo indicativo de la importancia de la presencia de direcciones web para ayudar al clasificador en su proceso de asignación de clases. En cuarta posición se encuentra el sistema que incluye características de opinión, como son el número de emoticonos positivos y negativos, el número de palabras positivas y negativas, y en el caso de alguna de esas palabras cuente con letras repetidas el vocablo cuenta doble. Antes de proseguir, se requiere precisar que estas cuatro configuraciones son las que probablemente mejor resultado pueden obtener, a tenor de los resultados que se han obtenido con los datos de entrenamiento. Como se verá más adelante, el orden en su clasificación particular se ve alterada cuando los sistemas generados con esas configuraciones se aplican a los datos de test. Los resultados que se obtuvieron en la competición se muestran en la Tabla 4.11.

Los resultados obtenidos en la competición (Tabla 4.11) son a priori algo desconcertantes si se comparan con los resultados alcanzados en la evaluación previa, la cual fue empleada para guiar la selección de la configuración del sistema más apropiada para alcanzar unos mejores resultados en la competición. Pero el desconcierto se centra exclusivamente

EXP	Clases	Precisión
EXP2	P+	81,00%
	P	87,00%
	NEU	0,00%
	N	0,00%
	N+	0%
	NONE	99,31%
	General	35,28%
EXP6	P+	4,69%
	P	27,00%
	NEU	0,00%
	N	12,00%
	N+	37,00%
	NONE	96,48%
	General	35,65%
EXP4	P+	58,00%
	P	47,00%
	NEU	0,00%
	N	1,00%
	N+	4,00%
	NONE	98,65%
	General	34,97%
EXP19	P+	60,99%
	P	94,00%
	NEU	38,0%
	N	33,61%
	N+	32,00%
	NONE	71,52%
	General	54,68%

Tabla 4.11: Resultados obtenidos en la edición 2012 de la competición organizada en el TASS.

en la discordancia con los resultados de la evaluación previa. La hipótesis de partida se basaba en que cuanta más características relacionadas con la interpretación de la orientación semántica de los distintos elementos que conforman el *tweet* se incluyeran, mejor sería el comportamiento de la clasificación emprendida por el algoritmo de clasificación. Para validar la hipótesis se diseñaron diversos sistemas y se evaluaron mediante validación cruzada. Los resultados que se obtuvieron no confirmaban la hipótesis, dado que la primera configuración que incluye información de opinión (EXP19) obtuvo el cuarto mejor resultado. Por contra, la evaluación emprendida por la organización del taller invierte el orden obtenido en la evaluación previa. Teniendo en cuenta la diferencia considerable entre los tamaños del conjunto de entrenamiento (7219) y el de test (60 798), puede considerarse que los *tweets* que conforman el entrenamiento no son los suficientes para la elaboración de un modelo de aprendizaje automático, que represente de la manera más adecuada posible la información de opinión que conllevan los *tweets*.

Lo que sí reafirma, tanto los experimentos de evaluación previa como los realizados por la organización del TASS, es la conclusión alcanzada en (Martínez-Cámara et al., 2015), es decir, que la representación de los *tweets* como vectores de *unigramas* a los que previamente se le ha aplicado un proceso de normalización morfológica (*stemming*) es la más beneficiosa para la clasificación de la opinión. Además, como apuntan los resultados obtenidos en la edición de 2012 de TASS, se puede concluir que, la inclusión de información relacionada con la orientación de los elementos que constituyen el *tweet*, como pueden ser emoticonos, la cantidad de palabras positivas y negativas e incluso tener en cuenta características lexicográficas utilizadas por las personas para transmitir intensidad, como es la repetición de caracteres, ayuda a mejorar la calidad de la clasificación de la polaridad.

4.5. Conclusión

En el presente capítulo, se ha expuesto el estudio de la aplicación de técnicas de aprendizaje automático supervisado a la identificación de la polaridad de textos de opinión escritos en español. El estudio ha diferenciado los dos tipos de textos identificados en el Capítulo 3: textos largos y textos cortos. Tras la descripción de la experimentación, es momento de exponer las conclusiones a las que se ha llegado:

1. La experimentación ha demostrado que los textos largos y cortos requieren de un tratamiento diferente.

2. Con una representación basada en vectores de *unigramas* se obtienen buenos resultados en ambos tipos de textos.

3. Los textos largos prefieren TF-IDF para medir la relevancia de los *unigramas*. Por contra, los textos cortos alcanzan unos mejores resultados cuando se emplea la frecuencia relativa (TF) para medir la importancia de los *unigramas*.

4. Las diferencias también llegan a los métodos de reducción de características. Con los textos largos se consiguen mejores resultados cuando son eliminadas las *stopwords* y no se emplea un *stemmer*, mientras que en los textos cortos ocurre todo lo contrario, los resultados más favorables se alcanzan cuando se utiliza el *stem* de los *unigramas* y no se eliminan las *stopwords*.

5. En los dos tipos de texto, SVM es el algoritmo que logra mejores resultados de clasificación.

5

Aprendizaje No Supervisado

5.1. Introducción

Hasta ahora se ha intentado ofrecer soluciones al problema de determinar la polaridad de una opinión siguiendo una estrategia de aprendizaje supervisado. El aprendizaje supervisado se basa en el estudio de las características de un objeto, documento o ejemplo, con la intención de describir mediante una función matemática la relación existente entre las características del objeto y la característica clase. Como ya se ha indicado en varias ocasiones, para emprender dicho estudio se requiere de la disposición de un conjunto de datos, en nuestro caso, una colección de documentos, cuyo conjunto de características cuente con la clase a la que pertenece el documento. En ocasiones, no es sencillo disponer de una colección de documentos con su respectiva clase, con la que poder construir un modelo matemático que posibilite la clasificación de nuevos documentos. Esta situación es más frecuente de lo que se piensa, por no decir que es un escenario más que real. Para entender esta afirmación es recomendable que el lector se platee la siguiente cuestión: ¿es factible la generación de un modelo matemático monolítico para la clasificación de la polaridad de textos que se publican en Internet teniendo presente la evolución constante del lenguaje, la variabilidad y diversidad de temáticas, cada una con un registro de lenguaje muy específico? Pues la verdad, es que la viabilidad de ese modelo matemático monolítico es más que dificultosa.

En las situaciones en las que las colecciones de datos no cuentan con una característica que indique la clase a la que pertenece el documento, no es posible la construcción del modelo matemático que permita la clasificación de nuevos ejemplos. En esos casos se tiene que estudiar las peculiaridades propias de cada documento con la intención de descubrir la posible clase a la que pertenece. Ésto último se hace cuando se conocen a priori las clases o categorías de los documentos, pero en muchas ocasiones ni es posible saber de antemano los grupos que se pueden distinguir en un corpus. Éstas son las situaciones en las que se debe aplicar aprendizaje no supervisado. Las matemáticas y la estadística principalmente proporcionan dos métodos para emprender la construcción de sistemas no supervisados, que son el análisis de componentes principales y el *clustering*[1]. En el ámbito del PLN, a las dos técnicas mencionadas se añade la definición de reglas en función de las peculiaridades lingüísticas de los documentos.

La aplicación de técnicas no supervisadas requiere en muchas ocasiones del aprovechamiento del conocimiento que aportan recursos lingüísticos

[1]La descripción del análisis de componentes principales y del *clustering* excede el propósito de la presente memoria, por lo que al lector interesado se le recomienda la lectura de (James et al., 2013).

externos. En el ámbito del AO se pueden distinguir dos tipos de recursos con información de opinión, los cuales se presentan a continuación:

- Listas de palabras de opinión: En inglés existe un amplio abanico de listas de opinión, que sólo pueden ofrecer un listado diferenciado de palabras positivas y negativas, como es el caso de la lista de opinión compilada por Bing Liu (Hu & Liu, 2004), o proporcionar incluso la categoría morfológica de los términos, con la intención de identificar el sentido con el que se está utilizando el término en cuestión, como es el caso del lexicón MPQA (Wilson et al., 2005b). En español la variedad no es tan amplia, pero se debe destacar el léxico iSOL (Molina-González et al., 2013) que se describirá en el Capítulo 6.

- Bases de conocimiento léxicas: Las bases de cocimiento léxicas proporcionan mayor información que las listas de palabras, dado que no se limitan a indicar si una palabra es positiva o negativa, sino que proporcionan un valor de polaridad, que normalmente suele estar asociado a la probabilidad de que el término se esté empleando con un sentido positivo, negativo u objetivo. Su mayor capacidad informativa se debe principalmente a que se han construido sobre la base de conocimiento de conceptos relacionados semánticamente WordNet (Miller, 1995). WordNet se puede considerar como un grafo en el que los nodos son conceptos o sentidos, y las aristas que los unen son relaciones semánticas. Hay que prestar atención a la afirmación de que los nodos de WordNet son sentidos o conceptos, es decir, no son términos propiamente dichos, sino los usos de los términos en función de su significado. Verbigracia, el vocablo *bank* en inglés[2] tiene dos usos o sentidos preponderantes, que son el de extensión de terreno en el fondo de un río o lago y el de institución financiera. Por lo tanto, *bank* no va a tener una entrada en WordNet, sino que al menos va a tener dos, que se corresponden con los dos sentidos expuestos.

  La naturaleza de WordNet ha permitido generar bases de datos de conceptos similares, en las que a cada nodo se le ha asignado un valor o conjunto de valores de polaridad. De este tipo de recursos para AO en inglés se pueden encontrar SentiWordNet (Baccianella et al., 2010), WordNet-Affect (Strapparava & Valitutti, 2004) o Q-WordNet (Agerri & García-Serrano, 2010). Para el español no existen actualmente ningún recurso similar a los tres citados.

[2]Se emplea como ejemplo la traducción al inglés del vocablo banco, debido a que WordNet es un recurso lingüístico para inglés.

Los recursos lingüísticos de opinión, por sí mismos, proporcionan una información valiosa, pero, para que esa información sea verdaderamente útil, y mejore la clasificación de la polaridad, deben ser correctamente empleados, y adecuadamente combinados en el caso de que se quiera aprovechar la información de varios de ellos. Con la intención de no requerir de un conjunto de datos en el que cada ejemplo cuente con la clase a la que pertenece, y con el empeño de aprovechar idóneamente la información que proporcionan diversos recursos lingüísticos de opinión, se ha definido un método de clasificación de la polaridad basado en la expansión del significado de los términos presentes en un texto. Ese método, al igual que se ha realizado con las técnicas de aprendizaje supervisado descritas en el Capítulo 4, se ha evaluado sobre textos largos y cortos. A continuación se va a desgranar el sistema de clasificación que se ha diseñado. Posteriormente, se van a indicar los distintos recursos que se han integrado en el algoritmo de clasificación y, por último, se van a detallar las evaluaciones que se le han aplicado.

5.2. Clasificación de la polaridad por expansión del significado

Se ha desarrollado un método de clasificación de la polaridad que se caracteriza fundamentalmente por su naturaleza modular, y por permitir la integración de herramientas externas de PLN y recursos lingüísticos. La libertad en el uso, tanto de herramientas de PLN, como de recursos lingüísticos, permite en cada momento emplear los algoritmos más adecuados en función del tipo de documento que se quiera clasificar, así como del idioma en el que se encuentre escrito, proporcionando de esta manera un carácter multilingüe al método definido. Para posibilitar la compresión de la naturaleza modular del método se van a describir sus cinco módulos principales:

1. **Limpieza de los datos**: Dependiendo de los textos con los que se esté trabajando, puede que sea necesario la aplicación de un proceso de limpieza de los mismos, antes de proceder a la identificación de la polaridad. Por ejemplo, en la Sección 4.4 se indica que los textos a clasificar, en ese caso *tweets*, tienen que ser preparados antes de ser clasificados, porque están constituidos por palabras mal formadas. El módulo de limpieza es opcional, y se puede adaptar en función de los textos que se estén tratando.

2. **Análisis morfológico**: Una vez que se han preparado los textos

para ser procesados, se les aplica un análisis morfológico, el cual está constituido por: un procedimiento de *tokenización*, la identificación de las funciones morfológicas de los términos y la determinación del lema de cada una de las palabras.

3. **Desambiguación**: En varias ocasiones se ha indicado que el AO es una tarea del PLN altamente dependiente del significado con el que se están empleando los términos. En los métodos supervisados, esa necesidad puede cubrirse con el uso intensivo de datos, pero en un método no supervisado, en el que se quiera afinar el sentido específico con el cual se están empleando las palabras, es necesario aplicar un proceso de desambiguación. Éste es el módulo fundamental del método, y sobre el que se sustenta el módulo de expansión del significado. Debido a su importancia, se dedica una sección completa (Sección 5.2.1) a la explicación del algoritmo de desambiguación que se ha evaluado en los diversos experimentos que se han realizado.

4. **Expansión**: Este módulo está íntimamente relacionado con el de desambiguación, y su fin es incrementar la información semántica presente en un texto por medio de la inclusión de conceptos relacionados. La Sección 5.2.1 expondrá con mayor detalle el funcionamiento de este módulo.

5. **Cálculo de la polaridad**: Una vez que se tienen identificados todos los conceptos es momento de obtener su polaridad. Mediante el aprovechamiento de bases de conocimiento de opinión como SentiWordNet y Q-WordNet se obtiene el valor final de polaridad del documento.

De una manera gráfica la arquitectura del método se puede visualizar en la Figura 5.1.

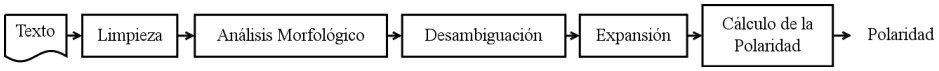


Figura 5.1: Arquitectura del modelo no supervisado de clasificación de la polaridad propuesto.

A continuación se van a describir los módulos fundamentales del método propuesto, como son el de desambiguación y el de expansión.

5.2.1. Desambiguación

La tarea de desambiguación se define como el proceso de determinar el sentido concreto que se está empleando de una palabra polisémica (Manning & Schütze, 1999c). Desde una perspectiva más pragmática, la desambiguación se puede considerar como un problema de clasificación, en el cual el objetivo es asignar a cada palabra, en función de su contexto, una etiqueta semántica de las posibles del diccionario semántico que se haya elegido como referencia. Por lo tanto, el resultado de la desambiguación va a depender enormemente del inventario de conceptos elegido. En la bibliografía relacionada con la desambiguación se pueden distinguir diversas fuentes de conceptos o sentidos, siendo las más relevantes las siguientes:

1. **Diccionario**: La mayoría de los primeros algoritmos de desambiguación (Lesk, 1986; Walker & Amsler, 1986) estaban fundamentados en el uso de diccionarios en los que se identificaban los usos o sentidos de los términos.

2. **Bases de datos conceptuales**: La fuente de conceptos con mayor predicamento en el seno de la comunidad investigadora es sin ninguna duda WordNet (Miller, 1990, 1995). WordNet, al igual que la materia, está conformado por mínimas unidades, que en lugar de átomos, reciben la denominación de conceptos o sentidos. Los átomos por si solos no constituyen un objeto material perceptible, sino que los átomos tienen que enlazarse para constituir un elemento material tangible, y dependiendo del enlace, el elemento resultante tiene una naturaleza u otra. Los átomos o conceptos de WordNet tienen un comportamiento similar, dado que la valía de WordNet es el conocimiento que aportan los enlaces semánticos que se establecen entre las unidades conceptuales que se conocen como *synsets*. Los enlaces semánticos de WordNet posibilitan representar la base de datos léxica como un grafo, lo cual aprovechan los algoritmos de recorrido de grafos para colegir la máxima cantidad de información posible.

3. **Etiquetas de dominio**: La desambiguación no se limita únicamente a la determinación del sentido con el que se emplea un término, sino también con la identificación de la temática sobre la que versa un fragmento de texto. Por ende, no es raro encontrarse con inventarios de etiquetas temáticas en lugar de conceptuales, como el que se puede construir a partir del diccionario *Longman Dictionary of Contepmporary English* (Procter, 1978).

4. **Corpus multilingüe**: También se suelen emplear como etiquetas conceptuales las diversas traducciones que un mismo término en un idioma puede tener en otra lengua diferente. Ejemplos de sistemas que aprovechan el poder discriminatorio de corpus paralelos multilingües a la hora de seleccionar el sentido que más adecuadamente se ajusta al uso de una palabra se pueden encontrar en (Gale et al., 1992a; Ng et al., 2003).

5. **Inventarios específicos**: Colecciones de etiquetas semánticas que se confeccionan manualmente para resolver problemas muy concretos.

6. **Pseudopalabras**: Para evaluar un proceso de desambiguación, o para construir un sistema de aprendizaje supervisado para desambiguar un fragmento de texto, se requiere de texto de referencia en el que cada término tenga asociado la etiqueta semántica que le corresponde. La generación de *corpora* representativo con etiquetas semánticas es una tarea bastante ardua, de manera que no abunda este tipo de recurso lingüístico. Con el fin de aliviar la carga de la generación de colecciones de datos de referencia para desambiguación, se idearon las pseudopalabras. Una pseudopalabra es un término formado por la combinación de dos palabras. Por ejemplo, de la palabra puerta y plátano se puede generar la pseudopalabra *puerta-platano*, y en la colección de documentos se sustituyen todas las apariciones de los vocablos plátano y puerta por el nuevo término originado. De esta manera simple se ha generado un conjunto de datos ambiguo, considerándose el corpus original como no ambiguo, y por tanto el modelo de referencia. Representantes de esta filosofía son (Gale et al., 1992b; Schütze, 1992a).

7. **Conjunto de sentidos automáticamente inducidos**: Estos conjuntos de conceptos son los que se obtienen en los procesos de desambiguación no supervisada basados en *clustering*. Algunos trabajos que intentan inferir etiquetas semánticas de fragmentos de texto son (Schütze, 1992b, 1998; Pantel & Lin, 2002).

Aunque la presente memoria no se centra en la desambiguación de texto[3], va al menos a mencionar que, como en la mayor parte de las tareas de aprendizaje automático, se puede seguir una estrategia supervisada o no supervisada. Los sistemas de desambiguación basados en aprendizaje supervisado no son más que clasificadores que emplean unas características

[3]Un estudio más amplio sobre desambiguación se puede encontrar en (Manning & Schütze, 1999c; Agirre & Edmonds, 2006; Yarowsky, 2010)

determinadas para emprender la asignación de una etiqueta semántica a cada término. Esta afirmación se valida fácilmente repasando la literatura relacionada con desambiguación, dado que lo que se encuentra es la adaptación de algoritmos de clasificación a la tarea de desambiguación. Mientras que en los que se sigue un enfoque no supervisado, lo que se persigue es la adaptación de métodos de *clustering*.

Anteriormente se ha indicado que una fuente de conocimiento para los procesos de desambiguación son las bases de conocimiento léxicas, y que su estructura de enlaces semánticos permiten representarlas como un grafo. Los grafos que se generan a partir de bases de conocimiento han dado lugar al estudio de la aplicación de algoritmos de recorrido de grafos para escrudiñar el conocimiento que atesoran, y que con un procesamiento superficial no se puede obtener. Dependiendo de los autores, los algoritmos de desambiguación fundamentados en el recorrido de los grafos de bases de conocimiento léxicas, como es WordNet, los consideran no supervisados, como es el caso de lo autores del algoritmo UKB (Agirre & Soroa, 2009), que se detallará más adelante, sin embargo otros más puristas no le otorgan la cualidad de no supervisados sino de mínimamente supervisados, dado que, según su entender, el uso de un recurso elaborado manual o semiautomáticamente implica el aprovechamiento de conocimiento confeccionado previamente, imposibilitando por tanto la catalogación como métodos sin supervisión (Yarowsky, 2010). Nuestra opinión no es tan purista, nos alineamos con los autores de UKB, y consideramos que el uso de recursos lingüísticos en desambiguación no implica una supervisión, ya que dichos recursos no se pueden aplicar directamente para la elaboración de un modelo estadístico basado en la asociación de términos y sus respectivas etiquetas semánticas, de igual modo que un corpus etiquetado, y, además, porque dichos recursos no pueden aplicarse directamente, sino que requieren de un tratamiento dependiente del problema en los que se estén aplicando para que su concurso sea valioso.

El algoritmo de desambiguación que se ha empleado en los experimentos de clasificación de la polaridad no supervisados es UKB[4] (Agirre & Soroa, 2009). UKB es un algoritmo de desambiguación que se cataloga dentro de los algoritmos basados en recorrido de grafos. Entre los algoritmos de recorrido de grafos se encuentran los conocidos como de camino aleatorio, o en inglés *random walk*. El camino aleatorio es una formulación matemática de una trayectoria que consiste en tomar sucesivos pasos aleatorios. El lector comprenderá mejor a qué se refiere el camino aleatorio cuando se le mencione un ejemplo de algoritmo conocido de este tipo, y que no es otro

[4] `http://ixa2.si.ehu.es/ukb/`

que PageRank (Page et al., 1999), el algoritmo donde descansa el potencial de Google. UKB aprovecha la capacidad de PageRank para determinar la relevancia de un nodo dentro de un grafo para identificar los nodos de WordNet más relevantes en función de un conjunto de términos de entrada, o dicho de otro modo, UKB, empleando PageRank, trata de descubrir los sentidos de los términos que forman un texto de entrada en función del contexto en el que se encuentran. Para entender UKB se requiere la compresión de PageRank, por lo que en las siguientes líneas se va a intentar repasar el afamado algoritmo.

PageRank

PageRank es un algoritmo de recorrido de grafos, cuya aplicación más conocida es la de clasificación por relevancia de páginas web independiente de la consulta de búsqueda que se haya realizado. PageRank está basado en una medida de reputación o prestigio en redes sociales, de manera que el valor de PageRank de cada web se puede interpretar como su valor de prestigio. PageRank se sustenta sobre la idea de que la relevancia o importancia de una web i es la suma de la relevancia de todas las páginas que apuntan a dicha web. De la afirmación anterior se pueden deducir varios hechos: el PageRank de una página se divide entre todas las páginas a las que apunta; cuantos más enlaces reciba una web más relevante será; y cuanto más importantes sean las páginas que apuntan a una web más prestigiosa será la página apuntada.

La ideas expuestas anteriormente, propias de la teoría de análisis de redes sociales (Wasserman & Faust, 1994), requieren de una fuerte formulación matemática para poder implementarlas algorítmicamente, y de esa manera aplicarlas en un entorno real. PageRank considera la web como un grafo dirigido $G = (V, E)$ donde V es un conjunto de vértices, nodos o páginas web, y E es el conjunto de aristas del grafo, que se corresponden con los enlaces entre páginas web. El PageRank de una web o su valor de relevancia vendrá determinado por la ecuación:

$$P(i) = \sum_{(j,i)\in E} \frac{P(j)}{O_j} \tag{5.1}$$

donde O_j es el grado de salida o número de páginas a las que apunta la web j.

Matemáticamente, con lo que se cuenta es con un sistema de n ecuaciones lineales, siendo n el número total de páginas web ($n = |V|$). Con una expresión matricial se entenderá mejor el sistema de ecuaciones en

las que se fundamenta el cálculo de PageRank. Sea P un vector columna $n-dimensional$ de valores de PageRank,

$$P = (P(1), P(2), ..., P(n))^T$$

. Sea A la matriz de adyacencia del grafo con los valores:

$$A_{ij} = \begin{cases} \frac{1}{O_i} & \text{si } (i,j) \in E \\ 0 & \text{En otro caso} \end{cases} \tag{5.2}$$

Teniendo en cuenta las dos ecuaciones anteriores, la Ecuación 5.1 se podría expresar matricialmente de la siguiente manera:

$$P = A^T P \tag{5.3}$$

Al lector que cuente con ciertos conocimientos de álgebra lineal, la Ecuación 5.3 le habrá recordado a la ecuación característica de un sistema de vectores propios o autovectores, donde P es un autovector con autovalor igual a 1. Si se observa con cierta atención la Ecuación 5.3 es posible percatarse de que se trata de una ecuación recursiva, que debe resolverse mediante un algoritmo iterativo. Ese algoritmo iterativo requiere de que el grafo cumpla con el requisito de que tiene que ser estocástico, irreducible y no periódico, de que 1 tiene que ser el máximo valor del autovalor, y el vector de PageRank P tiene que ser el autovector. No solamente al álgebra lineal hay que recurrir para resolver la ecuación 5.3, sino también al análisis numérico, porque en el método de las potencias es donde se encuentra la solución para obtener la resolución numérica al problema.

Pero antes de llegar a la solución numérica, se deben cumplir las condiciones que impone el álgebra lineal, es decir, que el grafo debe ser estocástico, irreducible y no periódico. Para la condición estocástica se va a acudir a la base teórica de las cadenas de Markov (Grimmett & Stirzaker, 1989). En una cadena de Markov, cada página web, o un nodo del grafo de WordNet, se considera un estado, mientras que los enlaces y relaciones se interpretan como transiciones de un estado a otro. Una persona que navega a través de Internet o recorre un grafo va saltando de nodo en nodo de manera totalmente aleatoria. Anteriormente se ha dicho que cada nodo tiene un determinado número de enlaces, relaciones o transiciones que parten de él hacia otros nodos, O_i. Por consiguiente, la probabilidad de saltar del estado i a cualquiera de los estados enlazados va a ser igual a $1/O_i$. Se asume que la persona que navega por Internet, o que se encuentra recorriendo el grafo, siempre sigue un enlace y nunca retrocede. De este modelo estocástico se deriva que la matriz A de adyacencia es:

$$\begin{pmatrix} A_{11} & A_{12} & . & . & . & A_{1n} \\ A_{21} & A_{22} & . & . & . & A_{2n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ A_{n1} & A_{n2} & . & . & . & A_{nn} \end{pmatrix}$$

A_{ij} representa la probabilidad de transición del estado i al estado j. La definición de dicha probabilidad se puede consultar en la Ecuación 5.2.

Teniendo en cuenta las definiciones anteriores, si se tiene una distribución inicial de probabilidad (p_0) en la que se indique la probabilidad de que el recorrido comience en unos determinados estados p_0 = *(p_0(1), p_0(2), ..., p_0(n))*^{T} y una matriz de adyacencia como la definida anteriormente (A), entonces se tendría que:

$$\sum_{i=1}^{n} p_0(i) = 1 \tag{5.4}$$

$$\sum_{j=1}^{n} A_{ij} = 1 \tag{5.5}$$

En el caso ideal que se está planteando la ecuación 5.5 se cumple, pero en el caso de la Web no tiene lugar, debido a que no todas las páginas tienen enlaces a otras páginas, incluso no es extraño encontrarse con webs que no tienen ningún tipo de enlace. En el caso de WordNet no se da esta situación, dado que todas los conceptos que se recogen tienen algún tipo de relación semańtica que sale de ellos. Se podría pensar que los hipónimos son nodos hoja en el gran grafo de WordNet, pero no hay que olvidar que todo hipónimo está relacionado con su hiperónimo. Por consiguiente, WordNet cumple con la condición estocástica que impone el álgebra lineal para solucionar el sistema de ecuaciones de PageRank. Pero ¿cómo se solventa este problema en el contexto de Internet? De una manera muy sencilla, las páginas web que no tienen enlaces a otras webs se les asigna un enlace hacia cada una de las páginas web existentes, o dicho de otro modo, a los nodos que tienen grado de salida igual a cero se les crea una arista por cada nodo existente en el grafo. De esta manera la fila de la matriz de adyacencia correspondiente a ese tipo de nodos pasará de tener solo valores cero, a contar en cada posición con el valor $1/n$, siendo n igual al número de nodos del grafo. Con esta transformación, ya sí se consigue que todos las filas sumen 1, y por ende la matriz sea estocástica.

Por el teorema de la ergodicidad de las cadenas de Markov, una cadena

de Markov definida por una matriz de adyacencia estocástica, como es el caso de la matriz A, tiene una distribución de probabilidad estacionaria si A es irreducible y no periódica, o dicho de otro modo, si el grafo asociado a la matriz A es fuertemente conectado y no existen ciclos. ¿Qué significa que la matriz tenga una distribución de probabilidad estacionaria? Quiere decir, que la ecuación recursiva 5.3 tras una serie de iteraciones convergerá en un valor constante de probabilidad. Matemáticamente esta afirmación se define como:

$$\lim_{k\to\infty} p_k = \pi \tag{5.6}$$

La Ecuación 5.6 significa que $p_k{=}p_{k+1}{=}\pi$, y como consecuencia $\pi{=}A^T\pi$. π es el vector principal de A^T con autovalor igual a 1. En la formulación de PageRank π es tomado como el vector P, por lo que la Ecuación 5.6 nos retrotrae a la Ecuación 5.3, es decir a $P{=}A^TP$.

Hasta ahora se tiene que el grafo de adyacencia es estocástico, y por el teorema de la ergodicidad de las cadenas de Markov se tiene que la ecuación 5.3 converge en un autovector, siempre y cuando el grafo sea irreducible y no periódico. Un grafo irreducible es aquel que se encuentra fuertemente conectado, definiéndose este estado como: dado el grafo $G{=}(V, E)$ se dice que es un grafo fuertemente conectado si y solo si entre cada par de nodos $u, v \in V$ existe un camino que une u y v.

El grafo dirigido en el que se proyecta la Web no es un grafo fuertemente conectado, ya que no todos los pares de páginas están unidos por un conjunto de aristas que conforman un camino, o dicho de otro modo, no desde todas las páginas se puede llegar al resto. En WordNet sucede algo similar, no desde todos los conceptos se puede alcanzar el resto de conceptos.

La tercera condición para la resolución del sistema de ecuaciones es que el grafo sea no periódico. Un grafo no periódico se define como aquel que carece de nodos periódicos. Un nodo o estado i es periódico, con periodo $k > 1$, siempre que el mínimo valor de todos los caminos que conectan el estado i consigo mismo tengan una longitud que sea múltiplo de k. Si un estado no es periódico, la condición a la que se quiere llegar, entonces k es igual a 1. Dado que ya se sabe que tanto el grafo de Internet como el de WordNet son reducibles, entonces también serán periódicos, porque como existen parejas de nodos que no se encuentran conectados, entonces es probable la existencia de ciclos de longitud mayor a 1 entre un nodo y sí mismo.

Los dos problemas anteriores se pueden resolver con una sola operación, y que consiste en añadir a cada nodo una arista que lo una con el resto de nodos del grafo con una probabilidad de transición mínima controlada por

un parámetro *d*. La matriz de adyacencia resultante es ya una matriz que representa un grafo estocástico, irreducible, y no periódico. Por lo tanto, si se recorre el grafo resultante caben dos opciones:

1. Con probabilidad *d*, se puede saltar de un nodo a otro unido por una arista no añadida artificialmente.

2. Con probabilidad *1-d*, se puede ir de un nodo a cualquier o otro a través de una arista añadida artificialmente.

Con las transformaciones emprendidas en la matriz de adyacencia la ecuación 5.3 quedaría de la siguiente manera:

$$P = ((1-d)\frac{E}{n} + dA^T)P \tag{5.7}$$

donde E es ee^T, y e es a su vez un vector columna en el que todos sus elementos son todos unos. Ésto hace que E sea una matriz cuadrada *nxn* de unos. Este hecho va a posibilitar realizar de nuevo otra transformación para simplificar la ecuación. Si se fija la atención en la Ecuación 5.7 y teniendo en cuenta que E es ee^T, es posible percatarse de que se tiene el producto e^TP. Si se extiende dicho producto, se tiene que e^TP = *P(1) + P(2) + ... + P(n)*, el cual, si se tiene en mente que P sigue una distribución de probabilidad estacionaria que converge en π con autovalor igual 1, y se multiplica ambos miembros de la ecuación por n, entonces la Ecuación 5.7 se queda de la siguiente forma:

$$P = (1-d)e + dA^TP \tag{5.8}$$

y para el valor de PageRank para cada nodo se definiría de la siguiente manera:

$$P(i) = (1-d) + d\sum_{j=1}^{n} A_{ji}P(j), \tag{5.9}$$

lo que es equivalente a la fórmula inicial de PageRank 5.1 para cada nodo del grafo:

$$P(i) = (1-d) + d\sum_{(j,i)\in E} \frac{P(j)}{O_j} \tag{5.10}$$

Por último decir, que el parámetro d es el conocido como factor de amortiguamiento, en inglés *damping factor*, el cual toma valores entre 0 y 1, y normalmente es igual a 0,85.

UKB

Una vez comprendido cómo actúa PageRank para determinar la relevancia de un nodo de un grafo, es momento de retornar a la descripción de UKB. Ya se tiene una base de conocimiento representable con un grafo, WordNet, y asimismo se cuenta con un algoritmo que tiene la capacidad de generar conocimiento a partir del recorrido sobre un grafo. Una primera posibilidad de desambiguación sería la de aplicar PageRank al grafo, es decir, a WordNet, y obtener una versión del mismo en la que todos los nodos tuvieran asociado su valor de PageRank. La desambiguación a partir de una versión de WordNet en la que todos sus nodos tienen ya asociado su valor de PageRank es muy sencilla, ya que para una palabra sólo se tendrían que consultar los conceptos o *synsets* asociados a la misma y tomar aquel que tuviera un valor más alto de PageRank. Aunque sencillo, la operación que se acaba de explicar es una desambiguación independiente del contexto, la cual proporcionaría siempre el mismo resultado indistintamente del texto en el que se sitúen los vocablos cuyo sentido concreto se pretende descubrir.

En desambiguación el contexto de un término es capital para la correcta obtención de su sentido, por lo tanto, como bien se describe en (Agirre & Soroa, 2008), una posible acción para combinar la información que proporciona el contexto, con la posibilidad de usar una versión de WordNet en la que todos sus nodos se encuentren ponderados por su valor de PageRank, consiste en la generación de un subgrafo conformado por aquellos sentidos relacionados con los términos del texto en el que se encuentran las palabras a desambiguar. Pero esta solución limita enormemente el conocimiento que se puede generar a partir de WordNet, cuando se cuenta con un algoritmo como PageRank con la capacidad de poder trabajar con el grafo completo.

Si no se quiere cercenar la capacidad de WordNet para proporcionar información al proceso de desambiguación, se debe idear un método de recorrido del grafo subyacente en WordNet para introducir el contexto de cada uno de los términos cuyo sentido se quiere obtener. Cabe preguntarse, ¿con las dos herramientas que hasta ahora se están considerando, WordNet y PageRank, se puede emplear el contexto de los términos? Para responder a la pregunta, el lector debe retrotraerse a la Ecuación 5.8 y fijarse en el componente e de la Ecuación. Si se recuerda e es un vector columna en el que todos sus elementos tienen como valor 1, de manera que al comienzo de la ejecución del algoritmo se asigna a todos los nodos la misma probabilidad de que en un recorrido aleatorio pueda interrumpirse el itinerario y saltar a un nodo cualquiera. Haveliwala (2002) propone que el vector e no tenga una distribución uniforme, de manera que se pueda asignar una mayor

probabilidad a unos determinados nodos y así sesgar hacia esos nodos el vector final de valores de PageRank. Por ejemplo, si en ese vector e solo a una de sus componentes se le asigna el valor 1 y al resto el valor cero, entonces se concentra toda la probabilidad de llegar aleatoriamente a dicho nodo, es decir, todo salto aleatorio fuera del recorrido siguiendo las aristas del grafo tendrá como destino el nodo en cuestión. Debido a la naturaleza de PageRank, la preponderancia que se le ha asignado al nodo se difunde por lo nodos que se encuentren en su vecindad, ya que su valor de relevancia se reparte entre todos los nodos a los que apuntan sus aristas. La consecuencia del sesgo introducido es que, el vector P final puede ser interpretado como el valor de relevancia de cada uno de los nodos del grafo en función del nodo de mayor probabilidad inicial. Pues bien, ¿se podría sesgar el recorrido de WordNet al contexto de una palabra para así descubrir el sentido con la que se está empleando? Pues sí, ese es el fundamento de UKB, explicado originalmente en (Agirre & Soroa, 2009) y más extensamente en (Agirre et al., 2014).

La versión de PageRank sesgada a unos determinados nodos es denominada por los autores de UKB como PageRank Personalizado, en inglés *Personalized PageRank*. Realmente, la orientación hacia los términos del contexto no los realiza UKB modificando los valores iniciales del vector e, sino insertando en el grafo de WordNet las palabras del contexto del término a desambiguar en forma de nodos apuntando a cada uno de lo conceptos asociados a ellos. De esta manera, lo que se consigue es asignar una mayor relevancia a los conceptos de los términos presentes en el contexto, y tras varias iteraciones del algoritmo sobresaldrá el sentido más adecuado en función del contexto de cada uno de los términos. En este caso, el vector P de valores de PageRank se puede ver como una medida de la relevancia estructural de los conceptos de la base de conocimiento (WordNet) en función de un contexto que se toma como entrada. En la Figura 5.2 se puede visualizar la acción de insertar las palabras del contexto[5] a desambiguar en el grafo de WordNet.

Una vez modificado el grafo de WordNet para otorgar una mayor relevancia a todos los conceptos asociados a los términos presentes en el contexto, el proceso pasa a estar en manos de PageRank. Tras varias iteraciones, el algoritmo asocia a cada término una lista de conceptos ordenados en orden decreciente en función del valor de relevancia asignado por el algoritmo. Por consiguiente, todos los conceptos que se sitúen en la

[5] Al ser UKB un algoritmo de desambiguación que tiene en cuenta el contexto, a la hora de desambiguar un término también se incorpora al proceso el contexto que le acompaña en el texto.

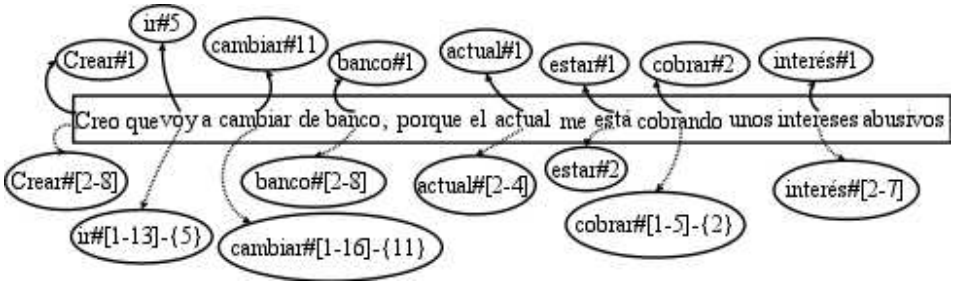


Figura 5.2: Porción del grafo de WorNet en español para ilustrar la incorporación del contexto en el proceso de desambiguación. Las aristas dibujadas con línea continua señalan a los conceptos desambiguados por UKB, mientras que las representadas por línea discontinua relacionan los términos con el resto de conceptos recogido en la versión española de WordNet empleada.

primera posición de la lista se corresponderán con los sentidos específicos de cada una de las palabras que se están empleando en el contexto considerado. En la Tabla 5.1 se pueden consultar los *synsets* concretos asociados con los términos de la Figura 5.2.

Término	***Synset***	**Lema desambiguado**	**PageRank**
creo	01617192-v	crear#1	0,00305471
voy	02685951-v	ir#5	0,00208047
cambiar	02256354-v	cambiar#11	0,00516669
banco	08420278-n	banco#1	0,00262062
actual	01731351-a	actual#1	0,00262062
está	02655135-v	estar#1	0,0105103
cobrando	02256354-v	cobrar#2	0,00516669
intereses	05682950-n	interés#1	0,00293488

Tabla 5.1: Resultado del proceso de desambiguación correspondiente a la Figura 5.2.

Si se observa la Tabla 5.1 se comprueba fácilmente el buen hacer de UKB, ya que únicamente yerra en el primer concepto. Si se analiza el procesamiento previo a la desambiguación se puede colegir que el error no es debido principalmente a UKB. El algoritmo de desambiguación requiere que las palabras se encuentren lematizadas, y que se haya identificado la categoría morfológica con la que actúan en la oración. Pues bien, en el caso

de la primera palabra de la oración del ejemplo, “Creo”, el lematizador[6] le ha otorgado como lema “crear” en lugar de “creer”, de manera que erroneamente el *synset* asociado a “crear” ha estado sobrevalorado desde la primera iteración de UKB sobre el grafo de WordNet. Por ende, no se puede acusar a UKB de haber realizado mal su trabajo, sino que el lematizador le ha inducido a cometer una equivocación.

5.2.2. Expansión

La novedad de la experimentación, cuyos resultados se van a detallar en las secciones subsiguientes, no se encuentra en la aplicación de un algoritmo de desambiguación basado en grafos, sino que se encuentra en aprovechar UKB para, una vez obtenidos los sentidos de cada uno de los términos, aplicar de nuevo el algoritmo sobre el grafo de WordNet, pero en lugar de asignar una mayor relevancia inicial a todos los conceptos de cada uno de los términos, en esta ocasión solamente se asigna una mayor relevancia a los conceptos identificados previamente. El resultado de este proceso es una versión de WordNet en el que los conceptos están ordenados según un valor de importancia dependiente del significado del contexto que se está teniendo en cuenta. Como consecuencia, los conceptos más preponderantes serán aquellos que estén asociados con los reconocidos en el proceso de desambiguación. Si como bien indica la teoría esos conceptos están íntimamente relacionados, entonces también deben tener una polaridad similar a los sentidos de la oración, o mejor dicho, del contexto que se está teniendo en cuenta. Por tanto, esos sentidos pueden emplearse para añadir carga de polaridad al contexto y hacer más evidente la orientación de la opinión o emoción que se pueda estar expresando. Ante esta afirmación cabe preguntarse, ¿cuál es el número de conceptos óptimo a usar para extender adecuadamente el significado del contexto en estudio? La respuesta a esta pregunta es cardinal para la elaboración del clasificador de la polaridad, de manera que se describirán en detalle posteriormente.

Como conclusión debe resaltase, que los sistemas de clasificación de la polaridad que se han construido de manera no supervisada se apoyan en la desambiguación de los términos que aparecen en un documento, y en la incorporación de conceptos relacionados a los identificados, con el fin de incrementar la carga de polaridad de los documentos y, de esta manera, facilitar al clasificador la asignación de la clase de polaridad correcta.

[6]Se ha empleado Freeling (Padró & Stanilovsky, 2012) para realizar el procesamiento lingüístico requerido para el ejemplo.

5.3. Recursos linguísticos

5.3.1. *Multilingual Central Repository*

Hasta el momento se ha estado describiendo el algoritmo de desambiguación UKB, que como se ha indicado es un algoritmo que está basado en el recorrido aleatorio del grafo subyacente de una base de conocimiento. La base de conocimiento mencionada en la exposición anterior es WordNet, que como el lector sabrá, es una base de datos léxica en inglés a nivel de concepto, los cuales se encuentran relacionados por medio de relaciones semánticas. Por consiguiente, si WordNet es una base de datos de conceptos en inglés, ¿cómo ha sido empleada para una experimentación sobre opiniones escritas en español? Realmente no se ha utilizado WordNet, sino una versión en español del mismo que se encuentra en el repositorio *Multilingual Central Repository*[7] (MCR) (Atserias et al., 2004; Agirre et al., 2012).

La elección de MCR no estuvo motivada porque fuera la única versión española de WordNet, ya que hasta la fecha la comunidad investigadora cuenta con dos versiones diferentes de WordNet en español. La primera de ellas, y la más conocida, principalmente por ser la más veterana, es EuroWordNet[8] (Vossen et al., 1998). Denominar a EuroWordNet versión española de WordNet, puede inducir al pensamiento de que dicha versión española está conformada por la traducción de todos los conceptos que engloba WordNet, dicho de otra manera, que EuroWordNet está constituido por 117 659 *synsets* al igual que WordNet. Pero EuroWordNet no cubre ni la cuarta parte de la variedad semántica de WordNet, ya que simplemente está constituido por 23 370 *synsets*, un 19,86 % del total. Mientras tanto, la versión de MCR cuenta, según la última publicación de sus autores (Agirre et al., 2012) con 59 227 *synsets*, lo que se traduce en un 50,33 % del total. La justificación de la elección de MCR es evidente, es significativamente mucho más completo que EuroWordNet, por lo que es más probable que los experimentos que se emprendan con MCR sean más precisos y tengan una mayor cobertura que los que se realicen con EuroWordNet.

5.4. Experimentación sobre textos largos

El método no supervisado de clasificación de la polaridad propuesto (ver Sección 5.2) se ha implementado de diferente manera en función de las necesidades, y se ha evaluado su efectividad sobre diferentes tipos de

[7] `http://adimen.si.ehu.es/web/MCR/`
[8] `http://www.illc.uva.nl/EuroWordNet/`

documentos. La presente sección se circunscribe a la evaluación que se ha llevado a cabo sobre textos largos.

5.4.1. Textos escritos en inglés

El método propuesto depende enormemente de la existencia de recursos para llevar a cabo las tareas de desambiguación y expansión. A la altura en la que nos encontramos de la memoria, no es raro volver a mencionar la desventaja en la que se encuentra el español, en lo que a la disponibilidad de recursos lingüísticos se refiere si se compara con el inglés. Por consiguiente, comenzar con una evaluación de la efectividad del método sobre textos en inglés es mucho más sencillo que iniciarla por el español.

La primera evaluación del método sobre textos largos se encuentra descrita en (Montejo Ráez et al., 2012). En dicho trabajo se adapta la arquitectura para la clasificación de la opinión en citas periodísticas. El corpus elegido para la ocasión es el *English Sentiment Quotes* corpus[9] (ESQ)[10], el cual se encuentra descrito por primera vez en (Balahur et al., 2010). El corpus está compuesto por 1590 citas en inglés extraídas automáticamente de diversas fuentes periodísticas publicadas en Internet. Cada una de las citas, o de los fragmentos del texto del corpus, son una opinión sobre una determinada entidad, estando etiquetada manualmente la orientación de dicha opinión por cuatro expertos. De los 1590 documentos que conforman la colección de datos, solamente se han tenido en cuenta para la experimentación los 427 que tienen carga subjetiva, y en las que la orientación de la opinión está consensuada por los anotadores.

Como ya se ha indicado, ESQ es un corpus constituido por citas periodísticas, lo que supone que los textos se encuentran redactados por profesionales, y por lo tanto se encontrarán bien formados. Por ende, en este caso no va a ser necesario la aplicación del módulo de limpieza. Los que sí es menester, es el concurso del módulo de análisis morfológico, con el fin de preparar los textos para su ulterior procesamiento por el módulo de desambiguación. Los analizadores elegidos para llevar a cabo las tareas de *tokenización*, análisis morfológico, y obtención de lemas, son los proporcionados por la librería de PLN escrita en Python, NLTK[11] (Loper & Bird, 2002).

Una vez que se ha dividido cada documento en una colección de términos, que se ha identificado su función morfológica en el texto, y que

[9] `http://islrn.org/resources/574-735-957-886-6/`

[10] Las siglas ESQ no son una denominación estándar del corpus, sino que se han empleado para facilitar su referencia en esta memoria.

[11] `http://www.nltk.org/`

se ha determinado su raíz léxica, el siguiente paso es precisar el sentido concreto con el que se está empleando cada término. Ese es el objetivo del módulo de desambiguación. El módulo de desambiguación desarrollado hace uso del algoritmo UKB, el cual ha sido descrito en la Sección 5.2.1. El algoritmo UKB requiere de la selección de una base de conocimiento representable en modo de grafo, donde pueda encontrar la información necesaria para realizar su función, que no es otra que la identificación del concepto subyacente en cada palabra. La base de conocimiento elegida fue WordNet, principalmente porque es la empleada por los autores de UKB, y por ser la base de conocimiento léxica más usada por la comunidad investigadora en PLN.

Tras la desambiguación viene el módulo de expansión, con el que se pretende incrementar la carga semántica del fragmento de texto. El módulo de expansión también se fundamenta en el empleo del algoritmo UKB, ya que su filosofía basada en el aprovechamiento del algoritmo de PageRank, posibilita la extracción de conceptos relacionados a los identificados previamente por el módulo de desambiguación. El módulo de expansión, al igual que le ocurre al algoritmo K-NN (ver Sección 4.2.4), requiere de un estudio previo que permita dilucidar el número de conceptos óptimo con los que expandir el significado. Pero antes de mostrar el estudio que se llevó a cabo, es preciso hablar del módulo de cálculo de la polaridad.

Una vez que se cuenta, no sólo con los sentidos con los que se están empleando los términos, sino también con la expansión del significado, el siguiente paso es el cálculo de la polaridad, para lo cual se emplea la base de conocimiento de opinión SentiWordNet. La clase de polaridad (Positivo, Negativo) se obtiene como resultado de la ecuación 5.11.

$$p = \frac{\sum_{\forall s \in c} r_s (swn_s^+ - swn_s^-)}{|s|} \tag{5.11}$$

donde s representa cada uno de los *synsets* obtenidos del proceso de desambiguación y de expansión; r es el conjunto de pesos PageRank de cada uno de los *synsets* calculados por el algoritmo UKB; y por último, $swn_s{}^+$ y $swn_s{}^-$ se refieren a los valores de polaridad obtenidos de SentiWordNet de cada uno de los *synsets* identificados por UKB durante el proceso de desambiguación y expansión del significado. La polaridad final del documento se obtiene como la suma de todas las polaridades de cada uno de los contextos, que en este caso, se corresponden con las oraciones del texto que se está considerando.

$$p_{total} = \sum_{i=1}^{n} p_i \tag{5.12}$$

Una vez expuesta la implementación concreta del método definido de clasificación de la polaridad basado en expansión del significado para la clasificación de citas periodísticas, ahora es momento de evaluar el algoritmo. Del algoritmo se van a evaluar dos elementos clave, como son la expansión, y si realmente es beneficioso ponderar la polaridad de los conceptos por su valor de PageRank (ver Ecuación 5.11).

La expansión se evaluó ejecutando el clasificador con distintos tamaños de expansión, *e*, en concreto con los valores comprendidos en el conjunto *e* ∈ *[0-50]*. Cuando *e* es igual a 0, el significado del *tweet* no es ampliado con conceptos adicionales, por lo que para la obtención de la polaridad sólo se utilizan los *synsets* identificados por UKB. Se comprobó que se obtenía un máximo con *e=2*, y a partir de *e=6* el rendimiento del clasificador comienza a decrecer y a estabilizarse entorno a un 59 % de F1. En la Figura 5.3 se muestran los 30 primeros valores de expansión, y se puede comprobar gráficamente lo afirmado anteriormente.

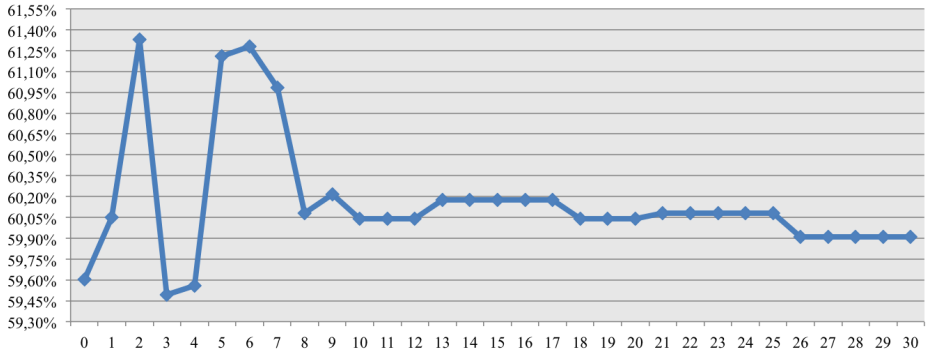


Figura 5.3: Evolución del valor de F1 con diferentes tamaños de expansión.

En la figura se puede observar una mejora repentina de los resultados, llegando F1 a su valor máximo cuando *e=2*. A esa subida acelerada, le sigue una bajada pronunciada, que es continuada por un nuevo repunte en el rendimiento del clasificador. Dicha oscilación puede ser responsabilidad de la relación de antonimia de WordNet. Se debe recordar que UKB es un algoritmo basado en el recorrido del grafo que representa una base de conocimiento, en este caso WordNet. WordNet está conformado por una gran variedad de relaciones semánticas entre los distintos conceptos que

recoge, siendo una de esas relaciones la de antonimia. Por lo tanto, en el proceso de identificación de los conceptos relacionados, es muy probable la inclusión de conceptos antónimos que enturbien el proceso de clasificación de la polaridad. El siguiente punto de inflexión se haya en *e=6*, pero el empeoramiento de los resultados no es tan acusado como a partir de *e=2*. Desde ese momento (*e=6*), el comportamiento del clasificador se estabiliza y no vuelve a experimentar ni mejoría ni retroceso de los resultados. La Tabla 5.2 complementa la representación gráfica de la evaluación del tamaño de expansión.

e	**Precisión**	***Recall***	**F1**
0	59,01%	59,29%	59,60%
1	60,39%	59,72%	60,05%
2	61,99%	60,68%	61,33%
3	60,12%	58,88%	59,49%
4	60,22%	58,91%	59,56%
5	61,86%	60,58%	61,21%
6	61,92%	60,65%	61,28%
7	61,67%	60,32%	60,98%
8	60,76%	59,42%	60,08%
9	60,89%	59,56%	60,22%
10	60,69%	59,40%	60,04%

Tabla 5.2: Evaluación del tamaño de expansión óptimo.

El siguiente aspecto del algoritmo a evaluar es, si conviene la ponderación del valor de polaridad de cada concepto que se obtiene a partir de SentiWordNet con su nivel de PageRank. Para ello se ha ejecutado el algoritmo con dos versiones de la Ecuación 5.11: la primera sin la consideración del parámetro r, y la segunda tal y como se encuentra definida.

Los datos que se muestran en la Figura 5.4 y en la Tabla 5.3 son bastantes reveladores, la ponderación por el valor de PageRank es beneficioso para el proceso de clasificación. El provecho de la ponderación no sólo se manifiesta en la mejora de la clasificación, sino además en la estabilización del comportamiento del clasificador, dado que se reduce la disparidad entre los valores de e. Éstos dos efectos positivos no se muestran cuando el tamaño de la expansión es menor o igual a 2, lo cual, es probable que esté motivado por la inclusión en el proceso de expansión de relaciones semánticas no valiosas para la clasificación de la polaridad.

El corpus ESQ ha sido empleado en otros trabajos (Balahur et al., 2010;

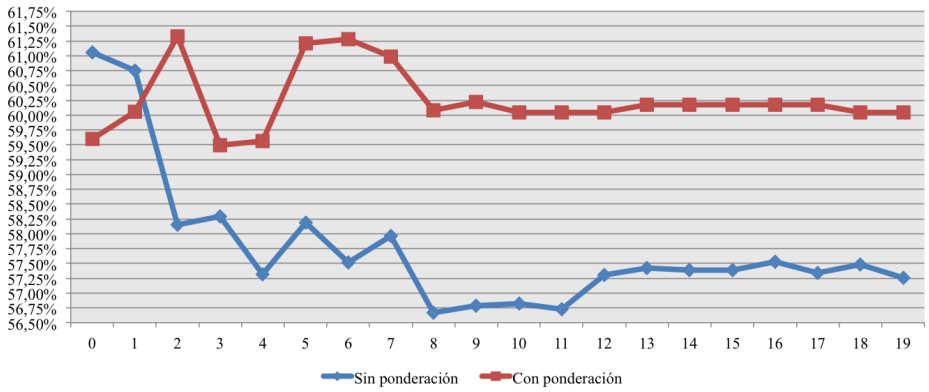


Figura 5.4: Evolución de la influencia de la ponderación de los valores de polaridad con PageRank.

e	**F1 sin ponderación**	**F1 con ponderación**	**Mejora**
0	61,06%	59,60%	-2,39%
1	60,76%	60,05%	-1,16%
2	58,15%	61,33%	5,45%
3	58,29%	59,49%	2,05%
4	57,31%	59,56%	3,90%
5	58,18%	61,21%	5,18%
6	57,51%	61,28%	6,55%
7	57,96%	60,98%	5,21%
8	56,67%	60,08%	6,02%
9	56,78%	60,22%	6,04%
10	56,82%	60,04%	5,67%

Tabla 5.3: Evaluación de la influencia de la ponderación por el valor de PageRank.

Boldrini et al., 2012), pero no se puede realizar una comparación correcta entre esas dos experimentaciones y la expuesta aquí, porque aunque se ha empleado el mismo corpus, en cada experimentación se empleó una selección distinta de documentos. Anteriormente se ha indicado que la evaluación sólo se iba a realizar sobre 427 documentos, selección no realizada por los autores de los dos trabajos citados. Por lo tanto, al realizarse la evaluación con distintos datos, no es posible una comparación real entre los distintos clasificadores que se han aplicado al corpus ESQ.

5.4.2. Textos escritos en español

La naturaleza modular del método de clasificación de la polaridad basada en la expansión del significado que se está presentando, permite sin esfuerzo emplearlo en un idioma distinto al inglés. Lo único que se debe hacer es cambiar la base de conocimiento empleada por el módulo de desambiguación y de expansión, y que ésta siga el estándar de identificación de conceptos de WordNet. Por consiguiente, la adaptación del método para clasificar textos en español sólo requiere el esfuerzo de buscar una base de conocimiento léxica similar a WordNet pero en español. Ésa adaptación, y el resultado de la evaluación, que fueron publicados en (Martínez Cámara et al., 2013c), se van a describir en los siguientes párrafos.

El corpus seleccionado para la evaluación sobre textos en español es la versión española del corpus *SFU Review Corpus*[12] (SFU) (Brooke et al., 2009). El corpus SFU es un corpus comparable en inglés (Taboada & Grieve, 2004) y español sobre opiniones en ocho dominios distintos: libros, coches, ordenadores, lavadoras, hoteles, cine, música y teléfonos. Se trata de un corpus balanceado conformado por 400 opiniones, correspondiendo a cada dominio 50 (25 positivas y 25 negativas). Las opiniones en español fueron descargadas de la web de opiniones Ciao.es[13].

Una vez que se cuenta con el corpus en español, el siguiente paso se corresponde con la búsqueda de una base de conocimiento léxica para el módulo de desambiguación y de expansión. La base de conocimiento léxica compatible con WordNet para español más adecuada, hasta donde nuestro conocimiento alcanza, es MCR, la cual fue descrita en la Sección 5.3.1. A pesar de ser la más adecuada, hay que resaltar que sólo cubre el 50 % de WordNet, de manera que muchos conceptos se quedarán sin identificar debido a las limitaciones del propio recurso. MCR no sólo traslada a español los conceptos que recoge WordNet, sino también las relaciones semánticas existentes entre ellos. En la experimentación sobre el corpus ESQ se ha indicado que la posible intromisión de la relación de antonimia, o de otras relaciones no deseables, han provocado que los resultados no fueran tan buenos como se esperaban. Por este motivo, se ha evaluado el comportamiento del método con un grafo de MCR que incluye la relación de antonimia, y otra versión en la que no se incluye.

En (Martínez Cámara et al., 2013c) se pone de manifiesto que los recursos no son sólo los que se pueden cambiar en el algoritmo propuesto, sino también la fórmula empleada por el método de clasificación de la polaridad. Las fórmulas que se emplearon fueron las siguientes:

[12] `https://www.sfu.ca/~mtaboada/download/downloadCorpusSpa.html`
[13] `http://www.ciao.es/`

$$p_{total} = \frac{1}{|c|} \sum_{i=1}^{|c|} \frac{\sum_{\forall s \in c} r_s (swn_s^+ - swn_s^-)}{|s|} \tag{5.13}$$

$$p_{total} = \frac{1}{|c|} \sum_{i=1}^{|c|} \frac{\sum_{\forall s \in c} r_s \, f(p_s)}{|s|} \tag{5.14}$$

$$f(p_s) = \begin{cases} swn_s^+ & \text{si } swn_s^+ > swn_s^- \\ swn_s^- & \text{si } swn_s^+ <= swn_s^- \end{cases}$$

$$p_{total} = \frac{1}{|c|} \sum_{i=1}^{|c|} \frac{\sum_{\forall s \in c} r_s \, f(p_s)}{|s|} \tag{5.15}$$

$$f(p_s) = \begin{cases} 1 & \text{si } s \in \text{[palabras positivas]} \\ -1 & \text{si } s \in \text{[palabras negativas]} \\ swn_s^+ & \text{si } swn_s^+ > swn_s^- \\ swn_s^- & \text{si } swn_s^+ <= swn_s^- \end{cases}$$

La Ecuación 5.13 es la misma que la evaluada anteriormente (ver Ecuaciones 5.11 y 5.12), pero ponderando el valor de la polaridad por el número de contextos. Mientras, las ecuaciones 5.14 y 5.15 introducen nuevos parámetros en el cálculo de la polaridad. La Ecuación 5.14 no minora el valor de polaridad mayor con el menor, sino que simplemente asigna el máximo valor de polaridad a cada *synset*. La Ecuación 5.15 incorpora un nuevo actor en el cálculo de la polaridad, una lista de palabras de opinión en español. La lista que se ha empleado es iSOL (Molina-González et al., 2013), que se explicará con detalle en el Capítulo 6. Según la Ecuación 5.15, si la palabra que se está considerando se encuentra recogida como positiva o negativa en iSOL, entonces se le asigna el valor 1 o -1, según corresponda; mientras que si no está, se aplica la Ecuación 5.14. En relación a las ecuaciones, indicar finalmente, que al igual que en el caso anterior, se ha considerado como contexto una oración.

La combinación de las tres ecuaciones con la evaluación de la consideración o no de la relación de antonimia, dio lugar a las siguientes configuraciones del clasificador:

1. wnet_ant+_eq1_es: Caso base de la experimentación, que se corresponde con la aplicación de la ecuación 5.13.

2. wnet_ant-_eq1_es: Igual que el caso anterior, pero en este caso no

se tiene en cuenta la relación de antonimia.

3. wnet_ant+_eq2_es: En esta configuración sí se emplea una versión del grafo de MCR con la relación de antonimia. Para el cálculo de la polaridad se utiliza la ecuación 5.14.

4. wnet_ant-_eq2_es: Igual que la configuración anterior, pero en este caso no se tiene en cuenta la relación de antonimia.

5. wnet_ant+_eq3_es: El módulo de desambiguación y expansión emplean una versión de MCR en la que sí se incluye la relación de antonimia, y la polaridad se calcula con la ecuación 5.15.

6. wnet_ant-_eq3_es: Lo mismo que en el caso anterior, pero en este caso no se tiene en cuenta la relación de antonimia.

En la experimentación sobre textos en inglés se evaluó el tamaño de la expansión y la influencia de la ponderación del valor de polaridad con el nivel de PageRank. Al demostrarse la influencia positiva de la ponderación con el valor de PageRank, en este caso no se va a comprobar. Lo que sí se va a continuar evaluando es el tamaño de expansión adecuado, ya que dicho parámetro depende del problema que se esté tratando. Además, en este caso se están utilizando tres ecuaciones diferentes para el cálculo de la orientación de la opinión, así como dos versiones distintas del grafo de MCR, uno con la relación de la antonimia, y otro sin ella. De manera que será preciso en cada caso determinar el tamaño de expansión más adecuado. La Figura 5.5 superpone el resultado obtenido por cada una de las 6 configuraciones anteriormente indicadas.

En la Figura 5.5 se pueden ver dos comportamientos bien diferenciados, el provocado por la primera y segunda ecuación, y el producido por la tercera. Centrando el análisis en el primer comportamiento, se puede apreciar que la eliminación de la base de conocimiento de la relación de antonimia influye positivamente en el proceso de clasificación, independientemente del tamaño de la expansión. Como se puede ver en la Tabla 5.4, cuando la antonimia es incluida en el grafo, la Ecuación 5.13 reporta un resultado mínimamente superior al de la Ecuación 5.14, en concreto de un 0,07%. Pero cuando la antonimia no es tenida en cuenta, es la Ecuación 5.14 la que resulta tener un mejor comportamiento, que se traduce en una diferencia de un 0,39%. Siendo las diferencias entre los resultados tan insignificantes para la elección de una ecuación u otra, habrá que considerar otro parámetro, que en este caso es el tamaño de la expansión. Cuando no se tiene en cuenta la antonimia, la Ecuación 5.13 alcanza su máximo resultado con

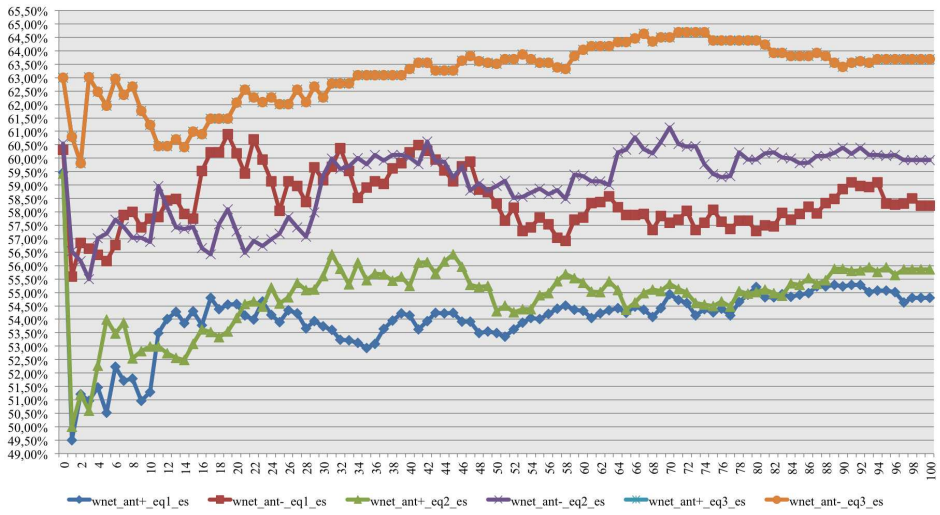


Figura 5.5: Evaluación del tamaño de expansión más adecuado para cada una de las configuraciones planteadas.

un expansión de 19 *synsets* adicionales. Mientras que cuando la Ecuación 5.14 es la empleada, el mejor resultado se obtiene con una expansión de 70 *synsets*. Por lo tanto, si se quiere primar el tiempo de ejecución, la Ecuación 5.13 es la que debe ser elegida, pero si se le da más importancia al resultado de la clasificación entonces debe ser elegida la Ecuación 5.14.

Configuración	***e***	**Precisión**	***Recall***	**F1**
wnet_ant+_eq1_es	0	65,42 %	54,50 %	59,46 %
wnet_ant-_eq1_es	19	64,39 %	57,75 %	60,89 %
wnet_ant+_eq2_es	0	68,03 %	52,75 %	59,42 %
wnet_ant-_eq2_es	70	64,62 %	58,00 %	61,13 %
wnet_ant+_eq3_es	71	65,91 %	63,50 %	64,68 %
wnet_ant-_eq3_es	71	65,91 %	63,50 %	64,68 %

Tabla 5.4: Mejores resultados obtenidos por la 6 configuraciones estudiadas del algoritmo de clasificación de la polaridad basado en expansión del significado.

El segundo comportamiento es el provocado por la Ecuación 5.15, que como se puede apreciar en la Figura 5.5, tanto se tenga en cuenta la antonimia o no, se consiguen los mejores resultados independientemente de los *synsets* adicionales que se inserten en el proceso de clasificación de

la polaridad. Si se recuerda, la Ecuación 5.15 combina SentiWordNet y la lista de palabras de opinión en español iSOL.

Los resultados que reportan las Ecuaciones 5.14 y 5.15 se diferencian en un 5,81 % en favor de la Ecuación 5.15, lo cual parece indicar que la combinación de SentiWordNet e iSOL ha sido bastante positiva. Pero para comprobar adecuadamente la superioridad de la Ecuación 5.15, se hace necesario compararla con un clasificador basado en léxico[14], siendo en este caso iSOL el léxico seleccionado. El clasificador que usa iSOL proporciona unos resultados que alcanzan un 64,33 % de F1, lo cual se traduce en una diferencia de 0,54 % con respecto al resultado que ofrece la Ecuación 5.15. La diferencia entre los dos clasificadores es nimia, por lo que la Ecuación 5.15 no es tan buena como aparentaba, ya que un clasificador más simple proporciona unos resultados semejantes. Por tanto, cabe concluir, que el método de expansión del significado para textos largos en español requiere todavía de trabajo y análisis para que realmente sea beneficioso su uso para la clasificación de la polaridad de textos largos en español.

5.5. Experimentación sobre textos cortos

Una vez descrito el comportamiento del método de clasificación de la polaridad con textos largos, es momento de comprobar su desempeño con textos cortos. Al igual que en el Capítulo 4, la fuente de textos cortos seleccionada fue Twitter. Ya se ha mencionado en varias ocasiones que uno de los obstáculos en la identificación de la opinión en *tweets* es su longitud, dado que en 140 caracteres es complicado que haya contexto suficiente que ayude al proceso de identificación de la orientación de la opinión. Un método, como el que se está presentando, basado en la extensión del significado del mensaje que se está tratando, puede ayudar considerablemente a la clasificación de la polaridad sobre *tweets*. Por consiguiente, se emprendió una evaluación sobre *tweets* escritos en inglés, y sobre *tweets* escritos en español.

5.5.1. Textos escritos en inglés

La disponibilidad de corpus de *tweets* en inglés para la comunidad investigadora no es muy elevada, debido principalmente a la restricción de la distribución de colecciones de *tweets*[15], que Twitter impuso cuando

[14] La descripción de la experimentación empleando clasificadores de opinión basado en léxico se realizará en el Capítulo 6.

[15] Debe precisarse que se pueden proporcionar conjuntos de identificadores de *tweets*, los cuales pueden usarse para descargar el texto de cada mensaje. Éste es el mecanismo

presentó la versión 1.1 de su API[16]. Pero, esa restricción no afectaba a los *tweets* que se habían descargado con anterioridad, como es el caso de los que forman parte del primer corpus de *tweets* en inglés, el *Stanford Twitter Corpus* (STS) (Go et al., 2009). Este corpus, además de ser el primero que se puso a disposición de la comunidad investigadora, es el que se suele tomar como referencia para evaluar los sistemas de clasificación de la polaridad en *tweets* escritos en inglés. El corpus STS se compone de una colección para entrenamiento de 1,6 millones de *tweets* categorizados según un enfoque de etiquetado impuro[17], y otra colección para test. La colección para test es mucho más reducida, dado que se compone solamente por 177 *tweets* negativos, y 182 *tweets* positivos etiquetados manualmente. En (Montejo-Ráez et al., 2014) se empleó el corpus STS para evaluar la clasificación de la polaridad basada en extensión del significado sobre textos cortos.

Si se recuerda, el método de expansión del significado, que se está describiendo, cuenta con varios módulos que se pueden utilizar dependiendo de la naturaleza de los datos que se van a procesar y de las características del problema que se pretende resolver. El primero de los módulos que componen el flujo de trabajo es el relativo a la limpieza o preparación de los datos. En el caso de los textos largos, como se ha explicado anteriormente, no se ha requerido el módulo de limpieza, debido a que los textos que se han procesado están bien formados. Por contra, los textos que provienen de Twitter no suelen estar bien formados, por lo que en este caso sí es preciso aplicar el módulo de limpieza. El preprocesamiento que se le ha aplicado a los *tweets* consta de las siguientes operaciones:

1. Al centrarse exclusivamente la evaluación del método al texto presente en los *tweets*, se estimó oportuno la eliminación de las direcciones web por no aportar información al proceso de clasificación de la polaridad.

2. Aunque pueda parecer inverosímil, en algunos *tweets* pueden aparecer rutas a ficheros. Normalmente esas rutas se corresponden con URL mal formadas. Estos elementos se eliminaron por no aportar información relevante al proceso de identificación de la orientación de la opinión.

3. Eliminación de todos los emoticonos que puedan aparecer en los *tweets*.

que emplea la tarea de AO en Twitter que se organiza en el taller SemEval.

[16] `https://dev.twitter.com/overview/terms`

[17] Debe recordarse que el etiquetado impuro es aquel que se sustenta en el significado de los emoticonos presentes en el *tweet*.

4. Borrado de las etiquetas y entidades HTML.
5. Eliminación de menciones, es decir, borrado de todas las expresiones del tipo *@[A-Za-z0-9_-]$^+$*.
6. Borrado de la onomatopeya de risa.
7. Eliminación de los caracteres no alfanuméricos.

Tras la depuración de los *tweets*, viene la aplicación del módulo correspondiente con el análisis morfológico. El análisis morfológico, en este caso, ha estado conformado por: un proceso de *tokenización*, otro de *pos-tagging* y finalmente por uno de *lematización*. La salida del análisis morfológico constituye la entrada del módulo de desambiguación, en el cual, al igual que en el caso de textos largos, se ha empleado el algoritmo UKB. El algoritmo UKB no toma un texto y lo desambigua, sino que recibe como entrada un conjunto de términos a los que se le llama contexto. En la experimentación realizada sobre textos largos, esos contextos han coincido con las oraciones de los documentos que se han procesado. Debido a la característica corta longitud de los *tweets*, se han evaluado dos formas distintas de generar los contextos, por un lado se ha tomado como contexto cada una de las oraciones que aparecen en los *tweets*, y por otro se ha considerado como contexto el *tweet* completo. Tras la desambiguación, es el turno del módulo de expansión, para lo cual también se ha utilizado el algoritmo UKB. Tanto para la desambiguación, como para la expansión, se ha comprobado el comportamiento de cuatro versiones distintas del grafo subyacente en WordNet. Esas cuatro versiones se corresponden con el grafo que incluye todas las relaciones, con aquel en el que la relación de antonimia no es considerada, con el conformado por todas las relaciones semánticas más las que surgen de la inclusión de las glosas[18], y por último, el resultante de la inserción de las glosas sin tener en cuenta la relación de antonimia.

Al último módulo le atañe el cálculo de la polaridad. Se ha evaluado el comportamiento de la ecuación 5.13 sobre *tweets* escritos en inglés. Si se recuerda, dicha ecuación pondera el valor de la polaridad por el número de contextos que se han construido para el proceso de desambiguación y de expansión ($|c|$). En la evaluación que se está describiendo, el parámetro $|c|$ puede tomar el valor 1, cuando se toma el *tweet* completo como contexto, o puede coincidir con el número de oraciones que se hallen en el mensaje. Asimismo se ha experimentado con la influencia de la ponderación del valor de polaridad con el nivel de PageRank del *synset*.

[18]Una glosa es la definición que WordNet asocia a cada uno de los *synsets* que lo conforman.

Expuesta la adaptación del algoritmo al caso específico de la clasificación de la polaridad de *tweets* escritos en inglés, antes de mostrar los resultados, y con ánimo de facilitar la compresión de la evaluación que se ha emprendido, se van a listar las distintas configuraciones del algoritmo que han sido descritas en los párrafos anteriores:

1. wnet_ant+_tweet_contexto: Configuración que se corresponde con la consideración del grafo de WordNet con la relación de antonimia, y se ha tomado como contexto para la desambiguación y expansión el *tweet* completo. En esta configuración no se ha ponderado la diferencia de la polaridad de cada *synset* por su valor de PageRank.

2. wnet_ant+_glosas_tweet_contexto: Idéntica configuración a la anterior, pero en este caso se incluye al grafo de WordNet las relaciones que surgen de las glosas.

3. wnet_ant-_tweet_contexto: Igual que wnet_ant+_tweet_contexto pero eliminando del grafo de WordNet la relación de antonimia.

4. wnet_ant-_glosas_tweet_contexto: Configuración similar a wnet_ant+_glosas_tweet_contexto, pero sin incluir en el grafo de WordNet la relación de antonimia.

5. wnet_ant+_tweet_contexto_pagerank: De nuevo se considera el grafo de WordNet con la relación de antonimia, el *tweet* como único contexto, pero ahora sí se pondera el valor de la diferencia de polaridad de cada *synset* por su nivel de PageRank.

6. wnet_ant+_glosas_tweet_contexto_pagerank: Igual que wnet_ant+_tweet_contexto_pagerank, pero insertando en el grafo de WordNet la información que proporcionan las glosas.

7. wnet_ant-_tweet_contexto_pagerank: Configuración semejante a wnet_ant+_tweet_contexto_pagerank, pero se elimina del grafo de WordNet la relación de antonimia.

8. wnet_ant-_glosas_tweet_contexto_pagerank: Situación en la que a wnet_ant-_tweet_contexto_pagerank se le incluye la información que proporcionan las glosas de WordNet.

9. wnet_ant+_oraciones_contexto: Es idéntico que la configuración wnet_ant+_tweet_contexto, pero en este caso se identifican las oraciones que componen el *tweet*, y se hace corresponder cada oración con un contexto. Ésto hace que en la fórmula de clasificación de

la polaridad el valor del parámetro $|c|$ coincida con el número de oraciones del *tweet*.

10. wnet_ant-_oraciones_contexto: Idéntico que wnet_ant+_oración_-contexto, pero sin introducir en el grafo de WordNet la relación de antonimia.

11. wnet_ant+_glosas_oraciones_contexto: Similar a wnet_ant+_oración_contexto pero incluyendo en el grafo las glosas de WordNet.

12. wnet_ant-_glosas_oraciones_contexto: Configuración idéntica a wnet_-ant+_glosas_oración_contexto pero sin incluir a la relación de antonimia.

13. wnet_ant+_oraciones_contexto_pagerank: Igual que wnet_ant+_-oración_contexto pero ponderando la diferencia de polaridad por el valor de PageRank.

14. wnet_ant+_glosas_oraciones_contexto_pagerank: Similar a wnet_-ant+_oración_contexto_pagerank, pero incluyendo la información que aportan las glosas.

15. wnet_ant-_oraciones_contexto_pagerank: Configuración semejante a wnet_ant+_oración_contexto_pagerank pero sin insertar en el grafo la relación de antonimia.

16. wnet_ant-_glosas_oraciones_contexto_pagerank: Idéntico a wnet_-ant-_oración_contexto_pagerank, pero además se incluyen la información que proporcionan las glosas.

Como se ha indicado en el la sección sobre textos largos (ver Sección 5.4), el método de clasificación que se propone depende de el número de conceptos con los se extiende el significado del mensaje. Por lo tanto, las anteriores configuraciones listadas se han evaluado con un rango de expansión que oscila entre 0 y 100 *synsets* adicionales. La Figura 5.6 muestra el comportamiento de las 16 configuraciones con una expansión $e \in [0..100]$.

Si se mira con detenimiento la Figura 5.6 se puede observar, que al contrario de lo que ocurría con la experimentación sobre textos largos en español, la relaciones semánticas del grafo de WordNet no son tan determinantes para el resultado de la clasificación. Cuando se incluye al grafo de WordNet la información procedente de las glosas, el rendimiento del algoritmo aumenta considerablemente, haciendo que esto sea un indicativo claro de que las glosas son un componente adecuado para la expansión

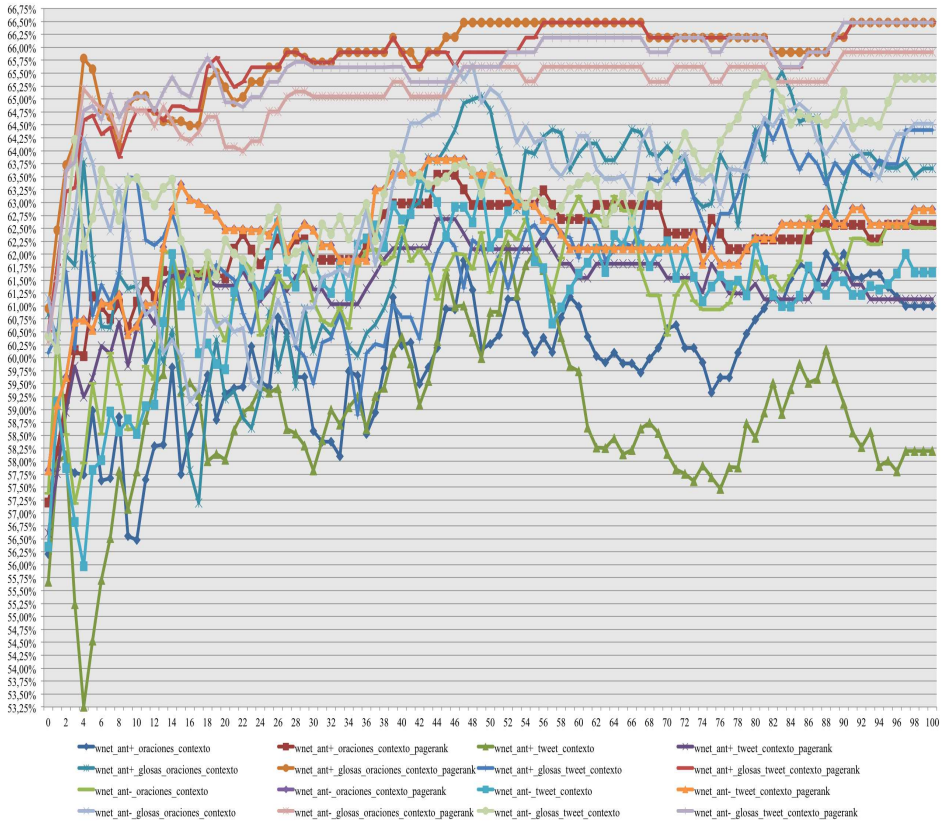


Figura 5.6: Evolución de los resultados obtenidos por cada una de las configuraciones del método de clasificación que se propone.

del significado del mensaje que transporta un *tweet*. Debido al elevado número de configuraciones del algoritmo que se están evaluando, la Figura 5.6 parece algo saturada, por lo que no permite vislumbrar otros detalles que posibiliten la identificación de la configuración idónea, o de simplemente analizar los elementos que facilitan el descubrimiento de la polaridad de un *tweet*. Como en esta ocasión, la toma en consideración o no de la relación de antonimia no es decisiva, se van a mostrar dos gráficas, una en la que se considera la relación de antonimia (Figura 5.7) y otra en la que no se tiene en cuenta (ver Figura 5.8).

En la dos figuras (5.7 y 5.8) se puede observar un comportamiento muy claro, y coincidente con el obtenido sobre textos largos. Dicho comportamiento es el que aporta la ponderación de los valores de polaridad por el valor de PageRank asociado a cada uno de los conceptos que se

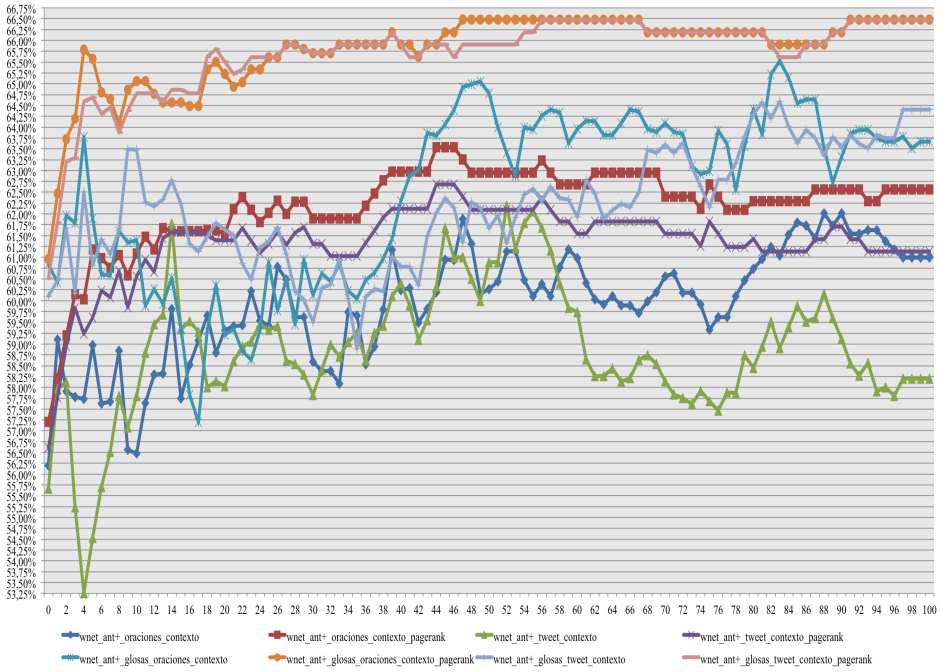


Figura 5.7: Evolución de los resultados obtenidos por las configuraciones en las que se tiene en cuenta la relación de antonimia.

han empleado en el proceso de expansión. La incorporación de PageRank en la fórmula de cálculo de la polaridad estabiliza el comportamiento del algoritmo, ya que las diferencias de los resultados que se obtienen con cada expansión son menos pronunciadas e incrementa el rendimiento del algoritmo. Debe recordarse, que ese mismo comportamiento también tuvo lugar en la experimentación con textos largos. Asimismo, cuando se emplea PageRank y se consideran las relaciones de antonimia de WordNet, se necesitan menos conceptos adicionales para la obtención del máximo resultado por parte del algoritmo, como se verá con más detalle en la tabla de resultados (ver Tabla 5.5).

También se ha evaluado la influencia del contexto de desambiguación a la calidad de la clasificación. Como se puede apreciar en las figuras no existe una diferencia clara entre identificar la oraciones que constituyen el *tweet*, o tomar como contexto el *tweet* completo, aunque cuando la antonimia no se incorpora en el proceso de desambiguación, el uso del *tweet* completo como contexto es ligeramente más beneficioso que la identificación de las oraciones.

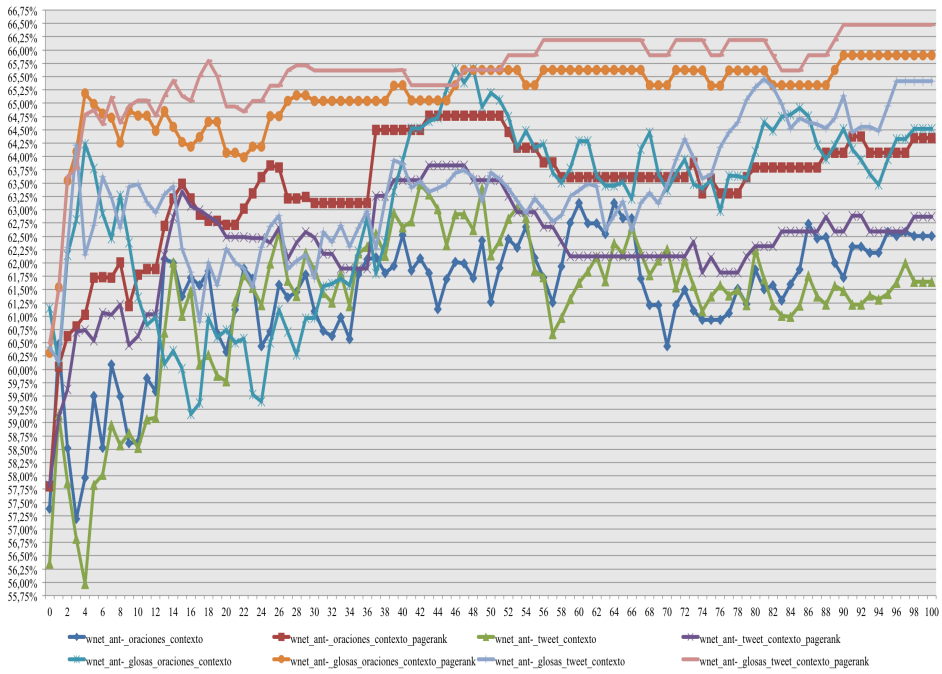


Figura 5.8: Evolución de los resultados obtenidos por las configuraciones en las que no se tiene en cuenta la relación de antonimia.

Analizado el comportamiento general de las diversas configuraciones evaluadas del método a través de las Figuras 5.6, 5.7 y 5.8, es momento de comprobar si las aserciones anteriores, a partir de la información contenida en dichas figuras, se corresponden con los mejores resultados que se han obtenido. Para ello la Tabla 5.5 muestra los resultados obtenidos por cada una de las configuraciones, tanto cuando no se emplea ningún *synset* adicional, como con el tamaño de expansión que mejor resultado obtiene.

Se ha afirmado anteriormente que la ponderación con PageRank siempre es beneficiosa, y la tabla de resultados corrobora dicha afirmación. Siempre que se emplea el valor de PageRank el algoritmo ofrece unos valores superiores de F1 y *Accuracy* que su versión sin PageRank. También se indicaba que la consideración de la antonimia no es determinante, pudiéndose confirmar este hecho con la observación de que la configuración que obtiene un mayor resultado de F1, un 66,48%, tiene en cuenta la antonimia, mientras que la siguiente configuración mejor obtiene un 66,47% de F1 sin incluir la relación de antonimia. La diferencia es nimia, por consiguiente no se puede afirmar que sea adecuada la retirada de

la antonimia. Pero, como se ha manifestado anteriormente, cuando la antonimia es tenida en cuenta, menos *synsets* adicionales son necesarios para alcanzar el máximo resultado por parte del algoritmo. Por ende, para la clasificación de la polaridad de *tweets* en inglés puede no ser recomendable retirar del grafo de WordeNet la relación de antonimia. Si la relación de antonimia no es determinante, cuando no se mira el tamaño de la expansión, la manera de construir el contexto de desambiguación tampoco es decisivo. A pesar de esto, parece más recomendable, a tenor de lo recogido por la Tabla 5.5, la identificación de las oraciones que conforman un *tweet*, que el uso del propio *tweet* como contexto.

Comparación de Resultados

Como se ha indicado anteriormente, una de las razones por las que se seleccionó el corpus STS para llevar a cabo la evaluación del método basado en expansión del significado es que se trata de un corpus ampliamente empleado por la comunidad investigadora, lo cual permite comparar los resultados que se han obtenido. El primer trabajo en el que se usa el corpus como conjunto de datos de evaluación es en el propio artículo donde se presenta, es decir, en (Go et al., 2009). Los autores comprueban tanto la validez de distintos conjuntos de características para representar la información contenida en los *tweets*, como varios algoritmos de aprendizaje supervisado. En cuanto a las características, los autores se detienen a analizar el potencial de representación de *unigramas*; *bigramas*; *unigramas* y *bigramas* y las anteriores características junto con la categoría morfológica. En cuanto a los algoritmos, se centran en la experimentación con SVM, Naïve Bayes, y Máxima Entropía. El mejor resultado se alcanzó representado los *tweets* como un conjunto de *unigramas* y *bigramas*, y empleando como algoritmo Máxima Entropía.

Bifet & Frank (2010) también presentan una evaluación de la clasificación de la polaridad en el corpus STS. Los autores se centran en la evaluación de algoritmos de clasificación en tiempo real o de datos en línea[19], como son Naïve Bayes multimodal (Lewis & Gale, 1994; Mitchell, 1997), Gradiente Estocástico Descendente (*Stochastic Gradient Descendent*) (Kushner & Yin, 1997) y el algoritmo conocido como Árbol Hoeffding (*Hoeffding Tree*) (Domingos & Hulten, 2000).

[19]Los algoritmos de procesamiento de datos en línea o en tiempo real son clasificadores preparados para la clasificación de conjunto de datos cuya distribución tiene una alta tasa de variación. No es el objetivo de esta memoria la descripción de este paradigma de clasificación, de manera que se recomienda la lectura de (Muthukrishnan, 2005) para aquellos lectores que estén interesados en este tipo de clasificación.

Experimento	***e***	**Precisión**	***Recall***	**F1**	***Accuracy***
wnet_ant+_oraciones_-	0	67,04%	48,76%	56,20%	48,46%
contexto	90	63,32%	60,39%	62,01%	60,44%
wnet_ant+_oraciones_-	0	67,25%	49,76%	57,20%	49,86%
contexto_pagerank	44	65,42%	61,74%	63,53%	61,83%
wnet_ant+_tweet_con-	0	66,63%	47,80%	55,66%	47,91%
texto	52	64,39%	60,13%	62,19%	60,16%
wnet_ant+_tweet_con-	0	66,60%	49,21%	56,60%	49,30%
texto_pagerank	44	64,55%	60,90%	62,67%	61,00%
wnet_ant+_glosas_ora-	0	71,69%	52,84%	60,84%	52,92%
ciones_contexto	83	67,52%	63,64%	65,52%	63,50%
wnet_ant+_glosas_oracio-	0	70,98%	53,40%	60,95%	53,48%
nes_contexto_pagerank	47	68,12%	64,91%	66,48%	64,90%
wnet_ant+_glosas_-	0	71,74%	51,70%	60,09%	51,81%
tweet_contexto	83	66,50%	62,79%	64,59%	62,67%
wnet_ant+_glosas_tweet_-	0	70,80%	52,83%	60,51%	52,92%
contexto_pagerank	66	68,14%	64,88%	66,47%	64,90%
wnet_ant-_oraciones_-	0	68,84%	49,19%	57,38%	49,30%
contexto	60	65,66%	61,20%	63,12%	61,28%
wnet_ant-_oraciones_con-	0	68,43%	50,02%	57,80%	50,13%
texto_pagerank	43	66,82%	62,84%	64,77%	62,95%
wnet_ant-_tweet_con-	0	68,65%	47,78%	56,34%	47,91%
texto	42	65,91%	61,20%	63,47%	61,28%
wnet_ant-_tweet_con-	0	68,45%	50,04%	57,82%	50,13%
texto_pagerank	43	65,75%	62,02%	63,83%	62,11%
wnet_ant-_glosas_ora-	0	72,03%	53,12%	61,15%	53,20%
ciones_contexto	46	68,05%	63,41%	65,64%	63,23%
wnet_ant-_glosas_oracio-	0	70,23%	52,84%	60,31%	52,92%
nes_contexto_pagerank	90	67,54%	64,34%	65,90%	64,34%
wnet_ant-_glosas_-	0	72,10%	51,98%	60,41%	52,08%
tweet_contexto	81	67,38%	63,62%	65,45%	63,50%
wnet_ant-_glosas_tweet_-	0	70,74%	52,84%	60,49%	52,92%
contexto_pagerank	90	68,13%	64,89%	66,47%	64,90%

Tabla 5.5: Resultados de la evaluación del sistema no supervisado de clasificación de la polaridad sobre *tweets* en inglés.

Speriosu et al. (2011) presentan un algoritmo de clasificación basado en un método de propagación de etiquetas (*Label Propagation*) (Zhu & Ghahramani, 2002). El método consiste en la combinación, por medio de un algoritmo semisupervisado de propagación de etiquetas, de información procedente de varias fuentes de conocimiento formadas por emoticonos, *tweets*, los usuarios de los *tweets*, *hashtags*, *unigramas* y *bigramas*.

Saif et al. (2012) intentan resolver el problema de la diversidad de referencias a un mismo concepto en Twitter mediante la representación del texto de los *tweets* a través de un conjunto de características semánticas, y un conjunto de características extraídas mediante un método de identificación de temas. Para la identificación de las características semánticas, los autores emplean un algoritmo de *clustering* de entidades. Por medio de esta técnica, el espacio de características se ve disminuido considerablemente, debido al agrupamiento de entidades en torno a etiquetas semánticas. Para la identificación de las características que representan el tema del texto, los autores emplean un método basado en un modelo de tema-sentimiento conjunto (*joint sentiment-topic*) (Lin & He, 2009). Como algoritmo de clasificación se utiliza Naïve Bayes.

Como se muestra en la Tabla 5.6, dos son las grandes diferencias entre el método aquí presentado y los descritos en los párrafos anteriores. La primera es el tipo de aprendizaje empleado en cada método. Mientras que el método propuesto es no supervisado, en los existentes en el estado del arte el tipo de aprendizaje es supervisado o semisupervisado. Como era de esperar, la otra diferencia se encuentra en el resultado, el método que se expone obtiene unos resultados inferiores, debido principalmente a que no utiliza ningún modelo de datos que se encuentren previamente etiquetados.

5.5.2. Textos escritos en español

La investigación con *tweets* en español se circunscribe a la participación en la edición de 2013 del taller TASS. Como ya se ha indicado, TASS es el principal foro de evaluación de sistemas de clasificación de la polaridad de textos en español publicados en Twitter en España e Iberoamérica. La participación en la edición de 2013 consistió en el desarrollo de un sistema similar al que se ha venido explicando hasta ahora en este capítulo, pero en lugar de apostar por la extensión del significado del mensaje subyacente al *tweet*, trata de obtener la polaridad mediante el aprovechamiento de la información de opinión que proporcionan SentiWordNet, Q-WordNet e iSOL. A continuación se desarrollará la participación en la edición de 2013 de TASS, la cual puede encontrarse en (Martínez Cámara et al., 2013a).

La arquitectura del sistema desarrollado es similar a la presentada

Artículo	**Aprendizaje**	**Algoritmo**	**Características**	***Accuracy***
(Go et al., 2009)	Supervisado	Máxima Entropía	*Unigramas* + *Bigramas*	83,00 %
(Bifet & Frank, 2010)	Supervisado	Naïve Bayes Multinominal	*Unigramas*	82,45 %
(Speriosu et al., 2011)	Semisupervisado	Propagación de etiquetas	*Unigramas* + *Bigramas* + Características propias de Twitter	84,70 %
(Saif et al., 2012)	Supervisado	Naïve Bayes	Semánticas + Tema-Sentimiento	86,30 %
(Montejo-Ráez et al., 2014)	No supervisado	Expansión de significado	*Unigramas*	64,90 %

Tabla 5.6: Comparación entre el método propuesto y otros algoritmos presentes en el estado del arte.

en la Figura 5.1, pero sin el proceso de expansión del significado. El procesamiento comienza por el módulo de limpieza, el cual se fundamenta en la aplicación de un corrector ortográfico y por la expansión de abreviaturas. La inclusión de estos dos procesamientos se debió a la propia naturaleza del sistema. Si en el sistema de clasificación supervisado, que se explicó en la Sección 4.4, era importante que las palabras del *tweet* estuvieran bien formadas, en este caso es aún más relevante, porque la incorrecta escritura de una palabra puede provocar que no se encuentre en uno de los tres recursos de opinión que se han incluido en el sistema, y por tanto, inducir a error al clasificador. Como corrector ortográfico se empleó GNU Aspell[20], que no es otro que la base de los correctores ortográficos que se incluyen en los sistemas GNU Linux. Aspell se caracteriza porque permite personalizar su comportamiento mediante la inclusión de diccionarios, verbigracia, se puede incluir un diccionario de palabras a considerar como correctas que no forman parte del diccionario de español. Esta característica es muy útil porque posibilita indicar al corrector que los vocablos propios de la jerga de Twitter, como puede ser *timeline*[21], no los considere como faltas de ortografía. En cuanto a las abreviaturas, se definió manualmente

[20] `http://aspell.net/`

[21] El *timeline* de un usuario es el propio perfil del usuario en Twitter, en el cual aparecen ordenados cronológicamente los *tweets* que ha ido publicando.

un diccionario de abreviaturas con su correspondiente expansión.

El siguiente módulo se corresponde con el análisis morfológico. Dicho análisis, como en los sistemas descritos anteriormente, se responsabiliza de la identificación de los términos que constituyen el *tweet*, de la identificación de la categoría morfológica de las palabras y de la extracción del lema de cada una de ellas. En este caso también se ha incluido un proceso de identificación de signos de exclamación y de sonrisas con el fin de incluir la información que aportan sobre la polaridad del mensaje en el módulo de clasificación de la polaridad. Para la desambiguación se ha empleado el ya más que descrito algoritmo UKB. En este caso, al carecer el sistema de módulo de expansión, UKB solo se ha ejecutado para determinar el concepto exacto al que se refiere cada palabra en el texto.

El clasificador de la polaridad se construye a partir del uso conjunto de SentiWordNet, Q-WordNet e iSOL, así como de la presencia de algunos elementos léxicos en el *tweet*, como es el caso de las admiraciones, *emoticonos* y de las onomatopeyas de risas. Siempre que una palabra de opinión está acompañada por un signo de admiración, su polaridad se incrementa en 0,1 unidades. A las expresiones que representan sonrisas se le asigna siempre el valor de polaridad 0,75. En cuanto a los *emoticonos* se ha intentado graduar la polaridad que transmiten. A partir de los *emoticonos* listados en Wikipedia[22] se ha elaborado una lista de 147 *emoticonos*, los cuales tienen asociados cuatro valores distintos de polaridad: _VERY_-POSITIVE_ (+1), _POSITIVE_ (+0,75), _NEGATIVE_ (-0,75) y _-VERY_NEGATIVE_ (-1).

Cinco módulos de clasificación de la polaridad diferentes fueron evaluados con el conjunto de entrenamiento del Corpus General del TASS, con el ánimo de encontrar el más apropiado para someterlo a evaluación en el seno del taller. Esos cinco módulos de la clasificación de la polaridad que se evaluaron fueron:

POL1: En este caso, se suma para cada término la polaridad positiva, negativa y neutra que proporciona cada recurso. Se debe recordar que iSOL es una lista de palabras positivas y negativas, de manera que siempre asignará cero como valor de neutralidad. En el caso de que el término sea un *emoticono* se le asigna su polaridad en función de la intensidad de opinión que transmita. En el caso de que una admiración acompañe a una palabra con polaridad, el valor de opinión de ésta se verá incrementada en 0,1 o -0,1, en función si es positiva o negativa. Si la palabra con polaridad tiene letras

[22] `http://es.wikipedia.org/wiki/Anexo:Emoticonos`

repetidas, se sobreentiende que el autor del *tweet* tiene la intención de expresarla con una mayor intensidad, por lo que se incrementa su valor de polaridad en 0,05. Por último, se suman las polaridades de todos los términos que constituyen el *tweet*, se normalizan para que el valor de polaridad siempre oscile en un rango de valor mínimo 0 y máximo de 1, y dependiendo de si la clasificación es en 6 clases o 4 clases se consideran unos umbrales u otros para la asignación de la clase final.

POL2: Se trata de un clasificador similar al anterior, pero la polaridad final del *tweet* se divide entre el número total de palabras que tienen polaridad.

POL3: En este caso se simplifica el cálculo de la polaridad de cada palabra, si la suma de las polaridades que cada uno de los tres recursos asigna a cada palabra es positiva, entonces a la palabra se le asigna un valor 1 de polaridad, si es negativa, entonces -1, si por contra es neutra su polaridad es cero. La polaridad del *tweet* se obtiene a partir de la suma de las polaridades de cada una de las palabras, y dividiendo la dicha suma entre el número de palabras que tienen polaridad.

POL4: Igual que POL3, pero dividiendo por el número total de palabras que forman el *tweet*.

POL5: Igual que POL3, pero considerando el valor de polaridad aportado por cada recurso.

Cada uno de los cinco módulos de cálculo de la polaridad descritos originan un sistema de clasificación distinto, de manera que fue necesario la selección de uno de ellos para generar la clasificación final que fue enviada como participación en el taller. Con tal fin, se tomó el conjunto de entrenamiento del Corpus General del TASS, y se ejecutaron los cincos algoritmos resultantes del desarrollo de los anteriores módulos de clasificación descritos. El taller siempre evalúa los sistemas en un entorno de clasificación de 6 y 4 clases. La evaluación interna que se llevó a cabo para la elección del sistema a enviar se realizó en un entorno de clasificación de cuatro clases. La Tabla 5.7 muestra los resultados que se obtuvieron en la evaluación interna, y la Figura 5.9 permite observar de un vistazo el sistema que devolvió los mejores resultados.

Con total claridad, como se puede comprobar de una manera más nítida en la Figura 5.9, el módulo de clasificación de la polaridad que mejores resultados devuelve es POL1. Hay dos elementos que pueden ser

Algoritmo	Precisión	*Recall*	F1	*Accuracy*
POL1	52,24%	51,80%	52,02%	53,80%
POL2	50,41%	45,94%	48,07%	45,80%
POL3	49,30%	48,42%	48,86%	50,87%
POL4	50,77%	40,74%	45,21%	38,53%
POL5	49,35%	49,07%	49,21%	50,94%

Tabla 5.7: Resultado de la evaluación interna de los módulos de clasificación de la polaridad.

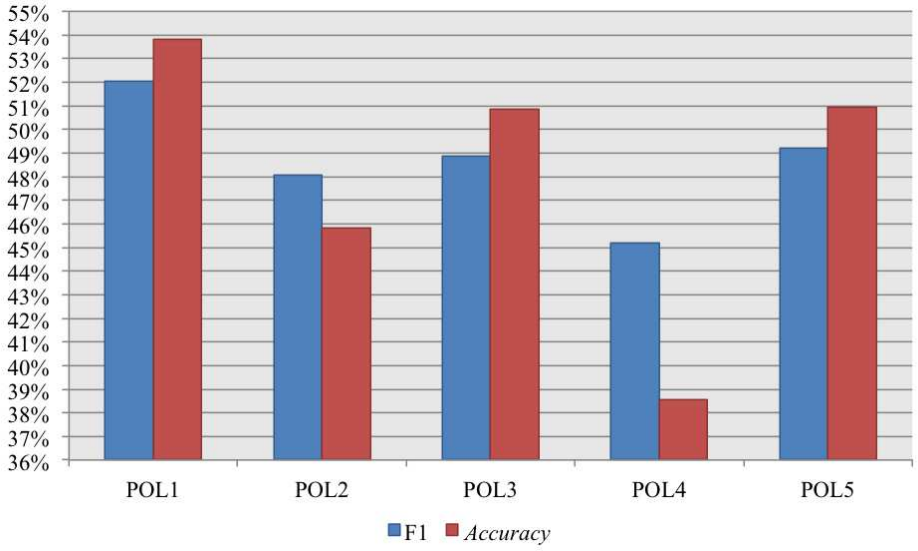


Figura 5.9: F1 y *Accuracy* obtenidos por los módulos de clasificación de la polaridad desarrollados.

los responsables por los que POL1 tiene un mejor comportamiento. Por un lado, en POL1 los tres recursos empleados tienen la misma importancia, dado que para cada término se suma la polaridad positiva y negativa indicada por cada recurso, se restan las dos sumas, y si el resultado es positivo entonces la palabra es positiva, y si es menor que cero, entonces la polaridad de la palabra es negativa. Por otro lado, en POL1 no se minora el valor de polaridad del *tweet* mediante su división entre el número total de palabras, o número de palabras con polaridad. Estas dos afirmaciones llevan a pensar que la combinación de recursos lingüísticos es beneficiosa para la

clasificación de la polaridad, lo cual se estudiará en el Captítulo 7, y que no es adecuado ponderar el valor de polaridad por el número de palabras del texto. Puede ocurrir que un mensaje con una alta carga de opinión, dicha carga esté determinada por un único vocablo o término, mientras que un texto constituido por varias palabras con polaridad puede que no trasmita una opinión muy definida, siendo esto principalmente causado por la manera específica en la que esas palabras se han insertado en el discurso del texto.

	4 clases			**6 clases**		
	Precisión	*Recall*	F1	Precisión	*Recall*	F1
SINAI-EMML	40,9 %	40,9 %	40,9 %	31,4 %	31,4 %	31,4 %
Mejor	68,6 %	68,6 %	68,6 %	61,6 %	61,6 %	61,6 %
Peor	23 %	23 %	12,6 %	12,6 %	12,6 %	12,6 %
Media	52,97 %	52,95 %	52,96 %	43,27 %	43,25 %	43,25 %

Tabla 5.8: Resultados oficiales obtenidos en la edición 2013 del TASS.

La evaluación interna facilitó la elección de la configuración con la que participar en la edición de 2013 de TASS. Los resultados oficiales que se consiguieron con el conjunto de test del Corpus General del TASS, que se debe recordar que está compuesto por 68 798 *tweets*, se muestran en la Tabla 5.8.

El análisis de los resultados evidenciaron tres deficiencias del clasificador que se presentó:

1. **Tratamiento de la negación**: La negación es un fenómeno lingüístico muy importante en AO, dado que puede modificar el significado de una palabra o del mensaje completo que se quiere transmitir a través del *tweet*. El clasificador presentado carece de un módulo que permita la identificación de partículas negativas y su posterior interpretación, por lo que los *tweets* compuestos por palabras con polaridad negadas no se han clasificado de manera correcta.

2. **Clasificación de la subjetividad**: Una de las clases que se debe asignar a los *tweets* es NONE, la cual hace referencia a la ausencia de significado subjetivo o de opinión. El clasificador presentado sólo considera como signo de ausencia de opinión, la no presencia de palabras indicadoras de opinión, y no introduce ninguna característica indicativa de objetividad. Por tanto, el clasificador tiene un alto nivel de errores en la clasificación de la clase NONE.

3. **Clasificación de la ironía**: La ironía también está presente en los *tweets* de la colección de test del corpus General del TASS. Al igual que ocurre con la negación, no incluir ningún módulo responsable de la detección e interpretación de la ironía lleva al clasificador a errar en la clasificación de *tweets* irónicos.

5.6. Conclusión

En el presente capítulo se ha abordado la resolución de la clasificación de la polaridad a través de una metodología de aprendizaje no supervisado. Se ha presentado un método fundado en la expansión del significado de los sentidos presentes en un documento, el cual se ha evaluado tanto en textos largos y cortos, así como escritos en inglés y en español.

Cuando los documentos son de los que aquí hemos venido a llamar largos y están escritos en inglés, los resultados que se han alcanzado demuestran que la inclusión de conceptos relacionados en un texto es fructífero para la determinación de su polaridad. Asimismo, la Tabla de resultados 5.3 exhibe el buen hacer de la ponderación de la polaridad con PageRank, ya que mejora el resultado y reduce las diferencias entre los resultados que devuelve el algoritmo con cada expansión.

Cuando los documentos largos está escritos en español los resultados no son tan positivos, ya que como se ha visto en la Sección 5.4.2, los resultados superan por muy poco los obtenidos con la lista de palabras indicadora de opinión iSOL. Dos pueden ser los posibles motivos de este comportamiento:

1. La elección de los contextos de desambiguación: Los autores de UKB (Agirre & Soroa, 2009; Agirre et al., 2014) recomiendan construir contextos conformados por al menos 20 palabras para la desambiguación. En el trabajo que se ha descrito, se consideraron las oraciones como contextos, independientemente de la longitud de las mismas. Ante esta diferencia, cabe preguntarse ¿habrá influido la construcción de los contextos en la calidad de la desambiguación realizada por UKB? Esta pregunta deberá guiar las experimentaciones que continúen haciéndose con el mismo clasificador sobre textos largos escritos en español.

2. La elección de la base de conocimiento léxica y de opinión: En cuanto a la selección de la base de conocimiento léxica poco se puede hacer, porque MCR es la actual versión de referencia de WordNet para español. En relación al uso de SentiWordNet, cabe preguntarse si SentiWordNet es la base de conocimiento de opinión más apropiada,

de manera que como trabajo futuro se debe enmarcar la evaluación con otras bases de conocimiento de opinión como Q-WordNet (Agerri & García-Serrano, 2010) o WordNet Affect (Strapparava & Valitutti, 2004).

La experimentación llevada a cabo sobre textos cortos ha evidenciado el buen hacer de la expansión del significado cuando el idioma es el inglés. Además, al igual que ocurría con los textos largos, se ha puesto de manifiesto que la ponderación del valor de polaridad con el puntuación de PageRank es beneficioso para el comportamiento del algoritmo de clasificación. Se evaluó la conveniencia de la exclusión de la relación de antonimia del grafo de WordNet, pero los resultados demostraron que su eliminación es una rémora para el comportamiento del método de clasificación.

Por último, la evaluación de textos cortos escritos en español nos dice que la combinación de recursos lingüísticos de opinión es beneficiosa para la clasificación de la polaridad. Esta lectura de los resultados sirvió de acicate para el inicio de una serie de experimentaciones enfocadas a estudiar los beneficios que se pueden obtener de la combinación, no sólo de recursos, sino también de diferentes métodos de clasificación. Dichos experimentos serán el centro de atención del Capítulo 7.

6

Recursos para el Análisis de Opiniones en Español

6.1. Introducción

En el Capítulo 5 se ha evidenciado la necesidad de los clasificadores basados en aprendizaje no supervisado de usar recursos lingüísticos, para adquirir la información necesaria que les permita discernir el nivel de opinión de un texto. Por contra, los métodos basados en aprendizaje supervisado, en ocasiones, se les otorga a priori una capacidad de clasificación superior a la que realmente demuestran. Principalmente, esta creencia engreída encuentra apoyo, por un lado, en la evidencia empírica de que alcanzan mejores resultados que los algoritmos de aprendizaje no supervisado y, por otro, en el convencimiento de que por tener la capacidad de extraer conocimiento de grandes cantidades de datos, no requieren la ayuda de la información que pueden aportar al proceso de clasificación recursos lingüísticos externos. En el contexto de la evaluación de una metodología de adaptación de un clasificador de polaridad de textos en inglés a español, Brooke et al. (2009) ponen de manifiesto que los algoritmos independientes del idioma, como son los algoritmos basados en aprendizaje supervisado, aunque proveen buenos resultados base, siempre se ven mejorados con la incorporación de información dependiente del idioma, dicho de otro modo, de información procedente de recursos lingüísticos externos.

También ha sido resaltado en varias ocasiones que la mayor parte de la investigación en AO presta únicamente atención al inglés, y por ende, la mayoría de los algoritmos y recursos lingüísticos están exclusivamente preparados para el tratamiento del inglés. En el párrafo anterior ha quedado patente el requisito de la disposición de recursos lingüísticos para mejorar la calidad de los métodos de clasificación. En (Brooke et al., 2009) también se enuncia tres estrategias a seguir para emprender la clasificación de opiniones en un idioma distinto al inglés, en ese de opiniones escritas en español, o de adaptar algoritmos preparados para el inglés a otro idioma. La primera recomendación se corresponde con la aplicación de métodos independientes del idioma que requieren escasos recursos lingüísticos. Un ejemplo claro de esta primera estrategia son los métodos basados en aprendizaje automático, ya que su único requisito es la disposición de una colección de documentos en la que tiene que estar identificada la orientación de la opinión del documento. Ejemplos de esta estrategia ya se han visto en esta memoria (Martínez Cámara et al., 2011b,a; Martínez-Cámara et al., 2015). El segundo enfoque recomienda seguir aprovechando los métodos y recursos existentes en inglés, por lo que se hace necesario la traducción de los textos a clasificar a este idioma, para así poder clasificarlos con un algoritmo preparado para el inglés. Ejemplos siguiendo esta segunda

sugerencia se encontrarán en el Capítulo 7. Como tercera estrategia, se propone la generación de recursos en el idioma en cuestión, como tarea previa a la clasificación. Este es el enfoque que siguen los experimentos que se irán progresivamente describiendo en el presente capítulo.

La investigación en AO en español no se caracteriza por una prolija cantidad de recursos lingüísticos de opinión. Esta realidad llevó a generar tanto un corpus de opiniones en el dominio de los establecimientos hoteleros, como a la preparación de una lista de palabras de opinión. El desarrollo de métodos basados en el aprovechamiento de recursos lingüísticos trae consigo el problema del dominio de clasificación. El AO es una tarea que presenta un alto grado de dependencia al dominio de los documentos que se quieren clasificar. Ampliamente es conocido el ejemplo de la palabra "predecible": decir que el comportamiento de un coche es predecible es positivo, mientras que la calificar a una película de predecible es más bien negativo. Por tanto, el desarrollo de metodologías que pretendan introducir un cierto grado de automatismo a la adaptación de los recursos a un dominio específico, también va a ser objeto de este capítulo.

6.2. Generación de recursos

En función del uso que se le dé a un recurso se pueden distinguir, entre los que se emplean para evaluar sistemas y los que constituyen un elemento más del propio sistema de clasificación. Los recursos de evaluación normalmente se corresponden con colecciones de documentos, en las que al menos, los documentos tienen asociada la clase de opinión a la que pertenecen. El lector podría estar pensando que las colecciones de documentos también se utilizan como base de los métodos de aprendizaje automático, y por ende, son parte del sistema. Dejando a un lado la calidad de la colección de documentos y su capacidad de representación del lenguaje, dicho lector debería preguntarse ¿la calidad de la clasificación final depende más de la colección de documentos, o de la capacidad del algoritmo de clasificación de escrudiñar el corpus a través de sus características, y encontrar las diferencias oportunas que permitan la asignación de la clase correcta? En nuestro caso la respuesta es clara, la calidad de la clasificación tiene un mayor grado de dependencia de la capacidad del algoritmo, que de la colección de documentos. Por tanto, las colecciones de documentos etiquetadas, además de ser vitales para la construcción de métodos de clasificación supervisados, son fundamentales para la evaluación de algoritmos supervisados, no supervisados o híbridos.

Por otro lado, se encuentran los recursos que aportan información a un

proceso de clasificación. Por este motivo, a estos recursos se les considera que constituyen una parte esencial del algoritmo de clasificación. En la Sección 5.1, se indica que los principales tipos de recursos en el ámbito del AO que constituyen una fuente de información son las listas de palabras de opinión y las bases de conocimiento léxicas.

En la presente sección se van a describir dos recursos lingüísticos que se han generado para formar parte y evaluar los métodos que se están describiendo en este capítulo. Primeramente se va a presentar una lista de palabras de opinión, iSOL, la cual, como se ha podido leer en el Capítulo 5, se leerá en el actual y en el Capítulo 7, es piedra angular de la mayoría de los métodos de clasificación que se han evaluado. Tras iSOL, se expondrá COAH, un corpus de opiniones en el dominio de establecimientos hoteleros. COAH, además de añadir una colección de opiniones en español al reducido conjunto que existe en la actualidad, está permitiendo la evaluación de los distintos métodos de clasificación que se están describiendo aquí en el dominio hostelero, así como está facilitando el estudio y desarrollo de métodos de adaptación de la clasificación a un dominio concreto.

6.2.1. iSOL

Como bien señala Bing Liu en (Liu, 2012b), en la bibliografía relacionada con AO, se encuentra tres metodologías para la generación de listas de palabras de opinión: manual, basada en diccionario y en corpus. Como el lector se podrá imaginar, la construcción de una lista de palabras manualmente es una tarea bastante tediosa y cara, en cuanto a coste personal y tiempo se refiere. Por tanto, la confección manual de una lista de palabras es siempre una labor que se procura evitar a toda costa. Los métodos automáticos tienen la ventaja de que permiten ahorrar tiempo y horas de trabajo, pero en su contra juega el hecho de que es necesaria la revisión de la lista generada por el sistema automático. Por consiguiente, el enfoque manual se restringe a la validación y corrección de los métodos automáticos.

Los métodos de generación de listas de opinión basados en diccionario son aquellos que aprovechan la información subyacente en un diccionario o base de conocimiento. El método comienza con la selección manual de un conjunto reducido de palabras claramente positivas y negativas. Tomando como referencia ese grupo de términos, se inicia un proceso iterativo de búsqueda de palabras relacionadas en el diccionario que se haya elegido. Como es de suponer, del diccionario o de la base de conocimiento se intentará extraer el máximo provecho de las relaciones de sinonimia y de antonimia. Cada consulta que se realice al diccionario aumentará el tamaño

de la lista de términos, y el proceso se detendrá cuando no se encuentren más palabras que añadir. Por último, la lista compilada iterativamente requerirá de una revisión manual para descartar aquellos términos que no transmitan claramente una opinión. Ésta es la metodología seguida por Hu & Liu (2004), y que dio como resultado la lista de palabras de opinión BLOL[1]: *Bing Liu Opinion Lexicon*, en español, Léxico de Opinión de Bing Liu. El diccionario que Hu & Liu (2004) utilizaron para la generación de la lista de opinión fue WordNet, y las relaciones semánticas que aprovecharon fueron las de sinonimia y antonimia.

Kim & Hovy (2004) también aplican una metodología basada en diccionario para construir una lista de palabras de opinión, y además, prácticamente idéntica a la emprendida en (Hu & Liu, 2004). Pero cuando Kim & Hovy (2004) se encuentran revisando manualmente la lista de términos de opinión construida automáticamente, caen en la cuenta de que existe un conjunto de palabras que aparecen frecuentemente, tanto en el corpus de vocablos positivos como en el de negativos. Algunos de esos términos no tienen polaridad, o simplemente expresan un sentimiento neutro, y otros expresan una determinada polaridad aunque aparezcan en las dos listas. Debido a esta situación, los autores se plantean asignar niveles de intensidad a las palabras de opinión, y en función de si superan un determinado umbral se consideran positivas o negativas. Los autores definen una medida de intensidad fundada en el número de sinónimos positivos o negativos que se encuentran en WordNet de una palabra dada. De esta manera intentan refinar la metodología basada en diccionario y aliviar la revisión manual.

WordNet sigue siendo explotado para la generación de un léxico de opinión en (Hassan & Radev, 2010). En dicho trabajo los autores aplican un algoritmo de camino aleatorio de Markov (*Markov random walk*) sobre un grafo de palabras relacionadas. Para construir el grafo de palabras relacionadas se elabora un subgrafo a partir de la relación de sinonimia e hiperonimia. Los autores definen una medida de distancia entre una palabra, o mejor dicho de un nodo de ese subgrafo, y un conjunto de nodos relacionados. La medida de distancia mide el número de nodos que tiene que recorrer un caminante aleatorio hasta llegar al subconjunto de nodos considerado. Como se podrá imaginar el método divide el subgrafo entre un conjunto de nodos semilla positivos y otro conjunto de nodos semilla

[1]No es fácil encontrar un acrónimo en la bibliografía para referirse a este léxico de opinión, de manera que se debe dejar claro que BLOL no es un acrónimo que se pueda encontrar en los trabajos que emplean este recurso. Se ha decidido emplear un acrónimo formado por las iniciales de uno de sus autores y de las palabras *opinion lexicon*, para facilitar la referencia a este recurso en la memoria.

negativos. Si la palabra en estudio tiene una menor distancia del conjunto positivo entonces se considera como positiva, y en caso contrario como negativa.

Turney & Littman (2003) siguen una filosofía similar a la desarrollada en (Turney, 2002), en donde, no debe olvidarse, se obtiene la orientación de la opinión de un texto como la diferencia de la Información Mutua Puntual, en inglés *Pointwise Mutual Information* (PMI), de las palabras que constituyen el texto que se está analizando. PMI es la medida elegida por Turney & Littman (2003) para medir el grado de asociación de un conjunto de palabras en estudio con un reducido grupo de palabras semilla, cuya orientación semántica está claramente definida. Aquellas palabras cuya diferencia entre el PMI con el conjunto de términos semilla positivos sea mayor que con el de términos negativos serán consideradas como positivas, y en caso contrario serán añadidas al conjunto de palabras negativas.

SentiWordNet (Esuli & Sebastiani, 2006) es otro ejemplo de recurso de opinión generado siguiendo un enfoque basado en diccionario. Sus autores toman como vocablos semilla los definidos en (Turney & Littman, 2003), y aprovechando las relaciones de sinonimia y antonimia de WordNet generan dos conjuntos de *synsets*: positivos y negativos. Para la construcción del conjunto de *synsets* objetivos seleccionan aquellos *sysnsets* que no se han considerado como positivos o negativos, y que sus términos no se encuentran catalogados como positivos o negativos en General Inquierer Stone et al. (1966). Cada *synset* seleccionado es representado por un vector, en el que cada una de sus dimensiones se corresponde con los *unigramas* de su glosa correspondiente, ponderadas por su valor de TF-IDF. Los autores siguen la hipótesis de que tendrán glosas similares aquellos conceptos con semejante orientación semántica. El modelo vectorial definido es el que emplean los autores para alimentar un comité de clasificadores, que son los que determinan la probabilidad de ser positivo, negativo u objetivo de cada un de los *synsets* de WordNet.

En (Maks et al., 2014) se expone un método también basado en el aprovechamiento de WordNet para la generación de listas de opinión. En este caso, los autores no se limitan a producir una lista de opinión en inglés, sino también en español, francés, italiano y holandés.

Los métodos basados en corpus emplean una colección de datos, en lugar de un diccionario, para incrementar un conjunto inicial de vocablos semilla. Hatzivassiloglou & McKeown (1997) son los precursores de aprovechar una colección de opiniones como fuente de información para incrementar el tamaño de una lista de palabras de opinión. Según los autores, los vocablos relacionados por una conectiva, o tienen la misma orientación

semántica, o la opuesta. Verbigracia, si un término semilla está relacionado con otro a través de la conectiva *and*[2] (y) o la conectiva *or* (o) tendrán una misma orientación semántica. Por contra, si esos términos están relacionado por las conectivas *but* (pero), *either-or* (uno o lo otro) o *neither-nor* (ni uno ni otro), entonces se considerarán que tienen una orientación contrapuesta. Hatzivassiloglou & McKeown (1997) denominan a esta asunción "consistencia de opinión", en inglés *sentiment consistency*. Debido a que confiar en exceso en las conectivas puede llevar a error, los autores intentan aminorar el número de equivocaciones mediante la construcción de un grafo que aglutina las palabras consideradas como positivas y negativas. Posteriormente aplican un algoritmo de *clustering*, con el fin de dirimir finalmente los vocablos que son positivos y negativos.

El problema de intentar incrementar la cobertura de un conjunto de palabras de opinión mediante un enfoque basado en corpus radica en que, dentro de un documento, una palabra puede tener desemejantes orientaciones semánticas en función del contexto en el que se encuentre. Por ejemplo, en una revisión sobre un determinado teléfono móvil, se puede decir que su batería tiene una larga duración, lo cual es positivo; mientras que de la cámara de fotos del mismo teléfono se puede manifestar que hay que esperar un largo periodo tiempo hasta que la cámara termina de enfocar, lo cual es negativo. Ésto indica, que el análisis propuesto por Hatzivassiloglou & McKeown (1997) hay que llevarlo a nivel de contexto.

Ding et al. (2008) con el ánimo de tratar de identificar la orientación semántica de las palabras en función del contexto, aplican el método propuesto en (Hatzivassiloglou & McKeown, 1997), pero en lugar de tomar solo los vocablos relacionados por las conectivas, también extraen los aspectos o entidades que acompañan a los vocablos susceptibles de tener orientación semántica. Por tanto, en el ejemplo anterior, la salida del sistema propuesto en (Ding et al., 2008) habría sido, *[duarción_bateria, larga]* y *[cámara_enfoque, larga]*. El lector ya estará pensando que en dicho trabajo se ha dado también solución al problema del AO a nivel de aspectos o entidad, ya que aparentemente se identifican los aspectos que aparecen en un documento y se encuentran las palabras que manifiestan un parecer sobre dichos aspectos. Pero nada más lejos de la realidad, dado que el método es sólo aplicable siempre y cuando exista una lista de aspectos preestablecida.

Otro ejemplo de métodos basados en corpus se encuentra en (Castellucci et al., 2015). Castellucci et al. (2015) presentan un método independiente del idioma basado en un modelo semántico distribuido, en inglés *distri-*

[2]El trabajo que se está describiendo trata opiniones escritas en inglés, por lo que las conectivas que se van mencionar son palabras inglesas.

butional semantics, para la generación de listas de opinión a partir de un conjunto de *tweets*. Los modelos actuales de semántica distribuida se fundamentan principalmente en la extracción de relaciones semánticas de los términos a partir de la coocurrencia de los mismos. Por tanto, el método que se propone, primeramente requiere de un conjunto de documentos sobre los que construir el modelo de semántica distribuida. Para ello, los autores recopilan un conjunto de 20 millones de *tweets* durante los últimos meses de 2014. Una vez que se tienen los datos, debe construirse el modelo distributivo, para lo cual los autores emplean el método conocido como *word2vect* (Mikolov et al., 2013). Dicho método permite representar cada uno de los términos de los 20 millones de *tweets* como vectores de 250 dimensiones o características.

Con el modelo construido todavía no es posible crear el lexicón, ya que es necesario la generación de un modelo de clasificación. Para ello, del conjunto de *tweets* descargados seleccionan aquellos que contienen emoticonos positivos, emoticonos negativos, y los que su último *token* se corresponde con una dirección web. Los *tweets* con emoticonos positivos se toman como positivos, los que tienen emoticonos negativos como negativos, y los que concluyen con una *url* como neutros. Como se puede comprobar los autores emplean el mismo método de etiquetas impuras que (Go et al., 2009) para la generación de un conjunto de *tweets* con etiquetas de opinión. Con este corpus de *tweets* etiquetados y con el modelo distributivo generado anteriormente, se construye un modelo vectorial, que sirve para entrenar un algoritmo de aprendizaje automático supervisado, en concreto SVM. Para generar la lista final de términos, se toman las palabras que conforman el conjunto inicial de 20 millones de *tweets* y se clasifican con el clasificador que se ha construido.

En resumen, la estrategia basada en diccionario es recomendable aplicarla cuando se tiene como objetivo la generación de una lista de palabras de opinión independientes del dominio, ya que es una metodología sencilla y permite de una manera ágil tener una lista de palabras de opinión. En su contra juega el hecho de que los términos resultantes son independientes del dominio, y una opinión es altamente dependiente del dominio. Los métodos basado en corpus requieren de un conjunto de opiniones etiquetadas, lo cual complica en cierta manera el proceso para generar una lista de opinión, o para adaptarla al domino de la colección de opiniones que se está considerando. Por tanto, la puesta en práctica de una estrategia u otra dependerá del problema o del objetivo que se pretenda alcanzar.

Una lista de palabra de opinión está limitada por el número de palabras

que la conforman, lo cual constituye su principal rémora, que no es otra que una condicionada cobertura del lenguaje. Dicho de otra manera, un sistema basado exclusivamente en el uso de una lista de palabras de opinión solamente considerará como palabras portadoras de opinión aquellas que se encuentren en la lista, mientras que el resto de vocablos transmisores de opinión presentes en los documentos que no estén recogidos en la lista, no serán considerados como términos con opinión. Pues, a pesar de este problema evidente, las listas de opinión son recursos bastante útiles, como ya se ha visto principalmente en el Capítulo 5, para la determinación de la orientación de la opinión de un documento, o para la inserción de información en un algoritmo de clasificación, como se podrá ver en el Capítulo 7.

Debido a la valía de una lista de opinión en el ámbito de la clasificación de la opinión y a la escasez de este tipo de recurso en español, al menos durante el desarrollo de la investigación que recoge la presente memoria, se decidió la idoneidad de compilar una lista de palabras de opinión en español. iSOL, *improved Spanish Opinion Lexicon*, en español "Lexicón de Opiniones en Español mejorado", es el nombre por el que se le conoce a la lista, la cual fue presentada en (Molina-González et al., 2013).

Teniendo en cuenta los distintos métodos de generación de listas de opinión, para la elaboración de un lexicón de opinión independiente del contexto se debería haber empleado un diccionario o una base de conocimiento léxica en español, o haber tomado un corpus de opiniones independiente del dominio, lo cual es harto complicado, o una colección de opiniones formada de un número considerable de dominios para que pudiera ser considerado como independiente del dominio. El único diccionario que alberga el rico vocabulario de la lengua española es el ofrecido por la Real Academia Española[3] (RAE). Dicho diccionario es un diccionario en el sentido más estricto del término, es decir, es un conjunto ordenado de vocablos acompañados por la definición de cada una de sus acepciones. En el diccionario de la RAE los términos no se encuentran relacionados, por lo que no es una fuente útil para descubrir palabras semejantes a un conjunto de ellas preestablecidas. Descartada la estrategia basada en diccionario, queda encontrar un corpus lo suficientemente representativo que posibilite la construcción de una lista de palabras de opinión independientes a cualquier dominio. Si se realiza una revisión en el estado del arte de AO en español, se puede comprobar que hay dos corpus que sobresalen en la investigación. Por un lado se encuentra el corpus SMR, del cual ya se ha hablado en el Capítulo 4, y por otro lado la versión española del corpus SFU, el

[3] `http://www.rae.es/recursos/diccionarios/drae`

cual ha sido empleado en las experimentaciones descritas en el Capítulo 5. El corpus SMR es una colección de opiniones circunscritas en el dominio del cine, lo cual lo invalida para la construcción de una lista de opinión independiente del dominio. El corpus SFU, como se ha indicado en el Capítulo 5, está conformado por 8 dominios diferentes, lo cual podría hacer que fuera independiente del dominio. Pero, cuando se contabiliza la cantidad de documentos correspondientes a cada dominio, 50 opiniones por dominio, se llega al convencimiento de que no es lo suficientemente representativo como para elaborar un lexicón de opiniones.

Debido a las razones expuestas en el párrafo anterior, era complicado seguir de una manera ortodoxa la teoría de generación de listas de opinión para AO. Banea et al. (2008) se platean si es posible generar colecciones de documentos, y por extensión recursos lingüísticos, en un idioma distinto al inglés a partir de la traducción automática de recursos que se encuentran en inglés. La experimentación de Banea et al. (2008) demuestra que, al menos, la traducción de un corpus de opiniones en inglés de manera automática a rumano y a español es perfectamente viable, ya que sólo tiene lugar un reducido empeoramiento de los resultados. Los positivos resultados obtenidos por Banea et al. (2008) animaron a buscar una lista de opinión en inglés que sirviera de base para la generación de iSOL. Se tomó como referencia el léxico BLOL, debido a que se encuentra disponible para investigación en la web de su autor, y que es ampliamente conocido en la comunidad investigadora. Como se ha mencionado anteriormente, BLOL es una lista de palabras de opinión que se construyó siguiendo un enfoque basado en diccionario. Los autores tomaron un reducido conjunto de términos de opinión como semilla, y fueron incrementando iterativamente dicho conjunto mediante la inclusión de palabras tomadas de WordNet semejantes a las iniciales. Actualmente[4] BLOL está conformado por 4783 palabras negativas y 2006 palabras positivas.

Una vez elegida la lista de palabras de opinión de referencia en inglés, el siguiente paso fue diseñar el proceso de traducción. La manera más efectiva de emprender la traducción de 6789 palabras es la de emplear un traductor automático de los existentes en la Web. El elegido fue el traductor Reverso[5], debido principalmente a que fue el que menos problemas provocó durante el proceso de traducción. Cada palabra de BLOL se convirtió en una consulta al traductor, el cual devolvía como respuesta varias posibles traducciones. Como heurística se decidió siempre tomar la primera de las traducciones.

[4]Se dice actualmente porque los autores no han cesado de añadir vocablos a BLOL desde el momento que se presentó por primera vez.

[5]`http://www.reverso.net/`

Una vez realizada la traducción automática, se requería de una revisión manual con el fin de depurar el trabajo del traductor. De la lista de palabras resultantes se eliminaron 1068 palabras negativas y 364 palabras positivas porque estaban repetidas. Esto es debido a que varios términos de BLOL se correspondían con la misma palabra en español. En la Tabla 6.1 se recopilan algunos ejemplos de palabras que dieron lugar a la misma traducción.

Palabra en BLOL	**Traducción**
Bogus, disingenuous, dud, false, phony, spurious, untrue	Falso
Castigate, chasten, chastise, penalize, punish	Castigar
Crabby, glum, ill-tempered, moody, peevish, sullen	Malhumorado
Absurd, absurdness, farcical, ludicrous, preposterous	Absurdo
Gaily, jolly, joyfully, joyously, merrily	Alegremente
Beautifully, gloriously, marvelously, splendidly, wonderfully	Maravillosamente
Bright, lustrous, shiny, sparkling, twinkly	Brillante
Affordable, economical, low-cost, low-priced, thrifty	Económico

Tabla 6.1: Ejemplo de palabras de BLOL que tienen un misma traducción en español.

El proceso de traducción también tuvo como resultado un conjunto de palabras no reconocidas por el traductor, y otras tantas, que debido a que no están bien escritas en BLOL, al traductor le resultó imposible encontrar su versión en español. Hay que decir, que las palabras mal escritas, o con faltas de ortografía, fueron añadidas por los autores adrede porque son frecuentes en los textos que se publican en Internet. De la lista de palabras negativas, 435 vocablos fueron eliminados por estas razones, y de las positivas 159. En la Tabla 6.2 se recoge una pequeña muestra de este conjunto de términos.

El resultado de la primera depuración de la traducción de BLOL por parte de Reverso fue un conjunto de palabras de opinión formado por 1483 palabras positivas y 3280 términos negativos, lo cual hace un total de 4763 palabras. A esta primera lista se le asignó la denominación de SOL[6].

La formación de SOL se puede considerar que de manera transitiva ha seguido una estrategia basada en diccionario, dado que BLOL fue

[6] `http://sinai.ujaen.es/sol/`

Palabra mal escrita	**Palabra no reconocida**
Assult	*Bonny*
Goood	*Fav*
Prospros	*Pettifog*
Slooow	*Bumpping*
2-faces	*Jollily*
Danken	*Brainiest*
Jutter	*Prik*

Tabla 6.2: Palabras de BLOL mal escritas y no reconocidas por el traductor automático.

compilada según indica tal metodología. Pero la traducción no estaba del todo depurada, y además, todavía era necesario resolver el problema de las palabras con distintas formas para masculino y femenino, y para singular y plural. Se comenzó revisando la traducción elaborada por Reverso, y se mejoraron algunas de las traducciones realizadas. Una muestra de traducciones corregidas se encuentra en la Tabla 6.3.

Palabra en BLOL	**Traducción de Reverso**	**Corrección**
Brainless	Sin cerebro	Descerebrado
Aimless	*Sin rumbo*	Desorientado
Arrogantly	*Con arrogancia*	Arrogantemente
Deadlock	*Punto muerto*	Estancado
Worthless	*Sin valor*	Devaluado
Fashionable	*A la moda*	Moderno

Tabla 6.3: Muestra de traducciones de Reverso corregidas.

En español la mayoría de los adjetivos cuentan con cuatro formas, una para el masculino singular, otra para el masculino plural, y las dos correspondientes para el femenino. Por consiguiente, cada adjetivo inglés le van a corresponder cuatro palabras en español. La mejor manera de solucionar este problema, que Reverso no resolvía, era incluyendo manualmente las derivaciones de aquellas palabras que lo necesitaran. En la Tabla 6.4 se recogen algunos ejemplos de palabras cuyas derivaciones tuvieron que incluirse manualmente.

Tras añadir las correcciones indicadas, la adición de las derivaciones de genero y número, la inclusión a propósito de palabras con faltas de ortografía e incluso la anexión de palabras claramente con polaridad en

Palabra en BLOL	Derivaciones incluidas
Good	Bueno, buena, buenos, buenas
Famous	Famoso, famosa, famosos, famosas
Pretty	Guapo, guapa, guapos, guapas
Ugly	Feo, fea, feos, feas
Aching	Dolido, dolida, dolidos, dolidas
Bad	Malo, mala, malos, malas

Tabla 6.4: Palabras de BLOL que se corresponden con varios términos en español.

español cuya versión inglesa no está en BLOL, se obtuvo una lista de palabras con 2509 términos positivos y 5626 términos negativos, lo cual hace un total de 8135 palabras. Como ya se ha indicado anteriormente, a esta lista se le asignó el nombre de iSOL[7]. La Tabla 6.5 muestra el tamaño de BLOL, SOL e iSOL para facilitar su comparación.

Lista	Palabras positivas	Palabras negativas	Total
BLOL	2006	4783	6789
SOL	1483	3280	4763
iSOL	2509	5626	8135

Tabla 6.5: Resumen del número de palabras que conforman a cada lista.

6.2.2. COAH

Ante la escasa variedad de colecciones de opiniones en español, se estimó que sería una buena aportación a la comunidad investigadora un nuevo corpus de opiniones. Asimismo, se intentó que la nueva colección de opiniones estuviera enmarcada en un dominio sobre el que no existiera ya un corpus disponible para investigación, y que dicho dominio fuera realmente interesante en un entorno de aplicación. El dominio que reunía esas condiciones es el de los establecimientos hoteleros, o de una manera más simple, el dominio de hoteles[8].

Determinado el dominio, la siguiente elección se centró en la fuente de donde obtener dichas opiniones de establecimientos hoteleros. Actualmente,

[7] `http://sinai.ujaen.es/isol/`

[8] El lector puede estar pensando que el nombre correcto del dominio es turismo, pero realmente no es así. El turismo engloba más aspectos que los hoteles, verbigracia, aglutina a restaurantes, oferta cultural, transporte, ocio...

Booking[9] y TripAdvisor[10] son los dos principales portales de búsqueda y contratación de servicios hosteleros a nivel internacional. La facilidad y simplicidad con la que estos portales de acceso a la información hostelera permiten que una gran cantidad de usuarios satisfagan sus necesidades de información, podría considerarse que es la única razón de su éxito tan espectacular. Ya se comentaba en el Capítulo 1, que el proceso personal de tomar una decisión no carece de complejidad, y que las personas solemos buscar cualquier ayuda para que dicho esfuerzo sea aliviado de alguna manera. La experiencia o la opinión de personas de nuestro entorno de confianza, o simplemente experiencias a las que se les otorgue una mínima credibilidad, es siempre una ayuda a la que asirse para decidirse finalmente por un servicio hostelero u otro. Pues bien, unos de los valores añadidos de los dos portales anteriormente citados, es que han sabido construir una comunidad de usuarios que publican sus experiencias, y por ende, sus opiniones de los hoteles en los que se han alojado. Por consiguiente, esta característica convierte a TripAdvisor y Booking en los portales preferentes de donde extraer las opiniones.

Reducido a dos los posibles orígenes de los datos, cabe preguntarse si emplear ambos o sólo uno de ellos. Los dos portales ofrecen abundantes opiniones, pero ¿cuál de ellos es más fiable? ¿Las opiniones están constituidas por texto, o simplemente por puntuaciones? No es una perogrullada afirmar que interesa una fuente cuyas opiniones sean de usuarios que realmente hayan pernoctado en el hotel, y que esté conformada por un texto cuya longitud permita su análisis e identificación de la posición de su autor con respecto al establecimiento. De los dos portales, Booking es el único que trata de asegurarse que las opiniones que se publican se correspondan con usuarios que sí se han hospedado en el establecimiento que han contratado a través de la plataforma, mediante la invitación a publicar la opinión una vez que ha transcurrido la estancia. Por contra, TripAdvisor no toma esta precaución, por lo que no se puede asegurar que las opiniones que están en TripAdvisor se correspondan con usuarios reales de los hoteles. Parece que esto convierte a Booking en la fuente perfecta, empero, las opiniones de Booking tienen un escaso contenido textual al estar principalmente formadas por puntuaciones de diversas categorías, como pueden ser la limpieza, comodidad, ubicación, servicios, personal, relación calidad precio, y la calidad de la conexión inalámbrica a Internet. Aunque sean más fiables las opiniones de Booking, su naturaleza las inutiliza para la composición de una colección destinada al estudio de técnicas y

[9] `http://www.booking.com/`
[10] `http://www.tripadvisor.es/`

metolodologías de AO. Las opiniones de TripAdvisor, además de estar constituidas por puntuaciones numéricas al igual que Booking, sí cuentan con una longitud suficiente para llegar a ser útiles para el AO. Por tanto, TripAdvisor fue la fuente escogida para la recopilación de opiniones.

Hasta ahora es conocido el dominio, la fuente, y para evitar la confusión del lector se va a explicitar que la lengua en la que deben estar escritas las opiniones es el español. Conocido todo esto, el siguiente elemento a determinar es la procedencia geográfica de los hoteles de los que se van a obtener las opiniones. Se podría haber dejado este elemento al albur del extractor de opiniones en español de TripAdvisor, pero se pretendió buscarle además de un interés investigador, un posible interés empresarial. Andalucía es la región más turística de España, ya que es la que cuenta con una mayor extensión de costa, y está bañada por el océano Atlántico y el mar Mediterráneo. Además, Andalucía es una región rica en parajes naturales, y más abundante aún en patrimonio histórico. Por tanto, no sólo era relevante la elaboración y puesta a disposición de una colección de opiniones en español desde un punto de vista investigador, sino que también era importante para transferir las técnicas de AO en estudio al sector hotelero del sur de España. Ésta fue la razón que llevó a recuperar exclusivamente opiniones en español, publicadas en TripAdvisor sobre hoteles asentados en Andalucía. El resultado de la recopilación de opiniones fue COAH[11] que en inglés significa *Corpus of Opinions of Andalusian Hotels*, y en español Corpus de Opiniones de Hoteles de Andalucía. La Tabla 6.6 recoge todas las características de COAH.

El corpus COAH se presenta en formato XML. Cada documento del corpus está formado por cuatro campos: identificador (`<coah:id>`), puntuación de opinión (`<coah:rank>`), titular de la opinión (`<coah:abstract>`) y la opinión propiamente dicha (`<coah:review>`). La opiniones en TripAdvisor están representadas por una escala de 5 estrellas, de manera que la puntuación de opinión puede ser un valor perteneciente al conjunto [1-5]. Para que el lector se haga una idea de las opiniones recogidas en COAH el extracto Código 6.1 muestra una de las opiniones de COAH.

6.3. Clasificación de la polaridad

El fin de la generación de recursos es su aprovechamiento, en el caso de las listas de palabras de opinión para el desarrollo de sistemas de clasificación de la polaridad, y en el caso de los corpus etiquetados para la evaluación de sistemas de clasificación, o con el uso conjunto de algoritmos

[11] `http://sinai.ujaen.es/coah/`

Número de opiniones	1816
Etiquetas de opinión	Desde 1 (muy negativo) hasta 5 (muy positivo)
Número de unidades lingüísticas	268 715
Número de palabras	236 024
Número de palabras únicas	153 470
Diversidad léxica	0,65023
Número de caracteres	1 372 737
Número de caracteres sin espacios	1 135 306
Número de nombres	55 113
Número de verbos	39 772
Número de adjetivos	19 349
Número de adverbios	16 026
Número de lemas	236 024
Número de lemas únicos	137 907
Diversidad de lemas	0,58429
Número de sentidos	104 047
Número de sentidos únicos	75 902
Número de sentencias	104 047
Longitud media de sentencia	22,88384
Número medio de nombres	0,2335
Número medio de verbos	0,1685
Número medio de adjetivos	0,08197
Número medio de adverbios	0,06789
Número medio de nombres por opinión	30,34856
Número medio de verbos por opinión	21,90088
Número medio de adjetivos por opinión	10 65473
Número medio de adverbios por opinión	8 82488

Tabla 6.6: Características de COAH.

de aprendizaje automático para la construcción de modelos estadísticos. Pues bien, en la presente sección se va a explicar el aprovechamiento de los recursos descritos para la construcción de sistemas de clasificación de la polaridad.

Una manera sencilla de evaluar una lista de palabras de opinión es mediante el desarrollo de un sistema de clasificación de la polaridad a nivel de documento. Eso es precisamente lo que se hizo en (Molina-González et al., 2013) y se va a intentar describir en los siguientes párrafos.

Para la evaluación de iSOL se empleó el corpus de opiniones en español

```
<coah:hotel_review xmlns:coah=``http://sinai.ujaen.es/
    coah''>
        <coah:id>4</coah:id>
        <coah:rank>5</coah:rank>
        <coah:abstract>Por segunda vez, inmejorable</
            coah:abstract>
        <coah:review>Hemos vuelto por segunda vez a
            este estupendo hotel de Granada y nos ha
            vuelto a maravillar: magnífica relación
            calidad-precio, mucha limpieza, y buena
            localización. Hemos aparcado el coche en la
             puerta y hasta la salida no lo hemos
            necesitado, ya que con el autobús llegas al
             centro enseguida. Para hacer la reserva no
             tuve que adelantar nada, y eso que sólo
            era una noche y les mareé con la reserva.
            El trato es el que considero correcto:
            amable y solícito pero sin empalagar. Además, nos tocó una cama de latex en la que
            dormimos como en casa. Si vuelvo a Granada,
             repito.
        </coah:review>
</coah:hotel_review>
```

Código 6.1: Extracto del corpus COAH

Spanish Movie Reviews (SMR) del cual ya se ha hablado en el Capítulo 4. Se desarrolló un sistema de clasificación de la polaridad a nivel de documento para la identificación de dos niveles de opinión, positiva y negativa. SMR es un corpus en el que las opiniones están etiquetadas con cinco niveles distintos de polaridad, por lo que para poder emplearlo en el sistema de clasificación binaria fue necesario la reducción de las cinco clases a dos. Para ello, primeramente se eliminaron 1253 opiniones catalogadas como neutras, dado que el objetivo del sistema es la identificaciones de opiniones claramente orientadas hacia una posición u otra. Asimismo, las opiniones etiquetadas con un valor 5, el cual representa la máxima intensidad de positividad, se consideraron similares a las catalogadas con un valor 4, el cual significa positivo. Con las opiniones clasificadas como negativas, que son las que están asociadas con el valor uno y dos se hizo lo mismo, es decir, se las tomó como negativas. Por tanto, el corpus pasó de tener

3875 opiniones, a tener 2625, de las cuales 1274 son positivas y 1351 son negativas.

Un sistema de clasificación basado en lista de palabras de opinión tiene que estar regido al menos por una regla que determine la orientación de la opinión o del texto en función de las listas de opinión. El comportamiento del clasificador que se elaboró estuvo gobernado por la Ecuación 6.1.

$$\text{opinión(t)} = \begin{cases} positiva & Si \forall p \in t, A = \{p \in iSOL^{+}\}, B = \{p \in iSOL^{-}\}, \\ & |A| \geq |B| \\ negativa & Si \forall p \in t, A = \{p \in iSOL^{+}\}, B = \{p \in iSOL^{-}\}, \\ & |B| > |A| \end{cases} \tag{6.1}$$

donde t es una opinión del corpus, p es una palabra de cualquiera de las opiniones, iSOL^{+} es el conjunto de palabras positivas de iSOL y iSOL^{-} representa al conjunto de palabras negativas de iSOL.

En la Sección 6.2.1 se comentó que iSOL es el resultado de un proceso de construcción de una lista de palabras de opinión siguiendo una estrategia basada en diccionario. Dicho proceso generó como resultado intermedio la lista de palabras de opinión SOL, que también se utilizó para evaluar su potencial de clasificación. La Tabla 6.7 recoge los resultados que se obtuvieron con las dos listas de opinión. Las medidas de evaluación que se han empleado son las ya descritas en el Capítulo 4.

	Macro-Precisión	**Macro-*Recall***	**Macro-F1**	***Accuracy***
SOL	56,15 %	56,00 %	56,07 %	56,23 %
iSOL	62,22 %	61,47 %	61,84 %	61,83 %

Tabla 6.7: Resultados alcanzados con SOL e iSOL sobre el corpus SMR.

Como se puede apreciar la diferencia entre SOL e iSOL es considerable, lo cual demuestra que la revisión manual de la traducción automática de BLOL, así como la inclusión de nuevas palabras con sus respectivas derivaciones, ha resultado ser muy positivo para incrementar el poder de clasificación de la lista de palabras de opinión. Se podría pensar que los resultados no son aceptables, pero hay que tener en cuenta que simplemente se han contado el número de palabras positivas y negativas de iSOL que aparecen en los textos, y no se han incluido en el sistema otras técnicas como pueden ser la interpretación de la negación o de los intensificadores. Otro modo de comprobar que el resultado alcanzado con iSOL es aceptable, es

comparando su comportamiento con el obtenido por otros sistemas sobre el mismo corpus. SMR es un corpus conocido por la comunidad investigadora, por lo que no son escasos los sistemas que se han evaluado empleando SMR como corpus. La Tabla 6.8 muestar la comparación con otros sistemas.

	Tipo de aprendizaje	**Macro-P**	**Macro-R**	**Macro-F1**	***Accuracy***
(Cruz et al., 2008)	No supervisado	N/D	N/D	N/D	69,50%
	Supervisado	N/D	N/D	N/D	77,50%
(del Hoyo et al., 2009)	Híbrido	N/D	N/D	N/D	80,86%
(Malvar-Fernández & Pichel-Campos, 2011)	Supervisado	77,00%	77,00%	77,00%	N/D
(Martínez Cámara et al., 2011b)	Supervisado	86,84%	86,67%	86,75%	86,74%
(Martín-Valdivia et al., 2013)	Híbrido	88,58%	88,57%	88,57%	88,57%
iSOL	No supervisado	62,22%	61,47%	61,84%	61,83%

Tabla 6.8: Comparación de los resultados obtenidos por iSOL con otros sistemas aplicados al corpus SMR.

Como se puede observar, a excepción de la experimentación llevada a cabo por los autores de SMR, ningún clasificador evaluado con SMR realiza un tipo de aprendizaje no supervisado. Por ende, es lógico que el resultado de nuestro sistema sea inferior a los supervisados. Centrando la comparación entre iSOL y (Cruz et al., 2008), hay que indicar que Cruz et al. (2008) emplean un subcorpus aún más reducido que el utilizado en esta evaluación, por lo que la comparación no es del todo fidedigna. Además, hay que destacar que Cruz et al. (2008) desarrollan un sistema similar al expuesto en (Turney, 2002). La complejidad de dicho sistema es mucho mayor que la del aquí descrito, ya que se fundamenta en el uso de un sistema de recuperación de información de documentos web. Por consiguiente, esta primera evaluación de la lista de palabras iSOL es una primera señal de su valía para la identificación de la orientación de la opinión de un texto.

La comparación que se muestra en la Tabla 6.8 se limita a confrontar el rendimiento de diversos sistemas de disímil naturaleza. Dicha comparación manifiesta que iSOL proporciona unos resultados aceptables cuando se le confronta con sistemas supervisados e híbridos, pero ¿proporcionaría iSOL un rendimiento superior que otro sistema idéntico pero que emplee otra lista de opinión en español? Como ya se ha repetido en varias ocasiones durante la memoria, el español no es prolífico en cuanto a recursos para el AO, pero a pesar de ello, es posible encontrar recursos disponibles para la comunidad investigadora. Un ejemplo es la lista de opinión SEL (Sidorov et al., 2013), la cual está conformada por 2036 palabras etiquetadas con una categoría de emoción. El sistema desarrollado para la evaluación de iSOL está preparado para el uso de una lista de opinión con dos tipos de palabras, positivas y negativas. Por tanto, para poder utilizar SEL en el mismo sistema es preciso proyectar el espacio que constituyen las categorías de emociones de SEL en un espacio definido por las dimensiones positivo y negativo. Para ello las categorías de SEL "alegría" y "sorpresa" se tomaron como positivas, y las categorías "miedo", "tristeza", "enfado", y "disgusto" se consideraron como negativas. SEL asigna un valor denominado PFA, el cual se puede interpretar como de probabilidad de pertenencia de la palabra o término a una de las categorías emocionales que tiene en cuenta. Dos han sido los sistemas que se han desarrollado con SEL, el primero, en el que se tienen en cuenta todas las palabras, y un segundo en el que sólo se han considerado aquellas cuyo valor de PFA sea superior a 0,2. La determinación del valor 0,2 fue totalmente arbitraría, ya que la única voluntad era la de intentar escoger solamente aquellos términos que representaran con claridad la categoría emocional a la que pertenecen en SEL. La Tabla 6.9 recoge los resultados alcanzados por iSOL, SEL y la versión de SEL constituido únicamente por la palabras con un valor de PFA superior a 0,2.

	Macro-Precisión	**Macro-*Recall***	**Macro-F1**	***Accuracy***
SEL	52,40 %	51,62 %	52,00 %	52,49 %
SEL (PFA >0,2)	52,56 %	51,81 %	52,18 %	52,64 %
iSOL	62,22 %	61,47 %	61,84 %	61,83 %

Tabla 6.9: Comparación entre iSOL y SEL.

iSOL supera a SEL en el corpus SMR, tanto cuando se emplean todas las palabras, como cuando se intentan reducir aquellas que tienen un valor

reducido de pertenencia a alguna de las categorías emocionales de SEL. Debe remarcase también, que el procedimiento seguido para filtrar en cierta manera SEL ha sido adecuado, ya que mejora levemente los resultados conseguidos cuando se emplean todas las palabras de SEL.

Parece que iSOL, una lista de palabras de opinión independiente del dominio, obtiene unos resultados aceptables en el dominio del cine. Pero para afirmar que iSOL es un recurso adecuado para la construcción de sistemas de clasificación de polaridad en español, es necesario evaluarlo con al menos otro conjunto de opiniones en un dominio diferente. Por consiguiente, una nueva evaluación de iSOL se llevó a cabo con el corpus COAH, que como ya se ha indicado, se trata de un conjunto de opiniones circunscritas en el dominio de hoteles. Esta evaluación se encuentra publicada en (Molina-González et al., 2014). Para la evaluación se empleó el mismo sistema de clasificación que en la evaluación con SMR, alternándose únicamente el conjunto de documentos a clasificar. Los resultados que se obtuvieron se recogen en la Tabla 6.10.

	Macro-Precision	**Macro-*Recall***	**Macro-F1**	***Accuracy***
SOL	84,70 %	75,22 %	79,68 %	82,24 %
iSOL	91,61 %	83,25 %	87,23 %	88,46 %

Tabla 6.10: Resultados alcanzados con SOL e iSOL sobre el corpus COAH.

Como se puede comprobar, con la simple visualización de las Tablas 6.7 y 6.10, sobre un dominio tan distinto del cine, como es el de los establecimientos hoteleros, el sistema de clasificación ha seguido el mismo patrón de comportamiento, es decir, iSOL proporciona un rendimiento mayor que el de SOL. Por ende, se tiene una evidencia adicional de la valía de iSOL para la tarea de AO en español.

Para validar la calidad de iSOL en el dominio del cine se ha comparado con otros sistemas de naturaleza distinta al desarrollado, y con otra lista de palabras de opinión en español. En este caso, es complicado encontrar otro sistema que haya empleado COAH para evaluar su efectividad, debido principalmente a su relativa reciente publicación y puesta a disposición a la comunidad investigadora[12]. Por tanto, con la intención de llevar a cabo una comparación, se ha desarrollado primeramente un sistema basado en aprendizaje automático, y por otro lado se ha utilizado el mismo sistema basado en lista de palabras de opinión, pero empleando la lista SEL como

[12]En el periodo de redacción de la memoria COAH no había sido empleado en una evaluación distinta a la realizada por sus autores.

recurso lingüístico de opinión. El sistema supervisado se ha construido con SVM, dado que es uno de los métodos de aprendizaje automático más empleado en el ámbito del AO. Las opiniones se han representado como vectores de *unigramas*, y TF-IDF se ha utilizado como medida para valorar la relevancia de cada *unigrama* en el conjunto del corpus. En este caso, la lista SEL no ha sido filtrada y se ha utilizado completamente. La Tabla 6.11 muestra la comparación entre los tres sistemas.

	Macro-Precisión	**Macro-*Recall***	**Macro-F1**	***Accuracy***
SEL	81,72 %	69,00 %	74,82 %	78,16 %
iSOL	91,61 %	83,25 %	87,23 %	88,46 %
SVM	95,22 %	93,14 %	94,17 %	94,82 %

Tabla 6.11: Comparación entre SEL, iSOL y SVM sobre el corpus COAH.

De nuevo iSOL proporciona un resultado superior al ofrecido por SEL, constatando en un dominio diferente que es más apropiada que SEL para la identificación de la opinión en textos escritos en español. En cuando a la diferencia con respecto al resultado alcanzado por SVM se puede decir que era de esperar, ya que los algoritmos basados en aprendizaje supervisado suelen obtener unos resultados superiores a los que alcanzan los métodos fundados en aprendizaje no supervisado. En cualquier caso, debe resaltarse que la diferencia no es excesiva, y que incluso es aceptable una cierta perdida de exactitud en la clasificación, si se puede con eso aprovechar una ventaja de los sistemas no supervisados, como es la de no requerir de una colección de datos etiquetada para su aplicación.

6.4. Adaptación al dominio

Si por algo se caracteriza el AO es por su elevada dependencia al dominio sobre el que versan los textos en estudio. Célebre es el ejemplo de la palabra “impredecible”, que inserta en una opinión sobre vehículos es muchas más que negativa, porque a nadie le interesa tener un coche con un comportamiento impredecible, mientras que si el término se encuentra presente en una crítica de cine, es muy probable que la dificultad de intuir el desarrollo del argumento de una película sea considerado como positivo por el autor de la crítica. Por consiguiente, en el contexto del AO es relevante que los sistemas estén preparados para adaptarse a los diferentes dominios sobre los que pueden tratar los textos a clasificar.

Jiang & Zhai (2007) identifica dos necesidades a la hora de desarrollar técnicas de adaptación al dominio en el ámbito del PLN. Por un lado se encuentra el desarrollo de métodos de adaptación de la función de clasificación, ya que características que podían representar una clase en un dominio, en otro pueden estar simbolizando la clase opuesta; y por otro lado, el tener en cuenta la adaptación a las peculiaridades de los ejemplos de cada uno de los dominios, como puede ser la variación en la distribución del vocabulario.

El desarrollo de un sistema de adaptación del dominio no implica tener en cuenta los dos elementos remarcados por Jiang & Zhai (2007), como por ejemplo sí se tienen en Xia et al. (2013), dado que la adaptación de la función de etiquetado, o los módulos del sistema encargados de acomodarse a la naturaleza de los datos puede ser suficiente. Por ende, se ha estudiado el desarrollo de sistemas basados en el uso de lista de palabras de opinión adaptadas a cada dominio en estudio, lo cual se corresponde con la adaptación a las peculiaridades de los ejemplos correspondientes a cada dominio.

En la Sección 6.2.1 se comentaba que tres son los métodos para la generación de listas de palabras de opinión: manual, basado en diccionario y basado en corpus. En la misma sección se indicaba que una desventaja del método basado en diccionario, siempre y cuando el dominio sea un parámetro importante a tener en cuenta en el problema en cuestión, era que su carácter genérico impedía que los sistemas se pudieran adaptar al dominio de las opiniones que se están considerando; mientras que por otro lado, el método basado en corpus tenía como ventaja la facilidad de adaptarse a un dominio, necesitando para ello únicamente de un corpus en el ámbito del dominio que se quería tener en cuenta. Por tanto, la propia definición de los métodos obligó a que la estrategia a seguir para la adaptación de una lista de palabras de opinión fuera la basada en corpus.

El método basado en corpus se fundamenta en el uso de una colección de documentos para extraer términos característicos de la colección, o palabras similares o relacionadas con un conjunto de términos semilla. Si se sigue este proceder para insertar información de un determinado corpus a una lista de palabras de opinión, entonces se consigue adaptar la lista de palabras a las peculiaridades de las instancias o documentos que constituyen el corpus, debido a que lo que se tiene en cuenta es el propio léxico del corpus. Si el resultado de esta metodología es un lista de palabras de opinión adaptada a las características de un corpus, y ese corpus es de un dominio determinado, entonces de manera transitiva se puede manifestar que se ha obtenido una lista de palabras de opinión adaptada al dominio del corpus. Por

consiguiente, siguiendo una estrategia de generación de listas de palabras de opinión se puede adaptar dicha lista a un dominio determinado, y por extensión al sistema que se basa en el uso de la lista, satisfaciendo de esta manera una de las dos necesidades para la adaptación al dominio de sistemas de PLN destacadas por (Jiang & Zhai, 2007).

La teoría parece indicar que se puede adaptar un lexicón de opinión siguiendo un enfoque basado en corpus, por lo que en Molina-González et al. (2014) se comenzó a estudiar la manera más adecuada para adaptar iSOL a diversos dominios. Al seguir un enfoque basado en corpus, es imperante la necesidad de disponer de al menos una colección de datos centrada en una temática específica. Se decidió trabajar con la versión española del corpus SFU, el cual se describió en la Sección 5.4.2, aunque debe recordarse que aglutina opiniones de ocho dominios diferentes: libros, coches, ordenadores, lavadoras, hoteles, películas, música y teléfonos. Al igual que globalmente el corpus está balanceado, también tiene esa propiedad a nivel de dominio, ya que cada uno de los dominios están constituidos por 50 opiniones, de las cuales 25 son positivas, y 25 negativas. Los comentarios originales descargados estaban puntuados por los usuarios en una escala de opinión de 1 a 5. Para la evaluación se requería de un corpus con dos clases, por lo que se tuvo que considerar las opiniones puntuadas con uno y dos puntos como negativas, descartar las que estaban catalogadas con un nivel 3 de opinión, y tomar las marcadas con los valores 4 y 5 como positivas.

Debido a que en español, y por lo menos hasta donde alcanza nuestro conocimiento, no existe una colección de opiniones para cada uno de los dominios que componen el corpus SFU, obliga que éste primer estudio de adaptación de iSOL a un dominio determinado, se tenga que hacer sobre el mismo corpus SFU. Por tanto, para la generación de las versiones de iSOL adaptadas a cada dominio y para la posterior evaluación se ha dividido el corpus en dos, un primer conjunto de 30 opiniones por dominio (15 positivas y 15 negativas) para la generación de las nuevas listas, y 20 opiniones (10 positivas y 10 negativas) se han reservado para la evaluación.

Una vez que se tiene el conjunto de datos sobre los que obtener los términos propios del dominio, la siguiente acción es la de definir el método de selección de los términos característicos de cada dominio. En (Du et al., 2010) se asume que en un documento positivo (negativo) es más probable que esté formado por un mayor número de palabras positivas (negativas) que negativas (positivas), de manera que una palabra con una mayor frecuencia de aparición en documentos positivos (negativos) que negativos (positivos) será más probable que sea positiva (negativa) que negativa (positivas). Tal asunción se tomó como verosímil, y sentó la base del método

de selección de términos a incluir en iSOL. Como primer método de elección de términos, se definió una regla que consideraba una palabra positiva si el cociente de su frecuencia de aparición en todos los documentos de una de las clases del corpus (positiva o negativa) entre su frecuencia de ocurrencia en todos los documentos de la clase opuesta (negativa o positiva) era superior a un determinado umbral, entonces se añadía al subconjunto correspondiente de iSOL. La Ecuación 6.2 representa matemáticamente la regla de selección. Como se puede observar en la ecuación, la regla se basa en la suma de la frecuencia de cada palabra en cada uno de los documentos de una clase, o dicho de otro modo, en la suma de la frecuencia local de cada término, de manera que a esta regla se le denominó "Local".

$$
\begin{aligned}
frec^{+}(w) &= \sum_{\forall\, d \in corpus^{+}} frec(w) \\
frec^{-}(w) &= \sum_{\forall\, d \in corpus^{-}} frec(w) \\
\text{polaridad(w)} &= \begin{cases} positiva & Si\,(frec^{-} = 0 \wedge frec^{+} \geq t) \vee \\ & (\dfrac{frec^{+}}{frec^{-}} \geq t) \\ negativa & Si\,(frec^{+} = 0 \wedge frec^{-} \geq t) \vee \\ & (\dfrac{frec^{-}}{frec^{+}} \geq t) \end{cases}
\end{aligned}
\tag{6.2}
$$

donde t es un umbral para determinar el nivel de apertura del filtro de palabras. Si el valor de t es reducido, entonces pequeña será la diferencia de frecuencia entre positivo y negativo, y por tanto mayor será el número de palabras que se seleccionarán. En cambio, cuanto mayor sea el valor de t, más elevada será la exigencia para elegir una palabra del corpus.

La Ecuación 6.2, a priori, puede conducir a seleccionar términos que realmente no determinan la polaridad de un documento en el dominio en cuestión. Dicho de otro modo, si un término es muy frecuente en un documento de una de las clases del corpus, pero es inexistente en el resto de documentos de la misma clase, cabe preguntarse ¿es ese término verdaderamente representativo de la clase de los documentos a los que pertenece? No es extraño pensar que ese tipo de palabras sesgan la regla local de selección de términos. Con la intención de evitar este tipo de situación se definió una regla similar, pero en este caso, la frecuencia de una palabra en un documento no puede ser superior a uno, de manera que la frecuencia máxima de una palabra en una clase es igual al número

de documentos de dicha clase. Esta manera de considerar la frecuencia proporciona una visión global del número de ocurrencias de una palabra, por lo que se le llamó "Global". Para facilitar la comprensión de esta nueva regla se le recomienda ver la Ecuación 6.3.

$$
\begin{aligned}
frec^{+}(w) &= \sum_{\forall\, d \in corpus^{+} \wedge\, w \in d} 1 \\
frec^{-}(w) &= \sum_{\forall\, d \in corpus^{-} \wedge\, w \in d} 1 \\
\text{polaridad(w)} &= \begin{cases} positiva & Si\,(frec^{-} = 0 \wedge frec^{+} \geq t) \vee \\ & (\dfrac{frec^{+}}{frec^{-}} \geq t) \\ negativa & Si\,(frec^{+} = 0 \wedge frec^{-} \geq t) \vee \\ & (\dfrac{frec^{-}}{frec^{+}} \geq t) \end{cases}
\end{aligned}
\tag{6.3}
$$

El lector ya estará cayendo en la cuenta que las Ecuaciones 6.2 y 6.3 seleccionan palabras que directamente pueden transmitir opinión, y palabras que aparentemente proyectan un hecho objetivo. Y así es, ambas ecuaciones posibilitan la inclusión de términos que explícitamente no manifiestan opinión. Para entender con mayor claridad la afirmación anterior, las Tablas 6.12 y 6.13 muestran algunas palabras que se han seleccionado de cada dominio siguiendo la Ecuación 6.2.

Como puede observarse en las Tablas 6.12 y 6.13 se han seleccionado términos que directamente no transmiten una opinión, pero su uso más frecuente en una clase que en otra puede estar escondiendo la transmisión de una opinión; verbigracia, que en el dominio de los coches la palabra taller aparezca con mayor asiduidad en los textos marcados como negativos es bastante significativo, porque, aunque puede estar debido a un sinfín de causas, el destacar en una opinión que se ha tenido que llevar el coche al taller no es positivo. Por tanto, es probable que la palabra taller esté trasmitiendo de manera implícita una opinión negativa en el domino de los coches. Esto obliga a recordar que en la Sección 2.3.2 del Capítulo 2 se definen las oraciones implícitas como enunciados objetivos que transmiten una opinión regular o comparativa, y que normalmente se corresponde con la descripción de una situación deseable o no deseable. Tomando de nuevo el ejemplo del taller, tener que llevar el coche al taller no es una situación nada deseable.

Las nuevas listas generadas a partir de la inclusión de información de

Palabra	Dominio	Frec. en opiniones positivas	Frec. en opiniones negativas
Consumo	Coches	10	1
Maletero	Coches	6	0
Menú	Hoteles	4	0
Minibar	Hoteles	5	0
Temperatura	Lavadoras	12	1
Capacidad	Lavadoras	7	2
Recuerdos	Libros	10	1
Introducción	Libros	5	1
Conectividad	Teléfonos	6	1
Navegación	Teléfonos	6	0
Ritmos	Música	8	1
Sonidos	Música	8	1
Rendimiento	Ordenadores	13	1
Plataforma	Ordenadores	11	0
Escena	Películas	19	2
Estreno	Películas	5	0

Tabla 6.12: Palabras consideradas como positivas según la Ecuación 6.2.

cada dominio en iSOL reciben el nombre de eSOL[DOMINIO]Local, cuando se emplea la Ecuación 6.2 para la selección de las palabras a incluir, y reciben la denominación de eSOL[DOMINIO]Global, cuando es la Ecuación 6.3 la que gobierna la elección de los términos a añadir. El número de palabras añadidas para cada dominio por cada una de las ecuaciones, y el tamaño nuevo resultante de cada una de las nuevas listas se recogen en las Tablas 6.14 y 6.15.

Para poder seleccionar los términos a añadir y poder construir las nuevas listas de palabras de opinión es obligatorio definir el parámetro t de las fórmulas Local y Global. Aunque lo ideal sería plantear un estudio, de igual manera que se hace con el algoritmo KNN (ver Sección 4.2.4), para determinar qué valor de t posibilita generar unas mejores listas Local y Global. En esta primera evaluación no se ha llevado a cabo tal estudio, eligiéndose 3 como valor de t.

Una vez generada una lista por cada dominio, el siguiente proceso fue la elaboración de un sistema de clasificación de la polaridad similar al que se empleó en la evaluación de iSOL, y que está definido en la Ecuación 6.1. En este caso, se le aplicó una pequeña modificación a la ecuación, la cual consiste en darle mayor relevancia a la clase negativa, de manera que, al

Palabra	Dominio	Frec. en opiniones positivas	Frec. en opiniones negativas
Taller	Coches	2	19
Sensor	Coches	0	5
Manchas	Hoteles	0	4
Moqueta	Hoteles	1	6
Acero	Lavadoras	0	3
Cocina	Lavadoras	1	8
Serie	Libros	2	9
Ritmo	Libros	0	5
Cobertura	Teléfonos	0	8
Carga	Teléfonos	0	5
Remix	Música	1	6
Versiones	Música	0	4
Pantalla	Ordenadores	0	4
Computadora	Ordenadores	0	8
Tráiler	Películas	1	6
Saga	Películas	0	9

Tabla 6.13: Palabras consideradas como negativas según la Ecuación 6.2.

Dominio	Nº palabras positivas	Nº palabras negativas
Coches	28 (2537)	36 (5662)
Hoteles	24 (2533)	15 (5641)
Lavadoras	18 (2527)	22 (5648)
Libros	29 (2538)	36 (5662)
Teléfonos	42 (2551)	36 (5662)
Música	43 (2552)	26 (5652)
Ordenadores	51 (2560)	25 (5651)
Películas	58 (2567)	29 (5655)

Tabla 6.14: Tamaño de las nuevas listas siguiendo la heurística Local.

contrario de lo fijado en la Ecuación 6.1, si el número de términos positivos y negativos presentes en las opiniones es el mismo, entonces el sistema asignará como clase negativo en lugar de positivo.

Antes de mostrar los resultados alcanzados por cada una de las listas generadas, es menester mostrar previamente los resultados obtenidos por iSOL en cada uno de los dominios, dado que el sistema basado en iSOL

Dominio	Nº palabras positivas	Nº palabras negativas
Coches	28 (2537)	34 (5660)
Hoteles	21 (2530)	15 (5641)
Lavadoras	18 (2527)	17 (5643)
Libros	27 (2536)	29 (5655)
Teléfonos	35 (2544)	33 (5659)
Música	37 (2548)	21 (5647)
Ordenadores	51 (2560)	21 (5647)
Películas	39 (2548)	25 (5651)

Tabla 6.15: Tamaño de las nuevas listos siguiendo la heurística Global.

es el que se toma como caso base. La Tabla 6.16 muestra los resultados alcanzados por iSOL.

Dominio	Macro-P	Macro-R	Macro-F1	*Accuracy*
Coches	81,25 %	70,00 %	72,21 %	70,00 %
Hoteles	85,71 %	80,00 %	82,76 %	80,00 %
Lavadoras	56,67 %	55,00 %	55,82 %	55,00 %
Libros	70,83 %	70,00 %	70,41 %	70,00 %
Teléfonos	77,78 %	60,00 %	67,74 %	60,00 %
Música	43,33 %	45,00 %	44,15 %	45,00 %
Ordenadores	56,67 %	55,00 %	55,82 %	55,00 %
Películas	55,49 %	55,00 %	55,25 %	55,00 %

Tabla 6.16: Resultados obtenidos por iSOL sobre cada uno de los dominios.

Como se puede ver en la Tabla 6.16 obtiene buenos resultados en algunos de los dominios, como es el caso de hoteles, coches y libros. Por otro lado, en algunos, como el de música, está por debajo del 50 % de *Accuracy*, lo cual no es nada positivo. Las Tablas 6.17 y 6.18 van a recoger los resultados obtenidos por cada una de las listas adatpadas a cada uno de los dominios, y en los que se espera obtener un mejor resultado que iSOL.

Aunque las tablas recogen todos los resultados, para facilitar su análisis, comprensión y comparación se han aglutinado en una gráfica (ver Figura 6.1). En dicha gráfica es fácil identificar varios comportamientos:

1. En los dominios de libros y teléfonos las dos nuevas listas de palabras de opinión adaptadas a esos dominios no han conseguido mejorar el comportamiento del sistema base. Este comportamiento indica

Dominio	Macro-P	Macro-R	Macro-F1	*Accuracy*	Mejora (F1)
Coches	85,71%	80,00%	82,76%	80,00%	10,04%
Hoteles	85,71%	80,00%	82,76%	80,00%	0,00%
Lavadoras	88,46%	85,00%	86,70%	85,00%	55,31%
Libros	70,83%	70,00%	70,41%	70,00%	0,00%
Teléfonos	77,78%	60,00%	67,74%	60,00%	0,00%
Música	59,80%	55,00%	57,30%	55,00%	29,78%
Ordenadores	23,68%	45,00%	31,03%	45,00%	-44,41%
Películas	50,00%	50,00%	50,00%	50,00%	-9,49%

Tabla 6.17: Resultados obtenidos por las listas eSOL[DOMINIO]Local sobre cada uno de los dominios.

Dominio	Macro-P	Macro-R	Macro-F1	*Accuracy*	Mejora (F1)
Coches	85,71%	80,00%	82,76%	80,00%	10,04%
Hoteles	83,33%	75,00%	78,95%	75,00%	-4,6%
Lavadoras	88,46%	85,00%	86,70%	85,00%	55,31%
Libros	70,83%	70,00%	70,41%	70,00%	0,00%
Teléfonos	77,78%	60,00%	67,74%	60,00%	0,00%
Música	59,80%	55,00%	57,30%	55,00%	29,78%
Ordenadores	59,80%	55,00%	57,30%	55,00%	2,64%
Películas	50,00%	50,00%	50,00%	50,00%	-9,49%

Tabla 6.18: Resultados obtenidos por las listas eSOL[DOMINIO]Global sobre cada uno de los dominios.

primeramente que las palabras insertadas en iSOL no han sido determinantes para corregir los fallos que se cometían con iSOL, de manera que para dichos dominios el sistema de clasificación de la polaridad debe complementarse con otras técnicas, que permitan la extracción de conocimiento de esos textos.

2. En los dominios de coches, lavadoras, música y películas los dos métodos de seleccionar los términos a incluir en iSOL obtienen los mismos resultados. A excepción del dominio de películas, en todos esos dominios se ha mejorado los resultados alcanzados por iSOL. La mejora con respecto al caso base, es una muestra de la validez del método de inclusión de términos dependientes del dominio. Debe ser

destacado el hecho de que en el dominio de las lavadoras la mejora que se ha producido es de un 55,31 %, lo cual es muy significativo. La igualdad de resultados entre las listas es una señal de que las palabras más frecuentes tienen una distribución homogénea entre todos los documentos del subconjunto del corpus que se ha empleado para generar las listas, de manera que el intento de corregir el posible sesgo, que alguna palabra con un gran número de ocurrencias en un documento pueda introducir en las nuevas listas, se queda en intento, porque en dichos dominios, al menos por los resultados que se han obtenido, parece que no se encuentran términos de tal naturaleza.

3. En el dominio de ordenadores es en el único en el que se produce una gran diferencia entre las dos ecuaciones diseñadas para la selección de los términos. Parece que en este caso las palabras que ha añadido la Ecuación 6.2 han generado un mayor número de errores que las propias de iSOL. A falta de un análisis más profundo, es muy probable que se haya producido un sesgo por palabras demasiado frecuentes en un documento y no en el resto. Por tanto, este comportamiento parece indicar la superioridad de la Ecuación 6.3 sobre la Ecuación 6.2.

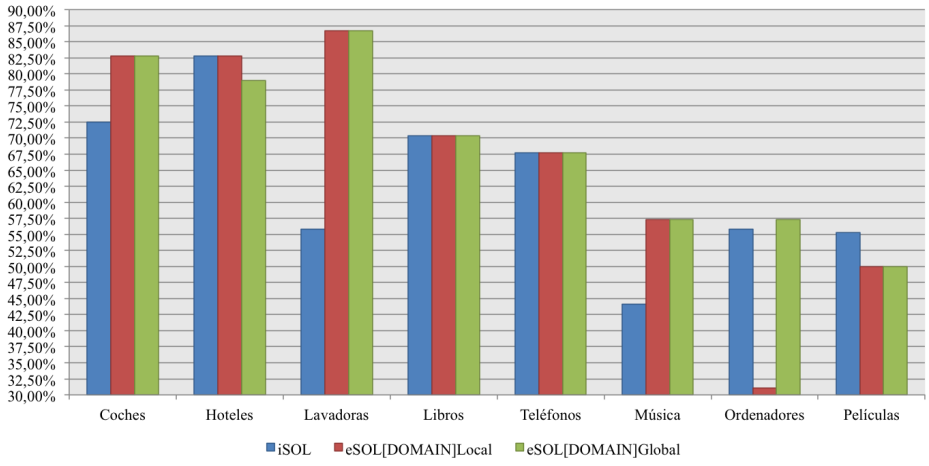


Figura 6.1: Comparación de resultados entre iSOL, eSOL[DOMAIN]Local y eSOL[DOMAIN]Global.

Los comportamientos que se manifiestan en la Figura 6.1, permiten concluir, por un lado, que la inclusión de información del dominio en una lista de palabras de opinión mejora la exactitud del clasificador en

la mayoría de los casos. Por otro lado, en lo que respecta a las ecuaciones 6.2 y 6.3, no existen evidencias suficientes para decantarse por una o por otra. De todas maneras, siempre habrá que tener en cuenta la diversidad del léxico del corpus que se emplea para adaptar la lista al dominio, ya que si es homogéneo en todos los documentos del corpus, puede que convenga más emplear la Ecuación 6.2, mientras que si el vocabulario no es homogéneo entonces será preferible el uso de la Ecuación 6.3, para evitar el sesgo que puedan introducir los términos muy frecuentes en un reducido número de documentos de la colección empleada para introducir información del dominio.

Podría pensarse que la evaluación anterior no es suficiente para validar el método basado en corpus descrito para la adaptación de listas de opinión a un dominio concreto, porque, aunque se han usado subcolecciones diferentes para la adaptación de las listas y para la evaluación de las mismas, se han empleado opiniones de un mismo corpus. Por tanto, aprovechando que se ha generado una lista de palabras adaptadas al dominio de hoteles, y que, como se ha descrito previamente, se cuenta con una colección de opiniones en el dominio hotelero, se ha evaluado la validez de la lista de opinión de hoteles generada con el corpus COAH.

Para la evaluación se tomó la lista de opinión adaptada al dominio de hoteles generada con la Ecuación 6.2 y llamada eSOLHotelesLocal, porque obtuvo mejores resultados que la obtenida con la Ecuación 6.3 (ver Tablas 6.17 y 6.18, y Figura 6.1). Con el objetivo de poder comparar el resultado con el alcanzado por iSOL en COAH, el algoritmo de clasificación estuvo gobernado por la Ecuación 6.1. Los resultados que se obtuvieron se recogen en la Tabla 6.19.

	Macro-P	**Macro-R**	**Macro-F1**	***Accuracy***
iSOL	91,61 %	83,25 %	87,23 %	88,46 %
eSOLHotelLocal	91,59 %	84,31 %	87,80 %	89,05 %

Tabla 6.19: Comparación entre iSOL y eSOLHotelLocal

Como se puede comprobar observando la Tabla 6.19, el léxico de opinión adaptado al dominio mejora levemente la exactitud del sistema. Por consiguiente, se puede concluir que, el método de identificación de términos relevantes de un dominio para incluirlos en una lista de palabras de opinión es válido porque obtiene mejores resultados. A pesar de la positiva conclusión, también debe indicarse que la mejora que se muestra en la Tabla 6.19 es limitada, lo cual es probable que esté debido a que se ha empleado un corpus pequeño, como es SFU para la adaptación al dominio de iSOL.

Por tanto, estos resultados han servido como punto de inicio de una línea de investigación en la tarea de adaptación de métodos de clasificación en un dominio determinado y al desarrollo de técnicas multidominio.

6.5. Conclusión

En el presente capítulo se ha descrito un resultado relevante de la investigación que se ha llevado a cabo, y que no es otro que la generación de recursos. iSOL y COAH son dos recursos que han aumentado el reducido catálogo de recursos en español para el AO. Asimismo, deben destacarse los buenos resultados alcanzados con iSOL, que ha sido evaluado con tres corpus distintos: SMR, COAH y SFU. Ese buen comportamiento demuestra que es posible mediante traducción automática la construcción de recursos en español, lo cual es un complemento al trabajo de (Banea et al., 2008), dado que en dicho estudio se manifiesta que, una estrategia posible para la generación de corpus de opiniones en español se corresponde con la traducción colecciones de opiniones que se encuentran en inglés a español.

Varias veces a lo largo de la memoria se ha indicado el alto grado de dependencia de la tarea de AO al dominio en el que se desarrollan los documentos, por lo que la adaptación de los sistemas de clasificación al dominio es fundamental para que puedan realizar correctamente su cometido. En el presente capítulo se ha descrito un método basado en corpus para la adaptación al dominio de iSOL. La evaluación se ha realizado sobre 8 dominios diferentes, y en la mayoría de ellos, como se ha podido ver, se han mejorado los resultados. Ésto nos indica el acierto en la definición del método de adaptación, lo cual no es óbice para que actualmente se siga trabajando en mejorar el método de selección de términos propios del dominio en iSOL.

7

Combinación de Clasificadores

7.1. Introducción

En los capítulo anteriores se han descrito diversos métodos para la identificación de la polaridad tanto en inglés como en español. Entre esos métodos se distingue entre los que siguen una estrategia de aprendizaje supervisado y no supervisado, así como los que se basan principalmente en el aprovechamiento de la información que aportan recursos lingüísticos destinados al AO. Cada uno de los métodos tienen sus ventajas e inconvenientes, por lo que cabe preguntarse si sería posible su uso combinado para mejorar la exactitud de los sistemas individuales.

El buen hacer de la combinación de clasificadores en el ámbito de la minería de datos ya ha sido enunciada en (Dietterich, 2000) y (Rokach, 2005). Los resultados positivos alcanzados por métodos de combinación de clasificadores en minería de datos, llevaron a los investigadores del campo del PLN a comprobar si la utilidad de dichos métodos también se veía reflejada en las tareas propias del PLN. De esa curiosidad surgió el sistema que se describe en Belkin et al. (1993), el cual trata de resolver el problema que se plantea en la tarea de recuperación de información mediante la combinación de diferentes formulaciones booleanas de consultas sobre una base de datos de documentos. Los buenos resultados que se obtuvieron sobre el conjunto de datos de la competición de referencia en recuperación de información TREC, sirvieron como señal para que los investigadores de PLN comenzaran a considerar la combinación de clasificadores como una alternativa más para mejorar la exactitud de los sistemas. A partir de ese momento se empezaron a publicar descripciones de experimentaciones basadas en clasificadores combinados, que trataban de proponer una solución al problema del análisis sintáctico (Henderson & Brill, 1999), desambigüación (Pedersen, 2000) y reconocimiento de entidades sobre textos en inglés (Florian, 2002) y en español (Enríquez de Salamanca Ros, 2011).

El AO es otra tarea del PLN que no debe quedar al margen del estudio del uso de métodos de combinación de clasificadores para mejorar la exactitud de los sistemas individuales. Por ello, el presente capítulo va a estar centrado en la experimentación que se ha llevado a cabo para incrementar el potencial de los clasificadores de polaridad de textos en español, mediante su adecuada combinación. Sin embargo, antes de describir dichos métodos, la Sección 7.2 va a realizar una función propedéutica, para que el lector tenga un conocimiento básico de la teoría relacionada con la combinación de clasificadores.

7.2. Combinación de clasificadores

El área del aprendizaje automático ha dado abundantes métodos que permiten inferir información y conocimiento a partir de un conjunto de datos con una exactitud considerable. Aunque no ha cesado el estudio para la definición de nuevos métodos de clasificación, desde hace bastante tiempo se ha tenido en cuenta la combinación de los métodos de clasificación existentes para intentar mejorar la calidad de la información colegida. En (Buhlmann & Yu, 2003) se fija el año 1977, y más concretamente el trabajo (Tukey, 1977), como el momento a partir del cual se inicia el estudio de métodos de clasificación con la mirada puesta en mejorar los resultados que proporcionan los métodos de clasificación por separado. Esa mejora de los resultados se ha visto validada empíricamente en varios estudios relacionados con el aprendizaje automático (Domingos, 1996; Quinlan, 1996; Bauer & Kohavi, 1999). Aunque la experiencia parece indicar que es provechoso combinar diversos métodos de clasificación, es recomendable, primeramente, repasar las condiciones que se tienen que dar para que la combinación pueda reducir el error de los métodos de clasificación base, y, posteriormente, las razones por las que la combinación tiene una alta probabilidad de enriquecer el resultado final.

Según Hansen & Salamon (1990) existen dos condiciones necesarias y suficientes para que el uso combinado de varios métodos de clasificación proporcionen una mayor calidad de clasificación que la que alcanzan los métodos de clasificación por separado, independientemente del clasificador que se utilice:

- **Diversidad**: Se refiere a que los clasificadores base que componen el sistema combinado comentan errores diferentes a la hora de clasificar un nuevo ejemplo. Dicho de una manera más simple, que cuando un clasificador yerre, el resto acierte. El fundamento de la combinación es que los distintos clasificadores aporten puntos de vista desemejantes del mismo problema, lo cual se puede conseguir mediante el uso de diferentes subconjuntos de características por parte de cada uno de los clasificadores, la utilización de distintas subcolecciones del conjunto de datos, o el empleo de diferentes métodos de clasificación. Si la diversidad es una condición suficiente y necesaria, se podría pensar que si existe una gran diversidad entre varios métodos de inferencia, entonces está asegurado el incremento de la exactitud de la clasificación en un grado proporcional a la diversidad que hay entre los clasificadores base.

  El pensamiento anterior lleva a preguntarse si existen medidas que

cuantifiquen la diversidad entre varios métodos de clasificación. En (Kuncheva & Whitaker, 2003) se realiza un profundo repaso de las distintas medidas estadísticas que se pueden utilizar para llevar a cabo un estudio del nivel de diversidad que existe entre varios métodos de clasificación. Al mismo tiempo, los autores realizan un estudio sobre si es cierta la correlación entre diversidad y disminución de los errores de clasificación. Las conclusiones de ese estudio no son muy positivas, ya que se pone de manifiesto que no existen evidencias claras sobre la correlación entre diversidad y exactitud de los métodos combinados de clasificación. A pesar de ello, no desacosejan el cálculo de la diversidad, ya que la existencia de diversidad no asegura matemáticamente un incremento al mismo nivel de la exactitud del sistema combinado, pero la no existencia de la misma sí que es un indicativo de que no se vaya a producir una mejora. Por ello, recomiendan el uso de alguna medida de la diversidad, y en especial el estadístico Q (Yule, 1900).

- **Precisión**: Los clasificadores base deben proporcionar una tasa de error inferior a la del clasificador aleatorio. En el caso de que no sea así, al clasificador combinado le será muy complicado superar el resultado del mejor clasificador base, por no decir casi imposible, debido a que no se dispone de información suficiente para extraer beneficios a través de la combinación.

Una vez conocidos los requisitos que deben cumplir un conjunto de clasificadores para formar parte de un sistema combinado, es momento de responder a la pregunta ¿por qué construir un sistema de clasificadores combinados? Dietterich (2000) es quien trata de responder a esa pregunta exponiendo tres razones:

1. **Estadística**: Desde el punto de vista estadístico la selección de un clasificador entraña un riesgo importante, porque a priori es complicado saber si va a tener un correcto proceder con los datos de entrenamiento. Por consiguiente, la combinación de clasificadores va a mitigar dicho riesgo, ya que se van a combinar diversos tratamientos de los datos que llevan a hipótesis distintas. La esperanza de los métodos de combinación radica en que la hipótesis resultante del uso conjunto de los clasificadores base se acerque más al objetivo.

2. **Computacional**: No es raro que los métodos de inferencia alcancen un máximo local durante el proceso de búsqueda del objetivo. La combinación de varios clasificadores puede ayudar a que los

clasificadores base salgan de su estancamiento en los máximos locales y se acerquen al máximo global.

3. **Representación**: Puede darse la situación de que la solución a la que se pretende llegar no se encuentre en el espacio de búsqueda de los clasificadores base. Esta tesitura es otra razón más para combinar los métodos de clasificación, con el fin de ampliar su espacio de búsqueda y así poder aproximarse al objetivo de la clasificación.

A las razones de Dietterich hay que añadirle la enunciada por (Freund et al., 2001), la cual remarca que por medio de la combinación de métodos de clasificación se puede reducir el sobreentranimiento.

7.2.1. Tipología

No son pocos los métodos disponibles que permiten la combinación de todos los elementos involucrados en un proceso de clasificación y, además, su naturaleza es bastante dispar. Por consiguiente, no es fácil encontrar un criterio para agrupar las diversas metodologías existentes en el estado del arte. En (Rokach, 2005) se expone una categorización fundada en la relación existente entre los clasificadores involucrados en la combinación. Dicha relación se ve influenciada por varios factores, lo cuales se van a detallar a continuación:

1. **Interrelación entre clasificadores**: En función de la manera en la que la acción de un clasificador influye en la de otro, se puede distinguir entre métodos de combinación secuenciales y concurrentes.

2. **Método de combinación**: La manera en la que se intentan conjuntar los distintos clasificadores también determina el sistema resultante. Estos métodos van desde la simple agrupación de las salidas de los clasificadores base, a métodos más sofisticados basados en aprendizaje automático, como puede ser el *stacking*.

3. **Generación de diversidad**: Como ya se ha indicado, el comportamiento no homogéneo de los clasificadores es un indicador del posible éxito del clasificador final. En un proceso de clasificación combinado se puede generar diversidad mediante la clasificación de distintas partes de la colección de datos, a través de la selección de distintos conjuntos de características, o variando los parámetros de los métodos de clasificación.

4. **Tamaño de la combinación**: El número de clasificadores involucrados es otro factor que determina el comportamiento del sistema de combinación.

Combinación secuencial

La combinación secuencial se refiere al proceso iterativo de clasificación en el que un algoritmo de inferencia toma como ventaja el conocimiento generado en las iteraciones anteriores. Un ejemplo de combinación secuencial lo constituye la técnica de *Boosting* (Schapire, 1990). Este método ejecuta repetidas veces un algoritmo de clasificación sobre distintas muestras del conjunto de datos de entrenamiento, y combina los clasificadores resultante en uno solo con un error inferior a los clasificadores generados en cada una de las iteraciones.

La implementación más conocida del método *boosting* es el algoritmo AdaBoost (Freund & Schapire, 1995). El fundamento de AdaBoost es ejecutar iterativamente un mismo algoritmo sobre un conjunto de datos ponderado en función de la dificultad de clasificación de cada uno de los ejemplos que lo componen. En la primera iteración, todos los ejemplos del conjunto de datos tienen asignado el mismo peso. Como es lógico, el proceso de clasificación de la primera iteración devuelve un conjunto de ejemplos bien clasificados, y otra colección de elementos cuya clase no ha sido identificada correctamente. Si un conjunto de datos no ha sido clasificado correctamente, puede ser debido a que por su naturaleza presente una mayor resistencia al clasificador y, por tanto, requiera de una atención especial por parte del clasificador. En la segunda iteración, el clasificador recibe el mismo conjunto de datos, pero en esta ocasión, los ejemplos que fueron mal clasificados en la primera iteración tendrán una mayor importancia que los que sí se clasificaron correctamente. El incremento de la importancia de los elementos mal clasificados y la reducción de los que son bien inferidos se irá repitiendo durante todo el proceso, con el objetivo de que el clasificador centre todos sus esfuerzos en descubrir la clase de los ejemplos más díscolos.

Los métodos basados en *Boosting* incrementan el resultado del sistema final por dos razones principalmente:

1. El clasificador final genera unos mejores resultados en el conjunto de entrenamiento que los clasificadores individuales que se han construido en el proceso iterativo.

2. La varianza del clasificador final es significativamente menor que la de los clasificadores que se han ido generando.

Pero el *Boosting* también tiene sus inconvenientes. Como bien señala (Quinlan, 1996), los algoritmos basados en *boosting* pueden originar como resultado un clasificador sobreentranado. Un número elevado de iteraciones puede llevar a que el algoritmo resultante esté sobreajustado a los datos de entrenamiento, de manera que sea muy probable que produzca un mayor número de errores que el clasificador base. Otro inconveniente de los métodos basados en *boosting*, aunque de menor importancia, es el inferior nivel de comprensibilidad del sistema final con respecto al clasificador base. A pesar de estos dos inconvenientes, Breiman (1996) cataloga al *boosting* como el avance más significativo en el campo de la clasificación automática en la década de los 90 del siglo XX.

Combinación concurrente

La combinación concurrente se asienta en la división del conjunto de datos original en varios subconjuntos, llevando a cabo un muestreo con o sin reemplanzamiento. Cada subconjunto se utiliza para el entrenamiento de distintos clasificadores, cuyo resultado final debe ser combinado. La combinación concurrente no impone ninguna restricción relacionada con el uso de un mismo método de inferencia para cada subconjunto generado, ni para el procedimiento de combinación final.

El ejemplo más representativo de combinación concurrente es el algoritmo Bagging (Breiman, 1994) (*bootstrap aggregation*). Este método de clasificación está diseñado principalmente para algoritmos de inferencia inestables, es decir, para clasificadores cuyo rendimiento varía considerablemente cuando se aplican pequeñas modificaciones a los datos de entrenamiento. Bagging mitiga dicha inestabilidad ejecutando el algoritmo sobre distintos muestreos con reemplazamiento del conjunto de datos original. La salida final del sistema se corresponde con la clase mayoritaria entre las devueltas por cada una de las ejecuciones, o dicho de otra manera, la combinación se articula a través de un sistema de voto mayoritario.

Métodos de combinación

La combinación secuencial y concurrente se fundamentan en la mejora del resultado de clasificación a través de la combinación de las ejecuciones del clasificador sobre distintas versiones del conjunto de datos original. Por contra, en la presente sección se va a tratar de describir sucintamente cómo combinar varios clasificadores.

Dentro de los métodos de combinación se encuentran métodos simples de composición de múltiples clasificadores y los metaclasificadores. Los

métodos de combinación simples están destinados a conjuntar algoritmos de clasificación que obtienen unos resultados comparables cuando se ejecutan sobre un mismo conjunto de datos. Los métodos de combinación simple se suelen caracterizar por sufrir en la clasificación de ejemplos atípicos (*outliers*). Los metaclasificadores, a pesar de adolecer de los problemas del aprendizaje agregado, como es el sobreentrenamiento y el ser costoso en relación al tiempo de entrenamiento, presentan una mayor capacidad de clasificación que los métodos de combinación simples.

Más adelante se describirá el estudio que se ha llevado a cabo en relación al uso de métodos de combinación en la tarea de clasificación de la polaridad sobre textos en español. En dichas experimentaciones se ha elegido un sistema de voto como representante de los métodos de combinación simple, y *stacking* como método de metaclasificación. Por este motivo, estos dos métodos se van a describir con algo de más profusión.

Clasificación por voto

El voto es una herramienta que utilizan las sociedades o un conjunto de ciudadanos para medir el nivel de consenso existente sobre una determinada situación, y que se emplea para la toma de decisiones. Un ejemplo de uso del voto es la democracia, la cual es un sistema de organización social que atribuye la titularidad del poder a los ciudadanos. Platón definió democracia como el gobierno de la multitud o gobierno de la mayoría, de manera que las decisiones que se tomaran en un Estado o en una sociedad estuvieran regidas por la voluntad de la mayoría. Este esquema de toma de decisiones que tiene lugar en una sociedad, se puede trasladar al contexto de un conjunto de clasificadores. Cada clasificador expone su preferencia sobre una determinada clase, al igual que un ciudadano expresa mediante el voto su predilección por un partido político. La decisión final del clasificador se corresponderá con la clase que tenga un mayor número de apoyos por parte de los clasificadores base, al igual que el partido gobernante se corresponde con aquel que ha obtenido un mayor número de votos.

Según se recoge en (Kuncheva, 2004), se distinguen tres modelos de consenso basado en voto: unanimidad, mayoría simple y pluralidad. El voto por unanimidad exige que todos los clasificadores base coincidan en su decisión, mientras que un sistema fundado en el esquema de mayoría simple sólo obliga a que la mitad más uno de los clasificadores se inclinen por la misma clase. Los sistemas basados en voto mayoritario asignan como clase final aquella que más votos haya obtenido por parte de los clasificadores base, sin la exigencia de que el número de clasificadores tenga que ser el de la mitad más uno. Desde una perspectiva matemática, un sistema de

voto constituido por L clasificadores, que tiene que asignar a una colección de datos una clase entre un conjunto de c clases posibles, la salida de cada clasificador base se puede caracterizar como un vector c-dimensional binario $[d_{i,1},...,d_{i,c}]^T \in \{0,1\}^c$, $i = 1,...,L$, donde $d_{i,j} = 1$ si el clasificador base D_i ha acertado en la clasificación. Por consiguiente, el funcionamiento de un sistema de voto estará regido por la fórmula 7.1.

$$\sum_{i=1}^{L} d_{i,k} = \max_{j=1}^{c} \sum_{i=1}^{L} d_{i,j} \tag{7.1}$$

Los sistemas de voto por mayoría pueden tener como resultado un empate, por lo que los sistemas deben prevenirse mediante la definición de un mecanismo de resolución de dichos empates. Con la intención de evitar los empates Xu et al. (1992) propone un voto mayoritario dirigido por umbrales. Primeramente el método considera una clase más, que se la asigna a todos aquellos ejemplos que han originado un empate. La ecuación que gobierna la decisión final del sistema de voto propuesto es:

$$clase = \begin{cases} w_k & \text{Si } \sum_{i=1}^{L} d_{i,k} \geq \alpha \cdot L, \\ w_{c+1} & \text{en otro caso} \end{cases} \tag{7.2}$$

donde $w_k \in c$ y w_{c+1} es la clase que se ha añadido al conjunto de clases , y que está destinada para los datos que dan lugar a empate. Aparentemente el método no aporta mucho, porque simplemente cataloga a los ejemplos complicados para los clasificadores con una nueva clase impostada. Pero la novedad radica en la *parametrización* del límite que marca la mayoría de votos, de manera que si una clase supera el umbral $\alpha \cdot L$, entonces dicha clase es la elegida, pero si ninguna clase sobrepasa el umbral entonces al ejemplo se le asigna la clase w_{c+1}. El valor de α oscila entre 0 y 1 ($0 < \alpha \leq 1$), y suele ser igual a $\frac{1}{2} + \varepsilon$, definiéndose en el intervalo $0 < \varepsilon < \frac{1}{L}$. Una situación particular de la Ecuación 7.2 tiene lugar cuando $\alpha = 1$, ya que en ese caso se correspondería a un sistema de votación por unanimidad.

Según la Ecuación 7.1, un sistema basado en la regla de voto por mayoría acertará, siempre y cuando $\lfloor \frac{L}{2} \rfloor + 1$ clasificadores acierten en la clasificación. Por tanto, la probabilidad de que el sistema de voto acierte ($P_{mayoritario}$) en la predicción se puede determinar según la ecuación:

$$P_{mayoritario} = \sum_{m=\lfloor \frac{L}{2} \rfloor + 1}^{L} \binom{L}{m} p^m (1-p)^{L-m} \tag{7.3}$$

donde p se corresponde con la probabilidad de éxito de los clasificadores base.

Según el teorema de Condorcet Jury (1785) (Shapley & Grofman, 1984), si $p > 0,5$, entonces el sistema de voto tendrá una probabilidad mayor de éxito:

1. Si $p > 0,5$, entonces $P_{mayoritario}$ en la Ecuación 7.3 es monótamente creciente:

$$P_{mayoritario} \to 1 \,\text{siempre y cuando}\, L \to \infty \tag{7.4}$$

2. Si $p < 0,5$, entonces $P_{mayoritario}$ en la Ecuación 7.3 es monótamente decreciente:

$$P_{mayoritario} \to 0 \,\text{siempre y cuando}\, L \to \infty \tag{7.5}$$

3. Si $p = 0,5$, entonces $P_{mayoritario} = 0,5$ para cualquier valor de L.

Metaclasificación

Los sistemas de voto deciden la clase a la que pertenece un ejemplo mediante la aplicación de una regla social, o dicho de otra manera, se otorga aquella clase que más clasificadores hayan considerado que es la apropiada para el ejemplo en estudio. Si mediante la aplicación de una regla social se puede inferir nuevo conocimiento a partir de la generalización realizada por un conjunto de clasificadores base, cabe preguntarse, si a esa clasificación base se le podría aplicar otro método de generalización que diera lugar a una inferencia superior. Con esa duda comienza Wolpert (1992) su artículo donde presenta su metodología de metaclasificación, *Stacked Generalization*, conocida como Stacking.

De una manera simple, el método de Stacking toma la salida de un conjunto de clasificadores, y constituye con ellas las características de entrada de un nuevo clasificador, el cual será el responsable de determinar la clase a la que pertenece el objeto en estudio. Desde una perspectiva más formal, el proceder que se debe seguir para preparar un metaclasificador de Stacking es el siguiente:

1. Tomar un conjunto de entrenamiento Z y un conjunto de clasificadores base D.

2. Generar k particiones del conjunto de datos Z, y para cada clasificador D_j, con $j = 1, \ldots, n$ y cada partición Z_i con $i=1, \ldots, n$:

   a) Entrenar el clasificador D_j con los datos que resultan de extraer de Z la partición Z_i (Z-Z_i).

   b) Una vez entrenado el clasificador D_j, ejecutarlo sobre el subconjunto Z_i.

3. El proceso anterior genera como resultado una nueva versión del conjunto de datos Z, que se llamará Z^s, y que estará conformado por el mismo número de ejemplos que Z, pero sus características serán las salidas de los clasificadores pertenecientes al conjunto D. Entrenar el metaclasificador con el conjunto de datos Z^s.

4. Para obtener el clasificador final, se tienen que entrenar los clasificadores base con todos los datos de la colección Z con el fin de obtener la versión definitiva de los clasificadores base.

Una vez preparado el clasificador, su *modus operandis* sería:

1. Dado un ejemplo x_i, aplicarle cada uno de los clasificadores base.

2. Generar las nuevas características de x_i a partir de las salidas de los clasificadores base ($x_i{}^s$).

3. Procesar $x_i{}^s$ con el metaclasificador y obenter la clase final.

7.3. Experimentación sobre textos largos

Llegado a este punto del capítulo, ya se conocen los fundamentos esenciales de los métodos de combinación de clasificadores. Por tanto, ya es momento de desarrollar la descripción de la investigación que se ha llevado a cabo con métodos de combinación de clasificadores, con el objetivo de continuar mejorando la precisión de los sistemas de clasificación de la polaridad. Pero los sistemas de clasificación de la polaridad que más perentoriamente necesitan de una mejora, son los que tratan sobre textos distintos a la lengua inglesa, debido, a lo que ya se ha repetido en varias ocasiones en la presente memoria, a la escasez de recursos lingüísticos. Por consiguiente, ante el potencial de clasificación que ofrecen los métodos de combinación, ¿podrían utilizarse clasificadores especializados en inglés para aumentar la capacidad de clasificación de métodos de inferencia centrados en otras lenguas?

La respuesta a la pregunta planteada se va a articular mediante la descripción de la investigación que se ha desarrollado en el ámbito de la clasificación de opiniones en dos idiomas distintos al inglés: árabe y español.

7.3.1. Textos no escritos en español

La primera tentativa de estudiar la idoneidad de los métodos de combinación de clasificadores, vino impulsada por la necesidad de mejorar la exactitud de la clasificación de la opinión en árabe. El árabe, a pesar de ser una lengua hablada por un número muy elevado de personas en el mundo, no ha sido objeto de estudio por parte de los investigadores de PLN hasta hace pocos años. El reciente nacimiento del interés investigador por el árabe, tiene como consecuencia la escasez de recursos lingüísticos, lo cual dificulta la investigación en este idioma. En el contexto de la investigación de la identificación de la orientación de la opinión, como ya se ha remarcado en varias ocasiones, es muy importante el aprovechamiento de la información que aportan los recursos lingüísticos para la identificación de la orientación de la opinión recogida en un texto. Por ende, la situación actual del árabe en la investigación en PLN, invita a estudiar si la combinación de clasificadores de la polaridad específicos para árabe con clasificadores especializados en inglés, que sí tienen la opción de servirse de recursos lingüísticos, es beneficioso para mejorar la exactitud de la clasificación de la opinión en árabe. Éste fue el objetivo de la experimentación publicada en (Perea-Ortega et al., 2013).

Con la mirada puesta en dar respuesta a la pregunta expuesta al inicio de la sección 7.3, se platearon dos sistemas de combinación por voto mayoritario, que se diferenciaban en el número de clasificadores base. Una primera combinación constituida por dos algoritmos de clasificación basados en aprendizaje automático, uno centrado en opiniones en árabe y otro en opiniones escritas en inglés; y otro conformado por tres métodos de inferencia base, los clasificadores anteriores más un clasificador de opiniones en inglés basado en el uso de un recurso lingüístico. Para llevar acabo la experimentación, era necesario contar con un corpus paralelo en árabe e inglés, así como de un recurso lingüístico de opinión en inglés. El corpus paralelo que se empleó fue OCA-EVOCA, que se describirá en el siguiente párrafo, y el recurso lingüístico de opinión en inglés que se usó fue SentiWordNet.

El corpus paralelo OCA-EVOCA está conformado por los *corpora* en árabe OCA[1] y su traducción automática al inglés EVOCA[2]. El corpus OCA está constituido por un conjunto de críticas de cine, que fueron descargadas de 15 sitios web especializados en la publicación de opiniones en lengua árabe de películas de cine. Se trata de un corpus balanceado formado por 250 opiniones positivas y 250 opiniones negativas. En (Rushdi-Saleh et al.,

[1] `http://sinai.ujaen.es/oca-corpus/`

[2] `http://sinai.ujaen.es/evoca-corpus/`

2011b) se encuentra una descripción mucho más profunda del proceso de construcción de OCA.

EVOCA (*English Version of OCA*) es la versión inglesa de OCA, la cual se generó a través del uso del traductor automático PROMPT-Online[3]. La traducción automática tuvo que salvar tanto problemas técnicos como lingüísticos. Los técnicos estuvieron relacionados con la limitación a 500 caracteres de cada petición de traducción. Los lingüísticos se centraron en la falta de correspondencia, en ocasiones, de la categoría morfosintáctica entre una palabra árabe y su traducción al inglés, así como en la correcta traducción de las oraciones irónicas y sarcásticas. Todos esos problemas se lograron superar, encontrándose una descripción pormenorizada de cada uno de ellos en el artículo de presentación de EVOCA (Rushdi-Saleh et al., 2011a).

Tres fueron los clasificadores base que se construyeron, dos basados en aprendizaje automático, y uno en el uso de una base de conocimiento (SentiWordNet). El algoritmo de aprendizaje automático seleccionado para los dos idiomas fue SVM, que como ya se ha indicado en viarias ocasiones ofrece un buen rendimiento en problemas de clasificación de la polaridad. La aplicación de un algoritmo de aprendizaje automático no consiste simplemente en tomar un texto y ejecutar el algoritmo, sino que precisa de una preparación y representación del texto. Se llevó a cabo un estudio profundo de la conveniencia, tanto en árabe como en inglés, de la eliminación de términos no representativos (*stopwords*), y de la aplicación de un *stemmer*. Las dos operaciones anteriores tienen como fin la de representar de una manera más clara la información subyacente en un texto, pero ello también puede conseguirse por medio del filtrado de los términos según un determinado criterio. Por tanto, se estudió los beneficios del filtrado por longitud de palabra, es decir, no se consideraron las palabras que tuvieran una longitud inferior a 4 caracteres.

También fue objeto de estudio la representación de los documentos, analizándose si convenía más emplear un conjunto de *unigramas*, *bigramas* o *trigramas*. Los algoritmos de aprendizaje automático requieren que las características, a través de las cuales se representa los objetos a clasificar, tengan asociadas una medida que valore la relevancia de la misma. En el Capítulo 4 ya se definieron 4 medidas de relevancia de las características que pueden representar a un texto, y de las cuales se han evaluado dos en esta experimentación: TF-IDF y TF.

Aunque el estudio completo se encuentra recogido en (Rushdi-Saleh et al., 2011b,a), en las Tablas 7.1 y 7.2 se van a mostrar los resultados

[3] `http://translation2.paralink.com/`

más relevantes obtenidos con el algoritmo SVM, estando los *n-gramas* ponderados por la métrica TF-IDF, ya que superó en todos los experimentos a TF.

Stopper	***Stemmer***	***n-gramas***	**Precisión**	***Recall***	***Accuracy***	**F1**
✓		1	86,99 %	94,80 %	90,20 %	90,73 %
✓	✓	1	86,14 %	88,00 %	86,80 %	87,06 %
✓		2	87,38 %	95,20 %	90,60 %	91,22 %
✓	✓	2	86,85 %	90,80 %	88,40 %	88,78 %
✓		3	86,55 %	96,40 %	90,65 %	91,22 %
✓	✓	3	87,21 %	91,20 %	88,80 %	89,16 %

Tabla 7.1: Estudio de configuración SVM con TF-IDF sobre el corpus OCA.

Antes de resaltar cual es la configuración de los algoritmos que han llevado a obtener el mejor resultado, debe ser destacada la reducida diferencia entre el mejor resultado alcanzado sobre OCA (91,22 % de F1 y 90,65 % de *Accuracy*) y el obtenido sobre EVOCA (90,87 % de F1 y 90,60 % de *Accuracy*). Este hecho es una validación más que se une a la original de (Banea et al., 2008) de que la traducción automática es una técnica a la que se puede recurrir para generar recursos en el ámbito del AO. Dicho lo cual, esos mejores resultados se obtuvieron, en el caso de OCA, eliminando las *stopwords*, no aplicando *stemmer* y empleando *trigramas* como característica de representación de las opiniones en árabe. Cuando el corpus es EVOCA, el mejor resultado se alcanza cuando no se eliminan las *stopwords*, no se aplica *stemming* y las opiniones se representan como un conjunto de *bigramas*.

Una vez comprobado el buen hacer de los dos algoritmos basados en aprendizaje automático, queda por conocer cómo calcula la polaridad el tercer clasificador base. Este tercer método de inferencia desarrolla un aprendizaje no supervisado, ya que se fundamenta en el cálculo de la orientación semántica de las palabras de los documentos. Para el cálculo de la orientación semántica de las palabras, el método de inferencia utiliza SentiWordNet. Pero SentiWordNet por sí solo no es capaz de identificar la orientación semántica de un documento, sino que requiere de una regla que combine adecuadamente las puntuaciones de polaridad que asigna a las palabras. Para este caso se ha utilizado la regla propuesta por Denecke (2008), la cual consiste en calcular tres valores de polaridad para cada documento. Dichos valores de polaridad se corresponden con los mismos que proporciona SentiWordNet: positivo, negativo y objetivo. El método de

Stop-per	*Stem-mer*	Longi-tud >3	*n-gramas*	Precisión	*Recall*	*Accuracy*	F1
			1	87,98%	92,40%	89,80%	90,06%
		✓	1	89,00%	92,00%	90,20%	90,39%
	✓		1	88,78%	90,40%	89,40%	89,48%
	✓	✓	1	88,20%	88,40%	88,20%	88,23%
✓			1	88,01%	90,00%	88,80%	88,82%
✓		✓	1	87,39%	90,40%	88,60%	88,75%
✓	✓		1	88,31%	86,80%	87,40%	87,31%
✓	✓	✓	1	86,89%	88,40%	87,40%	87,53%
			2	88,10%	94,00%	90,60%	90,87%
		✓	2	88,27%	90,80%	89,20%	89,40%
	✓		2	89,04%	92,00%	90,20%	90,39%
	✓	✓	2	89,51%	89,60%	89,40%	89,44%
✓			2	87,94%	89,20%	88,40%	88,40%
✓		✓	2	87,80%	90,40%	88,80%	88,93%
✓	✓		2	88,46%	88,40%	88,20%	88,26%
✓	✓	✓	2	88,12%	89,20%	88,40%	88,50%
			3	85,62%	94,40%	89,00%	89,59%
		✓	3	88,56%	92,00%	89,80%	90,01%
	✓		3	88,21%	93,20%	90,00%	90,35%
	✓	✓	3	88,74%	90,80%	89,40%	89,57%
✓			3	88,22%	90,00%	88,80%	88,91%
✓		✓	3	88,48%	91,60%	89,60%	89,83%
✓	✓		3	88,56%	88,40%	88,20%	88,25%
✓	✓	✓	3	89,32%	89,60%	89,20%	89,28%

Tabla 7.2: Estudio de configuración SVM con TF-IDF sobre el corpus EVOCA.

generación es muy sencillo, ya que consiste en calcular la media aritmética de las puntuaciones de positivo, negativo y objetivo, de todas aquellas palabras del documento que se está procesando en SentiWordNet. La clase de polaridad del documento se corresponderá con aquel valor de polaridad cuya media sea mayor. Pero ¿cómo se obtiene el valor de positivo, negativo y objetivo de cada palabra? Como es sabido, SentiWordNet no representa la polaridad de palabras, sino de los sentidos con los que se suelen utilizar las palabras, por lo que Denecke (2008) obtiene la media aritmética del valor de positivo, negativo y objetivo de todos los sentidos de la palabra en cuestión.

SentiWordNet solamente incluye los sentidos de las palabras que se corresponden con las cuatro categorías morfológicas: nombre, verbo, adjetivo y adverbio. Denecke (2008) tiene en cuenta en su experimentación

todos los sentidos de las palabras independientemente de la función morfológica que estén desarrollando dentro de una oración. En nuestro caso, se llevó a cabo un estudio de las categorías morfológicas que más podrían aportar al proceso de clasificación de la polaridad. Los resultados que se obtuvieron sobre EVOCA teniendo en cuenta distintas combinaciones de categorías morfológicas se muestran en la Tabla 7.3.

Categorías morfosintácticas	**Precisión**	***Recall***	**F1**
nombres	51,79%	81,20%	63,24%
adjetivos	56,57%	79,20%	66,00%
verbos	52,70%	86,00%	65,35%
adverbios	57,89%	17,60%	26,99%
adjetivos+nombres	55,35%	84,80%	66,98%
adjetivos+verbos	54,86%	83,60%	66,24%
adjetivos+adverbios	72,19%	48,80%	58,23%
nombres+verbos	52,73%	88,80%	66,17%
nombres+adverbios	66,21%	38,40%	48,61%
verbos+adverbios	62,07%	28,80%	39,34%
adjetivos+nombres+verbos	53,47%	86,40%	66,06%
adjetivos+nombres+adverbios	69,30%	59,60%	64,09%
nombres+verbos+adverbios	64,76%	54,40%	59,13%
adjetivos+verbos+adverbios	67,82%	5480%	60,62%
adjetivos+nombres+verbos+adverbios	64,48%	66,00%	65,74%

Tabla 7.3: Configuración del clasificador base fundado en el uso de SentiWordNet.

Como se puede apreciar en la tabla de resultados, el mejor comportamiento del clasificador se ha obtenido cuando se ha tenido en cuenta la polaridad de los nombres y los adjetivos de los documentos. Muy cerca de esta configuración se han quedado las que utiliza sólo adjetivos y verbos; y nombres y verbos. Por tanto, parece que en este caso los adverbios no son de mucha ayuda a la hora de identificar la polaridad. Si se compara con los resultados alcanzados por los otros dos clasificadores base, hay que destacar que existe una diferencia considerable de resultados, pero ésto era esperable, dado que los otros dos clasificadores son métodos que siguen un enfoque supervisado y éste último es no supervisado.

Una vez que se ha comprobado que los tres clasificadores base obtienen un resultado al que obtendría un método aleatorio (>50%), se puede emprender la combinación. La combinación se ha articulado mediante un sistema de voto mayoritario. Se han evaluado las siguientes combinaciones:

1. Voto_ocasvm_evocasvm: Se trata de un combinación de los dos clasificadores basados en aprendizaje automático. Al ser el número de clasificadores base par, es posible que la combinación dé lugar a empates. Para la resolución de los empates se han evaluado dos reglas distintas:

   a) Voto_ocasvm_evocasvm_neg: En esta combinación se le otorga una mayor preferencia a la clase negativa, ya que solamente se asigna la clase positivo a la opinión en estudio, si y sólo si los dos clasificadores base han considerado la opinión como positiva.

   b) Voto_ocasvm_evocasvm_pos: En este caso es la clase positiva la que tiene preferencia, porque siempre y cuando un clasificador base considere que una opinión es positiva, entonces el sistema de voto clasifica la opinión como positiva.

2. Voto_ocasvm_evocasvm_evocaswn: Combinación de los tres clasificadores base que se han descrito.

Una vez descrito el método de combinación de los clasificadores base, ya sólo queda mostrar los resultados que han alcanzado cada una de las combinaciones para comprobar si se ha mejorado el resultado de clasificación. La Tabla 7.4 recoge los resultados que se han obtenido.

Combinación	**Precisión**	***Recall***	**F1**
Voto_ocasvm_evocasmvneg	89,84%	92,00%	90,91%
Voto_ocasvm_evocasmvpos	84,83%	98,40%	91,11%
Voto_ocasvm_evocasmv_-evocaswn	85,66%	98,00%	91,42%

Tabla 7.4: Configuración del clasificador base fundado en el uso de SentiWordNet.

La Figura 7.1 permite de una manera gráfica comparar los resultados obtenidos por los clasificadores base y por las combinaciones evaluadas. Tanto en la tabla de resultados como en el gráfico se puede apreciar los buenos resultados que obtiene SVM tanto con el corpus OCA como con el corpus EVOCA. Pero esos buenos resultados que tienen individualmente no se transforman en una mejoría cuando se utilizan de manera conjunta. Ésto es un indicativo de que el comportamiento de ambos clasificadores es bastante homogéneo, por lo que no es posible un enriquecimiento mutuo y, como consecuencia, una mejora global de la calidad de la clasificación. Lo que sí ha tenido lugar es una mitigación de la perdida de información

que se produce tras la traducción automática, ya que la combinación obtiene mejores resultados que el clasificador base EVOCA_SVM. Pero la mejora global tiene lugar cuando se introduce en la combinación el algoritmo que aprovecha el conocimiento atesorado en SentiWordNet, que a pesar de sus no tan buenos resultados, introduce en la combinación la diversidad necesaria para que tenga lugar el repunte en la exactitud de la clasificación que se produce en el sistema de voto Voto_ocasvm_-evocasmv_evocaswn. Esta mejora, aunque mínima, es un indicativo que la combinación de clasificadores es un recurso que se puede emplear para mejorar la clasificación de la polaridad en lenguas, como en este caso el árabe, que no se caracterizan por una gran cantidad de recursos lingüísticos.

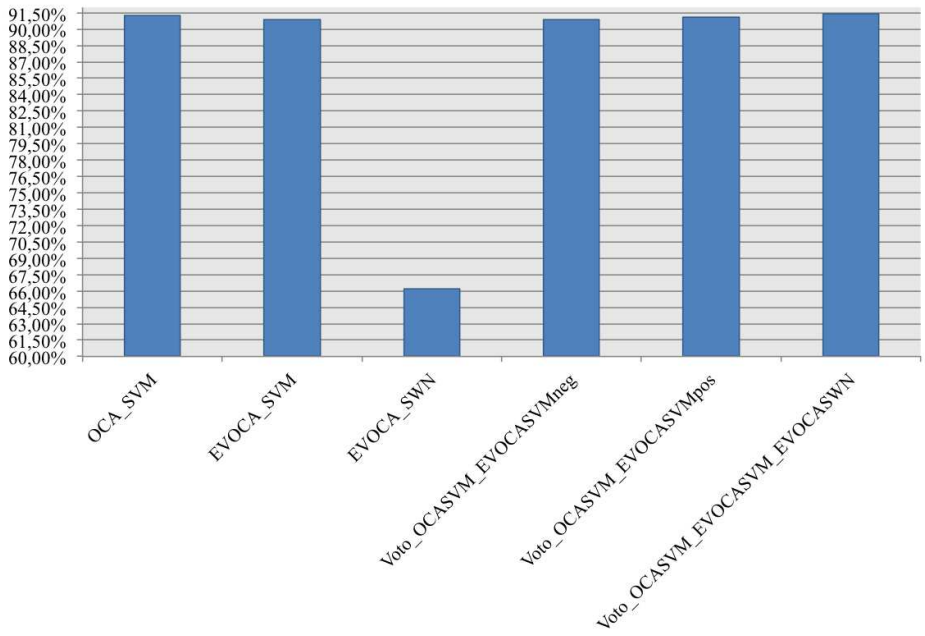


Figura 7.1: Comparación del valor de F1 alcanzado por los clasificadores base y los sistemas de voto.

La conclusión de esta experimentación debe ir unida a lo descrito en la Sección 7.2, es decir, que aunque no se puede decir que exista una correlación directa entre diversidad y mejora de la exactitud de la inferencia, sí es una condición necesaria para que se produzca. En la experimentación se ha visto como la combinación de dos clasificadores muy buenos, pero al parecer con resultados homogéneos, no ha mejorado al mejor resultado de

los dos. Cuando se ha insertado en la combinación un clasificador que sigue una filosofía totalmente distinta, el sistema de voto ha conseguido mejorar al clasificador base más exacto, lo cual indica que dicho clasificador sí ha introducido la diversidad necesaria para que el sistema de voto haya podido aumentar el espacio de búsqueda de la solución, y así poder encontrar la clase adecuada de aquellos ejemplos que los clasificadores base no han sabido clasificar correctamente.

7.3.2. Textos escritos en español

El principal objetivo de la investigación que se describe en esta memoria es el de proporcionar métodos de clasificación de la polaridad para el idioma de Cervantes. Viendo el resultado positivo que ha tenido la aplicación de un sistema de voto, que combinaba un clasificador de polaridad especializado en árabe con dos clasificadores de polaridad en inglés, ¿por qué no intentar los mismo en español? Como se ha visto en la sección 7.2, los métodos de combinación no se limitan a aplicar la regla social del voto para combinar la inferencia de varios métodos de clasificación, por lo que para el español también se ha estudiado el uso de un método de metaclasificación, en concreto el Stacking. En los subsiguientes párrafos se va a detallar la experimentación que se ha realizado en el ámbito de la combinación de clasificadores de polaridad sobre textos en español, que se ha publicado en (Martín-Valdivia et al., 2013) y (Martínez-Cámara et al., 2014b).

La principal razón que dio pie a realizar sobre español una experimentación similar a la llevada a cabo con árabe, es que el español se encuentra en una situación similar al árabe en cuanto a la escasez de recursos lingüísticos se refiere. Por tanto en (Martín-Valdivia et al., 2013) se trató de mejorar el sistema de clasificación de la polaridad supervisado descrito en la sección 4.3 y publicado en (Martínez Cámara et al., 2011a), con la incorporación en el proceso de clasificación de un clasificador de la polaridad supervisado y otro no supervisado de opiniones escritas en inglés.

Si el lector recuerda, la experimentación descrita en la Sección 4.3 se realizó sobre la colección de opiniones de cine SMR. Si la intención es la de incorporar información procedente de recursos lingüísticos en lengua inglesa, entonces se convierte en perentoria la necesidad de disponer de una colección paralela a SMR en inglés. Para alcanzar tal objetivo, se emprendió la traducción automática de SMR. La traducción se realizó empleando la API de traducción de Microsoft[4], que es la base de su producto de traducción Bing translate[5]. El resultado de la traducción fue el corpus

[4] `https://www.microsoft.com/translator/api.aspx`

[5] `https://www.bing.com/translator/`

MCE[6] (MuchoCine en inglés)[7].

La experimentación que se desarrolló sobre SMR en la Sección 4.3 no se realizó sobre el corpus completo, ya que no se tuvieron en cuenta las opiniones marcadas con una valoración 3, al considerarse que representan una opinión neutra. También a modo de recordatorio decir, que las opiniones catalogadas con un valor de opinión 1 y 2 fueron consideradas como negativas y las marcadas con 4 y 5 fueron tomadas como positivas. Con MCE se siguió la misma regla, por lo que al igual que con SMR, la experimentación con MCE se llevó a cabo con 1274 opiniones positivas y 1351 negativas. Observando las Tablas 4.3, 4.4 y 4.5 de la Sección 4.3 se puede comprobar que el algoritmo de clasificación más adecuado para descubrir la orientación de las opiniones de SMR es SVM, y que la mejor configuración de la clasificación está conformada por la representación de las opiniones como un vector de *unigramas*, al que se le han eliminado las palabras carentes de un significado representativo o *stopwords*, no se le ha aplicado *stemming* y la relevancia de esos *unigramas* se ha medido empleando TF-IDF. Para árabe se ha visto que la diferencia entre la clasificación de la versión original del corpus y la traducida es mínima, ¿ocurrirá lo mismo con opiniones escritas en español? La única manera de saberlo es replicando el mismo estudio que se llevó a cabo con SMR, pero en esta ocasión con MCE. En esta ocasión se va a aprovechar la experiencia adquirida con SMR, y sólo se va a analizar el comportamiento de SVM ante el uso de *stopper*, *stemming* y diferentes medidas de la importancia de los *unigramas* (TF-IDF, TF, TO, BTO). La Tabla 7.5 recoge los resultados obtenidos por SVM sobre el corpus MCE.

Observando las Tablas 4.4 y 7.5 se puede observar el mismo comportamiento que desarrolló la experimentación con árabe, es decir, una perdida limitada de exactitud en la clasificación de la colección de opiniones traducidas. Además, en español la coincidencia es algo mayor que en árabe, dado que tanto en SMR como con MCE los mejores resultados se han alcanzado con la misma configuración del clasificador, es decir, cuando no se utiliza ni *stopper*, ni *stemmer* y la medida de la importancia de los *unigramas* es TF-

[6] `http://sinai.ujaen.es/mce-corpus/`

[7] A modo aclaratorio, cuando se comenzó a trabajar con el corpus SMR, éste no tenía un nombre, por lo que se empleaba la denominación de la fuente, MuchoCine, como nombre del corpus. En las publicaciones que se corresponden con esta memoria, el corpus MuchoCine se encuentra referenciado por medio del acrónimo MC. Si la versión española era MC, el acrónimo directo para su versión inglesa fue MCE. En el momento que se comenzó a escribir la presente memoria el autor de MC le asignó un nombre, que como el lector ya sabe es *Spanish Movie Reviews* (SMR). Aunque para la versión española se ha preferido utilizar la nueva denominación, para la versión inglesa, al ser un desarrollo propio, se prefiere continuar con el nombre original de MuchoCine English (MCE).

	Stop	**Stem**	**Precisión**	***Recall***	**F1**	***Accuracy***
TF-IDF	✓	✓	85,02%	84,97%	84,99%	84,89%
	✓		87,04%	86,93%	86,98%	86,97%
		✓	85,86%	85,82%	85,84%	85,83%
			87,76%	87,69%	87,22%	87,69%
TF	✓	✓	84,23%	84,15%	84,19%	84,19%
	✓		85,54%	85,40%	85,47%	85,44%
		✓	78,53%	78,45%	78,49%	78,47%
			78,09%	78,03%	78,06%	78,06%
TO	✓	✓	84,63%	84,60%	84,61%	84,61%
	✓		84,97%	84,89%	84,93%	84,91%
		✓	76,51%	76,47%	76,49%	76,42%
			74,83%	74,74%	74,78%	74,67%
BTO	✓	✓	82,76%	82,67%	82,71%	82,70%
	✓		84,62%	84,51%	84,56%	84,53%
		✓	82,81%	82,73%	82,27%	82,74%
			84,42%	84,30%	84,36%	84,30%

Tabla 7.5: Resultados obtenidos por SVM sobre el corpus MCE.

IDF. En este caso, al igual que se hizo con la experimentación descrita en la Sección 4.3 se va a tomar el segundo mejor resultado porque la diferencia con el mejor es mínima, y al utilizarse *stopper* el número de características con las que tiene que trabajar el clasificador es menor, consiguiéndose así el incremento de la eficiencia de la clasificación.

El tercer clasificador base es un método de clasificación no supervisado, que al igual que en el caso de la experimentación con árabe, se fundamenta en el aprovechamiento de SentiWordNet para identificar la orientación de la opinión. En este caso también se utiliza la regla definida por Denecke (2008) para calcular el valor de polaridad de los documentos. En esta ocasión, también se ha llevado a cabo un estudio de qué combinación de categorías morfológicas aportan más al proceso de clasificación. Los resultados del tercer clasificador base sobre el corpus MCE se muestran en la Tabla 7.6.

La clasificación con SentiWordNet de opiniones en español traducidas al inglés sigue un patrón de comportamiento similar al de la traducción inglesa de opiniones árabes. En el caso de la traducción de textos en español, la combinación de categorías morfológicas que ha obtenido mejores resultados ha sido la de adjetivos y verbos, mientras que en el caso de

Categorías morfológicas	**Precisión**	***Recall***	**F1**	***Accuracy***
nombres	52,31%	82,81%	64,11%	55,01%
adjetivos	57,26%	78,96%	66,38%	61,18%
verbos	52,63%	89,40%	66,26%	55,81%
adverbios	56,17%	25,35%	34,94%	54,17%
adjetivos+nombres	56,46%	86,81%	68,42%	61,10%
adjetivos+verbos	56,69%	87,44%	68,79%	61,49%
adjetivos+adverbios	62,51%	57,06%	59,66%	62,55%
nombres+verbos	51,91%	89,80%	65,78%	54,67%
nombres+adverbios	57,81%	50,55%	53,94%	58,10%
verbos+adverbios	57,28%	41,68%	48,25%	56,61%
adjetivos+nombres+ verbos	55,20%	90,35%	68,53%	59,73%
adjetivos+nombres+ adverbios	61,61%	70,17%	65,61%	64,30%
nombres+verbos+ adverbios	58,60%	62,32%	60,40%	60,34%
adjetivos+verbos+ adverbios	62,25%	67,19%	64,63%	64,30%
adjetivos+nombres+ verbos+adverbios	61,01%	77,00%	68,08%	64,95%

Tabla 7.6: Configuración del clasificador base fundado en el uso de SentiWordNet.

la traducción de textos árabes, el mejor resultado se alcanzó cuando se consideraron conjuntamente adjetivos y nombres. Pero si se observa con más detenimiento los resultados, se puede comprobar que la combinación de adjetivos y nombres alcanza unos resultados similares a los proporcionados por la combinación de adjetivos y verbos.

Como se ha podido comprobar, el comportamiento de los clasificadores base con los corpus SMR y MCE es similar al que desarrollaron con los corpus OCA y EVOCA. Por lo tanto, cabe preguntarse si la combinación de los clasificadores sobre opiniones en español e inglés, producirá el mismo efecto positivo que en el caso de la combinación de clasificadores especializados en árabe e inglés. En esta ocasión se han evaluado dos métodos de combinación, por un lado un sistema de voto y por otro un sistema basado en Stacking. Al igual que con OCA, se han evaluado tres sistemas de voto, que son los siguientes:

1. Voto_smrsvm_mcesvm: Combinación de los clasificadores base que siguen una filosofía de aprendizaje supervisado. Al ser dos el número de clasificadores supervisados pueden producirse empates, de manera que para dichos casos se han definido dos reglas para decidir qué clase asignar al documento que se esté procesando:

   a) Voto_smrsvm_mcesvm_neg: Solamente se clasificará el documento como positivo si los dos clasificadores base consideran que es positivo. En caso contrario el documento es clasificado como negativo.

   b) Voto_smrsvm_mcesvm_pos: Solamente se requerirá que uno de los clasificadores considere el documento como positivo, para que la clase que devuelva el sistema de voto sea positivo.

2. Voto_smrsvm_mcesvm_mceswn: Combinación mediante voto de los tres clasificadores, los dos que emplean SVM como algoritmo de clasificación y el basado en el aprovechamiento de la información atesorada en SentiWordNet.

Los resultados obtenidos por cada uno de los tres sistemas de voto se recogen en la Tabla 7.7. Teniendo en cuenta la medida F1 se puede apreciar que tanto la combinación de los tres clasificadores, como la conjunción de solamente los que siguen un enfoque supervisado con la regla de desempate Voto_smrsvm_mcesmv_pos superan los resultados alcanzados por los tres clasificadores base. El método de inferencia que utiliza SentiWordNet, si se tiene en cuenta también el valor de F1 alcanzado, también se puede comprobar que supera a los tres clasificadores base. La diferencia de F1 entre Voto_smrsvm_mcesmv_pos y Voto_smrsvm_-mcesmv_mceswn se puede considerar insignificante, y si se tiene en cuenta el valor de *Accuracy*, Voto_smrsvm_mcesmv_mceswn es ligeramente superior a Voto_smrsvm_mcesmv_pos. Por lo que, se concluye que la combinación de un clasificador supervisado de opiniones en español, con un clasificador supervisado y no supervisado sobre el mismo dominio en inglés mediante un sistema de voto, permite incrementar ligeramente la clasificación de opiniones en español.

Además del desarrollo del sistema de voto, también se experimentó, como se ha indicado al comienzo de la sección, con un método de combinación basado en *stacking*. Como algoritmo de metaclasificación se han evaluado SVM (ver en la Sección 4.2.1), Naïve Bayes (ver en la Sección 4.2.2), BBR (ver en la Sección 4.2.3) y C4.5 (ver en la Sección 4.2.5). La metaclasificación por Stacking toma la salida de los clasificadores base y las

Combinación	**Precisión**	***Recall***	**F1**	***Accuracy***
Voto_smrsvm_mcesvm_-neg	85,51 %	88,93 %	87,19 %	87,31 %
Voto_smrsvm_mcesvm_-pos	80,03 %	98,43 %	88,28 %	87,31 %
Voto_smrsvm_mcesmv_-mceswn	81,60 %	96,08 %	88,25 %	87,58 %

Tabla 7.7: Resultados del sistema de voto sobre los corpos SMR y MCE.

convierte en las características de los documentos que tiene que procesar el metaclasificador. Dado que, tanto los dos clasificadores supervisados (SVM) como el no supervisado (SentiWordNet) devuelven como resultado la clase que asigna al documento y un valor de confianza de la decisión, se ha diseñado tres conjuntos de características disímiles:

1. Clases (CL): Las características con las que se representarán a los documentos que procesará el metaclasificador serán las clases asignadas por los clasificadores base. En este caso el número de características será tres, porque tres son los clasificadores que se están combinando.

2. Confianzas (CF): SVM es un algoritmo que además de devolver la clase que considera que pertenece un objeto, retorna un valor de confianza para cada una de las clases involucradas en el proceso de clasificación. La clase que decide como la propia del objeto que está clasificando es la que cuenta con un mayor valor de confianza. En el caso que nos atañe es una clasificación binaria, es decir, solamente se está trabajando con dos clases, positivo y negativo, por lo que conjuntamente los dos clasificadores supervisados proporcionan cuatro valores de confianza. El lector debe recordar, que el algoritmo que utiliza SentiWordNet aglutina los valores de polaridad de las palabras del documento siguiendo la regla definida por Denecke (2008). Por ende, a cada documento, el clasificador además de catalogarlo con una clase le asocia tres valores de polaridad: positivo, negativo y objetivo. Por tanto, los valores de confianza del método de clasificación basado en SentiWordNet se corresponderán con las puntuaciones de positivo, negativo y objetivo resultantes de aplicar la regla de Denecke (2008). Según esta configuración, el metaclasificador trabajará con documentos representados por siete características.

3. Clases y confianza (CL_CF): En este caso se genera un conjunto de características formado por las clases y los valores de confianza devueltos por cada uno de los tres métodos de clasificación. El tamaño de este conjunto de características es diez.

Los resultados obtenidos por cada uno de los cuatro metaclasificadores con cada conjunto de características se muestran en la Tabla 7.8. Como se puede apreciar, el metaclasificador construido sobre Naïve Bayes es el que mejores resultados ha obtenido entre los cuatro algoritmos que se han evaluado. Entre los tres conjunto de características, aquel conformado por las clases y los valores de confianza (CL_CF) es que el ha devuelto el mejores resultados. Si se compara con los resultados obtenidos por el mejor sistema de voto, también se puede comprobar que en todas las medidas de evaluación el metaclasificador por Stacking supera al sistema de voto. Este comportamiento se puede interpretar como que el Stacking debe ser tenido en cuenta como método de combinación entre clasificadores en el contexto del AO.

Metaclasificador	**Características**	**Precisión**	***Recall***	**F1**	***Accuracy***
SVM	CL	87,71 %	87,64 %	87,68 %	87,66 %
	CF	88,33 %	88,28 %	88,31 %	88,31 %
	CL_CF	87,71 %	87,64 %	87,68 %	87,66 %
NB	CL	87,81 %	87,82 %	87,82 %	87,81 %
	CF	88,36 %	88,34 %	88,35 %	88,34 %
	CL_CF	88,58 %	88,57 %	88,56 %	88,57 %
C4.5	CL	87,71 %	87,64 %	87,66 %	87,66 %
	CF	86,54 %	86,30 %	86,42 %	86,32 %
	CL_CF	86,54 %	86,30 %	86,42 %	86,32 %
BBR	CL	87,70 %	87,55 %	87,63 %	87,58 %
	CF	88,14 %	88,13 %	88,14 %	88,15 %
	CL_CF	88,29 %	88,24 %	88,27 %	88,27 %

Tabla 7.8: Resultados obtenidos por Stacking sobre los conjuntos de datos SMR y MCE.

Con el ánimo de ayudar a la comparación de los distintos métodos de combinación, y comprobar con una simple observación el mejor desempeño del Stacking, se presenta en la Figura 7.2 una comparación de los mejores resultados alcanzados con los métodos evaluados.

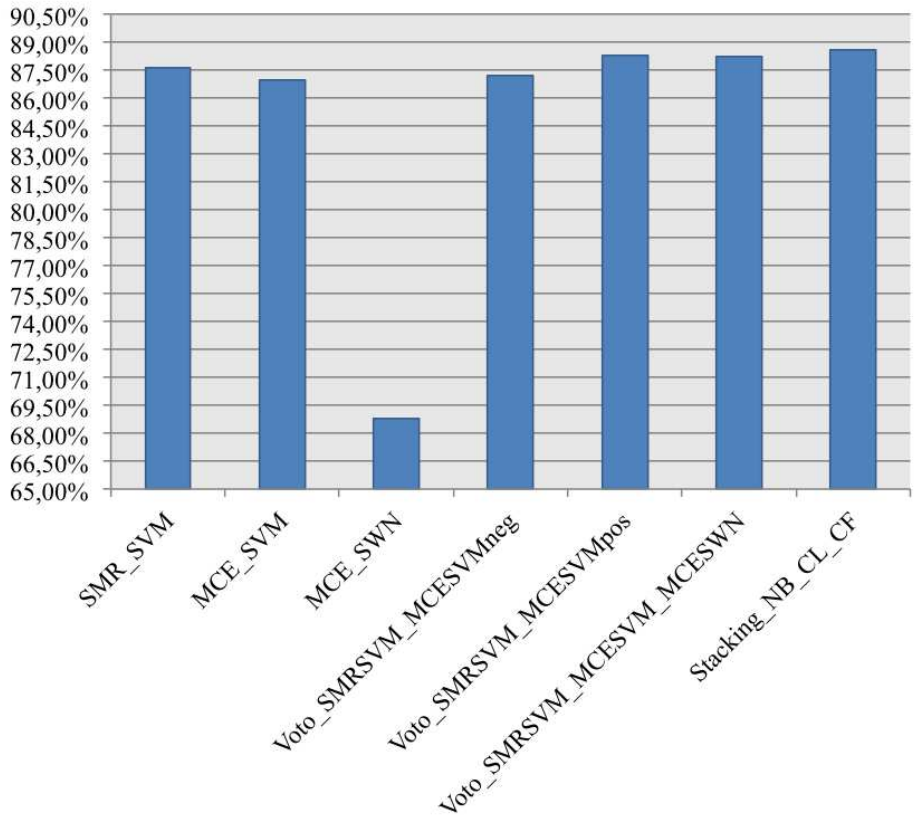


Figura 7.2: Comparación del valor de F1 alcanzado por los clasificadores base y los sistemas de combinación de clasificadores evaluados.

Los buenos resultados alcanzados en la combinación de un clasificador especializado en español y otro en inglés, parece que confirma la aserción de Banea et al. (2010) sobre que la subjetividad se mantiene entre lenguas, a pesar de que se pueda expresar de una manera disímil, y por tanto, es beneficioso para la clasificación de la polaridad la combinación de varios corpus paralelos en diversos idiomas.

En la anterior experimentación se combinaban dos algoritmos base supervisados con un método de clasificación no supervisado. El resultado fue la obtención de un resultado superior al alcanzado por cada uno de ellos por separado. La siguiente experimentación va a intentar diferenciarse de la anterior en el hecho de que ningún clasificador base va a seguir un enfoque supervisado, sino que se van a fundamentar en el uso de recursos lingüísticos

para el descubrimiento de la orientación de las opiniones. Tomando la afirmación de Banea et al. (2010) como punto de partida, a continuación se va a describir el estudio sobre la conveniencia de la combinación de la información proporcionada por recursos lingüísticos en inglés y español para la clasificación de opiniones escritas en español, el cual se encuentra publicado en (Martínez-Cámara et al., 2014b).

Tanto para inglés como para español se van a emplear dos recursos léxicos, uno a nivel de palabra o término y otro a nivel de sentido o concepto. Los recursos elegidos a nivel léxico para el español son la lista de palabras de opiniones iSOL, cuya descripción completa se encuentra en la sección 6.2.1, y para el inglés el lexicón de palabras de opinión BLOL, el cual también fue explicado en la Sección 6.2.1. En cuanto a los recursos a nivel de concepto se ha seleccionado tanto para inglés como para español SentiWordNet. SentiWordNet, como ya se indicó en el Capítulo 5, es un recurso orientado a proporcionar información a clasificadores de polaridad destinados al procesamiento de textos escritos en inglés. Pero, como también se comentó en el Capítulo 5, se puede aprovechar su arquitectura idéntica a la de WordNet para conectar SentiWordNet con otras versiones de WordNet en otros idiomas. En el Capítulo 5 ya se describió una versión de WordNet en español, la que forma parte de MCR. Por tanto, es totalmente viable el uso de SentiWordNet tanto para la clasificación de textos en inglés y en español. Sí debe recordarse que el número de *synsents* de la versión española de WordNet en MCR es sensiblemente inferior al número total de conceptos en WordNet, conteniendo solamente un 50,33 % del total de *synsets* de WordNet.

En resumen, se va a analizar el comportamiento de cuatro clasificadores, dos por idioma, de los cuales uno de ellos utilizará un recurso a nivel de palabra y el otro un recurso a nivel de concepto. Posteriormente, se analizará en cada idioma la combinación de la información de los recursos a nivel de palabra y concepto a través de un metaclasificador. Por último, se intentará comprobar que la combinación de dos clasificadores de polaridad especializados en dos lenguas distintas pueden mejorar, a través de un esquema de combinación de metaclasificación, la exactitud de los resultados obtenidos por los clasificadores base de textos en español. Si los resultados son positivos, entonces se habrá encontrado una muestra más de la veracidad de la aserción de Banea et al. (2010). Gráficamente los sistemas que se van a evaluar se muestran en las Figuras 7.3, 7.4, 7.5, 7.6 y 7.7.

Por lo que las figuras representan, siete son los sistemas que se van a evaluar. La evaluación requiere, al igual que la anterior experimentación, de un corpus paralelo en inglés y en español. Para satisfacer este requerimiento

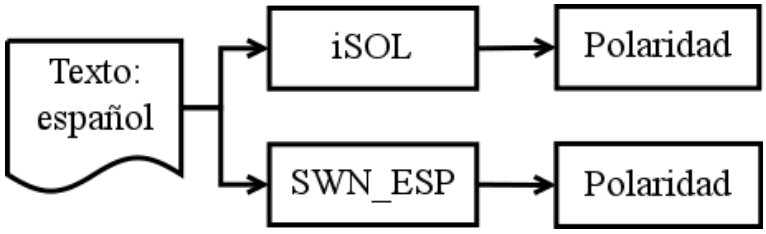


Figura 7.3: Clasificadores que clasifican opiniones en español con dos recursos lingüísticos distintos.

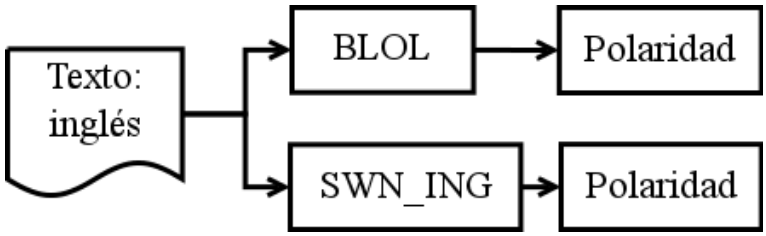


Figura 7.4: Clasificadores que clasifican opiniones en inglés con dos recursos lingüísticos distintos.

se seleccionó el mismo corpus que en el estudio anterior, es decir, el constituido por SMR y por su traducción al inglés MCE. Huelga remarcar, que SMR es la colección de opiniones que se usó para la evaluación de los sistemas que usan recursos en español, y que MCE es la fuente de datos de los sistemas que emplean recursos en inglés. Una vez conocidos tanto la colecciones de documentos utilizadas y lo recursos lingüísticos en los que se sustentan los clasificadores, es momento de presentar los resultados de los sistemas de clasificación sin ningún tipo de combinación, es decir, los que se representan en las Figuras 7.3 y 7.4. Para facilitar la forma de referenciar a los sistemas, cuyos resultados se recogen en la Tabla 7.11, se les va asignar los siguientes nombres:

- SMR_iSOL: Sistema que clasifica las opiniones del corpus SMR, por medio de un clasificador que se limita a contar el número de palabras positivas y negativas de iSOL que se encuentran en cada opinión. Si el número de términos positivos es superior, la opinión se cataloga como positiva, si la cantidad de palabras negativas sobrepasa al de positivas, entonces la opinión será negativa, y en caso de empate la opinión se tomará como positiva.

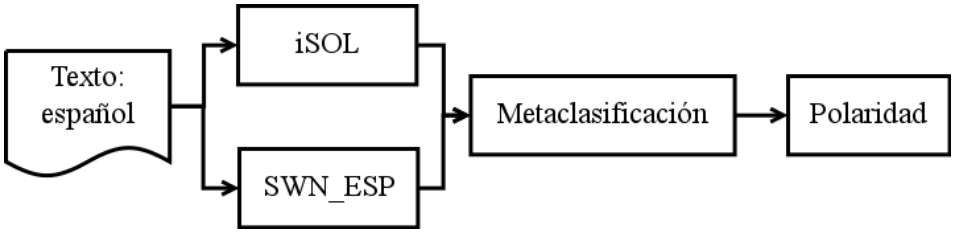


Figura 7.5: Combinación por metaclasificación de los clasificadores que utilizan recursos lingüísticos en español.

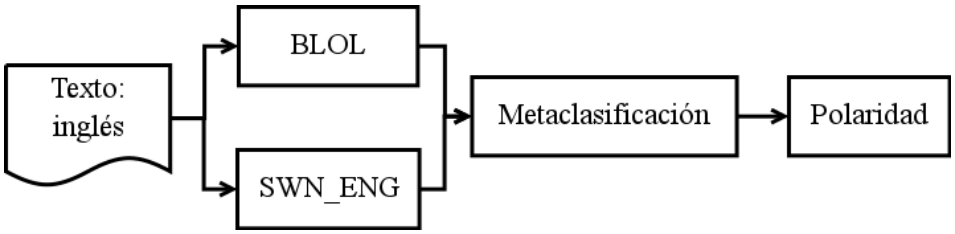


Figura 7.6: Combinación por metaclasificación de los clasificadores que utilizan recursos lingüísticos en inglés.

- SMR_SWN_ESP: Clasificador de las opiniones de SMR, que se encuentra asistido por la información de polaridad de la opinión que le proporciona SentiWordNet. Al igual que la experimentación explicada con anterioridad y publicada en (Martín-Valdivia et al., 2013), este clasificador sigue una filosofía similar a la expuesta en (Denecke, 2008), es decir, no se lleva a cabo ninguna operación de desambiguación para la determinación de los conceptos concretos con los que se emplean las palabras, sino que calcula la media aritmética de cada una de las puntuaciones de polaridad de cada sentido correspondiente a una palabra, para obtener el vector de polaridad de la misma. La polaridad del documento se obtiene mediante, de nuevo, el cálculo de la media aritmética de la polaridad de todas las palabras del documento, asignando como clase la polaridad con la mayor media aritmética.

  Al igual que se hizo en el estudio anterior, se ha emprendido un análisis de las categorías que proporcionan una información más precisa para la determinación de la polaridad. Los resultados de este análisis se muestran en la Tabla 7.9.

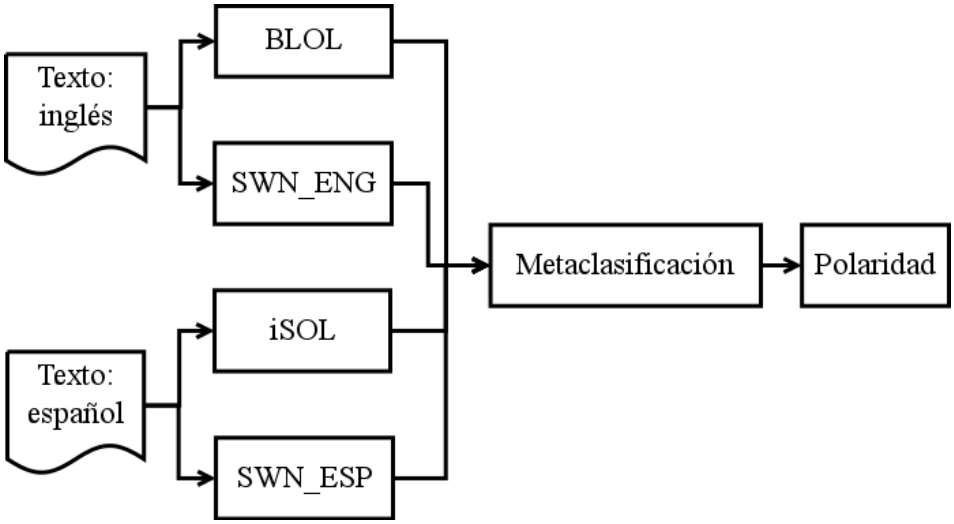


Figura 7.7: Combinación por metaclasificación de los clasificadores que emplean recursos en inglés y en español.

- MCE_BLOL: Sistema equivalente a SMR_iSOL, pero en esta ocasión el corpus es MCE, y la lista de palabras indicadoras de opinión es BLOL.
- MCE_SWN_ENG: Clasificador con idéntico funcionamiento que SMR_SWN_ESP, pero utilizando MCE como colección de documentos de opinión, y la versión completa de SentiWordNet. También se ha realizado un estudio del aporte de cada categoría morfológica a la clasificación de la opinión, mostrándose los resultados en la Tabla 7.10.

Antes de mostrar los resultados del estudio del nivel de calidad del clasificador en función de las categorías morfológicas consideradas, es preciso realizar una aclaración. En la experimentación anterior se desarrolló el mismo análisis sobre el corpus MCE y utilizando la misma regla para el cálculo de la polaridad de los documentos. Pero la diferencia entre un estudio y otro estriba en las medidas de evaluación utilizadas. En el primer estudio se calcularon la Precisión, el *Recall* y el F1, mientras que en el segundo se determinaron sus correspondientes Macromedias (*Macro-averaged*). El cambio se debe, a que este segundo estudio se llevó a cabo en una fase más avanzada de la investigación, en la que se quería emplear una medida que indicara con un mayor grado de exactitud la calidad

Categorías morfológicas	**Macro-P**	**Macro-R**	**Macro-F1**	***Accuracy***
nombres	51,27%	51,06%	51,16%	51,66%
adjetivos	62,06%	58,74%	60,36%	59,50%
verbos	51,48%	50,51%	50,99%	51,70%
adverbios	55,20%	54,11%	54,65%	54,78%
adjetivos+nombres	60,41%	56,60%	58,45%	57,49%
adjetivos+verbos	59,46%	54,18%	56,70%	55,28%
adjetivos+adverbios	63,68%	58,79%	61,14%	59,66%
nombres+verbos	48,56%	49,37%	48,96%	50,48%
nombres+adverbios	53,87%	52,91%	53,39%	53,64%
verbos+adverbios	55,43%	51,79%	53,55%	52,99%
adjetivos+nombres+verbos	58,93%	53,87%	56,29%	54,97%
adjetivos+nombres+adverbios	61,81%	56,93%	59,27%	57,87%
nombres+verbos+adverbios	50,13%	50,05%	50,09%	51,20%
adjetivos+verbos+adverbios	61,16%	54,41%	57,59%	55,54%
adjetivos+nombres+verbos+adverbios	59,19%	53,47%	56,19%	54,63%

Tabla 7.9: Resultados obtenidos por cada categoría morfológica en el sistema SMR_SWN_ESP.

del procesamiento de cada clase por parte del clasificador. Para ello se emplearon las Macromedias, cuya definición se pueden consultar en la Sección 3.4. Por tanto, la Tabla 7.10 muestra el desempeño del clasificador MCE_SWN_ENG medido con las Macromedias.

Si se comparan las Tablas de resultados 7.6 y 7.10 se podrá comprobar que los valores de Precisión, *Recall* y F1 no son iguales, mientras que los de *Accuracy* sí. La diferencia es una señal de que el clasificador se siente más cómodo trabajando con una de las clases que con la otra, de manera que no se refleja con exactitud el rendimiento global del sistema. La nueva medición ha provocado que sea una configuración distinta del clasificador la que devuelva unos mejores resultados de clasificación, la cual se corresponde con la utilización de todas las categorías morfológicas contempladas en SentiWordNet. En cuanto a la clasificación de opiniones en español con una

Categorías morfológicas	**Macro-P**	**Macro-R**	**F1**	***Accuracy***
nombres	55,58 %	54,31 %	54,94 %	55,01 %
adjetivos	61,69 %	60,78 %	61,23 %	61,18 %
verbos	57,59 %	54,95 %	56,24 %	55,81 %
adverbios	58,26 %	55,09 %	56,63 %	54,17 %
adjetivos+nombres	62,90 %	60,48 %	61,66 %	61,10 %
adjetivos+verbos	63,42 %	60,85 %	62,11 %	61,49 %
adjetivos+adverbios	63,20 %	62,82 %	63,01 %	62,55 %
nombres+verbos	56,11 %	53,76 %	54,91 %	54,67 %
nombres+adverbios	58,79 %	58,41 %	58,60 %	58,10 %
verbos+adverbios	58,14 %	57,13 %	57,63 %	56,61 %
adjetivos+nombres +verbos	62,45 %	58,96 %	60,66 %	59,73 %
adjetivos+nombres +adverbios	64,27 %	64,25 %	64,26 %	64,30 %
nombres+verbos +adverbios	60,39 %	60,39 %	60,39 %	60,34 %
adjetivos+verbos +adverbios	64,32 %	64,33 %	64,33 %	64,30 %
adjetivos+nombres+ verbos+adverbios	65,13 %	64,72 %	64,92 %	64,95 %

Tabla 7.10: Resultados obtenidos por cada categoría morfológica en el sistema MCE_SWN_ENG.

versión limitada de SentiWordNet[8], el mejor resultado se obtiene cuando sólo se tienen en cuenta los niveles de polaridad de adjetivos y adverbios. La diferencia entre el mejor valor de *Accuracy* entre MCE_SWN_ENG y SMR_SWN_ESP es de un 8,14 %, la cual no se puede considerar como importante, si se considera que entre la versión de SentiWordNet empleada en SMR_SWN_ESP y MCE_SWN_ENG es de un 50 %.

Habiendo visto que la diferencia entre los clasificadores basados en SentiWordNet no es abultada, es lógico preguntarse por el comportamiento de los clasificadores que tienen como fuente de conocimiento a iSOL y BLOL. Por consiguiente, en la Tabla 7.11 se recogen los resultados obtenidos

[8]Hay que tener siempre presente que para poder aplicar SentiWordNet en español se requiere emplear una versión de WordNet en español. La versión de WordNet en español que ofrece una mayor cobertura es la contenida en MCR, que como ya se ha indicado en varias ocasiones en esta memoria, contiene alrededor del 50 % de los *synsets* de WordNet.

por los sistemas SMR_iSOL y MCE_BLOL.

Clasificador	**Macro-P**	**Macro-R**	**Macro-F1**	***Accuracy***
SMR_iSOL	62,22%	61,47%	61,84%	61,83%
BLOL_MCE	61,92%	56,58%	59,13%	57,56%

Tabla 7.11: Resultados de los sistemas SMR_iSOL y MCE_BLOL.

Los resultados de los dos clasificadores fundados solamente en el uso de una lista de palabras se puede decir que son aceptables, dado que superan al que hubiera obtenido un sistema basado en la asignación de la clase más frecuente. En esta ocasión es el clasificador de opiniones en español el que presenta un mejor comportamiento. iSOL contiene un mayor número de palabras que BLOL, por lo que es más probable en una opinión encontrar palabras pertenecientes a iSOL que a BLOL. Este hecho se ve reflejado en el Macro-R que es más elevado en el sistema SMR_iSOL que en BLOL_-MCE. Teniendo únicamente en cuenta el *Accuracy*, la diferencia entre los dos clasificadores asciende a 7,42%, que es similar al 8,14% de desemejanza entre los dos clasificadores que usan SentiWordNet. Si se comparan los sistemas entre los que tratan un mismo idioma, se puede observar como para el español iSOL proporciona una información de más calidad que la versión limitada de SentiWordNet, ya que SMR_iSOL obtiene unos mejores resultados que SMR_SWN_ESP. En el caso del inglés es justamente lo contrario, SentiWordNet posibilita una mejor clasificación que BLOL. El comportamiento de los clasificadores ingleses se podría decir que es el esperado, dado que se espera más de una base de conocimiento de opinión, que de una lista de palabras, pero el menor rendimiento que proporciona la versión de SentiWordNet en español está motivado principalmente porque sólo contiene la mitad de los *synsets* de WordNet, lo que limita su capacidad de representación del lenguaje.

Los resultados que se recogen en las Tablas 7.9, 7.10 y 7.11 muestran unos clasificadores con un comportamiento similar, pero el emplear fuentes de conocimiento disímiles permite pensar que realizan una exploración distinta del espacio de soluciones, lo cual da pie a pensar que su combinación puede mejorar el resultado de la clasificación. La primera combinación que se ha estudiado ha sido la de los sistemas centrados en un único idioma, para evaluar si la combinación de recursos en un mismo idioma reporta una mejora de la calidad de la clasificación. Debido a los superiores resultados de la combinación por metaclasificación en el anterior estudio, para esta ocasión solamente se ha evaluado la combinación por metaclasificación, y en concreto siguiendo la metodología de Stacking. Como algoritmos

de metaclasificación se han evaluado SVM, Naïve Bayes y la regresión logística bayesiana (BBR), los cuales ya fueron evaluados en la anterior experimentación y definidos en el Capítulo 4. El principal fundamento del Stacking es tomar la salida de los clasificadores base como características de los ejemplos a clasificar. Teniendo ésto y la naturaleza de los clasificadores base en cuenta, varias son las configuraciones que se puede definir, las cuales son listadas a continuación:

1. SMR_iSOL_SWN_ESP_cls: Se toman como características exclusivamente la clases devueltas por los clasificadores base.

2. SMR_iSOL_SWN_ESP_punt: Se toman como características la salida de SMR_iSOL y los tres valores de polaridad calculados por SMR_SWN_ESP para el documento que se está procesando. Se debe recordar que SMR_SWN_ESP determina la clase a la que pertenece el documento en función de la puntuación obtenida por los tres niveles de polaridad considerados en SentiWordNet.

3. SMR_iSOL_SWN_ESP_cls_punt: Las características en este caso son las clases que los clasificadores SMR_iSOL y SMR_SWN_ESP consideran que pertenecen el documento en estudio, más los tres valores de polaridad calculados por SMR_SWN_ESP.

4. MCE_BLOL_SWN_ENG_cls: Los documentos a procesar por el metaclasificador se representarán por las clases que determinen los dos clasificadores base MCE_BLOL y MCE_SWN_ENG.

5. MCE_BLOL_SWN_ENG_punt: De igual manera que en SMR_iSOL_SWN_ESP_punt se toman como características la salida de MCE_BLOL y los tres valores de polaridad calculados por MCE_SWN_ENG.

6. MCE_BLOL_SWN_ENG_cls_punt: En esta ocasión, las características para representar los documentos que tienen que clasificar el metaclasificador son las clases que devuelven los dos clasificadores base, más los valores de polaridad calculados por MCE_SWN_ENG.

Los resultados obtenidos por cada una de las configuraciones definidas se recogen en la Tabla 7.12.

Los resultados muestran que para el caso del español la combinación es beneficiosa, dado que ninguna combinación arroja un resultado inferior a cualquiera de los dos clasificadores base. Si se analiza a nivel de algoritmo de metaclasificación, al igual que en la experimentación anterior, Naïve

Características	**Metaclasificador**	**Macro-P**	**Macro-R**	**Macro-F1**	***Accuracy***
SMR_iSOL_-SWN_ESP_-cls	SVM	62,26%	61,47%	61,86%	61,83%
	NB	63,94%	63,67%	63,80%	63,85%
	BBR	63,94%	69,67%	63,80%	63,85%
SMR_iSOL_-SWN_ESP_-punt	SVM	62,26%	61,47%	61,86%	61,83%
	NB	63,33%	62,64%	62,93%	62,93%
	BBR	62,75%	62,08%	62,41%	62,40%
SMR_iSOL_-SWN_ESP_-cls_punt	SVM	62,26%	61,47%	61,86%	61,83%
	NB	63,84%	61,90%	62,86%	62,44%
	BBR	63,77%	63,46%	63,61%	63,65%
MCE_-BLOL_-SWN_ENG_-cls	SVM	63,60%	60,85%	62,19%	61,48%
	NB	63,68%	62,26%	62,96%	62,70%
	BBR	63,68%	62,26%	62,96%	62,70%
MCE_-BLOL_-SWN_ENG_-punt	SVM	61,96%	60,07%	61,03%	60,65%
	NB	60,02%	60,07%	61,03%	60,65%
	BBR	62,76%	57,80%	60,18%	58,70%
MCE_-BLOL_-SWN_ENG_-cls_punt	SVM	63,60%	60,85%	62,19%	61,48%
	NB	63,45%	62,08%	62,76%	62,51%
	BBR	63,68%	62,26%	62,96%	62,70%

Tabla 7.12: Resultados de la combinación por metaclasificación de los clasificadores por idioma.

Bayes es el método que mejor aprovecha la información aportada por las salidas de los clasificadores base. Teniendo en cuenta los resultados, también debe remarcarse que la configuración que solamente utiliza las clases asignadas por los clasificadores base es la que llega a alcanzar unos mejores resultados. En cuanto al inglés, se puede observar que ninguna combinación ha ofrecido una clasificación de mayor calidad que la proporcionada por el mejor clasificador base, MCE_SWN_ENG. En el caso del español, los dos clasificadores base tienen un comportamiento menos dispar que los métodos de inferencia que toman como entrada textos en inglés. Asimismo, el comportamiento del clasificador base MCE_BLOL dista menos del comportamiento que experimentaría un clasificador basado en la selección

siempre de la clase más frecuente, que el peor clasificador base de textos en español, lo cual explica que la combinación de los dos clasificadores base de inglés no haya sido tan fructífera como el caso de los de español.

No debe difuminarse el objetivo principal de esta experimentación, que es sin duda la mejora de la calidad de la clasificación de la polaridad de textos en español mediante la combinación de recursos lingüísticos en español y en inglés, dado que como afirma Banea et al. (2010) la combinación de métodos de inferencia en distintos idiomas es positivo para la clasificación de la polaridad. Pues bien, la siguiente evaluación se corresponde con la comprobación de la afirmación anterior, es decir, con la combinación de los dos sistemas de clasificación que utilizan recursos en español (SMR_iSOL y SMR_SWN_ESP) con uno o con los dos clasificadores que trabajan con recursos en inglés (MCE_BLOL y MCE_-SWN_ENG, ver Figura 7.7) por medio de metaclasificación. A continuación se listan las configuraciones que se han evaluado:

1. SMR_iSOL_SWN_ESP_cls_MCE_BLOL: Combinación de SMR_-iSOL con SMR_SWN_ESP y MCE_BLOL.

2. SMR_iSOL_SWN_ESP_punt_MCE_BLOL: Las características del metaclasificador estarán constituidas por la clase identificada por SMR_iSOL, por las puntuaciones de polaridad calculadas por SMR_-SWN_ESP y por la clase determinada por MCE_BLOL.

3. SMR_iSOL_SWN_ESP_cls_punt_MCE_BLOL: En este caso cada opinión va a estar representada por la conjunción de las características de SMR_iSOL_SWN_ESP_cls_MCE_BLOL y SMR_-iSOL_SWN_ESP_punt_MCE_BLOL.

4. SMR_iSOL_SWN_ESP_cls_MCE_SWN_ENG_cls: Esta configuración consiste en la combinación de SMR_iSOL, la clase identificada por SWN_ESP y por la clase a la que SWN_ENG considera que pertenece el documento.

5. SMR_iSOL_SWN_ESP_punt_MCE_SWN_ENG_punt: Configuración similar a la anterior, pero esta ocasión no se consideran como características la clase generada por los clasificadores basado en SentiWordNet, sino la puntuación de polaridad que determinan.

6. SMR_iSOL_SWN_ESP_cls_punt_MCE_SWN_ENG_cls_punt: En esta ocasión la combinación consistirá en identificar las opiniones que

está representadas por la clase identificada por SMR_iSOL, SMR_-SWN_ESP y MCE_SWN_ESP, y por las puntuaciones de polaridad obtenidas por SMR_SWN_ESP y MCE_SWN_ESP.

7. SMR_MCE_cls: Combinación de los cuatro clasificadores, pero de los basados en SentiWordNet solo se toma la clase de polaridad que han determinado.

8. SMR_MCE_punt: Configuración similar a la anterior, pero en esta ocasión, en lugar de la clase identificada por los clasificadores basados en SentiWordNet, se toma la puntuación de polaridad.

9. SMR_MCE_cls_punt: Combinación de los cuatro clasificadores base, incluyendo tanto las clases determinadas por los sistemas que utilizar SentiWordNet como las puntuaciones de polaridad que han calculado.

Los resultados obtenidos por cada una de las configuraciones indicadas se muestran en la Tabla 7.13.

La combinación, pero ahora de un número mayor de recursos, ha vuelto a arrojar una mejora en la calidad de la clasificación. La configuración que ha posibilitado ese repunte de la calidad ha sido la que representa los datos de entrada del metaclasificador como vectores conformados por la clase asignada por los dos clasificadores especializados en español y por el método de inferencia que utiliza SentiWordNet en inglés, así como por los valores de puntuación obtenidos por los clasificadores SWN_ESP y MCE_SWN. En esta ocasión, Naïve Bayes vuelve a ser el método de metaclasificación que mejor comportamiento demuestra, confirmándose así su mejor rendimiento en el aprendizaje de los errores de los clasificadores base. Los resultados confirman la hipótesis de la que partíamos, que la combinación de varios idiomas es positivo para la clasificación de la polaridad. A modo de resumen y de conclusión, la Figura 7.8 muestra la evolución positiva de la clasificación de la polaridad en español conforme se han ido combinando recursos lingüísticos en español y, por último, la combinación de un clasificador especializado en opiniones en inglés.

7.4. Conclusión

En el presente capítulo se ha estudiado la conveniencia de la combinación de clasificadores para el AO en idiomas con pocos recursos lingüísticos. Tanto en la experimentación llevada a cabo con árabe, como la que se ha desarrollado sobre español, se ha podido comprobar que la

Características	**Metaclasificador**	**Macro-P**	**Macro-R**	**Macro-F1**	***Accuracy***
SMR_iSOL_-SWN_ESP_cls_-MCE_BLOL	SVM	62,26%	61,47%	61,86%	61,83%
	NB	63,55%	61,06%	62,28%	61,68%
	BBR	63,94%	69,67%	63,80%	63,85%
SMR_iSOL_-SWN_ESP_-punt_MCE_-BLOL	SVM	62,26%	61,47%	61,86%	61,83%
	NB	63,53%	62,08%	62,80%	62,55%
	BBR	63,09%	62,43%	62,76%	62,74%
SMR_iSOL_-SWN_ESP_cls_-punt_MCE_-BLOL	SVM	62,26%	61,47%	61,86%	61,83%
	NB	63,99%	62,10%	63,03%	62,63%
	BBR	64,13%	63,74%	63,93%	63,96%
SMR_iSOL_-SWN_ESP_cls_-MCE_SWN_-ENG_cls	SVM	64,40%	63,98%	64,19%	64,07%
	NB	64,89%	63,63%	64,25%	64,04%
	BBR	64,70%	64,50%	64,60%	64,53%
SMR_iSOL_-SWN_ESP_-punt_MCE_-SWN_ENG_punt	SVM	62,26%	61,47%	61,86%	61,83%
	NB	64,30%	63,89%	64,09%	64,11%
	BBR	63,27%	62,63%	62,95%	62,93%
SMR_iSOL_-SWN_ESP_cls_-punt_MCE_-SWN_ENG_cls_-punt	SVM	63,93%	62,88%	63,40%	63,23%
	NB	65,25%	64,34%	64,79%	64,68%
	BBR	64,12%	63,70%	63,91%	63,92%
SMR_MCE_cls	SVM	63,55%	62,70%	63,12%	63,01%
	NB	64,97%	63,35%	64,15%	63,81%
	BBR	64,57%	63,42%	63,99%	63,77%
SMR_MCE_punt	SVM	62,26%	61,47%	61,86%	61,83%
	NB	64,55%	63,56%	64,05%	63,92%
	BBR	63,47%	62,76%	63,11%	63,08%
SMR_MCE_cls_-punt	SVM	63,37%	62,55%	62,96%	62,89%
	NB	65,12%	64,34%	64,73%	64,65%
	BBR	64,39%	63,70%	64,04%	64,00%

Tabla 7.13: Resultados de la combinación por metaclasificación de los cuatro clasificadores base.

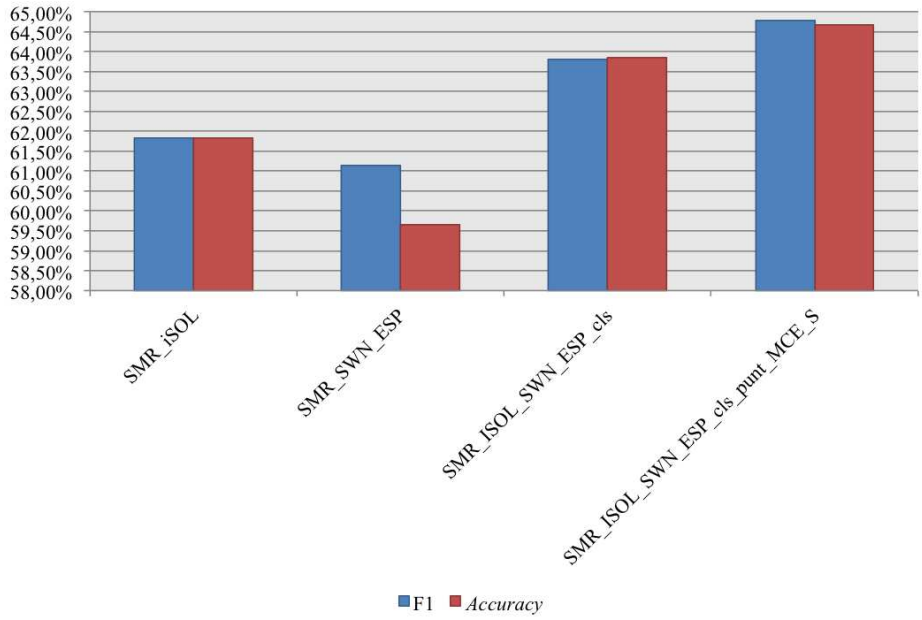


Figura 7.8: Comparación de los métodos de clasificación de la polaridad en español desarrollados.

conjunción de métodos de diferente naturaleza ha posibilitado la generación de la diversidad necesaria para que la combinación obtenga mejores resultados que los alcanzados por los clasificadores base. A su vez, se ha comprobado que la subjetividad se mantiene entre idiomas, de manera que la combinación de clasificadores especializados en distintos idiomas permite incrementar la calidad de la clasificación en una de las lenguas involucradas.

A modo de recordatorio y resumen, cuando el objetivo es clasificar opiniones en árabe, la metaclasificación que mejores resultados ha reportado ha sido la compuesta por un sistema de voto de tres clasificadores, dos basados en aprendizaje automático, de los cuales uno está especializado en árabe y otro en inglés, y un tercer clasificador fundado en el uso del lexicón de opinión SentiWordNet.

En el caso del español, se han combinado un mayor número de clasificadores, así como se ha tratado de desarrollar un metaclasificador conformado por clasificadores base no supervisados. Asimismo, con el español se ha estudiado la combinación por medio de Stacking. Cuando entre los clasificadores base se encuentran métodos de inferencia supervisados, el Stacking supera por poco un sistema de voto similar

al desarrollado para el árabe, lo cual es una señal de que la aplicación de Stacking puede mejorar todavía más la calidad de la clasificación en árabe. Pero debe decirse, que para que se produzca esa mejora, además de utilizar las clases de los clasificadores base como características de los datos de entrada del clasificador, también se ha empleado las puntuaciones de polaridad devueltas por uno de los métodos de inferencia base, concretamente el que aprovecha la información de SentiWordNet.

Siguiendo con el español, cuando tratamos de emplear sólo clasificadores no supervisados se obtienen, como era de esperar, unos resultados de menor calidad que los alcanzados cuando se usan clasificadores supervisados. Pero se repite el patrón de que el Stacking es el esquema de combinación más adecuado, y que Naïve Bayes es el método de inferencia más propio para la combinación. Asimismo, al igual que ocurría cuando se empleaban métodos supervisados, en este caso también es positiva la utilización de las puntuaciones de los clasificadores base que utilizan SentiWordNet.

8

Conclusions and Future Work

8.1. Conclusions

The former chapters have described the extensive research carried out during four years on the context of Sentiment Analysis (SA). The work have attempted to bridge the gap of research in the Spanish language, which, as it is indicated in Chapter 3, is the second native language all over the world. The thesis have been organized in several chapters, which each of them are linked to different strategies to discover the polarity of an opinion: supervised learning, unsupervised learning, opinion classification based on the use of sentiment lexicons and ensemble learning. The development of each strategy has allowed us to deeper know the problem of polarity classification, and also it has provide us the enough experience to define the subsequent steps on the study of opinions.

Chapter 4 is focused on the use of supervised learning to resolve the problem of the automatic classification of opinions. Two sort of texts have been considered in the study: long texts and short texts. The definition of each kind of text is in Chapter 3, but with the aim of reminding the difference among them, their definition will be again exposed. Long texts or long opinions are those that are composed by more than two sentences, or at least, they are longer than a tweet. On the other hand, short texts or short opinions are those that their length is up to two sentences, or at least, the length of a tweet.

The experiments made evident that long and short texts require a different processing. Although, the same model was utilised to represent the two kinds of texts, which is the vector space model of unigrams, dissimilar weighting scheme and different feature reduction methods were used. Concerning the weighting scheme, TF-IDF was the measure that allows to achieve better results with long texts. On the contrary, when short texts were the target of the classification, the relative frequency (TF) of each unigram was the most adequate weighting measure.

The problem of the vector space model lies in the number of features. If there is a great amount of unigrams (features), the space could be too big to be processed. Thus, systems usually take advantage of feature reduction methods with the aim of enabling the computational treatment of the space model. As it was said in Chapter 4, the erasing of stopwords and the stemming of the unigrams is not a features reduction method orthodoxly speaking. However, one of the results of the application of those operations is the diminution of the number of unigrams in the vector space model. In Information Retrieval is very common to jointly use the two methods, but in SA is not so common. For example, the work (Pang et al., 2002), which is considered a reference for supervised classification systems, presents a

classification system that does not remove stopwords and does not use a stemming. We thought that the analysis of the convenience of the removing of stopwords and the application of a stemmer was pertinent. The results reached (see Tables 4.4 and 4.10) show that the application of at least one method allows to match the performance achieved by the same system that does not use those methods. Therefore, it is very positive that a vector space model with a less dimensionality achieves a similar result than another with more dimensions.

The analysis reveals that for long text is only positive the removal of the stopwords, meanwhile the application of a stemmer reduces the performance of the classifier. On the other hand, the removal of stopwords in short text is a hindrance for the performance of the system. The stemming of the unigrams of short texts allows to improve the results of the system.

The last conclusion is related to the machine learning algorithm. SVM is the algorithm that reaches better results with the two kind of texts, which is concordant with the SA literature.

The Chapter 5 is concerned on the study of unsupervised methods for the polarity classification of texts written in English and Spanish. The main idea underlying the proposal is the enrichment of the meaning of the concepts of a text with related concepts. To achieve that goal, a modular and multilingual polarity classification system was designed. The main module of the system is the one responsible of the expansion of the meaning of a concept. The expansion module is based on the use of PageRank to retrieve the connected concepts to a given one in a lexical database (WordNet). This idea is similar to the new methodologies based on semantic vector space models (Turney & Pantel, 2010; Mikolov et al., 2013), which attempt to represent the meaning of a word by a vector of features.

The proposal was evaluated on English and Spanish long texts, and with short texts written in English. Firstly, the evaluation with English texts has shown that the incorporation of related concepts is positive for the automatic determination of the polarity of the texts. Despite the good results achieve with English texts, two elements should be study:

1. The retrieval process of similar concepts requires the definition of a context. A context is a set of words that expresses a meaning. In the evaluation, the sentences of the documents were considered as individual contexts. However, it is possible that the meaning of a word can be influenced by the words settled in previous or following sentences. So, the definition of multi-sentences contexts should be studied.

2. SentiWordNet was the source of opinion information. Although

SentiWorNet is one of the most used sentiment lexical database, other sentiment lexicons should be studied, such as Q-WordNet or WordNet-Affect. In Chapter 7 is shown that the combination of classifiers is good for polarity classification, and in the same line, the unsupervised experiments developed with short Spanish texts have demonstrated that the combination of sentiment resources is positive to the identification of the polarity of the documents. Thus, the combination of several lexical resources should be studied, with the aim of improving the representation of the polarity information of each concept.

The good results reached with English texts are not replicated by the Spanish long texts. On our opinion, the main reason to not reach a similar performance is due to the lack of a complete Spanish version of WordNet. As it is indicated in Chapter 5, the Spanish WordNet of MCR was used to accomplished the experiments. The problem of MCR is that only covers the 50% of the synsets of WordNet, so a great amount of Spanish concepts are not taking into account.

The final conclusion drawn from the unsupervised experiments with Spanish short texts is that the combination of several linguistic resources is beneficial for polarity classification, because the combination allow to measure the semantic orientation of a word from different perspectives.

As it has remarked along the present thesis, the lack of linguistic resources for SA in Spanish is an important obstacle. So the main result of Chapter 6 is the generation of iSOL and COAH. iSOL is a list of Spanish opinion bearing words and COAH is a corpus of Spanish reviews in the domain of hotels.

iSOL is the result of the translation of BLOL[1], the manual correction of translation errors and the posterior manual incorporation of Spanish words. The good results reached by iSOL are the proof that the machine translation is a good option for the generation of sentiment resources in other languages than English, such as Spanish, as Banea et al. (2008) indicated.

Moreover, the research carried out in the context of domain adaptation can be also read in Chapter 6. A methodology based on the variance among frequent words in positive and negative reviews was presented. Although, the methodology can be considered straightforward, it allows to append to iSOL common words in reviews of a specific domain. Among the common words are terms that apparently do not express any sentiment or opinion, but in the target domain, those words have a sentimental meaning.

[1] In Chapter 6 the acronym BLOL was defined as the sentiment lexicon developed by Bing Liu.

Therefore, the incorporation of those words gives to iSOL the capacity to identify implicit opinions (see Section 2.3.2). The methodology was assessed in several domains and reached good results in most of them.

Chapter 7 is focused in the application of ensemble classifiers in SA. The main conclusions drawn after the experiments are:

1. The combination of classifiers specialised in the classification of opinions in different languages is positive to the polarity classification of reviews in a specific language, such as Spanish or Arabic.

2. In the literature there are several methods to combine classifiers. In Chapter 7 was studied voting systems and Stacking. The system based on Stacking reached better results than those ones based on voting, but the difference among the results is so small that is difficult to say that Stacking is an ensemble method more appropriate for polarity classification than voting. Thus, we conclude that both methods should be considered for the improvement of the quality of polarity classification systems.

To conclude this section, the most important contributions of the present thesis to the field of SA will be enumerated:

1. The study and development of methods for the classification of Spanish reviews.

2. The demonstration that long and short texts should be treated differently. One of the differences is related to the weighting metric to represent the relevance of unigrams, meanwhile TF-IDF works better in long texts, the relative frequency of the unigrams performs better in short texts.

3. When the reviews are long, and the classifier is supervised, it is advisable to remove the stopwords and not stemming the words. On the other hand, if the texts are short, it is preferable to apply a stemmer to the text and not eliminate the stopwords.

4. The design, the development and the evaluation of a modular, multilingual and unsupervised method for polarity classification based on the extension of the meaning of words of a document. The extension of the meaning enables the aggregation of related concepts to the ones in the text, so the polarity classification module can take advantage from the concepts of the words in the text and the related concepts to the previous ones.

The results reached with the method encourage us to continue studying new linguistic resources for SA, which can improve the performance of the polarity classifier. Furthermore, the method can be integrated in a system based on vector space models, because the related concepts to the ones of the text may be considered as features, so that the subsequent supervised classifier can work with a larger set of features that represent in a better way the meaning of each word.

5. The relevance of the availability of linguistic resources for SA has been mentioned several times in the preceding chapters, so this thesis has also devoted efforts to the generation of linguistic resources for the research on SA in Spanish. In the context of long texts, we have compiled COAH, which is a corpus of Spanish reviews in the hotel domain. For the study of polarity classification techniques for short texts, we have published COST, which is a corpus of Spanish tweets based on noisy labels. Finally, we have followed a dictionary-approach to generate a list of opinion bearing words that is called iSOL.

6. Domain adaptation is a challenge in the context of SA, so the treatment of this problem has also been studied during the development of the underlying research of the present thesis. A lexical method was evaluated, which is oriented to the domain adaptation of list of opinion bearing words.

   The good results obtained with domain-adapted versions of iSOL encourage us to continue researching on domain-adapted word selection methods.

7. The lack of linguistic resources in Spanish can be counteract with the integration of resources in other languages into a polarity classification system. In the present thesis we have shown that the combination by means of the use of ensemble classifiers is a good strategy for the incorporation of the information of linguistic resources in other languages into the classification process in Spanish.

8.2. Future work

The work that will continue this thesis has as goal the combination of the two branches of Natural Language Processing: linguistic and statistical approach. I am convinced that the progress of the computational treatment of Natural Language Processing is in the correctly combination of the two approaches. Thus, in the context of the supervised learning, the study will

be focused on the analysis of the semantic vector space model as model to represent deeply the sentiment meaning of words. This study will be accompanied by an analysis of feature reduction methods with the objective that the semantic vector space model could be executed in a production environment. The research will also pay attention to the study of how to embed linguistic information to the semantic vector space model. One of the goal to reach is the treatment of the negation and intensifiers, which are two very determinant linguistic phenomenon in SA.

I do not want to forget the generation of new sentiment resources. I think that linguistic resources are essential for every task of Natural Language Processing and specially for SA. Thus, I would like to continue improving iSOL by means the integration of different sentiment lexicons and with the incorporation of domain information.

Natural Language Processing is now moving in the direction of the semantic processing of text, and SA is also moving in that way. Currently, that path is starring by the study of semantic concepts, which can be formed by several words. This research line is mainly personalised by Eric Cambria (Cambria & Hussain, 2012), and I would like to adapt and improve their proposal with Spanish texts.

8.3. Relevant publications

The research publications developed during the present thesis are exposed in this section.

8.3.1. Papers on international journals

- **Combining resources to improve unsupervised sentiment analysis at aspect-level**. (2015). Salud M. Jiménez-Zafra, M. Teresa Martín-Valdivia, Eugenio Martínez-Cámara, L. Alfonso Ureña-López. *Journal of Information Science. In press.* ISSN: 0165-5515. DOI: 10.1177/0165551515593686

  Impact factor: 1.158

  Position 58 of 139 in the category Computer Science, Information Systems

- **Language Technologies applied to Document Simplification for Helping Autistic People**. (2015). Eduard Barbu, M. Teresa Martín-Valdivia, Eugenio Martínez-Cámara, L Alfonso Ureña-López. *Expert Systems with Applications 45(12): 5076-5086.* ISSN: 0957-4174. DOI: 10.1016/j.eswa.2015.02.044

Impact factor: 2.240

Position 29 of 123 in the category Computer Science, Artificial Intelligence

- **A Spanish Semantic Orientation Approach to Domain Adaptation for Polarity Classification**. (2015). M. Dolores Molina-González, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López. *Information Processing & Management 51(4):520-531*. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2014.10.002

  Impact factor: 1.265

  Position 53 of 139 in the category Computer Science, Information Systems

- **Polarity Classification for Spanish Tweets Using the COST Corpus**. (2015). Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López, Ruslan Mitkov. *Journal of Information Science 41(3):263-272*. ISSN: 0165-5515.

  DOI: 10.1177/0165551514566564

  Impact factor: 1.158

  Position 58 of 139 in the category Computer Science, Information Systems

- **Integrating Spanish Lexical Resources by Meta-Classifiers for Polarity Classification**. (2014). Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, M. Dolores Molina-González, José M. Perea-Ortega. *Journal of Information Science 40(4):538-554*. ISSN: 0165-5515. DOI: 10.1177/0165551514535710

  Impact factor: 1.158

  Position 58 of 139 in the category Computer Science, Information Systems

- **A knowledge-based Approach for Polarity Classification in Twitter**. (2014). Arturo Montejo-Ráez, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López. *Journal of the Association for Information Science and Technology (previously JASIST) 65(2):414-425*. ISSN: 2330-1635 (1532-2882). DOI: 10.1002/asi.22984

  Impact factor: 1.846

  Position 29 of 139 in the category Computer Science, Information Systems

- **Ranked WordNet Graph for Sentiment Polarity Classification in Twitter**. (2014). Arturo Montejo-Ráez, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López. *Computer Speech & Language 28(1):93-107*. ISSN: 0885-2308. DOI: 10.1016/j.csl.2013.04.001

  Impact factor: 1.753

  Position 47 of 123 in the category Computer Science, Artificial Intelligence

- **Sentiment Analysis in Twitter**. (2014). Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López, Arturo Montejo-Raéz. *Natural Language Engineering 20(1):1-28*. ISSN: 1351-3249. DOI: 10.1017/S1351324912000332

  Impact factor: 0.639

  Position 101 of 123 in the category Computer Science, Artificial Intelligence

- **Semantic Orientation for Polarity Classification in Spanish Reviews**. (2013). M. Dolores Molina-González, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia, José M Perea-Ortega. *Expert Systems with Applications 40(18):7250-7257*. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2013.06.076

  Impact factor: 1.965

  Position 30 of 121 in the category Computer Science, Artificial Intelligence

- **Improving Polarity Classification of Bilingual Parallel Corpora Combining Machine Learning and Semantic Orientation Approaches**. (2013). José M. Perea-Ortega, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López, Eugenio Martínez Cámara. *Journal of the American Society for Information Science and Technology 64(9):1864-1877*. ISSN: 1532-2882. DOI: 10.1002/asi.22884

  Impact factor: 2.230

  Position 17 of 135 in the category Computer Science, Information Systems

- **Sentiment Polarity Detection in Spanish Reviews Combining Supervised and Unsupervised Approaches**. (2013). María-Teresa Martín-Valdivia, Eugenio Martínez-Cámara, Jose-M. Perea-Ortega, L. Alfonso Ureña-López. *Expert Systems with Applications*

40(10):3934-3942. ISSN: 0957-4174 DOI: 10.1016/j.eswa.2012.12.084

Impact factor: 1.965

Position 30 of 121 in the category Computer Science, Artificial Intelligence

8.3.2. Papers on national journals

- **CRiSOL: Base de Conocimiento de Opiniones para el Español**. (2015). M. Dolores Molina González, Eugenio Martínez Cámara, M. Teresa Martín Valdivia. *Procesamiento del Lenguaje Natural, Volume 55, pp. 143-150*. ISSN: 1135-5948.

  SCImago Journal Rankings (SJR): 0.270

  Position 340 of 505 in the category Computer Science Applications

- **TASS 2014-The Challenge of Aspect-based Sentiment Analysis**. (2015). Julio Villena Román, Eugenio Martínez Cámara, Janine García Morera, Salud M Jiménez Zafra. *Procesamiento del Lenguaje Natural, Volume 54, pp. 61-68*. ISSN: 1135-5948.

  SCImago Journal Rankings (SJR): 0.270

  Position 340 of 505 in the category Computer Science Applications

- **Tratamiento de la Negación en el Análisis de Opiniones en Español**. (2015). Salud M. Jiménez Zafra, Eugenio Martínez Cámara, M. Teresa Martín Valdivia, M. Dolores Molina González. *Procesamiento del Lenguaje Natural, Volume 54, pp. 37-44*. ISSN: 1135-5948.

  SCImago Journal Rankings (SJR): 0.270

  Position 340 of 505 in the category Computer Science Applications

- **eSOLHotel: Generación de un lexicón de opinión en español adaptado al dominio turístico**. (2015). M. Dolores Molina González, Eugenio Martínez Cámara, M. Teresa Martín Valdivia, Salud M. Jiménez Zafra. *Procesamiento del Lenguaje Natural, Volume 54, pp. 21-28*. ISSN: 1135-5948.

  SCImago Journal Rankings (SJR): 0.270

  Position 340 of 505 in the category Computer Science Applications

- **Proyecto FIRST (Flexible Interactive Reading Support Tool): Desarrollo de una herramienta para ayudar a personas con autismo mediante la simplificación de textos**. (2014). María-Teresa Martín Valdivia, Eugenio Martínez Cámara, Eduard Barbu, L. Alfonso Ureña López, Paloma Moreda, Elena Lloret. *Procesamiento del Lenguaje Natural, Volume 53, pp. 143-146*. ISSN: 1135-5948.

  SCImago Journal Rankings (SJR): 0.270

  Position 340 of 505 in the category Computer Science Applications

- **TASS-Workshop on Sentiment Analysis at SEPLN**. (2013). Julio Villena-Román, Sara Lana-Serrano, Eugenio Martínez-Cámara, José Carlos González-Cristóbal. *Procesamiento del Lenguaje Natural, Volume 50, pp. 37-44*. ISSN: 1135-5948.

  SCImago Journal Rankings (SJR): 0.235

  Position 365 of 500 in the category Computer Science Applications

- **SINAI en TASS 2012**. (2013). Eugenio Martínez Cámara, Miguel Ángel García Cumbreras , M. Teresa Martín Valdivia , L. Alfonso Ureña López. *Procesamiento del Lenguaje Natural, Volume 50, pp. 53-60*. ISSN: 1135-5948.

  SCImago Journal Rankings (SJR): 0.235

  Position 365 of 500 in the category Computer Science Applications

- **Detección de la Polaridad en Citas Periodísticas: Una Solución No Supervisada**. (2012). Arturo Montejo Ráez, Eugenio Martínez Cámara, M. Teresa Martín Valdivia, L. Alfonso. Ureña López. *Procesamiento del Lenguaje Natural, Volume 49, pp. 149-156*. ISSN: 1135-5948.

  SCImago Journal Rankings (SJR): 0.219

  Position 376 of 493 in the category Computer Science Applications

- **MarUja: Prototipo de Asistente Virtual para la Carta de Servicios del Servicio de Informática de la Universidad de Jaén**. (2011). Eugenio Martínez Cámara, L. Alfonso Ureña López, José M. Perea Ortega. *Procesamiento del Lenguaje Natural, Volume 47, pp. 319-320*. ISSN: 1135-5948.

- **Técnicas de Clasificación de Opiniones Aplicadas a un Corpus en Español**. (2011). Eugenio Martínez Cámara, Valdivia

Martín, M. Teresa, José Manuel Perea Ortega, L. Alfonso Ureña López. *Procesamiento del Lenguaje Natural, Volume 47, pp. 163-170.* ISSN: 1135-5948.

8.3.3. Papers on international conferences

- **Improving Spanish Polarity Classification Combining Different Linguistic Resources**. (2015). Eugenio Martínez-Cámara, Fermín L. Cruz, M. Dolores Molina-González, M. Teresa Martín-Valdivia, F. Javier Ortega, L. Alfonso Ureña-López. *Natural Language Processing and Information Systems Volume 9103 of the series Lecture Notes in Computer Science pp. 234-245.* ISSN: 978-3-319-19580-3. DOI: 10.1007/978-3-319-19581-0_21

  Conference: International Conference on Applications of Natural Language to Information Systems (NLDB).

  Location: Passau, Germany.

- **Cross-Domain Sentiment Analysis Using Spanish Opinionated Words**. (2014). M. Dolores Molina-González, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López. *Natural Language Processing and Information Systems Volume 8455 of the series Lecture Notes in Computer Science pp. 214-219.* ISSN: 978-3-319-07983-7. DOI: 10.1007/978-3-319-07983-7_28

  Conference: International Conference on Applications of Natural Language to Information Systems (NLDB).

  Location: Montpellier, France.

- **Combining Supervised and Unsupervised Polarity Classification for non-English Reviews**. (2013). José M Perea-Ortega, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia, L Alfonso Ureña-López. *Natural Language Processing and Information Systems Volume 7817 of the series Lecture Notes in Computer Science pp. 63-74.* ISSN: 978-3-642-37255-1. DOI: 10.1007/978-3-642-37256-8_6

  Conference: International Conference on Intelligent Text Processing and Computational Linguistics (CICLING).

  Location: Samos, Greece.

- **Opinion Classification Techniques Applied to a Spanish Corpus**. (2011). Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia, L Alfonso Ureña-López. *Natural Language Processing and*

Information Systems Volume 6716 of the series Lecture Notes in Computer Science pp 169-176. ISSN: 978-3-642-22326-6. DOI: 10.1007/978-3-642-22327-3_17

Conference: International Conference on Applications of Natural Language to Information Systems (NLDB).

Location: Alicante, Spain.

- **MarUja: Virtual Assistant Prototype for the Computing Service Catalogue of the University of Jaén**. (2011). Eugenio Martínez-Cámara, L Alfonso Ureña-López, José M Perea-Ortega. *Natural Language Processing and Information Systems Volume 6716 of the series Lecture Notes in Computer Science pp 309-312*. ISSN: 978-3-642-22326-6. DOI: 10.1007/978-3-642-22327-3_45

  Conference: International Conference on Applications of Natural Language to Information Systems (NLDB).

  Location: Alicante, Spain.

8.3.4. Papers on national conferences

- **Detección de la Polaridad de *Tweets* en Español**. (2011). Eugenio Martínez Cámara, Miguel Á. García Cumbreras, M. Teresa Martín Valdivía, L. Alfonso Ureña López. *Actas del Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*

  Conference: Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural.

  Location: Huelva, Spain.

8.3.5. Papers on international workshops

- **Corpus-based Approach for Sentiment Lexicon Domain Adaptation**. (2015). Eugenio Martínez-Cámara, M. Dolores Molina-González, L. Alfonso Ureña-López, M. Teresa Martín-Valdivia. In *Workshop on Replicability and Reproducibility in Natural Language Processing: adaptive methods, resources and software at IJCAI 2015*

  Location: Buenos Aires, Argentina

- **Ensemble Classifier for Twitter Sentiment Analysis**. (2015). Eugenio Martınez-Cámara, Yoan Gutiérrez-Vázquez, Javi Fernández, Arturo Montejo-Ráez, Rafael Munoz-Guillena. In *Proceedings of the Workshop on NLP Applications: Completing the Puzzle co-located*

with the 20th International Conference on Applications of Natural Language to Information Systems (NLDB 2015)

Location: Passau, Germany.

- **SINAI: Syntactic Approach for Aspect Based Sentiment Analysis**. (2015). Salud M. Jiménez-Zafra, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 730-735.* Association for Computational Linguistics

  Location: Denver, Colorado, United States of America.

- **SINAI: Voting System for Twitter Sentiment Analysis**. (2014). Eugenio Martínez-Cámara, Salud M. Jiménez-Zafra, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 572-577.* Association for Computational Linguistics

  Location: Dubling, Ireland.

- **SINAI: Syntactic Approach for Aspect-Based Sentiment Analysis**. (2014). Salud M. Jiménez-Zafra, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pp. 730-735.* Association for Computational Linguistics

  Location: Dubling, Ireland.

- **SINAI: Machine Learning and Emotion of the Crowd for Sentiment Analysis in Microblogs**. (2013). Eugenio Martínez-Cámara, Arturo Montejo-Ráez, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 402-407.* Association for Computational Linguistics

  Location: Atlanta, Georgia, United States of America.

- **Bilingual Experiments on an Opinion Comparable Corpus**. (2013). Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, M. Dolores Molina-González, L. Alfonso Ureña-López. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity,*

Sentiment and Social Media Analysis, pp. 87-93. Association for Computational Linguistics

Location: Atlanta, Georgia, United States of America.

- **Random Walk Weighting over SentiWordNet for Sentiment Polarity Detection on Twitter**. (2012). Arturo Montejo-Ráez, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, pp. 3-10*. Association for Computational Linguistics

  Location: Jeju, South Korea.

8.3.6. Papers on national workshops

- **SINAI-EMMA: Vectores de Palabras para el Análisis de Opiniones en Twitter**. (2015). Eugenio Martínez Cámara, Miguel Ángel García Cumbreras, M. Teresa Martín Valdivia, L. Alfonso Ureña López. In *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN co-located with 31st SEPLN Conference (SEPLN 2015)*

  Location: Alicante, Spain.

- **Desafíos del Análisis de Sentimientos**. (2014). Salud M. Jiménez Zafra, Eugenio Martínez Cámara, M. Teresa Martín Valdivia, L. Alfonso Ureña López. In *Actas de las V Jornadas de la Red en Tratamiento de la Información Multilingüe y Multimodal (TIMM 2014)*

  Location: Cazalla de la Sierra, Sevilla, Spain.

- **SINAI-ESMA: An Unsupervised Approach for Sentiment Analysis in Twitter**. (2014). Salud M. Jiménez Zafra, Eugenio Martínez Cámara, M. Teresa Martín Valdivia, L. Alfonso Ureña López. In *Actas del Taller de Análisis de Sentimientos en la SEPLN, TASS*

  Location: Gerona, Spain.

- **SINAI-EMML: Combinación de Recursos Lingüísticos para el Análisis de la Opinión en Twitter**. (2013). Eugenio Martínez Cámara, Miguel Ángel García Cumbreras, M. Teresa Martín Valdivia, L. Alfonso Ureña López. In *Actas del Taller de Análisis de Sentimientos en la SEPLN, TASS*

Location: Madrid, Spain.

- **SINAI at Twitter-Normalization 2013**. (2013). Arturo Montejo Ráez, Manuel Carlos Díaz Galiano, Eugenio Martínez Cámara, M. Teresa Martí Valdivia, Miguel Ángel García Cumbreras, L. Alfonso Ureña López. In *Proceedings of the Tweet Normalization Workshop co-located with 29th Conference of the Spanish Society for Natural Language Processing (SEPLN 2013)*

  Location: Madrid, Spain.

- **Análisis de Opiniones en la Web 2.0**. (2011). Eugenio Martínez Cámara, L. Alfonso Ureña López M. Teresa Martín Valdivia. In *IX Jornadas Doctorales Andaluzas*

  Location: Mengibar, Jaén, Spain.

- **Análisis de Sentimientos**. (2011). Eugenio Martínez Cámara, L. Alfonso Ureña López, M. Teresa Martín Valdivia. In *IV Jornadas TIMM*

  Location: Torres, Jaén, Spain.

Bibliografía

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12:1–12:34. ISSN 1046-8188. doi:10.1145/1361684.1361685.

Agerri, R. & García-Serrano, A. (2010). Q-wordnet: Extracting polarity from wordnet senses. En N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias, editores, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta. ISBN 2-9517408-6-7.

Agirre, A. G., Laparra, E., Rigau, G., & Donostia, B. C. (2012). Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. En *GWC 2012 6th International Global Wordnet Conference*, página 118.

Agirre, E. & Edmonds, P. (2006). *Word Sense Disambiguation. Algorithms and Applications*, tomo 33 de *Text, Speech and Language Technology*. Springer Netherlands. ISBN 978-1-4020-4808-1.

Agirre, E., López de Lacalle, O., & Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84. doi:10.1162/COLI_a_00164.

Agirre, E. & Soroa, A. (2008). Using the multilingual central repository for graph-based word sense disambiguation. En B. M. Nicoletta Calzolari,

Khalid Choukri & J. Mariani, editores, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA), Marrakech, Morocco. ISBN 2-9517408-4-0.

Agirre, E. & Soroa, A. (2009). Personalizing Pagerank for word sense disambiguation. En *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, páginas 33–41. Association for Computational Linguistics, Stroudsburg, PA, USA.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., & Vossen, P. (2004). The meaning multilingual central repository. En *GWC 2012 6th International Global Wordnet Conference*. Brno: Masaryk University.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. En N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias, editores, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta. ISBN 2-9517408-6-7.

Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., & Belyaeva, J. (2010). Sentiment analysis in the news. En N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias, editores, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta. ISBN 2-9517408-6-7.

Banea, C., Mihalcea, R., & Wiebe, J. (2010). Multilingual subjectivity: Are more languages better? En *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, páginas 28–36. Association for Computational Linguistics, Stroudsburg, PA, USA.

Banea, C., Mihalcea, R., Wiebe, J., & Hassan, S. (2008). Multilingual subjectivity analysis using machine translation. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, páginas 127–135. Association for Computational Linguistics, Stroudsburg, PA, USA.

Banfield, A. (1982). *Unspeakable Sentences*. Routledge and Kegan Paul, Boston.

Bauer, E. & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Mach. Learn.*, 36(1-2):105–139. ISSN 0885-6125. doi:10.1023/A:1007515423169.

Bayes, M. & Price, M. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions*, 53:370–418. doi:10.1098/rstl.1763.0053.

Belkin, N. J., Cool, C., Croft, W. B., & Callan, J. P. (1993). The effect multiple query representations on information retrieval system performance. En *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, páginas 339–346. ACM, New York, NY, USA. ISBN 0-89791-605-0. doi:10.1145/160688.160760.

Benamara, F., Chardon, B., Mathieu, Y., & Popescu, V. (2011). Towards context-based subjectivity analysis. En *Proceedings of 5th International Joint Conference on Natural Language Processing*, páginas 1180–1188. Asian Federation of Natural Language Processing, Chiang Mai, Thailand.

Berger, A. L., Pietra, V. J. D., & Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71. ISSN 0891-2017.

Bestgen, Y., Fairon, C., & Kerves, L. (2004). Un baromètre affectif effectif: Corpus de référence et méthode pour déterminer la valence affective de phrases. En *Journées internationales d'analyse statistique des donnés textuelles (JADT)*, páginas 182–191.

Bifet, A. & Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. En *Proceedings of the 13th International Conference on Discovery Science*, DS'10, páginas 1–15. Springer-Verlag, Berlin, Heidelberg. ISBN 3-642-16183-9, 978-3-642-16183-4.

Boldrini, E., Balahur, A., Martínez-Barco, P., & Montoyo, A. (2010). Emotiblog: A finer-grained and more precise learning of subjectivity expression models. En *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, páginas 1–10. Association for

Computational Linguistics, Stroudsburg, PA, USA. ISBN 978-1-932432-72-5.

Boldrini, E., Balahur, A., Martínez-Barco, P., & Montoyo, A. (2012). Using emotiblog to annotate and analyse subjectivity in the new textual genres. *Data Mining and Knowledge Discovery*, 25(3):603–634. ISSN 1384-5810. doi:10.1007/s10618-012-0259-9.

Breiman, L. (1994). Bagging predictors. Informe Técnico 421, University of California, Berkeley, California, USA.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140. ISSN 0885-6125. doi:10.1023/A:1018054314350.

Brooke, J., Tofiloski, M., & Taboada, M. (2009). Cross-linguistic sentiment analysis: From english to Spanish. En *Proceedings of the International Conference RANLP-2009*, páginas 50–54. Association for Computational Linguistics, Borovets, Bulgaria.

Buhlmann, P. & Yu, B. (2003). Boosting with the l2 loss: Regression and classification. *Journal of the American Statistical Association*, 98:324–339.

Cambria, E. & Hussain, A. (2012). *Sentic Computing*, tomo 2 de *SpringerBriefs in Cognitive Computation*. Springer Netherlands. ISBN 978-94-007-5069-2. doi:10.1007/978-94-007-5070-8.

Cambria, E., Speer, R., Havasi, C., & Hussain, A. (2010). SenticNet: A publicly available semantic resource for opinion mining.

Carbonell, J. G. (1979). *Subjective Understanding: Computer Models of Belief Systems*. Tesis Doctoral, University of Yale.

Carter, D. & Inkpen, D. (2015). Inferring aspect-specific opinion structure in product reviews using co-training. En A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, tomo 9042 de *Lecture Notes in Computer Science*, páginas 225–240. Springer International Publishing. ISBN 978-3-319-18116-5. doi:10.1007/978-3-319-18117-2_17.

Castellucci, G., Croce, D., & Basili, R. (2015). Acquiring a large scale polarity lexicon through unsupervised distributional methods. En C. Biemann, S. Handschuh, A. Freitas, F. Meziane, & E. Métais, editores, *Natural Language Processing and Information Systems*, tomo

9103 de *Lecture Notes in Computer Science*, páginas 73–86. Springer International Publishing. ISBN 978-3-319-19580-3. doi:10.1007/978-3-319-19581-0_6.

Chang, C.-C. & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chatman, S. (1978). *Discourse: Narrative Structure in Fiction and Film*. Cornell University Press, Ithaca, New York, EE.UU.

Chaudhuri, A. (2006). *Emotion and reason in consumer behavior*. Elsevier Butterworth-Heinenmann.

Chenlo, J. M. & Losada, D. E. (2014). An empirical study of sentence features for subjectivity and polarity classification. *Information Sciences*, 280(0):275 – 288. ISSN 0020-0255. doi:http://dx.doi.org/10.1016/j.ins.2014.05.009.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. Prager, New York, USA.

Cohn, D. (1978). *Transparent Minds: Narrative Modes for Representating Consciousness in Fiction*. Princeton University Press, Princeton, New Jersey, EE.UU.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297. ISSN 0885-6125. doi:10.1023/A:1022627411411.

Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27. ISSN 0018-9448. doi:10.1109/TIT.1967.1053964.

Cruz, F. L., Troyano, J. A., Enríquez, F., & Ortega, J. (2008). Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento del Lenguaje Natural*, 41:73–80. ISSN 1989-7553.

Cruz, F. L., Troyano, J. A., Pontes, B., & Ortega, F. J. (2014). Ml-senticon: Un lexicón multilingüe de polaridades semánticas a nivel de lemas. *Procesamiento del Lenguaje Natural*, 53(0):113–120. ISSN 1989-7553.

Dale, R., Somers, H. L., & Moisl, H., editores (2000). *Handbook of Natural Language Processing*. Marcel Dekker, Inc., New York, NY, USA, first edición. ISBN 0824790006.

Das, S. & Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. En *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. En *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, páginas 519–528. ACM, New York, NY, USA. ISBN 1-58113-680-3. doi:10.1145/775152.775226.

del Hoyo, R., Hupont, I., Lacueva, F. J., & Abadía, D. (2009). Hybrid text affect sensing system for emotional language analysis. En *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*, AFFINE '09, páginas 3:1–3:4. ACM, New York, NY, USA. ISBN 978-1-60558-692-2. doi:10.1145/1655260.1655263.

Denecke, K. (2008). Using sentiwordnet for multilingual sentiment analysis. En *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, páginas 507–512. doi:10.1109/ICDEW.2008.4498370.

Deng, L. & Wiebe, J. (2015). Mpqa 3.0: An entity/event-level sentiment corpus. En *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, páginas 1323–1328. Association for Computational Linguistics, Denver, Colorado.

Díaz Galiano, M. C. (2011). *Recuperación de información multimodal basada en integración de conocimiento*. Tesis Doctoral, Universidad de Jaén.

Dietterich, T. G. (2000). Ensemble methods in machine learning. En *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, páginas 1–15. Springer-Verlag, London, UK, UK. ISBN 3-540-67704-6.

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. En *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, páginas 231–240. ACM,

New York, NY, USA. ISBN 978-1-59593-927-2. doi:10.1145/1341531.1341561.

Doležel, L. (1973). *Narrative Modes in Czech Literature*. University of Toronto Press, Toronto, Ontario, Canadá.

Domingos, P. (1996). Using partitioning to speed up specific-to-general rule induction. En *Proceedings of the AAAI-96 Workshop on Integrating Multiple Learned Models*, páginas 29–34. AAAI Press.

Domingos, P. & Hulten, G. (2000). Mining high-speed data streams. En *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, páginas 71–80. ACM, New York, NY, USA. ISBN 1-58113-233-6. doi:10.1145/347090.347107.

Du, W., Tan, S., Cheng, X., & Yun, X. (2010). Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. En *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, páginas 111–120. ACM, New York, NY, USA. ISBN 978-1-60558-889-6. doi:10.1145/1718487.1718502.

Duda, R. O. & Hart, P. E. (1973). *Pattern recognition and scene analysis*. Wiley, New York.

Enríquez de Salamanca Ros, F. (2011). *Combinación de Sistemas mediante Aprendizaje Automático en Tareas de Procesamiento de Lenguaje Natural*. Tesis Doctoral, Universidad de Sevilla.

Esuli, A. & Sebastiani, F. (2006). SentiWordnet: A publicly available lexical resource for opinion mining. En *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA), Genoa, Italia.

Fernández Vítores, D. (2015). *El español: una lengua viva*. Instituto Cervantes.

Fersini, E., Messina, E., & Pozzi, F. (2014). Sentiment analysis: Bayesian ensemble learning. *Decision Support Systems*, 68(0):26–38. ISSN 0167-9236. doi:http://dx.doi.org/10.1016/j.dss.2014.10.004.

Florian, R. (2002). Named entity recognition as a house of cards: Classifier stacking. En *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, páginas 1–4. Association

for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1118853.1118863.

Fodor, J. D. (1979). The linguistic description of opaque contexts. En *Outstanding Dissertations in Linguistics*, tomo 13. Garland, New York and London.

Freund, Y., Mansour, Y., & Schapire, R. E. (2001). Why averaging classifiers can protect against overfitting. En *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*.

Freund, Y. & Schapire, R. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. En P. Vitányi, editor, *Computational Learning Theory*, tomo 904 de *Lecture Notes in Computer Science*, páginas 23–37. Springer Berlin Heidelberg. ISBN 978-3-540-59119-1. doi:10.1007/3-540-59119-2_166.

Gabrilovich, E. & Markovitch, S. (2004). Text categorization with many redundant features: Using aggressive feature selection to make svms competitive with c4.5. En *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, páginas 41–. ACM, New York, NY, USA. ISBN 1-58113-838-5. doi:10.1145/1015330.1015388.

Gale, W., Church, K., & Yarowsky, D. (1992a). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5-6):415–439. ISSN 0010-4817. doi:10.1007/BF00136984.

Gale, W. A., Church, K. W., & Yarowsky, D. (1992b). Work on statistical methods for word sense disambiguation. Informe técnico, Association for the Advancement of Artificial Intelligence, Menlo Park, CA.

Gamon, M. (2004). Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. En *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1220355.1220476.

Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304. doi:10.1198/004017007000000245.

Ghose, A., Ipeirotis, P., & Sundararajan, A. (2007). Opinion mining using econometrics: A case study on reputation systems. En *Proceedings*

of the 45th Annual Meeting of the Association of Computational Linguistics, páginas 416–423. Association for Computational Linguistics, Prague, Czech Republic.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. Informe Técnico CS224N, Stanford University, Stanford, USA.

Greene, S. & Resnik, P. (2009). More than words: Syntactic packaging and implicit sentiment. En *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, páginas 503–511. Association for Computational Linguistics, Stroudsburg, PA, USA. ISBN 978-1-932432-41-1.

Grimmett, G. & Stirzaker, D. (1989). *Probability and Random Processes*. Oxford University Press.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2006). Using KNN model for automatic text categorization. *Soft Computing*, 10(5):423–430. ISSN 1432-7643. doi:10.1007/s00500-005-0503-y.

Hansen, L. & Salamon, P. (1990). Neural network ensembles. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(10):993–1001. ISSN 0162-8828. doi:10.1109/34.58871.

Hassan, A. & Radev, D. (2010). Identifying text polarity using random walks. En *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, páginas 395–403. Association for Computational Linguistics, Stroudsburg, PA, USA.

Hatzivassiloglou, V. & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. En *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, páginas 174–181. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/976909.979640.

Haveliwala, T. H. (2002). Topic-sensitive Pagerank. En *Proceedings of the 11th International Conference on World Wide Web*, WWW '02, páginas 517–526. ACM, New York, NY, USA. ISBN 1-58113-449-5. doi: 10.1145/511446.511513.

Hearst, M. A. (1992). Text-based intelligent systems. Capítulo Direction-based Text Interpretation As an Information Access Refinement, páginas 257–274. L. Erlbaum Associates Inc., Hillsdale, NJ, USA. ISBN 0-8058-1189-3.

Henderson, J. C. & Brill, E. (1999). Exploiting diversity in natural language processing: Combining parsers. En *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, páginas 187–194. Association for Computational Linguistics, Maryland, EE.UU.

Hobbs, J. R. & Riloff, E. (2010). Information Extraction. En N. Indurkhya & F. J. Damerau, editores, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, second edición. ISBN 978-1420085921.

Hu, M. & Liu, B. (2004). Mining and summarizing customer reviews. En *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, páginas 168–177. ACM, New York, NY, USA. ISBN 1-58113-888-1. doi:10.1145/1014052.1014073.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Unsupervised learning. En *An Introduction to Statistical Learning*, tomo 103 de *Springer Texts in Statistics*, Capítulo 10, páginas 373–418. Springer New York. ISBN 978-1-4614-7137-0. doi:10.1007/978-1-4614-7138-7_10.

Jiang, J. & Zhai, C. (2007). Instance weighting for domain adaptation in NLP. En *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, páginas 264–271. Association for Computational Linguistics, Prague, Czech Republic.

Jiménez-Zafra, S. M., Martín-Valdivia, M. T., Martínez-Cámara, E., & Ureña-López, L. A. (2015). Combining resources to improve unsupervised sentiment analysis at aspect-level. *Journal of Information Science*. doi:10.1177/0165551515593686.

Jindal, N. & Liu, B. (2006). Identifying comparative sentences in text documents. En *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, páginas 244–251. ACM, New York, NY, USA. ISBN 1-59593-369-7. doi:10.1145/1148170.1148215.

Jindal, N. & Liu, B. (2008). Opinion spam and analysis. En *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, páginas 219–230. ACM, New York, NY, USA. ISBN 978-1-59593-927-2. doi:10.1145/1341531.1341560.

Joachims, T. (1998). Text categorization with suport vector machines: Learning with many relevant features. En *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, páginas 137–142. Springer-Verlag, London, UK, UK. ISBN 3-540-64417-2.

John, G. H. & Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. En *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, páginas 338–345. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. ISBN 1-55860-385-9.

Kang, H., Yoo, S. J., & Han, D. (2012). Senti-lexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, 39(5):6000–6010. ISSN 0957-4174. doi: http://dx.doi.org/10.1016/j.eswa.2011.11.107.

Kantrowitz, M. (2003). Method and apparatus for analyzing affect and emotion in text.

Keerthi, S., Duan, K., Shevade, S., & Poo, A. (2005). A fast dual algorithm for kernel logistic regression. *Machine Learning*, 61(1-3):151–165. ISSN 0885-6125. doi:10.1007/s10994-005-0768-5.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to platt's smo algorithm for svm classifier design. *Neural Comput.*, 13(3):637–649. ISSN 0899-7667. doi: 10.1162/089976601300014493.

Kim, S.-M. & Hovy, E. (2004). Determining the sentiment of opinions. En *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1220355.1220555.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26. ISSN 1532-2890. doi:10.1002/asi.20961.

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, first edición. ISBN 0471210781.

Kuncheva, L. I. & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207. ISSN 0885-6125. doi:10.1023/A:1022859003006.

Kuroda, S.-Y. (1973). Where epistemology, style and grammar meet: A case study from the japanese. En S. R. Anderson & P. Kiparsky, editores, *A Festschrift for Morris Halle*, páginas 377–391. Holt, Rinehart & Winston, New York, EE.UU.

Kuroda, S.-Y. (1976). Reflections on the foundations of narrative theory - from a linguistic point of view. En T. A. van Dijk, editor, *Pragmatics of Language and Literature*, páginas 107–140. North-Holland, Amsterdam.

Kushner, H. & Yin, G. G. (1997). *Stochastic approximation algorithms and applications*. Springer-Verlag New York, Inc., New York, NY, USA.

Langley, P., Iba, W., and, & Thompson, K. (1992). An analysis of bayesian classifiers. En *Proceedings of the Tenth National Conference on Artificial Intelligence*, AAAI'92, páginas 223–228. AAAI Press. ISBN 0-262-51063-4.

Laws, F. & Schätze, H. (2008). Stopping criteria for active learning of named entity recognition. En *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, páginas 465–472. Association for Computational Linguistics, Stroudsburg, PA, USA. ISBN 978-1-905593-44-6.

Lee, C.-C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9–10):1162 – 1171. ISSN 0167-6393. doi:http://dx.doi.org/10.1016/j.specom.2011.06.004. Sensing Emotion and Affect Facing Realism in Speech Processing.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. En *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, páginas 24–26. ACM, New York, NY, USA. ISBN 0-89791-224-1. doi:10.1145/318723.318728.

Lewis, D. D. (1992). *Representation and Learning in Information Retrieval*. Tesis Doctoral, University of Massachusetts, Amherst, Massachusetts, EE.UU.

Lewis, D. D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. En *Proceedings of the 10th European Conference on Machine Learning*, ECML '98, páginas 4–15. Springer-Verlag, London, UK, UK. ISBN 3-540-64417-2.

Lewis, D. D. & Gale, W. A. (1994). A sequential algorithm for training text classifiers. En *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, páginas 3–12. Springer-Verlag New York, Inc., New York, NY, USA. ISBN 0-387-19889-X.

Lin, C. & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. En *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, páginas 375–384. ACM, New York, NY, USA. ISBN 978-1-60558-512-3. doi:10.1145/1645953.1646003.

Liu, B. (2006–2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. ISBN 3540378812.

Liu, B. (2007). *Web Data Mining*, Capítulo Naïve Bayes Classification, páginas 87–91. Exploring Hyperlinks, Contents, and Usage Data. Springer-Verlag New York, Inc.

Liu, B. (2012a). *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers. ISBN 9781608458844. doi:10.2200/S00416ED1V01Y201204HLT016.

Liu, B. (2012b). *Sentiment Analysis and Opinion Mining*, Capítulo Sentiment Lexicon Generation. Morgan & Claypool Publishers. ISBN 9781608458844.

Loper, E. & Bird, S. (2002). NLTK: The natural language toolkit. En *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, páginas 63–70. Association for Computational Linguistics, Stroudsburg, PA, USA. doi: 10.3115/1118108.1118117.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, páginas 142–150. Association for Computational Linguistics, Stroudsburg, PA, USA. ISBN 978-1-932432-87-9.

Magnini, B. & Cavaglià, G. (2000). Integrating subject field codes into wordnet. En *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*. European Language Resources Association (ELRA), Athens, Greece. ACL Anthology Identifier: L00-1167.

Maks, I., Izquierdo, R., Frontini, F., Agerri, R., Vossen, P., & andoni Azpeitia (2014). Generating polarity lexicons with wordnet propagation in 5 languages. En N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis, editores, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland. ISBN 978-2-9517408-8-4.

Malvar-Fernández, P. & Pichel-Campos, J. R. (2011). Generación semiautomática de recursos de opinion mining para el gallego a partir del portugués y el español. En *Proceedings of the workshop on iberian cross-language natural language processing tasks (ICL 2011)*, páginas 59–63. ISSN 1613-0073.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA. ISBN 0521865719, 9780521865715.

Manning, C. D. & Schütze, H. (1999a). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA. ISBN 0-262-13360-1.

Manning, C. D. & Schütze, H. (1999b). *Foundations of Statistical Natural Language Processing*, Capítulo Text Categorization. MIT Press, Cambridge, MA, USA. ISBN 0-262-13360-1.

Manning, C. D. & Schütze, H. (1999c). *Foundations of Statistical Natural Language Processing*, Capítulo Word Sense Disambiguation. MIT Press, Cambridge, MA, USA. ISBN 0-262-13360-1.

Martín-Valdivia, M.-T., MartíNez-CáMara, E., Perea-Ortega, J.-M., & UreñA-LóPez, L. A. (2013). Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. *Expert Syst. Appl.*, 40(10):3934–3942. ISSN 0957-4174. doi:10.1016/j.eswa.2012.12.084.

Martín Valdivia, M. T., Ortiz Martos, A. J., Ureña López, L. A., & García Cumbreras, M. A. (2005). Detección automática de spam utilizando regresión logística bayesiana. *Procesamiento del Lenguaje Natural*, 35. ISSN 1989-7553.

Martínez Cámara, E., García Cumbreras, M. A., Martín Valdivia, M. T., & Ureña López, L. A. (2013a). SINAI-EMML: Combinación de recursos lingüísticos para el análisis de la opinión en Twitter. En A. Díaz Esteban, I. n. Alegria Loinaz, & J. Villena Román, editores, *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural. SEPLN 2013*, páginas 187–194. Sociedad Española para el Procesamiento del Lenguaje Natural, Daedalus, Grupo de investigación SINAI, Madrid, España.

Martínez Cámara, E., García Cumbreras, M. A., Martín Valdivia, M. T., & Ureña López, L. A. (2013b). SINAI en TASS 2012. *Procesamiento del Lenguaje Natural*, 50:53–60. ISSN 1989-7553.

Martínez-Cámara, E., Jiménez-Zafra, S. M., Martín Valdivia, M. T., & Ureña López, L. A. (2014a). SINAI: Voting system for Twitter sentiment analysis. En *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, páginas 572–577. Association for Computational Linguistics and Dublin City University, Dublin, Ireland.

Martínez-Cámara, E., Martín-Valdivia, M. T., Molina-González, M. D., & Perea-Ortega, J. M. (2014b). Integrating Spanish lexical resources by meta-classifiers for polarity classification. *Journal of Information Science*, 40(4):538–554. doi:10.1177/0165551514535710.

Martínez Cámara, E., Martín Valdivia, M. T., Molina González, M. D., & Ureña López, L. A. (2013c). Bilingual experiments on an opinion comparable corpus. En *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, páginas 87–93. Association for Computational Linguistics, Atlanta, Georgia.

Martínez Cámara, E., Martín Valdivia, M. T., Perea Ortega, J. M., & Ureña López, L. A. (2011a). Técnicas de clasificación de opiniones aplicadas a un corpus en español. *Procesamiento del Lenguaje Natural*, 47(0):163–170. ISSN 1989-7553.

Martínez Cámara, E., Martín-Valdivia, M. T., & Ureña López, L. (2011b). Opinion classification techniques applied to a Spanish corpus. En R. Muñoz, A. Montoyo, & E. Métais, editores, *Natural Language Processing and Information Systems*, tomo 6716 de *Lecture Notes in Computer Science*, páginas 169–176. Springer Berlin Heidelberg. ISBN 978-3-642-22326-6. doi:10.1007/978-3-642-22327-3_17.

Martínez Cámara, E., Martín Valdivia, M. T., & Ureña López, L. A. (2011c). Análisis de sentimientos. En L. A. Ureña López & F. Martínez Santiago, editores, *IV Jornadas TIMM*, páginas 61–63. Red Temática en Tratamiento de la Información Multilingüe y Multimodal, Jaén, España.

Martínez-Cámara, E., Martín-Valdivia, M. T., Ureña-López, L. A., & Mitkov, R. (2015). Polarity classification for Spanish tweets using the COST corpus. *Journal of Information Science*, In press.

Martínez-Cámara, E., Montejo-Ráez, A., Martín-Valdivia, M. T., & Ureña López, L. A. (2013). SINAI: Machine learning and emotion of the crowd for sentiment analysis in microblogs. En *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, páginas 402–407. Association for Computational Linguistics, Atlanta, Georgia, USA.

McCallum, A. & Nigam, K. (1998). A comparison of event models for naïve bayes text classification. En *AAAI-98 Workshop on learning for text categorization*, páginas 41–48. AAAI Press.

Meena, A. & Prabhakar, T. (2007). Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. En G. Amati, C. Carpineto, & G. Romano, editores, *Advances in Information Retrieval*, tomo 4425 de *Lecture Notes in Computer Science*, páginas 573–580. Springer Berlin Heidelberg. ISBN 978-3-540-71494-1. doi:10.1007/978-3-540-71496-5_53.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. En *Proceedings*

of Workshop at International Conference on Learning Representations (ICLR 2013).

Miller, G. A. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–312.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41. ISSN 0001-0782. doi:10.1145/219717.219748.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.

Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M.-T., & Perea-Ortega, J. M. (2013). Semantic orientation for polarity classification in Spanish reviews. *Expert Systems with Applications*, 40(18):7250 – 7257. ISSN 0957-4174. doi:http://dx.doi.org/10.1016/j.eswa.2013.06.076.

Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M. T., & Ure na López, L. A. (2014). A Spanish semantic orientation approach to domain adaptation for polarity classification. *Information Processing & Management*, (0):–. ISSN 0306-4573. doi:http://dx.doi.org/10.1016/j.ipm.2014.10.002.

Molina-González, M., Martínez-Cámara, E., Martín-Valdivia, M., & Ureña-López, L. (2014). Cross-domain sentiment analysis using Spanish opinionated words. En E. Métais, M. Roche, & M. Teisseire, editores, *Natural Language Processing and Information Systems*, tomo 8455 de *Lecture Notes in Computer Science*, páginas 214–219. Springer International Publishing. ISBN 978-3-319-07982-0. doi:10.1007/978-3-319-07983-7_28.

Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., & Ureña López, L. A. (2014). A knowledge-based approach for polarity classification in twitter. *Journal of the Association for Information Science and Technology*, 65(2):414–425. ISSN 2330-1643. doi:10.1002/asi.22984.

Montejo Ráez, A., Martínez Cámara, E., Martín Valdivia, M. T., & Ureña López, L. A. (2012). Detección de la polaridad en citas periodísticas: una solución no supervisada. *Procesamiento del Lenguaje Natural*, 49:149–156. ISSN 1989-7553.

Mooney, R. J. & Bunescu, R. (2005). Mining knowledge from text using information extraction. *SIGKDD Explor. Newsl.*, 7(1):3–10. ISSN 1931-0145. doi:10.1145/1089815.1089817.

Muthukrishnan, S. (2005). Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2):117–236. ISSN 1551-305X. doi:10.1561/0400000002.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., & Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter. En *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, páginas 312–320. Association for Computational Linguistics, Atlanta, Georgia, USA.

Nasukawa, T. & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. En *Proceedings of the 2Nd International Conference on Knowledge Capture*, K-CAP '03, páginas 70–77. ACM, New York, NY, USA. ISBN 1-58113-583-1. doi:10.1145/945645.945658.

Ng, H. T., Wang, B., & Chan, Y. S. (2003). Exploiting parallel texts for word sense disambiguation: An empirical study. En *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, páginas 455–462. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1075096.1075154.

O'Really, T. (2005). What is web 2.0. Design patterns and business models for the next generation of software. `http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html`. Consultado: 09/03/2015.

Padró, L. & Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. En N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis, editores, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey. ISBN 978-2-9517408-7-7.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The Pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Paltoglou, G. & Thelwall, M. (2010). A study of information retrieval weighting schemes for sentiment analysis. En *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, páginas 1386–1395. Association for Computational Linguistics, Stroudsburg, PA, USA.

Pang, B. & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. En *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1218955.1218990.

Pang, B. & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. En *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, páginas 115–124. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1219840.1219855.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135. ISSN 1554-0669. doi:10.1561/1500000011.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. En *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, páginas 79–86. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1118693.1118704.

Pantel, P. & Lin, D. (2002). Discovering word senses from text. En *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, páginas 613–619. ACM, New York, NY, USA. ISBN 1-58113-567-X. doi:10.1145/775047.775138.

Parrott, W. G. (2001). *Emotions in social psychology: Essential readings*. Psychology Press. ISBN 0863776825.

Pedersen, T. (2000). A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation. En *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, páginas 63–69. Association for Computational Linguistics, Stroudsburg, PA, USA.

Perea-Ortega, J. M., Martín-Valdivia, M. T., Ureña-López, L. A., & Martínez-Cámara, E. (2013). Improving polarity classification of bilingual parallel corpora combining machine learning and semantic orientation approaches. *Journal of the American Society for Information Science and Technology*, 64(9):1864–1877. ISSN 1532-2882. doi:10.1002/asi.22884.

Pérez-Rosas, V., Banea, C., & Mihalcea, R. (2012). Learning sentiment lexicons in Spanish. En N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis, editores, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey. ISBN 978-2-9517408-7-7.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. Informe Técnico MSR-TR-98-14, Microsoft Research.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). Semeval-2015 task 12: Aspect based sentiment analysis. En *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, páginas 486–495. Association for Computational Linguistics, Denver, Colorado.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). Semeval-2014 task 4: Aspect based sentiment analysis. En *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, páginas 27–35. Association for Computational Linguistics and Dublin City University, Dublin, Ireland.

Procter, P. (1978). *Longman Dictionary of Contemporary English*. Longman Group Ltd., Harlow, United Kingdon.

Quinlan, J. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106. ISSN 0885-6125. doi:10.1023/A:1022643204877.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. ISBN 1-55860-238-0.

Quinlan, J. R. (1996). Bagging, boosting, and c4.s. En *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 1*, AAAI'96, páginas 725–730. AAAI Press. ISBN 0-262-51091-X.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman, London.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. En *Proceedings of*

the ACL Student Research Workshop, ACLstudent '05, páginas 43–48. Association for Computational Linguistics, Stroudsburg, PA, USA.

Rijsbergen, C. J. V. (1979). *Information Retrieval*, Capítulo Chapter 7. Butterworth-Heinemann, Newton, MA, USA, 2nd edición. ISBN 0408709294.

Rokach, L. (2005). Ensemble methods for classifiers. En O. Maimon & L. Rokach, editores, *Data Mining and Knowledge Discovery Handbook*, páginas 957–980. Springer US. ISBN 978-0-387-24435-8. doi:10.1007/0-387-25465-X_45.

Rushdi Saleh, M., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña López, L. A. (2011). Experiments with SVM to classify opinions in different domains. *Expert Systems with Applications*, 38(12):14799 – 14804. ISSN 0957-4174. doi:http://dx.doi.org/10.1016/j.eswa.2011.05.070.

Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña López, L. A., & Perea-Ortega, J. M. (2011a). Bilingual experiments with an Arabic-English corpus for opinion mining. En *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, páginas 740–745. RANLP 2011 Organising Committee, Hissar, Bulgaria.

Rushdi-Saleh, M., Martín-Valdivia, M. T., Ureña López, L. A., & Perea-Ortega, J. M. (2011b). OCA: Opinion corpus for arabic. *J. Am. Soc. Inf. Sci. Technol.*, 62(10):2045–2054. ISSN 1532-2882. doi:10.1002/asi.21598.

Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A bayesian approach to filtering junk e-mail. En *Learning for Text Categorization: Papers from the 1998 workshop*, tomo 62, páginas 98–105.

Saif, H., He, Y., & Alani, H. (2012). Alleviating data sparsity for Twitter sentiment analysis. En *Making Sense of Microposts (#MSM2012)*, páginas 2–9.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620. ISSN 0001-0782. doi:10.1145/361219.361220.

Salton, G. & Yang, C. (1973). On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4):351–372. doi:10.1108/eb026562.

Saralegi, X. & San Vicente, I. (2013). Elhuyar at tass 2013. En *Proceedings of "XXIX Congreso de la Sociedad Española de Procesamiento del lenguaje natural". Workshop on Sentiment Analysis at SEPLN (TASS2013)*. Madrid, España. ISBN 978-84-695-8349-4.

Savoy, J. & Gaussier, E. (2010). Information retrieval. En N. Indurkhya & F. J. Damerau, editores, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, second edición. ISBN 978-1420085921.

Schapire, R. (1990). The strength of weak learnability. *Machine Learning*, 5(2):197–227. ISSN 0885-6125. doi:10.1007/BF00116037.

Schütze, H. (1992a). Context space. Informe técnico, Association for the Advancement of Artificial Intelligence, Menlo Park, CA.

Schütze, H. (1992b). Dimensions of meaning. En *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, Supercomputing '92, páginas 787–796. IEEE Computer Society Press, Los Alamitos, CA, USA. ISBN 0-8186-2630-5.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistic*, 24(1):97–123. ISSN 0891-2017.

Shapley, L. & Grofman, B. (1984). Optimizing group judgmental accuracy in the presence of interdependencies. *Public Choice*, 43(3):329–343. ISSN 0048-5829. doi:10.1007/BF00118940.

Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A., & Gordon, J. (2013). Empirical study of machine learning based approach for opinion mining in tweets. En I. Batyrshin & M. González Mendoza, editores, *Advances in Artificial Intelligence*, tomo 7629 de *Lecture Notes in Computer Science*, páginas 1–14. Springer Berlin Heidelberg. ISBN 978-3-642-37806-5. doi:10.1007/978-3-642-37807-2_1.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21. doi:10.1108/eb026526.

Speriosu, M., Sudan, N., Upadhyay, S., & Baldridge, J. (2011). Twitter polarity classification with label propagation over lexical links and the follower graph. En *Proceedings of the First Workshop on*

Unsupervised Learning in NLP, EMNLP '11, páginas 53–63. Association for Computational Linguistics, Stroudsburg, PA, USA. ISBN 978-1-937284-13-8.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.

Strapparava, C. & Valitutti, A. (2004). Wordnet Affect: an affective extension of wordnet. En *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA).

Taboada, M. & Grieve, J. (2004). Analyzing appraisal automatically. En *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text Stanford*, páginas 158–161.

Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770. doi: 10.1080/01621459.2012.734168.

Tan, S. & Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. *Expert Systems with Applications*, 34(4):2622 – 2629. ISSN 0957-4174. doi:http://dx.doi.org/10.1016/j.eswa.2007.05.028.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

Tong, R. M. (2001). An operational system for detecting and tracking opinions in on-line discussion. En *Proceedings of ACM SIGIR 2001 Workshop on Operational Text Classification (OTC)*.

Tukey, J. W. (1977). Exploratory data analysis. *Reading, Ma*, 231:32.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. En *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, páginas 417–424. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1073083.1073153.

Turney, P. D. & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346. ISSN 1046-8188. doi:10.1145/944012.944013.

Turney, P. D. & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188. ISSN 1076-9757.

Uspensky, B. (1973). *A Poetics of Composition*. University of California Press, Berkeley, Los Ángeles, EE.UU.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. ISBN 0387907335.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA. ISBN 0-387-94559-8.

Vargas Llosa, M. (2013). *La civilización del espectáculo*. Santillana Ediciones Generales S.L.

Villena-Román, J., Lana-Serrano, S., Martínez-Cámara, E., & González-Cristóbal, J. C. (2013). TASS - workshop on sentiment analysis at SEPLN. *Procesamiento del Lenguaje Natural*, 50(0):37–44. ISSN 1989-7553.

Villena Román, J., García Morera, J., García Cumbreras, M. A., Martínez Cámara, E., Martín Valdivia, M. T., & Ureña López, L. A., editores (2015a). *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN*, número 1397 en CEUR Workshop Proceedings. Aachen. ISSN 1613-0073.

Villena-Román, J., García-Morera, J., Lana-Serrano, S., & González-Cristóbal, J. C. (2014). TASS 2013 - a second step in reputation analysis in Spanish. *Procesamiento del Lenguaje Natural*, 52(0):37–44. ISSN 1989-7553.

Villena Román, J., Martínez Cámara, E., García Morera, J., & Jiménez Zafra, S. M. (2015b). TASS 2014 - the challenge of aspect-based sentiment analysis. *Procesamiento del Lenguaje Natural*, 54(0):61–68. ISSN 1989-7553.

Vossen, P., Bloksma, L., Climent, S., Martí, M. A., Taulé, M., J., G., Chugur, I., Verdejo, F., Escudero, G., Rigau, G., Rodríguez, H., Alonge, A., Bertagna, F., Marianelli, R., Roventini, A., & Tarasi, L. (1998). EuroWordnet subset2 for Dutch, Spanish and Italian. Informe Técnico EuroWordNet Deliverables LE-4003 D027 y D028, University of Amsterdam.

Walker, D. & Amsler, R. (1986). The use of machine-readable dictionaries in sublanguage analysis. *Analyzing Language in Restricted Domains*, páginas 69–83.

Wasserman, S. & Faust, K. (1994). *Social Network Analysis. Methods and Applications*. Cambridge University Press, Cambridge, United Kingdom. ISBN 9780521387071.

Wiebe, J. (1990). *Recognizing Subjective Sentences: A Computational Investigation of Narrative Text*. Tesis Doctoral, State University of New York, Buffalo, EE.UU.

Wiebe, J. & Bruce, R. (1995). Probabilistic classifiers for tracking point of view. En *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, páginas 181–187.

Wiebe, J. & Rapaport, W. (1986). Representing de re and de dicto belief reports in discourse and narrative. *Proceedings of the IEEE*, 74(10):1405–1413. ISSN 0018-9219. doi:10.1109/PROC.1986.13641.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistic*, 30(3):277–308. ISSN 0891-2017. doi:10.1162/0891201041850885.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210. ISSN 1574-020X. doi:10.1007/s10579-005-7880-9.

Wiebe, J. M. (1994). Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287. ISSN 0891-2017.

Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. En *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, páginas 246–253. Association for Computational Linguistics, Stroudsburg, PA, USA. ISBN 1-55860-609-3. doi:10.3115/1034678.1034721.

Wiebe, J. M. & Rapaport, W. J. (1988). A computational theory of perspective and reference in narrative. En *Proceedings of the 26th Annual Meeting on Association for Computational Linguistics*, ACL '88, páginas 131–138. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/982023.982039.

Wiebe, J. M. & Rapaport, W. J. (1991). References in narrative text. *Noûs*, 25(4):457–486.

Wilks, Y. & Bien, J. (1984). Beliefs, points of view and multiple environments. En *Proc. Of the International NATO Symposium on Artificial and Human Intelligence*, páginas 147–171. Elsevier North-Holland, Inc., New York, NY, USA. ISBN 0-444-86545-4.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005a). OpinionFinder: A system for subjectivity analysis. En *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, páginas 34–35. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1225733.1225751.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005b). Recognizing contextual polarity in phrase-level sentiment analysis. En *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, páginas 347–354. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1220575.1220619.

Wilson, T., Wiebe, J., & Hwa, R. (2004). Just how mad are you? finding strong and weak opinion clauses. En *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, páginas 761–767. AAAI Press. ISBN 0-262-51183-5.

Wilson, T. A. (2008). *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. Tesis Doctoral, University of Pittsburgh, Pittsburgh, EE.UU.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259. ISSN 0893-6080. doi:10.1016/S0893-6080(05)80023-1.

Xia, R., Zong, C., Hu, X., & Cambria, E. (2013). Feature ensemble plus sample selection: Domain adaptation for sentiment classification. *Intelligent Systems, IEEE*, 28(3):10–18. ISSN 1541-1672. doi:10.1109/MIS.2013.27.

Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181(6):1138 – 1152. ISSN 0020-0255. doi:http://dx.doi.org/10.1016/j.ins.2010.11.023.

Xiao-fei, Z., He-yan, H., & Ke-liang, Z. (2009). KNN text categorization algorithm based on semantic centre. En *Information Technology and Computer Science, 2009. ITCS 2009. International Conference on*, tomo 1, páginas 249–252. doi:10.1109/ITCS.2009.57.

Xu, L., Krzyzak, A., & Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE transactions on Systems, Man and Cybernetics*, 22(3):418–435.

Yang, Y. (1994). Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. En *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, páginas 13–22. Springer-Verlag New York, Inc., New York, NY, USA. ISBN 0-387-19889-X.

Yang, Y. (1995). Noise reduction in a statistical approach to text categorization. En *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, páginas 256–263. ACM, New York, NY, USA. ISBN 0-89791-714-6. doi:10.1145/215206.215367.

Yarowsky, D. (2010). Word sense disambiguation. En N. Indurkhya & F. J. Damerau, editores, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, second edición. ISBN 978-1420085921.

Yule, G. U. (1900). On the association of attributes in statistics: With illustrations from the material of the childhood society, &c. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 194:257–319. ISSN 02643952.

Zhang, L. & Liu, B. (2011). Identifying noun product features that imply opinions. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, páginas 575–580. Association for Computational Linguistics, Stroudsburg, PA, USA. ISBN 978-1-932432-88-6.

Zhang, P. & He, Z. (2015). Using data-driven feature enrichment of text representation and ensemble technique for sentence-level polarity classification. *Journal of Information Science*, 41(4):531–549. doi: 10.1177/0165551515585264.

Zhang, Y., Ji, D.-H., Su, Y., & Wu, H. (2013). Joint Naïve Bayes and lda for unsupervised sentiment analysis. En J. Pei, V. Tseng, L. Cao, H. Motoda, & G. Xu, editores, *Advances in Knowledge Discovery and Data Mining*, tomo 7818 de *Lecture Notes in Computer Science*, páginas 402–413. Springer Berlin Heidelberg. ISBN 978-3-642-37452-4. doi: 10.1007/978-3-642-37453-1_33.

Zhu, X. & Ghahramani, Z. (2002). Learning from labeled and unlabeled data with label propagation. Informe Técnico CMU-CALD-02-107, Carnegie Mellon University, Pittsburgh, PA, EE.UU.


VNIVERSITAS
GIENNENSIS