Universidad Politécnica de Madrid

Tesis Fin de Master

Expansión Supervisada de Léxicos Polarizados Adaptable al Contexto

Autor:
Eduardo C. Garrido Merchán

Supervisor:
Jesús Cardeñosa Lera

Grupo de Validación y Aplicaciones Industriales
Departamento de Inteligencia Artificial

19 de julio de 2015

Declaración de Autoría

Yo, Eduardo C. Garrido Merchán, declaro que esta tesis, titulada, 'Expansión Supervisada de Léxicos Polarizados Adaptable al Contexto' y el trabajo aquí presentado son mi propiedad. Yo confirmo que:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:
Eduardo César Garrido Merchán

Date:
19 de Junio de 2015

Dedicatoria

El valiente no es quién no siente miedo, sino aquel que conquista ese miedo.

Nelson Mandela.

Agradecimientos

UNIVERSIDAD POLITECNICA DE MADRID

Resumen

Escuela Técnica Superior de Ingenieros Informáticos
Departamento de Inteligencia Artificial

Master en Inteligencia Artificial

Expansión Supervisada de Léxicos Polarizados Adaptable al Contexto

by Eduardo C. Garrido Merchán

El análisis de opiniones es un área en la cual múltiples disciplinas han otorgado diferentes enfoques para elaborar modelos que sean capaces de extraer la polaridad de los textos analizados. En función del dominio o categoría del texto analizado, donde ejemplos de categorías son Deportes o Banca, estos modelos deben ser modificados para obtener un análisis de opinión de calidad. En esta tesis se presenta un modelo que pretende elaborar un análisis de opiniones independiente de la categoría a analizar y un extenso estado del arte sobre análisis de opiniones. Se propone un enfoque cuantitativo que hará uso de un léxico polarizado semilla como único recurso cualitativo del modelo. El enfoque propuesto hace uso de un corpus anotado de textos por polaridad y categoría y el léxico polarizado semilla para producir un modelo capaz de elaborar un análisis de opinión de calidad en las distintas categorías analizadas y expandir el léxico polarizado semilla con términos que se adecúan a las categorías procesadas.

UNIVERSIDAD POLITECNICA DE MADRID

Abstract

Escuela Técnica Superior de Ingenieros Informáticos
Departamento de Inteligencia Artificial

Master en Inteligencia Artificial

Expansión Supervisada de Léxicos Polarizados Adaptable al Contexto

by Eduardo C. Garrido Merchán

Sentiment analysis is an area in which multiple disciplines have given different approaches to make models that are able to extract the polarity of the analyzed texts. Depending on the domain or category of the analyzed text, where examples of categories are Sports or Banking, these models should be modified to obtain a good opinion analysis. This thesis presents a model that aims to develop a category independent opinion analysis model and a extensive sentiment analysis state of the art. A quantitative approach is proposed that will use a polarized lexicon as the only qualitative resource. The proposed approach uses an annotated corpus by polarity and category and a polarized lexicon seed to produce a model able to develop a good opinion analysis in the various categories analyzed and to expand the polarized lexicon seed with terms that fit the processed categories.

Índice general

Declaración de Autoría I

Resumen IV

Abstract V

Tabla de Contenidos VI

Listado de Figuras IX

Listado de Tablas X

1. Introducción 1

2. Estado del Arte 4
2.1. Introducción . . . 4
2.2. Métodos Tradicionales . . . 10
2.2.1. Introducción y Categorización de textos . . . 10
2.2.2. Psicología . . . 11
2.2.3. Heurísticas y algorítmica clásica . . . 12
2.2.4. Semántica y métodos avanzados . . . 17
2.3. Recursos Lingüísticos . . . 22
2.3.1. Introducción . . . 22
2.3.2. General Inquirer . . . 23
2.3.3. SenticNet . . . 24
2.3.4. Wordnet . . . 24
2.3.5. SentiWordNet . . . 25

2.3.6. LIWC: Linguistic Inquiry and Word Count . . . 26
2.3.7. MPQA Subjectivity Cues Lexicon . . . 26
2.3.8. EffectWordNet . . . 27
2.3.9. ConceptNet . . . 28
2.3.10. WordnetAffect . . . 28
2.3.11. Emotinet . . . 29
2.3.12. Bing Liu Opinion Lexicon . . . 29
2.3.13. Conclusiones . . . 30
2.4. Creación y expansión de léxicos polarizados y recursos . . . 31
2.4.1. Enfoques basados en Wordnet . . . 31
2.4.2. Enfoques basados en reglas y métodos de aprendizaje no supervisado . . . 34
2.4.3. Enfoques basados en modelos de optimización . . . 37
2.5. Métodos basados en Estadística y Aprendizaje Automático . . . 41
2.5.1. Introducción . . . 41
2.5.2. Características a extraer de los Conjuntos de Textos . . . 42
2.5.3. Métodos basados en Aprendizaje Supervisado . . . 50
2.5.3.1. Modelos empleados en Análisis de Opiniones . . . 50
2.5.3.2. Modelos de aprendizaje supervisado multietiqueta . . . 57
2.5.4. Métodos estadísticos y basados en Aprendizaje No Supervisado 58
2.5.4.1. Métodos Estadísticos . . . 59
2.5.4.2. Clustering . . . 61
2.6. Problemáticas específicas en Análisis de Opiniones . . . 65
2.6.1. Multilingüismo . . . 65
2.6.2. Negación . . . 67
2.6.3. Independencia de Contexto . . . 69
2.6.4. Extracción de Argumentos . . . 71
2.7. Tendencias explicadas por el Análisis de Opiniones . . . 72
2.8. Conclusiones . . . 75

3. Modelo **78**
3.1. Introducción y Motivación . . . 78
3.2. Trabajo previo y bases teóricas sobre las que se sustenta el Modelo . 82
3.3. Hipótesis . . . 85
3.4. Asunciones . . . 86
3.5. Restricciones y Limitaciones . . . 87
3.6. Objetivos . . . 88
3.7. Modelo Propuesto . . . 89

4. Experimentación **105**
4.1. Recursos Utilizados en los Experimentos . . . 106
4.1.1. Corpus de textos clasificados por categoría y anotados por polaridad . . . 106
4.1.2. Léxicos Polarizados Semilla . . . 107
4.1.3. Modelos de Aprendizaje Automático . . . 108
4.2. Diseño y Descripción de los Experimentos . . . 108
4.2.1. Experimento I . . . 109
4.2.2. Experimento II. Descripción . . . 110
4.2.3. Experimento III. Descripción . . . 112
4.2.4. Experimento IV. Descripción . . . 113
4.2.5. Experimento V. Descripción . . . 114
4.2.6. Experimento VI. Descripción . . . 116

5. Evaluación **119**
5.1. Introducción . . . 119
5.2. Experimentos . . . 119
5.2.1. Experimento I . . . 119
5.2.2. Experimento II . . . 122
5.2.3. Experimento III . . . 124
5.2.4. Experimento IV . . . 125
5.2.5. Experimento V. Descripción . . . 126
5.2.6. Experimento VI. Descripción . . . 128

6. Conclusiones **131**
6.1. Conclusiones . . . 131

7. Líneas Futuras **134**
7.1. Líneas Futuras . . . 134

Índice de figuras

2.1. Patrón de Flujo de Opinión. En este caso se modeliza que si la última parte tiene polaridad negativa la polaridad final del texto es negativa. 20
2.2. Porcentaje de acuerdo entre los términos polarizados de distintos recursos lingüísticos. 30
2.3. Cuatro formas de obtener *bad* desde *good* en tres saltos. 33
2.4. Resultados del procedimiento de expansión de léxico de Zhang [1]. . . 36
2.5. Hiperplano óptimo de separación creado por la L-SVM [2]. 51
2.6. Función de separación creada creado por la Soft-SVM con Kernel Gaussiano [3]. 54
2.7. Grafo construido por Chen [4] para la expansión de un léxico polarizado a un léxico polarizado multilingüe. 65
2.8. Serie temporal que refleja la positividad media de los dos candidatos a presidente en EEUU durante 2008. 73
2.9. Series temporales del Dow Jones (azul) y de uno de los rasgos del sistema de Bollen [5] (rojo). 74

3.1. Características de los distintos tipos de modelos aplicables en el análisis de opiniones. 80
3.2. Algoritmo de expansión del léxico y construcción de prototipos de modelos de análisis de opiniones. 103

5.1. Evolución de los resultados. 120
5.2. Evolución de los resultados. 123
5.3. Evolución de los resultados. 124
5.4. Evolución de los resultados. 125
5.5. Evolución de los resultados. 127
5.6. Evolución de los resultados. 127
5.7. Evolución de los resultados. 128
5.8. Evolución de los resultados. 129

Capítulo 1

Introducción

La lectura e interpretación automática de textos es una tarea de alta dificultad por parte de las máquinas. Esto es debido a que el lenguaje es ambiguo y una frase puede tener diferentes interpretaciones. Por ejemplo, considérese la frase: *El monitor me ha parecido muy bien*. En esta frase la palabra monitor puede tener dos significados: Monitor de gimnasio o pantalla. Sin información adicional, es imposible conocer cuál es el significado de monitor en esa frase. Pese a ello, el área conocida como procesamiento del lenguaje natural ha propuesto modelos para que las máquinas procesen el lenguaje.

El volumen de información a procesar ha crecido exponencialmente en los últimos años, con el crecimiento de internet y la aparición de las redes sociales. En los textos contenidos en estos medios, se encuentra información útil que puede explicar tendencias o ser tomada en cuenta como soporte a distintas decisiones. Una tendencia que puede ser explicada procesando texto en blogs o redes sociales es, por ejemplo, determinar si la campaña de marketing que se ha hecho sobre un producto ha sido o no efectiva.

Los textos que se procesan pueden tratar de cualquier tema o ser escritos con diferente estilo. Por ejemplo, el estilo empleado en Twitter no es el mismo que se emplea en un artículo periodístico. Los modelos que operan sobre lenguaje natural lidian con otros problemas como faltas de ortografía y errores sintácticos, multilingüismo, variedades geográficas, etc. Todo ello hace que cualquier problema perteneciente al procesamiento del lenguaje natural sea complejo. Aún así, se han conseguido resolver algunos de los problemas pertenecentes a este campo de forma existosa. Este progreso ha motivado a los investigadores del campo a ir afrontando problemas más avanzados.

Algunos de los problemas mencionados requieren analizar textos escritos en redes sociales. En las redes sociales, el lenguaje expresado por los usuarios no se limita a la descripción de hechos. Cuando un usuario se refiere a un producto, lo puede hacer desde una perspectiva objetiva o subjetiva. Por ejemplo, considérese el texto: *El coche de modelo X tiene una potencia de 80 CV*. Este texto expresa una característica del coche que es universal y que no depende del juicio que una persona hace sobre el coche. Se dice que este tipo de textos son objetivos, son relativos a un objeto en sí y no al modo de pensar o sentir de una persona en particular.

Por otro lado, considérese la frase: *El coche de modelo X tiene una potencia muy baja, no me parece que un coche así merezca la pena.* Al igual que en el texto anterior, el texto trata sobre una cualidad del coche: la potencia. Pero al contrario que en la primera frase, este texto denota un modo de pensar o sentir, en él el autor expresa una opinión. Se dice que este texto es subjetivo, identifica algo como propio de la forma de pensar o sentir de una persona.

Existen dos grandes categorías dentro de la forma de pensar o sentir de una persona. En el anterior texto, por ejemplo, la forma de pensar o sentir del autor es claramente negativa. En el texto presentado, el autor expone que la potencia del coche le parece bajísima y que por ello no merece la pena. Estas expresiones dotan al texto de lo que se conoce como una polaridad negativa. Se define por tanto el concepto de polaridad como la condición de lo que tiene propiedades o potencias opuestas. En este caso, la subjetividad expresada por el autor en el texto puede ser positiva o negativa.

Esta subjetividad que el autor expresa en un texto normalmente es conocida como opinión: Idea, juicio o concepto que se tiene sobre alguien o algo. Por tanto, para explicar si la campaña de marketing sobre un producto ha sido efectiva o no, se requiere un modelo que reconozca de entre los textos que se quieren analizar cuáles de ellos expresan una opinión y cuáles no. Una vez se ha efectuado esta tarea, se necesita además categorizar la polaridad del texto. El objetivo final del modelo es comprobar si la proporción de opiniones categorizadas como positivas es mayor que antes de ser lanzada la campaña de marketing.

A este tipo de modelos se los conoce como modelos de análisis de opiniones. Dada la alta complejidad que requieren estos modelos de procesamiento del lenguaje natural por lidiar con textos subjetivos, hasta la última década no existía prácticamente ningún modelo que efectuase análisis de opiniones de forma efectiva.

Actualmente, existen varios modelos de análisis de opiniones que resuelven la tarea de analizar opiniones eficientemente. No obstante, como se detallará en el estado

del arte, cada uno de los modelos de análisis de opiniones posee ventajas e inconvenientes. No existe un modelo de análisis de opiniones que sea el mejor en todos los dominios a analizar. Por ejemplo, un modelo determinado de análisis de opiniones es muy eficaz analizando opiniones de textos periodísticos pero no de críticas de cine. A este problema se le denomina dependencia del contexto o el dominio en el análisis de opiniones. Debido a que en cada tipo o dominio de textos el vocabulario y las expresiones utilizadas son muy diferentes, los modelos de análisis de opiniones fracasan en obtener resultados similares en dominios distintos. La implicación directa de esta problemática es que se necesitarán expertos en cada uno de los dominios a analizar para obtener modelos de análisis de opiniones.

En esta tesis fin de máster, se propondrá un modelo para la construcción de prototipos de modelos de análisis de opiniones capaces de obtener resultados aceptables en cualquier categoría. Para ello, se emplearán léxicos polarizados semilla, que se definirán en el estado del arte, y se modificarán mediante la técnica propuesta en esta tesis. Al final, se obtendrán léxicos polarizados generalizables a cualquier categoría de las consideradas. Estos léxicos no producirán las mismas medidas de evaluación que modelos de análisis de opiniones diseñados por expertos en cada una de las categorías a considerar. Pero, el modelo resultante, es general y obtiene resultados aceptables para un número indeterminado de categorías a analizar. En situaciones en las que se tiene un gran volumen de textos pertenecientes a un número indeterminado de categorías, este modelo se diferencia del resto en que, hasta que los expertos consigan analizar todas las categorías, se tendrá un modelo que analiza opiniones con efectividad en todas ellas. Dado que el número de categorías puede ser enorme, este modelo tiene el valor añadido de proporcionar una solución aceptable hasta que sean analizadas. A su vez, este modelo obtiene unos rasgos que pueden ser utilizados como punto de partida de los expertos, para que, cuando sea conveniente, se tenga la necesidad de construir un modelo más específico para cada una de las categorías.

Como ya se ha comentado, se necesita una base teórica multidisciplinar para lidiar con el análisis de opiniones. Por ello, a continuación, se presenta un extenso estado del arte sobre análisis de opiniones. En este estado del arte se comentarán, además de las técnicas en las que se basa el modelo más tarde propuesto, otros enfoques de áreas completamente distintas. Cada uno de los enfoques es analizado, explicando sus peculiaridades, y comentado, exponiendo desde una perspectiva personal las características del mismo. Se piensa que es importante revisar todas las áreas que estudian el análisis de opiniones, puesto que, únicamente desde el conocimiento de todas, se está en una posición capaz de proponer un modelo que no ha sido propuesto anteriormente por ningún autor, capaz de resolver una problemática no resuelta, al menos con eficacia, en el actual estado del arte.

Capítulo 2

Estado del Arte

2.1. Introducción

El análisis de opiniones de textos ha sido un campo de investigación muy estudiado en la última década, donde han surgido un gran número de publicaciones sobre este tema. Se trata de un área interdisciplinar donde se aplican técnicas pertenecientes a distintos campos de investigación para su resolución. Algunas de las áreas que se han dedicado al estudio del Análisis de Opiniones han sido la Computación, Lingüística, Estadística, Psicología entre otras. Esto es debido a que el análisis de opiniones contiene un alto carácter subjetivo, dado que una opinión es un dictamen o juicio que se forma de algo cuestionable.

Este dictamen o juicio es particular de cada escritor, que se expresa de forma distinta. Cada escritor, además de tener una forma particular de escribir, puede tener una opinión particular sobre lo que escribe. Se dice que la opinión tiene polaridad positiva si el escritor ha expresado una opinión positiva. Se dice que la opinión tiene polaridad negativa si el escritor ha expresado una opinión negativa. Existe una última casuística que se da cuando el escritor no expresa ninguna opinión en el texto, en este caso, se dice que el texto no contiene polaridad. Dado que estas opiniones son subjetivas y particulares de cada escritor, es un problema complicado clasificar textos en función de la polaridad que expresan estas opiniones. Por ello, este campo es tan atractivo para el mundo académico y tan interdisciplinar, puesto que cada área de las anteriormente mencionadas pretende demostrar que puede aportar métodos que son capaces de lidiar con la subjetividad.

Pero el interés sobre este campo no ha sido sólo académico, generando también una gran repercusión en la industria. El análisis de opiniones en textos puede ayudar a

capturar la opinión pública sobre eventos sociales, movimientos políticos, estrategias de compañías o preferencias sobre unos productos u otros. Estas u otras tendencias pueden tener una correlación con respecto a las opiniones expresadas en textos. Si los estudios demuestran que esto es cierto, el valor añadido de poseer un modelo de análisis de opiniones para una compañía es incalculable. Debido a este motivo, el análisis de opiniones posee una repercusión en el mundo empresarial importante, más allá del mundo académico. Cambria [6] estudia más en detalle el interés de la industria en este campo. El interés industrial es un motivo más por lo que es importante que un modelo de análisis de opiniones sea capaz de determinar de forma efectiva la polaridad de un texto, o clasificarlo como no polar.

Determinar este contenido subyacente en textos fue uno de los primeros trabajos que se desarrollaron en el campo del Análisis de Opiniones. A continuación se describe este enfoque:

En este trabajo, interesa determinar los rasgos específicos que determinan la opinión que los clientes tienen sobre un producto. De forma adicional, se quiere saber si estas opiniones son positivas o negativas. Para solucionar este problema, los autores identificaron en una primera fase los rasgos del producto de los cuáles un cliente tiene opiniones. Por ejemplo, en una cámara digital interesa saber si los clientes tienen buena o mala opinión de rasgos como la calidad de la foto o el tamaño. En una segunda fase, para cada uno de estos rasgos que ya se han identificado en el texto se elabora un conteo de las opiniones positivas y negativas que los clientes tienen de cada uno de ellos. El resultado final es un resumen en el cuál se indica el número de opiniones positivas y negativas de cada rasgo.

Para resolver esta tarea, se definen *palabras de opinión* que intentaron localizar cerca de los rasgos. En este trabajo, sólo se consideran adjetivos para estas *palabras de opinión*. Estas palabras de opinión están asociadas mediante reglas a polaridades positivas y negativas. Sólo se extrajeron las más comunes que aparecían en los textos que ellos consideraban, las que cumplían con un soporte mínimo que los autores definieron.

Por ejemplo, en la frase *The strap is horrible and gets in the way of parts of the camera you need access to*. En esta frase, el rasgo *strap* es identificado. Una vez se ha identificado, se busca el adjetivo más cercano. Este adjetivo es *Horrible*. Dado que se tiene en la base de conocimiento que horrible está asociado a una polaridad negativa, entonces esta frase refleja una opinión negativa sobre el rasgo *strap* de las cámaras.

El trabajo descrito, elaborado por Hu [7], generó una gran expectación en el campo académico, mostrando que es posible categorizar el carácter subjetivo expresado en opiniones. Cuenta con la desventaja de que un término puede ser ambiguo, y tener polaridad distinta en distintos dominios. Por ejemplo, el adjetivo *bajo* tiene polaridad positiva en la frase *coste bajo* pero polaridad negativa en la frase *beneficio bajo*. Este tipo de enfoques no contempla esta problemática y puede concluir en análisis de opiniones pobres. Se deben establecer dependencias entre los términos *coste, bajo y beneficio* asociadas a la polaridad definitiva o reglas lingüísticas para evitar esta problemática. Pese a ello, es un primer enfoque aceptable.

A partir de este trabajo, nacieron un gran número de enfoques que trataron de solucionar diversos problemas relativos al análisis de opiniones. Inicialmente, el análisis de opiniones fue categorizado como problema específico dentro del procesamiento del lenguaje natural, incluido como capítulo en reputados libros que tratan este área. En el libro sobre procesamiento del lenguaje natural escrito por Indurkhya [8], se puede encontrar un capítulo dedicado exclusivamente al análisis de opiniones. Se trata de un recurso adecuado para consulta sobre dudas que puedan surgir sobre un problema relevante al procesamiento del lenguaje natural o al análisis de opiniones.

El análisis de opiniones fue tratado también inicialmente como un problema particular de categorización de texto que podría ser resuelto mediante aprendizaje automático. Este enfoque pretende clasificar un texto en una serie de categorías definidas a priori en función a una serie de textos anotados en estas categorías de los que se conoce una serie de rasgos. Para ello, se utilizan diversos modelos de aprendizaje automático.

Un extenso estado del arte perteneciente a esa época que trata la categorización de textos usando aprendizaje automático y todos los modelos que se pueden usar y problemáticas que pueden aparecer es [9], en este estado del arte se puede consultar como tratar la categorización de textos con aprendizaje automático. Como ya se tratará más adelante en el estado del arte, el enfoque del aprendizaje automático o cuantitativo es muy eficiente si se cuenta con un gran corpus anotado de textos pero cuenta con el inconveniente de que si el corpus no tiene ni el volumen ni la calidad adecuados, los resultados del modelo de aprendizaje automático serán pobres. Por ello, si este enfoque es aplicado se debe contar con la seguridad de tener un corpus de volumen y calidad adecuados y unas características extraídas de los mismos relevantes para el problema, de lo contrario el clasificador obtenido será pobre.

En concreto, para análisis de opiniones, el planteamiento que se lleva a cabo es clasificar un texto como positivo, negativo o neutral teniendo un corpus anotado de textos en estas categorías. En cada uno de los textos anotados se extraen características de los mismos. Por ejemplo, una característica a extraer es la aparición de un término en el texto. Esta característica se modeliza como una variable booleana, cuyo valor será cierto si el término aparece y falso en caso contrario. Se expondrá un listado de las características extraíbles de los textos en un apartado de este estado del arte. Posteriormente, se entrena un modelo de aprendizaje automático usando este corpus anotado. Una vez entrenado, dado un texto entrante, se extraen los rasgos de este texto y el modelo predice la etiqueta que tendrá el nuevo texto. Este último enfoque ha servido como base para implementar sistemas recomendadores basados en los resultados de modelos de análisis de opiniones que emplean aprendizaje automático. Un artículo introductorio sobre esta metodología es el publicado por Schafer [10]. Es un artículo que se debe leer si no se conocen los conceptos básicos sobre Aprendizaje Automático y Sistemas Recomendadores, dado su estilo didáctico y su poca profundidad.

Pese a estar inicialmente englobado dentro del procesamiento del lenguaje natural y en concreto de la categorización de textos, muy pronto el número de áreas que estudiaron este problema creció, aumentando el número de publicaciones que exclusivamente trataban el análisis de opiniones. Hoy en día el número de publicaciones que tratan el análisis de opiniones es tan elevado e interdisciplinar que poco a poco el campo está independizándose y tratándose como problemática particular. Tanto es así, que en los últimos años se han publicado extensos revisiones del estado del arte del análisis de opiniones como [11] o [12], en los que se tratan diferentes enfoques que se pueden utilizar para el análisis de opiniones en detalle, como solucionar el análisis de opiniones por aprendizaje automático o de forma cualitativa, mediante reglas lingüísticas elaboradas por expertos aplicadas a diferentes dominios. Se tratan de revisiones extensas pero que deben ser leídas por el interesado en ser experto en el análisis de opiniones.

En este estado del arte se revisarán estos y otros enfoques más novedosos. De forma adicional, es importante mencionar que se ha publicado un tutorial muy citado sobre análisis de opiniones escrito por Potts, [13], en el que se introduce como tratar el análisis de opiniones mediante recursos lingüísticos o como expandir léxicos polarizados. Este último tutorial resulta de utilidad como introducción al análisis de opiniones para los iniciados en este campo aunque carece de profundidad en cada uno de los enfoques que cita.

Para conocer un mayor detalle, se dispone del libro sobre el análisis de opiniones escrito por Pang [11], en el cual se tratan los distintos problemas que se han encontrado en el análisis de opiniones, como clasificación de textos o extracción de términos que representan opiniones. Se trata de una publicación en la que se detalla con más profundidad los conceptos resumidos por Potts. Es especialmente interesante el capítulo dedicado en exclusiva a los rasgos que denotan opiniones en textos y para la industria el capítulo de las aplicaciones que pueden hacer uso del análisis de opiniones.

Inicialmente utilizado para determinar la polaridad en artículos de opinión o críticas sobre todo tipo de productos, el análisis de opiniones ha sufrido una revolución con el gran crecimiento de la web en los últimos años. Debido a esto, las técnicas empleadas en análisis de opiniones se han tenido que adaptar a la problemática de tener que procesar un gran número de textos existentes en la web o en redes sociales. En entornos con un gran volumen de textos la técnica a utilizar debe ser escalable, garantizando tiempos de ejecución adecuados procesando un gran número de textos.

Una publicación que ofrece una técnica escalable en grandes volúmenes de texto es la elaborada por Godbole [14]. Tras la expansión inicial de un léxico polarizado mostrada en este enfoque, para analizar la opinión global de un texto Godbole simplemente realiza un conteo de palabras que indican polaridad en los textos. Determina si un texto es subjetivo o no mediante el conteo de palabras que expresan subjetividad en un texto. Este enfoque cuenta con la bondad de ser muy rápido y poderse aplicar en cualquier dominio. Presenta la desventaja de no realizar un análisis preciso, y poder ofrecer resultados pobres en textos de tamaño reducido o pertenecientes a dominios específicos. En estos dominios es preferible realizar un análisis semántico mediante expertos o realizar un análisis mediante aprendizaje automático. Aún así, se trata de un enfoque que debe ser estudiado dada la rapidez de ejecución del mismo y la posibilidad de ser ampliado introduciendo otro tipo de técnicas.

Este estado del arte pretende clasificar y explicar los distintos enfoques más relevantes existentes en la literatura que lidian con el análisis de opiniones. Se inicia la exposición con un listado de los métodos tradicionales que se han empleado para resolver esta tarea, pertenecientes en su mayoría a la Lingüística Computacional. Se inicia este apartado con una introducción y con el rol de la psicología en la resolución de este problema. Se continúa describiendo algoritmos y heurísticas empleadas en el análisis de opiniones. Dado que las frases que exponen opiniones tienen contenido semántico, se cierra el apartado de métodos tradicionales exponiendo técnicas que tratan la semántica presente en los textos.

Un segundo apartado de este estado del arte comenta uno a uno los distintos recursos lingüísticos que se han elaborado para dar soporte a los modelos que tratan de analizar opiniones. Estos recursos han mostrado ser muy útiles en la resolución de esta tarea, pero no basta con utilizarlos para ejecutar un análisis de opiniones de calidad, necesitan ser expandidos con nuevos términos para mejorar su calidad. Este estado del arte incluye un apartado en exclusiva de modelos de creación y expansión de léxicos polarizados.

Todos los enfoques comentados pertenecen a las áreas de Lingüística Computacional, Psicología o Ciencia Computacional, pero como se ha dicho con anterioridad en esta introducción, el análisis de opiniones es un campo interdisciplinar en el que distintas áreas han ofrecido métodos para la resolución de esta tarea. En concreto, la Estadística y el Aprendizaje Automático supervisado y no supervisado han ofrecido enfoques que aportan modelos de análisis de opiniones de gran calidad. Se dedica una sección de este estado del arte a clasificar en diversas categorías estos modelos y a explicar su funcionamiento.

Dentro del análisis de opiniones han surgido problemas específicos que han impedido construir modelos que funcionan independientemente del texto que analicen. Este estado del arte comenta estos problemas, como son el análisis de distintos contextos, el multilingüismo o el análisis de distintos rasgos de una entidad. Para terminar el estado del arte, se comentan las distintas aplicaciones del análisis de opiniones como la resolución de tendencias y las distintas características que los modelos deben considerar en distintos contextos como Twitter o foros y blogs.

En cada apartado se comentan los modelos pertenecientes al tema que expone el apartado. Los artículos citados en este estado del arte emplean técnicas pertenecientes a varios apartados, como por ejemplo modelos que emplean semántica y clustering. Dado este hecho, se comenta en cada apartado la parte de los artículos que aplica en ese apartado y las otras partes se comentan en el resto de apartados para mantener la coherencia. Siguiendo con el ejemplo del artículo que contiene técnicas semánticas y de clustering, en el apartado de semántica se comentan los rasgos semánticos que se han tenido en cuenta y en el de clustering como se ha efectuado clustering teniendo en cuenta dichos rasgos. Se empieza con la descripción de los métodos tradicionales en el análisis de opiniones.

2.2. Métodos Tradicionales

2.2.1. Introducción y Categorización de textos

Como se ha comentado en la introducción, inicialmente se trató el problema del análisis de opiniones como categorización de textos, usando algoritmia y heurísticas tradicionales. También se usó aprendizaje automático, como ya se ha dicho que se puede consultar en el estado del arte de Sebastiani [9], enfoque que se detallará en la sección del estado del arte dedicada a este fin.

Por razones históricas, es necesario exponer un ejemplo de categorización de textos que emplee algoritmia y aprendizaje automático. Esto es debido a que el análisis de opiniones fue inicialmente tratado como un problema particular de la categorización de textos. Debido a esto, las técnicas empleadas en la categorización de textos son extrapolables al análisis de opiniones. Un ejemplo que se puede encontrar para la categorización de textos clásica se trata en el trabajo desarrollado por Melville [15], que se detalla a continuación. En este artículo se emplea un enfoque híbrido que combina léxico con aprendizaje automático para polarizar textos. En lo relativo a categorización clásica de textos de este modelo, este enfoque habla de clasificación de texto o léxica empleando un léxico que define la polaridad de las palabras, o léxico polarizado.

El método captura la frecuencia con la que aparecen estas palabras en el texto a analizar. De este modo define una probabilidad $P(+/D) = \frac{a}{a+b}$, donde a y b es el número de ocurrencias de términos positivos y negativos encontrados. Si la probabilidad $P(+/D) > t$, siendo t un umbral configurable, el documento es clasificado como positivo, de lo contrario es clasificado como negativo. Se trata de un método tradicional que categoriza textos en categorías en función al número de términos encontrados de cada categoría. Se trata de un método muy simple que presenta el problema de la ambigüedad, ya que una palabra puede ser positiva en unos dominios y negativa en otros. Pese a ello es una primera aproximación para la resolución eficiente de la categorización de textos. El resto del enfoque se explicará en la sección de aprendizaje automático de este estado del arte. Se tratan mas enfoques parecidos en el subapartado de heurísticas y algorítmica clásica. Dado que el análisis de opiniones es un caso particular de la categorización de textos que necesita conocimiento experto, es necesario conocerlo para poder aplicarlo en la clasificación. Por ello, en el siguiente apartado se mencionan los aportes que la psicología ha introducido en este tema.

2.2.2. Psicología

Debido al carácter subjetivo de las opiniones y a que estas son un reflejo de las emociones que los autores sienten cuando escriben, es necesario que la psicología aporte una serie de conceptos para elaborar un análisis de opiniones de calidad. Estudiar el concepto de emoción y de opinión bajo el punto de vista de los psicólogos es útil para averiguar que rasgos las definen. Estos rasgos pueden ser posteriormente extraídos de textos mediante algoritmia y servir como atributos para un modelo de aprendizaje automático. El trabajo más relevante del área es el escrito por Scherer [16], en el que define el concepto de emoción como: *Un episodio de cambios interrelacionados y sincronizados en los estados de todos o la mayoría de los cinco subsistemas orgánicos en respuesta a la evaluación de un evento que provoca un estímulo interno o externo que es entendido como relevante para las principales preocupaciones del organismo.* Estos cinco subsistemas son el procesamiento de la información, el soporte de los distintos sistemas del organismo, el sistema ejecutivo del organismo, el sistema motor o actuador del organismo y el monitor de sentimientos subjetivos. El concepto está expresado con un estilo muy críptico, por lo que se describe con mayor sencillez a continuación.

Explicado de forma más simple, se podría definir a la emoción como aquel cambio provocado por un evento que tiene repercusión en el organismo. Si se asume que la opinión es un reflejo de esta emoción, esta tendrá siempre asociado un objeto sobre el que el sujeto poseedor de la emoción quiere expresar su opinión. Adicionalmente, describirá la emoción que tiene de este objeto con una serie de características, que por norma general son adjetivos.

Una vez conocida esta casuística, se puede, por ejemplo, analizar textos mediante reglas lingüísticas en búsqueda de sujeto, objeto y como se relacionan. Si se saben que relaciones expresan una opinión, entonces se puede capturar la opinión de un sujeto sobre un objeto. Por ejemplo: *Javier está contento con el televisor.*. Reconocido el sujeto Javier, el objeto televisor y la relación que los une *estar contento con* se puede ejecutar la regla *contento(sujeto,objeto)-¿opinión positiva*. Poseer una base de conocimiento con múltiples reglas de este estilo consigue una gran precisión en el análisis de opiniones efectuado sobre textos. El problema radica en que existe un gran número de relaciones y su dependencia con el contexto es fuerte, por lo que se necesitan recursos que las contengan o personal que las estudie.

Al existir un gran número de categorías afectivas que simbolizan emociones que pueden ser descritas en opiniones, Scherer define en [16] un listado de estas categorías acompañadas con una serie de términos y raíces que las describen. Si estos términos

son encontrados junto al objeto de la opinión simbolizan la categoría afectiva de esta opinión. Por tanto, estos términos pueden ser utilizados para polarizar un texto, si se les asocia polaridad positiva o negativa adicionalmente a la categoría afectiva que simbolizan.

Por ejemplo, en lengua inglesa, Scherer sostiene que la categoría afectiva *Boredom* está expresada en una opinión si en el texto aparece cualquiera de los siguientes términos y raíces: *Bor*, ennui, indifferen*, languor*, tedi*, wear**. Si se conoce que la categoría *Boredom* es negativa, entonces se puede concluir que estos términos contribuyen a una polarización negativa de la opinión. En este trabajo de Scherer se incluyen distintas tablas con recursos léxicos muy utilizadas en análisis de opiniones.

El trabajo de Scherer es bastante completo para el analista de opiniones. Aun así, si se desea estudiar en más profundidad la visión de la psicología sobre el análisis de opiniones existen otras publicaciones citadas. Otros trabajos interesantes que se han efectuado sobre cómo influye la psicología en el lenguaje mediante el cual se expresan las opiniones son el de Pennebaker [17], en el que estudia como el lenguaje plasma las emociones o el de Tausczik [18] , en el que se detalla, con bastante profundidad, cómo se construye un léxico polarizado y los problemas encontrados en el análisis de opiniones, como por ejemplo discernir la ironía en un texto. Estos trabajos contienen conocimientos teóricos de alto valor pero carecen, en su mayoría, de aportar contenido práctico. Por este motivo, sentada la base teórica, es importante estudiar como traspasar este conocimiento a la práctica.

2.2.3. Heurísticas y algorítmica clásica

En este apartado se comentan algoritmos que se han utilizado en análisis de opiniones. En el apartado de introducción y categorización de textos se presentó el primer enfoque propuesto por Hu [7], basado en el conteo y que es bastante popular. Pese a su popularidad, se estudió que contiene altas carencias puesto que presenta la carencia de no lidiar con la ambigüedad o con los contextos. En este apartado se presentan enfoques similares que se han propuesto para el análisis de opiniones. Connor [19] presenta un trabajo que en una parte de su resolución define una puntuación diaria que refleja la polaridad de un tema utilizando conteo. Primero, se detecta un tema en los textos empleando palabras clave, una vez detectado se buscan palabras que representen polaridades negativas y positivas en los textos y se elabora un conteo de estas todos los días. La puntuación diaria es:

$$x_t = \frac{count_t(pos.word \wedge topic.word)}{count_t(neg.word \wedge topic.word)}$$

Donde $count_t(pos.word \wedge topic.word)$ es una variable cuyo valor será igual al número de veces que una palabra categorizada positivamente coocurre con una palabra clave. Del mismo modo, $count_t(neg.word \wedge topic.word)$ será una variable cuyo valor será igual al número de veces que una palabra categorizada negativamente coocurre con una palabra clave. La tendencia diaria sobre el tema será positiva en el día x_t si este número es positivo y de lo contrario será negativa. Como se puede ver, es un enfoque muy parecido al de Hu [7] pero añadiendo categorización de textos. Esto resuelve el problema del dominio, puesto que una vez detectado, se pueden emplear distintas palabras que expresan polaridad a cada dominio. Conlleva la problemática de tener que construir un léxico polarizado diferente para cada dominio. Este enfoque polariza a nivel de conjunto de documentos, pero también puede ser interesante polarizar a nivel de documento, sección, párrafo o frase. A continuación se expone un modelo que polariza a nivel de frase.

El uso de léxicos polarizados es común en los enfoques tradicionales. Estos enfoques primero expanden un léxico polarizado base y luego efectúan distintas heurísticas para clasificar con él. En un apartado posterior de este estado del arte se comentará como expandir y crear estos léxicos. En concreto, en el trabajo de Kim [20], primero se expande el léxico base asignando una probabilidad a cada término del léxico de pertenecer a categorías positivas y negativas. Después se localiza el objeto de la opinión. Hecho esto se proponen dos modelos para clasificar la región de texto en la cual se expresa la opinión, polarizando un documento a nivel de frase.

El primer modelo es una media armónica: Sea $p(c|w_i)$ la probabilidad de que la palabra polar w_i pertenezca a la categoría c y $n(c)$ el número de palabras en la región cuya categoría es c. Sea s la región de palabras, entonces la probabilidad de que una región de palabras pertenezca a una categoría es:

$$P(c|s) = \frac{1}{n(c)} \sum_{i=1}^{n} p(c|w_i)$$

$$if : argmaxp(c_j|w_i) = c$$

Es decir, en este modelo primero se normaliza si ninguna palabra del léxico implica que la probabilidad de que pertenezca a una categoría es 1. Después, la probabilidad de que una categoría esté expresada en una frase es igual al sumatorio de las

probabilidades de los términos polarizados de la frase entre el número de palabras en la región que simbolizan una categoría. Con este modelo se puede categorizar a nivel de frase. Se usa la media armónica dado que es más apta para las cantidades promedio, como estas probabilidades.

Otro modelo para polarizar a nivel de frase propuesto en [20] es la media geométrica:

$$P(c|s) = 10^{n(c)-1} \prod_{i=1}^{n} p(c|w_i)$$
$$if : argmaxp(c_j|w_i) = c$$

Pese a lo extraño de la expresión, la virtud de la media geométrica frente a la aritmética es que más apta para comparar elementos con rangos distintos y simboliza mejor las tasas de crecimiento. Esto hará que valores de probabilidad muy dispares sean tratados de forma más efectiva. Es útil considerar el resultado de distintos modelos en el análisis de opiniones puesto que ningún modelo es el más útil en todos los casos. Por ello, combinar el criterio de distintos modelos maximiza la probabilidad de que el resultado final sea el más adecuado. Esta es una metodología muy acertada introducida por Kim [20] al considerar en este caso concreto dos medias. Este criterio es escalable a considerar el resultado de distintos modelos de aprendizaje automático, conjuntos de reglas u otras soluciones.

Los enfoques tradicionales también pueden permitir incorporar rasgos significativos a los enfoques basados en aprendizaje automático. Desde un punto de vista personal, un modelo híbrido que combine soluciones cualitativas o basadas en la creación de reglas por lingüísticas y cuantitativas como modelos de aprendizaje automático es el enfoque más efectivo para lidiar con el análisis de opiniones. Esta opinión personal se defiende mediante dos argumentos. En primer lugar, por lo expuesto en el párrafo anterior con respecto a combinar criterios: Al obtener todos los criterios resultados superiores al criterio aleatorio, todos incorporan información útil. Cualquier criterio que no sea superado por otro en todos los casos, es decir que no esté dominado, debe ser incorporado. En segundo lugar, es conocido que los enfoques cualitativos carecen de la exhaustividad en la recuperación o *recall* de los enfoques cuantitativos y que los enfoques cuantitativos carecen de la precisión de los cualitativos. Por ello, combinar el criterio de ambos ofrece una solución de alta precisión y exhaustividad. Esta es la opinión personal por la que se defienden los modelos híbridos.

Una posible arquitectura de modelo híbrido es el que cuenta como atributos del enfoque cuantitativo con resultados de enfoques cualitativos. Se presenta un ejemplo de enfoque cualitativo que puede servir a un cuantitativo a continuación.

Con este objetivo, Martineau [21] propone una variante del algoritmo TF-IDF adaptado a análisis de opiniones para construir un vector de las frecuencias adaptadas de términos contenidos en un léxico polarizado. Este vector podrá ser luego usado como un vector de rasgos para los modelos de aprendizaje automático. El elemento del vector, en vez de modelizar una frecuencia con la que ocurre el término en el documento considerado con respecto a la de los otros documentos como en el TF-IDF, modeliza además como ocurre con más frecuencia en documentos anotados positivamente y negativamente.

El algoritmo propuesto, Delta TF-IDF, necesita un corpus de documentos anotado con las categorías positivas y negativas. Este algoritmo crea un vector de pesos de los distintos términos por como ocurren más en los documentos de una categoría frente a la otra. Mientras que un término ocurra más frecuentemente en una categoría que en la otra, mas puntuación en valor absoluto tendrá el peso de este término en el vector.

Si el término ocurre más en la categoría positiva tendrá un valor negativo, y si ocurre en la negativa tendrá un valor positivo. El logaritmo de la expresión que se muestra a continuación del Delta TF-IDF se emplea para que los términos distribuidos menos uniformemente tengan un mayor valor frente a los distribuidos prácticamente uniformes.

$$V_{t,d} = C_{t,d} * log_2(\frac{N_t}{P_t})$$

Donde $C_{t,d}$ es el número de veces que el termino t ocurre en el documento d. P_t es el número de documentos anotados positivamente donde aparece t y N_t es el número de documentos anotados negativamente donde aparece t. El resultado es el vector $V_{t,d}$ que muestra la puntuación asignada para cada término.

Martineau muestra en la evaluación como emplear un vector construido por el Delta TF-IDF mejora en aproximadamente un 6 % la precisión de una Support Vector Machine frente emplear un vector construido por el TF-IDF convencional en el corpus de documentos que considera. Por lo tanto, para análisis de opiniones, se utiliza este enfoque en vez del TF-IDF convencional para proporcionar la frecuencia de los términos de un léxico polarizado. Debido a los resultados conseguidos por este algoritmo, que mejoran las cifras obtenidas por modelos cuantitativos, y al sentido común que posee este enfoque, es un algoritmo recomendado para el análisis de opiniones.

De todos modos, el algoritmo TF, que simplemente construye un vector con la frecuencia de los términos encontrada en un texto y puede ser normalizado, es también

utilizado en enfoques como [22] en combinación con otras técnicas para efectuar análisis de opiniones. Es un rasgo incorporable a cualquier enfoque cuantitativo. En concreto, el enfoque presentado por Vechtomova [22], extrae los adjetivos dependientes de sustantivos que se saben polarizados y construye un vector de frecuencias con ellos. Es importante recordar la dependencia fuerte de la polaridad con respecto al dominio y la ambigüedad de los términos, al contener una sección de ellos múltiples acepciones.

A continuación se van a mostrar enfoques que hacen uso de recursos lingüísticos y las relaciones semánticas que estos proponen para polarizar un texto. Como puente entre estas técnicas y las vistas en este apartado, se cita brevemente a Ding [23], quién propone las bases de estos enfoques empleando reglas lingüísticas tradicionales basadas en las relaciones proporcionadas por estos recursos.

Un ejemplo sería el siguiente: Si se encuentra un sinónimo de un término polarizado positivamente en una frase, entonces la polaridad de la frase donde se ha encontrado se incrementará hacia la polaridad positiva. Lo contrario pasaría con un antónimo. Otro ejemplo sería tener dos frases conectadas por la conjunción coordinada *pero*, como se sabe que *pero* invierte la polaridad, entonces la segunda frase tendrá la polaridad inversa a la primera, etc.

Es especialmente importante contemplar problemáticas como la negación en estos enfoques, puesto que si no se contempla, el análisis de opiniones ofrecerá resultados incorrectos.

Se verá en el siguiente apartado que no sólo son importantes las relaciones semánticas entre los términos de una frase sino que también es importante distinguir entre la categoría gramatical de cada término, como introduce Taboada [24]. Mediante este enfoque se puede desambiguar las acepciones de cada término, lidiando con el citado problema de la ambigüedad. Además, en función a la categoría gramatical del término, a su semántica y a su relación con otros términos, se pueden tomar decisiones. Por ejemplo, en el caso de los adverbios se verá que su presencia puede contribuir a intensificar la polaridad del adjetivo del que son dependientes. Se trata de un análisis denso y que requiere un gran número de recursos, pero ofrece resultados muy precisos. Se exponen a continuación enfoques más avanzados que tratan la semántica subyacente en los textos para hacer análisis de opiniones y se detallan las problemáticas de estos enfoques.

2.2.4. Semántica y métodos avanzados

Como se introducía en el apartado anterior, la categoría gramatical de los términos estudiados y sus dependencias pueden aportar información útil al análisis de opiniones. En concreto, Dragut [25], estudia el rol de los adverbios en el análisis de opiniones. Dragut asume que mientras que unos términos como adjetivos tienen una polaridad inherente, otros como los adverbios actúan como intensificadores de dicha polaridad.

A este fin, Dragut pregunta a Amazon Mechanical Turk (AMT), mecanismo de crowdsourcing, que clasifiquen en las categorías fuertemente positivo (+2) a fuertemente negativo (-2) una serie de frases. Mide el grado de acuerdo entre los diferentes anotadores como $\frac{1}{|S(i)|}\sum_{j\in S(i)} ps_{ji}$. Donde S(i) es el set de frases anotadas por el anotador i y ps_{ji} es el porcentaje de anotadores que tienen la misma anotación para la frase j.

El estudio es que una parte de las frases eran idénticas a otras con la única diferencia de la existencia de un adverbio en ellas. Con esto Dragut elabora una tabla de tres columnas, en la primera de ellas se muestra el adverbio que se añade, en la segunda la diferencia de polaridad media de las frases que lo contenían y las que no y en la tercera si el adverbio hace revertir la polaridad.

Del estudio se extrae que aunque pocos adverbios hacen revertir la polaridad, prácticamente todos la intensifican, bien a negativo o a positivo. Por ejemplo, el adverbio *awfully* intensifica la polaridad negativa en 0.6 según la anterior escala y el adverbio *pretty* intensifica un 0.4 la intensidad positiva. Cabe destacar que una parte de los adverbios intensifican más la polaridad negativa o positiva de la frase. Por ejemplo *pretty* intensifica la polaridad negativa un 0.35 frente a un 0.4 de la polaridad positiva.

Se trata de un enfoque muy preciso, pero que contiene problemáticas. En cada dominio, es posible que la intensidad que modifica cada adverbio sea distinta. Para solucionar este problema, se tiene que analizar la repercusión de cada adverbio en cada dominio. La complejidad es por lo tanto exponencial. Se necesita el trabajo de lingüistas para analizar cualquier posible casuística y anotar la intensidad del adverbio en cada una de ellas. Dado que el número de posibles dominios es tan grande como los que se quieran analizar, es una tarea muy ardua. Por el contrario, la ganancia obtenida es únicamente un grado de intensidad en la polaridad de la opinión expresada. Es un enfoque que, en palabras del mundo industrial, es poco

rentable. Por esto, se recomienda que se polarice en valores nominales como positivo y no negativo y no en variables reales que expresen la intensidad de la polaridad.

Por otro lado, es un ejemplo de la importancia de las categorías gramaticales en el análisis de opiniones y un primer paso para construir un léxico polarizado que además de clasificar indique una intensidad de la polarización de un texto. Para más detalle, este concepto de modificador de la intensidad también aparece en el trabajo de Wilson [26], en el que elabora una lista de términos intensificadores junto a su categoría gramatical. Se enfatiza el hecho de que se considera una tarea muy ardua el categorizar a la polaridad en variables reales y no nominales.

Las categorías gramaticales no son el único aspecto importante a considerar en el análisis de opiniones desde el punto de vista semántico, las dependencias entre los términos de una frase también aportan rasgos al análisis de opiniones. Para identificar términos que expresen subjetividad, Jadhav [27] emplea UNL, una interlingua que está diseñada específicamente para representar datos semánticos extraídos de textos en lenguaje natural. Mediante distintas reglas que usan las relaciones que proporciona este lenguaje como *agt*, *obj* o *mod* entre términos, a los que denomina Universal Words (UW), identifica los términos subjetivos.

Una de estas reglas, por ejemplo, enuncia que si una UW es la fuente de la relación *agt*, entonces su polaridad es añadida a la polaridad total del texto. En la frase *'I like her'* la relación sería *agt(like,I)*, por lo que la polaridad de *'like'* se añadirá. La polaridad de cada término es buscada en un léxico polarizado que se ha construido a priori. El resto de reglas continúan decidiendo que términos añaden polaridad al texto y cuáles no, basándose en las dependencias que los términos, Universal Words, tienen entre sí.

El UNL contiene una estructura que aporta información de gran valor para el análisis de opiniones. Mediante esta interlingua y el análisis de un texto mediante la misma, es posible construir reglas de alta precisión. Además, al contener cada universal word un único significado, se resuelve el problema de la ambigüedad. Del mismo modo, si se detecta cada uno de los dominios, se resuelve el problema de la dependencia del dominio. La metodología consiste en construir diferentes reglas para cada dominio. Si el ámbito a resolver y el número de dominios es de un tamaño medio o pequeño, se trata de un enfoque óptimo de gran precisión.

No obstante, es posible que se necesite analizar las opiniones de un gran volumen de textos que pertenecen a cualquier dominio. Si se sigue la metodología anterior, se procede a categorizar primero un texto en un dominio y luego ejecutar las reglas de ese dominio. Se puede producir que el número de dominios sea tan alto que cada

texto pueda pertenecer a un dominio distinto. Por ejemplo, en redes sociales, se escribe sobre cualquier concepto. Si se quiere analizar la opinión en redes sociales de cualquier tema, la metodología citada es poco escalable.

Si el número de textos es muy elevado y el número de categorías indeterminado, el trabajo del analista de opiniones para resolver esta tarea con la metodología citada es inabordable. Es muy complicado y costoso construir una base de reglas que contenga cualquier posible casuística que se pueda expresar mediante el lenguaje. Desde el punto de vista personal, es inabarcable y poco escalable. Reglas que pueden afectar a determinadas situaciones no lo pueden hacer a todas, incluso en el mismo dominio.

Se añade un ejemplo. Supóngase que sólo se emplea un enfoque cualitativo. Considérese un dominio en el cuál se encuentra el término *bajo*. Se decide añadir la regla Bajo implica polaridad negativa. Se encuentra en el mismo dominio el texto *coste bajo*. En este contexto, la aparición de la palabra coste junto a bajo implica una polaridad positiva, contraria a la de bajo. Para contemplar esta casuística se añade la regla Coste AND Bajo implica polaridad positiva. Pero en otro texto del mismo dominio se analiza también *beneficio bajo*. La palabra *Beneficio* a solas contiene polaridad positiva pero acompañado de la palabra *bajo* la polaridad es negativa. Se añade a la base de conocimiento las reglas: Beneficio AND Bajo implica polaridad negativa y Beneficio implica polaridad positiva. Sin embargo, se encuentra posteriormente el texto *El beneficio bajo de la empresa competidora nos da ventaja para actuar*. Con la base de datos citada, se efectuaría un mal análisis de opiniones.

Las casuísticas que se pueden dar son ilimitadas, por ejemplo: *El beneficio bajo de la empresa competidora es irrelevante, puesto que el nuestro es menor*. Ningún modelo basado únicamente en reglas puede contemplar todas las posibles combinaciones. En cambio, si se dispone de un corpus anotado de extensión enorme, lo cual también es una desventaja puesto que es difícil generarlo, crearlo u obtenerlo, un modelo cuantitativo que considere rasgos cualitativos es capaz de encontrar todas las dependencias entre las distintas variables. El modelo de aprendizaje automático puede generar que la aparición simultánea de términos o reglas lingüísticas implica una polaridad que ninguna de las reglas por separado genera.

Por ello, se enfatiza que la precisión de estos modelos cualitativos debe ser combinada con el alcance que ofrecen los modelos cuantitativos. Los modelos cuantitativos son capaces de descubrir las dependencias entre un número indeterminado de variables. Si cada una de estas variables es el resultado de enfoques cualitativos, entonces si se podrá considerar prácticamente cualquier casuística. Aún así, de ser estudiada bien

cada regla, estas aportarán una precisión mayor a cualquier otro tipo de enfoque. Del mismo modo, aportan gran valor y pueden ser indispensables en modelos híbridos

La semántica no sólo está presente en las categorías gramaticales o en las dependencias de un término con otros sino en la posición de un término polarizado en el texto. Turney, [28], en el primer trabajo que se desarrolló sobre Análisis de Opiniones usando aprendizaje automático consideró la posición del término polarizado como rasgo a tener en cuenta. Distingue si el término aparece en la primera parte, mitad o final del texto, teniendo como hipótesis que si está presentado al final entonces su peso en la polarización del texto es mayor. También se tiene en cuenta como rasgo la categoría gramatical de los términos estudiados. Al final de este trabajo concluye que un mayor estudio de la estructura es necesario.

Este estudio ha sido realizado en mayor profundidad por Wachsmuth [29], quién quiere corroborar la hipótesis de que la clase del documento en lo relativo a análisis de opiniones está dada por su estructura además de por su contenido. De esta forma define y elabora un set de patrones de flujo de opinión. Un patrón de flujo de opinión se define como un vector cuyos elementos simbolizan la polaridad de las frases en las cuales se divide un texto que sigue un orden determinado. La siguiente figura muestra un ejemplo:

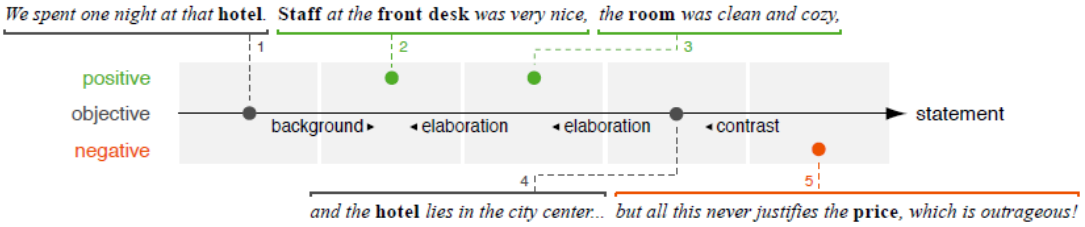

FIGURA 2.1: Patrón de Flujo de Opinión. En este caso se modeliza que si la última parte tiene polaridad negativa la polaridad final del texto es negativa.

En este trabajo se incorpora como un rasgo más a un aprendizaje supervisado estos flujos de opiniones. Los autores definen un set de posibles patrones. Posteriormente segmentan el texto para polarizar cada segmento por separado y construir el flujo del texto. Una vez hecho esto se comparan con los sets originales para ver si el flujo se parece a algún patrón. Dado que los flujos deben ser lo más parecidos a un patrón con los que se trabaja y lo más diferentes posibles del resto de los patrones se emplea clustering para averiguar a qué patrón pertenece el flujo. Dicho proceso de clustering se explicará en la sección del estado del arte dedicada a este tema.

El enfoque de la estructura se trata de un rasgo que no es evidente y que sin embargo refleja la realidad de la opinión expresada de una parte de los textos. En mi opinión, es insuficiente realizar un análisis de opiniones dependiendo únicamente de este rasgo, pero es complementario empleado conjuntamente con técnicas cualitativas o cuantitativas.

Otros enfoques, en vez de recurrir a estructurar el texto, analizar las dependencias entre términos o estudiar sus categorías gramaticales se basan en técnicas de carácter no supervisado o estadístico que determinan que términos contienen la misma semántica como la Latent Dirichlet Allocation (LDA), que serán explicadas en profundidad en el apartado de Métodos estadísticos y basados en Aprendizaje No Supervisado de esta memoria.

Por ejemplo, Maas [30] construye vectores de palabras semánticamente similares utilizando la técnica LDA y después anota a estos vectores con la semántica que reflejan. Estos vectores podrán contener términos que reflejan la positividad de una opinión o la negatividad de la misma y ser empleados para el análisis de opiniones. Este tipo de técnicas serán explicadas más adelante. Otro ejemplo de estas técnicas es el Latent Semantic Analysis (LSA), empleado por Mohtarami [31] para inferir la semejanza de las opiniones entre los sentidos subyacentes entre pares de palabras. Con estas semejanzas construye un vector de emociones humanas para cada sentido de las palabras que se encuentran en un texto con el cuál Mohtarami pretende categorizar las emociones subyacentes en los textos. Emplear estas técnicas junto a las cualitativas ofrece resultados óptimos.

A modo de conclusión de este apartado, se cita que el análisis semántico puede ser empleado para otras tareas relativas al análisis de opiniones como puede ser la extracción de las causas de una opinión [32]. En este trabajo, Neviarouskaya, utiliza un parser de dependencias entre los términos de una frase y mediante reglas lingüísticas extrae las causas. Por ejemplo, en la frase *'His tone scared her more than anything she could remember.'* el sujeto *'his tone'* es reconocido como la causa de la emoción *'Fear'* expresada por el verbo *'scared'*. En este trabajo se puede consultar una tabla con todas las reglas lingüísticas que Neviarouskaya considera para extraer las causas de las opiniones o emociones reconocidas. Este trabajo cuenta con el ya citado inconveniente que en determinados casos de uso presenta buenos resultados pero al ser consistente en un conjunto de reglas no es escalable a cualquier dominio ni aplicación en el lenguaje expresado en otra variedad geográfica o época. Sólo la combinación entre modelos cuantitativos y cualitativos puede resolver esta problemática.

Para lograr esta combinación se puede considerar el resultado de distintos criterios o hacer que los criterios vistos en este apartado y el anterior sirvan a los enfoques cuantitativos. Una metodología es que las reglas lingüísticas y enfoques cualitativos vistos en este apartado generen un corpus con polaridad anotada que, tras ser revisado por un lingüista y corregidos los fallos de estas reglas, puede ser introducido en un modelo de aprendizaje automático. Al ser automática la ejecución de reglas, si se tiene un volumen muy elevado de textos se pueden anotar todos los textos. Las dependencias existentes entre estas reglas y los términos polarizados que implican polaridades no consideradas por ellas solas, son resueltas por los modelos cuantitativos o de aprendizaje automático. El modelo que integra esta metodología consigue un análisis de opiniones sin problemáticas aparentes.

Contar con algoritmia y heurísticas no es suficiente para un modelo de análisis de opiniones. Sin recursos de los que extraer la información que estos enfoques necesitan, no se puede efectuar en la mayoría de los casos un análisis de opiniones. Por ejemplo, considérese que un texto es clasificado como positivo si en él se detecta un número mayor de términos positivos que negativos. Si no se tiene un recurso en el cual se almacenen los términos positivos y negativos, no se puede efectuar el análisis de opiniones. Por ello, estos recursos se detallan a continuación.

2.3. Recursos Lingüísticos

2.3.1. Introducción

Presentados los métodos tradicionales para analizar opiniones en textos se expondrán a continuación los recursos lingüísticos existentes en la literatura. Estos recursos lingüísticos se emplean como soporte para los métodos tradicionales y otros modelos de análisis de opiniones. Los recursos léxicos pueden ser simples listados grandes o recursos más elaborados en los que cada término está acompañado con información adicional de utilidad en el análisis de opiniones. Estos se han construido por expertos o por procesos semiautomáticos, pero siempre bajo una supervisión humana. Una particularidad de los principales recursos lingüísticos existentes en la literatura es que todos están escritos en lengua inglesa. Se puede encontrar un breve listado de estos recursos en el artículo escrito por Giouli [33]. En este apartado de la memoria se expande este listado.

En el análisis de opiniones, se emplean recursos lingüísticos como son los léxicos polarizados para ayudar a modelos cuantitativos y cualitativos a clasificar un texto

en distintas categorías. Por ejemplo, se emplea el léxico General Inquirer que se comenta a continuación para hacer un sencillo clasificador consistente en la siguiente lógica. Si se detectan un número mayor de palabras en el texto asociadas con polaridad positiva entonces el texto tiene polaridad positiva y tendrá polaridad negativa en caso contrario. Este clasificador es muy sencillo y tiene muchos problemas, pero se puede emplear una lógica similar para un modelo cuantitativo. Por ejemplo, considerar como rasgos de un modelo cuantitativo como una Máquina de Soporte Vectorial los términos recogidos en un léxico polarizado. Otro uso, en un modelo cualitativo, es crear una regla en la cual si se encuentra uno de los términos del recurso lingüístico acompañado por otro término definido en la regla, entonces el texto pertenecerá a una categoría determinada. Por todos estos usos, los recursos lingüísticos en el análisis de opiniones son de vital importancia. En este apartado de la memoria, se pretende elaborar un listado con los recursos lingüísticos más representativos en el análisis de opiniones, explicando sus particularidades.

2.3.2. General Inquirer

El recurso diseñado por Stone en Harvard en 1966 [34] fue el primer recurso lingüístico utilizado y su objetivo era producir un método para analizar un tema automáticamente. Por la época en la que se realizó, resulta imprescindible citarlo puesto que ha sido el precursor del resto de recursos. Hasta el 1990 no se elaboró una versión que no funcionase únicamente en los mainframes de IBM que soportasen el lenguaje PL/1, ahora está disponible para la investigación. El General Inquirer en su versión original contiene 1915 términos que denotan positividad y 2291 términos que denotan negatividad. Este recurso se ha ido revisando a lo largo de los años, existiendo más versiones que diferencian subcategorías de términos positivos y negativos. Por ejemplo, categorías recientemente construidas han sido: Placer, con 168 términos denotando esta categoría; Dolor, con 254 palabras denotando esta categoría o Excitación, con 166 términos. Las categorías recientemente añadidas al General Inquirer no sólo corresponden al análisis de opiniones sino que también tratan de reconocer otras categorías en el texto como lugares u objetos. Se trata de un recurso básico y que cuenta con el problema de que no contempla las múltiples acepciones de cada término y sólo tiene unigramas. Algunos conceptos sólo pueden ser expresados mediante términos que tienen más de una palabra. Estos conceptos, por tanto, no están contenidos. Pero dado que su obtención es libre, resulta un recurso que es útil para combinarlo con otros, pero no como único recurso.

2.3.3. SenticNet

SenticNet, del que se puede encontrar todo tipo de detalles en el artículo de Cambria [35], en su versión más reciente es un recurso que está basado en razonamiento del sentido común para la extracción de los aspectos a polarizar, además de aportar un diccionario para una vez extraídos los aspectos, polarizarlos. En su trabajo, Cambria sostiene que es más importante polarizar los rasgos del objeto reconocido en el texto en vez del texto en su conjunto. Por ejemplo, considérese la frase: *'I love the phone's touchscreen but its battery life is too short'*, SenticNet produce como salida: *'Touchscreen: +',' Battery: -'*. La base de conocimiento de SenticNet integra conocimiento de sentido común para extraer estos rasgos en forma de grafo. Este grafo contiene aproximadamente 25 millones de sentencias RDF con 2693200 nodos. 30000 son extraídos para categorías afectivas. Los rasgos son reconocidos y extraídos mediante dependencias reconocidas entre conceptos y también por clustering, concretamente por clustering aglomerativo y usando la distancia coseno. La polarización es efectuada por un diccionario. El recurso, además de polarizar, también ofrece otras salidas como un vector de categorías afectivas para el texto que ha procesado. Ejemplos de estas categorías son Placer, Atención o Sensibilidad. Además de considerar las categorías positivo y negativo, este recurso, al igual que el General Inquirer, también considera subcategorías positivas y negativas. Es un enfoque de mayor complejidad al General Inquirer pero que aporta un valor añadido a aplicaciones que necesiten las categorías citadas. La salida de SenticNet que polariza distintos sustantivos de una misma frase en categorías distintas es interesante para el problema del análisis de los aspectos de una opinión. Se trata por tanto de un recurso recomendado en distintos problemas específicos del análisis de opiniones y la computación afectiva.

2.3.4. Wordnet

Wordnet es probablemente el recurso lingüístico de uso general más utilizado en la lengua inglesa. Wordnet es un tesauro que estructura la información en grupos de sinónimos denominados Synsets, proporcionando un gran número de relaciones entre estos synsets como hiperonimia o meronimia. Estos grupos de sinónimos contienen acepciones de cada término.

Wordnet no puede ser usado en soledad como recurso lingüístico para el análisis de opiniones, pero si se puede emplear como soporte para incrementar la calidad de un modelo de análisis de opiniones. Esto es debido a que Wordnet proporciona relaciones entre términos como la sinonimia. La idea consiste en que si se tiene un

léxico polarizado base, entonces, si Wordnet detecta que una palabra que se lee es sinónimo de una que se tiene en el léxico base, está es considerada para polarizar. Esta es la idea básica sobre la que se sostienen una parte de los modelos de expansión de léxicos polarizados, que se comentarán más en profundidad en la próxima sección del estado del arte. En el ya mencionado anteriormente en este estado del arte enfoque de Kim, [20], este emplea Wordnet en el análisis de opiniones. Una vez ha reconocido el objeto sobre el cual la opinión se refiere, polariza la opinión a través de la información sobre relaciones de sinonimia encontrada en Wordnet y un léxico polarizado semilla.

Los detalles sobre el funcionamiento del algoritmo serán expuestos en la sección de Creación y expansión de léxicos polarizados y recursos del estado del arte. Este uso de los recursos léxicos no sólo aplica a Wordnet, puede aplicar a otros en función de sus relaciones. Enfatizar que este recurso es un añadido para los modelos de análisis de opiniones pero que no sirve en soledad para realizar análisis de opiniones. Además, la falta de coherencia encontrada al recorrer múltiples relaciones en Wordnet produce malos resultados en la expansión de léxicos polarizados, como se detallará más adelante. Por esta razón, desde un punto de vista personal se recomienda ser especialmente crítico con el uso de este recurso en el análisis de opiniones.

2.3.5. SentiWordNet

SentiWordNet, descrito en su totalidad en los artículos de Denecke [36] y Baccianella [37] , es un recurso lingüístico específicamente diseñado para dar soporte a las tareas de clasificación y minería de opiniones. Este recurso se puede usar de forma gratuita con fines dedicados a la investigación. Para desarrollar SentiWordNet se han anotado todos los synsets de WordNet de acuerdo a su grado de positividad, negatividad y neutralidad. Para ello se ha empleado un algoritmo semiautomático que se describe brevemente.

El algoritmo de anotación se divide en cuatro pasos. Primero se tiene un léxico semilla anotado y se expande a través de las relaciones de Wordnet. El segundo paso es entrenar un clasificador con todos los sets generadores para que anote en tres clases: Positiva, negativa y objetiva. El tercer paso es anotar las palabras con el clasificador que se van leyendo a través de un corpus. Un cuarto paso cambia parámetros y clasificadores para seguir anotando. Recientemente se ha empleado otro algoritmo para la redefinición de las puntuaciones y todas estas puntuaciones son supervisadas por criterio humano.

Dado que es un recurso generado a través de Wordnet de forma semiautomática, no se garantiza una precisión total en su contenido. Dado que estos recursos proporcionan rasgos de utilidad para enfoques cuantitativos, se recomienda que se elaboren manualmente por un equipo de expertos lingüistas. De no ser así, añadirán ruido a los enfoques cuantitativos y estos proporcionan resultados pobres. Dado que la tarea es ardua para los lingüistas, se recomienda que al menos todos los resultados de estos algoritmos automáticos sean revisados por los lingüistas.

2.3.6. LIWC: Linguistic Inquiry and Word Count

El recurso LIWC o Linguistic Inquiry and Word Count descrito en el artículo de Tausczik [18] es un recurso que incluye diccionarios para, siguiendo la descripción de los autores, programas que cuentan palabras que pertenecen a categorías de significado psicológico. Este recurso sirve para calcular el porcentaje de palabras que denotan emociones positivas y negativas en un texto.

El principal valor añadido de este recurso es que es producto de una investigación llevada a cabo por psicólogos, por lo que toman en cuenta rasgos como que un gran empleo de los pronombres de primera persona singular implican que el escritor de la opinión puede tener un estado psicológico depresivo. Los autores aportan en [18] una tabla con todas las categorías que incluyen en los diccionarios de este recurso. Por ejemplo, términos que denotan positividad como *Love, nice, sweet* o subcategorías de las categorías positiva y negativa como términos que denotan tristeza: *Crying, grief, sad.*

Un punto negativo de este recurso es que sólo tiene en cuenta unigramas, es decir términos cuyo significante es una palabra. Este recurso no identificaría por ejemplo que la expresión o ngrama *Like a dog with two tails* expresa positividad. Sin embargo, es un recurso creado por psicólogos, lo que garantiza amplia precisión en el espectro cubierto por su contenido.

2.3.7. MPQA Subjectivity Cues Lexicon

El Multi-Perspective Question Answering (MPQA) Subjectivity Cues Lexicon descrito en detalle en el artículo de Wilson [26] es una lista de términos que denotan subjetividad que se ha utilizado como léxico polarizado en un gran número de artículos pertenecientes al área de análisis de opiniones. Este léxico aporta para cada término las siguientes características: La categoría polar a la que pertenece el

término, positiva o negativa; la categoría gramatical del término, la intensidad de la subjetividad que este término expresa, cuyos valores son fuerte o débil; la longitud de palabras de la expresión y si el término es representado por su lexema o en su totalidad. Si estuviese representado por su lexema, entonces todos los términos que tienen ese lexema tienen la polaridad indicada en este recurso para ese lexema. En su versión para descarga, todas las expresiones de este léxico polarizado tienen una longitud de una palabra.

Sin embargo, este léxico polarizado se ha empleado en el trabajo de Wilson [26] para reconocer expresiones subjetivas. Los autores definen expresión subjetiva como aquella palabra o frase usada para expresar una opinión, emoción, evaluación, postura o especulación. Este recurso y los distintos corpus anotados que están disponibles junto al recurso están sujetos a los términos de la licencia GPU, por lo que su uso en investigación es permitido. Dada su facilidad de obtención, es un recurso que se recomienda utilizar, junto a otros, para fines de investigación como base en modelos de análisis de opiniones.

2.3.8. EffectWordNet

Este recurso, descrito en el artículo de Choi [38], es otra expansión de Wordnet que, al igual que SentiWordNet, fue creado anotando todos los synsets de Wordnet con etiqueta positiva y negativa. Partiendo de un léxico base, se construyó un grafo utilizando las relaciones existentes entre synsets de Wordnet y se propagó las etiquetas positivas y negativas a otros términos de forma semisupervisada. Estas técnicas de propagación y expansión de léxicos base se comentarán en el apartado de Creación y expansión de léxicos polarizados y recursos.

En este recurso, cada una de las acepciones de una palabra puede tener, por tanto, una etiqueta diferente. Por ejemplo, el término *purge*, tiene en su acepción de derrocar políticamente una etiqueta negativa pero en su acepción de liberación de un cargo una etiqueta positiva. De forma adicional se añade una última etiqueta, *null*, a aquellos términos o sentidos que no posean polaridad. El recurso contiene 3298 sentidos de términos polarizados con etiqueta positiva y 2427 sentidos de términos polarizados con etiqueta negativa. Adicionalmente presenta 5296 sentidos con etiqueta nula. Para cada uno de estos sentidos se adjunta una frase en la cual se expresa este sentido y una lista de significantes que expresan el significado. El recurso está disponible bajo licencia GPU por lo que su uso en investigación está permitido.

Se recomienda ser crítico con este recurso dado que esta generado mediante Wordnet, si se va a emplear en un modelo de análisis de opiniones, se recomienda también emplear otro recurso.

2.3.9. ConceptNet

Al igual que en el caso de Wordnet, ConceptNet es un recurso de ámbito general que no fue pensado para el análisis de opiniones pero que puede ser empleado para esta tarea. ConceptNet es una red semántica basada en la información de la base de datos Open Mind Common Sense (OMCS). OMCS es una gran base de conocimiento de sentido común cuya construcción ha sido efectuada por cientos de usuarios a través de su interfaz web. Por tanto, ConceptNet es un recurso empleado para el razonamiento de sentido común. Sus nodos son conceptos y las aristas que los conectan son aserciones de sentido común que se producen entre estos dos conceptos. Por ejemplo, *PrerequisiteOf(eat breakfast, wake up in the morning).*

Un ejemplo del uso de ConceptNet en el análisis de opiniones es el ofrecido por Mukherjee [39], que emplea ConceptNet para construir una ontología con todos los rasgos del producto que identifica en el texto. Estos rasgos son extraídos a través de las aserciones lingüísticas y de sentido común de ConceptNet. Una vez ha construido esa ontología, anota con una categoría polar todos los rasgos examinando más textos en búsqueda de expresiones que denoten subjetividad de los rasgos de este producto. Por ejemplo, del concepto cámara extrae de ConceptNet que el Flash es una parte de la cámara para posteriormente anotar la opinión del texto sobre el flash de la cámara. Este es un ejemplo de cómo cualquier recurso, pese a no ser diseñado específicamente para el análisis de opiniones, puede aportar valor al análisis de opiniones en un texto.

2.3.10. WordnetAffect

WordnetAffect, cuya descripción se puede encontrar en el artículo ofrecido por Strapparava [40], es un recurso que al igual que SentiWordnet y EffectWordNet esta creado a partir de anotar los synsets de Wordnet. Al contrario que los dos primeros recursos lingüísticos, WordnetAffect no se centra en las categorías positiva y negativa tan comunes en el análisis de opiniones sino que maneja una jerarquía de categorías afectivas. Ejemplos de las categorías que contempla este recurso son *Emotion, Mood, Trait, Cognitive State, Physical State, Edonic signal, Emotion-Eliciting situation,*

Emotional Response, Behaviour, Attitude y Sensation. Cada categoría tiene una lista de acepciones de términos que pertenecen a ella. Por ejemplo *anger* en su primera acepción pertenece a *Emotion*.

Inicialmente, los autores [40] cuentan con una base de datos léxica denominada AFFECT con 1903 términos relacionados directa o indirectamente con estados emocionales. A partir de las relaciones de sinonimia de Wordnet anotan más synsets con distintas de estas categorías. Los autores asumen que cualquier otra relación gramatical como por ejemplo la hiperonimia también expande la categoría, pero no en su totalidad. En su última versión WordnetAffect contiene 2874 synsets y 4787 términos anotados en las categorías *Emotion,Mood,Trait,Cognitive State, Physical State, Edonic signal, Emotion-Eliciting situation, Emotional Response, Behaviour, Attitude y Sensation*.

Se recomienda su uso, con precaución debido a estar generado desde Wordnet, para computación afectiva y problemas muy específicos relativos al análisis de opiniones. Para el problema de categorizar textos en las categorías positiva y negativa existen otros recursos que hacen que WordnetAffect sea prescindible.

2.3.11. Emotinet

Emotinet, recurso encontrado en los artículos de Balahur [41] e Ibanez [42], es una ontología empleada para la detección de emociones basada en conocimiento de sentido común. Emotinet representa el sentido común que se conoce de los conceptos, de la interacción entre ellos y la consecuencia afectiva que provoca esta interacción. Por ejemplo, Emotinet representa que una acción llevada a cabo por un agente puede repercutir en una emoción. También representa relaciones entre emociones. Si se conoce mediante otro recurso la polaridad de estas emociones, Emotinet es un recurso útil para un modelo de análisis de opiniones. Si se quiere categorizar textos en las categorías positiva y negativa, se recomienda usar otro recurso menos específico.

2.3.12. Bing Liu Opinion Lexicon

Para terminar la lista, otro recurso es el creado por Bing Liu que se encuentra en el artículo de Ding [23]. Bing Liu elaboró un léxico consistente en una lista de términos. Su uso es libre. En su última versión, incluye 2006 términos polarizados positivamente y 4783 términos polarizados negativamente. Ejemplos de características interesantes de este recurso son incluir términos expresados incorrectamente y

variantes morfológicas de los términos. Cuenta con las desventajas de los léxicos que sólo contienen términos de una palabra y no varias y de no considerar acepciones. Útil por su sencillez para los modelos cuantitativos, que resuelven el problema de las acepciones encontrando dependencias entre la aparición de un término del léxico polarizado y otros. Un modelo cuantitativo puede determinar que la coocurrencia de *bajo* y *coste* implica una polaridad positiva tras considerar un número lo suficientemente grande de textos anotados. Por este aspecto, los léxicos polarizados consistentes en unigramas son útiles para los modelos cuantitativos de análisis de opiniones.

2.3.13. Conclusiones

Se han listado una parte de los recursos lingüísticos más representativos en el área del análisis de opiniones. Dada la variedad de los recursos, incluso cubriendo la misma funcionalidad como es la de polarizar términos en las categorías positivo y negativo, es interesante poder ver una comparativa del rendimiento de los mismos en un experimento controlado. Potts, en su tutorial sobre análisis de opiniones [13], incluye una tabla en la cual se muestra el porcentaje de acuerdo entre distintos léxicos de entre los que se han expuesto. Los términos estudiados son los que estos léxicos tienen en común. Los resultados son los siguientes:

	MPQA	Opinion Lexicon	Inquirer	SentiWordNet	LIWC
MPQA	–	33/5402 (0.6%)	49/2867 (2%)	1127/4214 (27%)	12/363 (3%)
Opinion Lexicon		–	32/2411 (1%)	1004/3994 (25%)	9/403 (2%)
Inquirer			–	520/2306 (23%)	1/204 (0.5%)
SentiWordNet				–	174/694 (25%)
LIWC					–

FIGURA 2.2: Porcentaje de acuerdo entre los términos polarizados de distintos recursos lingüísticos.

El alto porcentaje de desacuerdo entre recursos, por ejemplo MPQA con respecto a SentiWordNet, se explica debido a que los recursos polarizan los términos por distintas acepciones. La desambigüación del sentido de los términos es por tanto un problema que presentan los léxicos considerados, pese a que una sección de ellos sí que tienen en cuenta las diferentes acepciones de los términos.

Otro problema existente con estos léxicos es que todos han sido construidos en lengua inglesa, por tanto, no se pueden utilizar para el análisis de opiniones de textos que no estén escritos en lengua inglesa.

Los léxicos presentados pretenden ser generalistas y aplicar a cualquier dominio sobre el que se tenga que efectuar el análisis de opiniones. Esto supone un problema, pues el análisis de opiniones es un problema fuertemente dependiente del contexto. Por ejemplo, en Twitter se usan emoticonos con intención de expresar una opinión positiva o negativa. La gran mayoría de los recursos presentados no contemplan el uso de emoticonos, lo que hace que su uso en redes sociales proporcione peores resultados. La falta de adaptabilidad de los recursos lingüísticos es un problema de los modelos que dependen únicamente de estos recursos para analizar opiniones.

Un problema adicional es el hecho de que la lengua es cambiante y las expresiones que hace tiempo denotaban positividad pueden no hacerlo en la actualidad. Incluso pueden haber surgido nuevas expresiones. Los recursos lingüísticos, al ser estáticos, no contemplan esta casuística.

Debido a los problemas mencionados, se hacen necesarios modelos que adapten los recursos estudiados a los textos que se pretenden estudiar. En la literatura existen más recursos que han surgido a partir de la expansión de los léxicos mencionados o de procesos de fusión de los léxicos que se han expuesto. El objetivo de estos modelos de expansión es hacer que el modelo se adapte a distintos contextos sobre los que trabaja. Los procedimientos para la expansión y creación de léxicos polarizados automática o semiautomática van a ser estudiados en la siguiente sección.

2.4. Creación y expansión de léxicos polarizados y recursos

En este apartado se comentan distintos enfoques para la creación y expansión de léxicos polarizados. Estos enfoques comienzan con un léxico polarizado semilla que expanden con nuevos términos mediante distintos procedimientos. Sólo se expone de las referencias citadas lo referente a la expansión de léxicos polarizados. El objetivo es exponer los distintos enfoques que se han considerado para expandir léxicos polarizados. Históricamente se pensó que los recursos lingüísticos por si solos serían capaces de analizar las opiniones con eficiencia. Posteriormente se descubre que depende del contexto, necesitan ser expandidos.

2.4.1. Enfoques basados en Wordnet

Un primer enfoque, ya citado en el apartado anterior, planteado por Kim [20], es emplear las relaciones del recurso Wordnet para expandir el número de términos del

léxico base. Kim parte con un léxico base de 23 verbos positivos, 21 verbos negativos, 15 adjetivos positivos y 19 adjetivos negativos. Kim obtuvo sinónimos de los verbos y los adjetivos a los que asignó la misma polaridad y antónimos de los adjetivos a los que invierte la polaridad. Con esto obtuvo 5880 adjetivos positivos, 6233 adjetivos negativos, 2840 verbos positivos y 3239 verbos negativos. Con este nuevo léxico ampliado, Kim define modelos para obtener una probabilidad de clasificar una nueva palabra, w, en una categoría polar c que puede ser positiva, neutral o negativa. Un ejemplo para calcular $P(c—w)$ sería el siguiente:

$$\underset{c}{argmax} P(c|w) = \underset{c}{argmax} P(c) \frac{\sum_{i=1}^{n} count(syn_i, c)}{count(c)}$$

Es decir, la palabra es asignada a la categoría que maximice la expresión planteada. La probabilidad a priori de una clase, $P(c)$ se calcula por máxima verosimilitud, es decir, del total de palabras del léxico, $P(c)$ es la proporción de palabras que pertenecen a esa categoría. La función $count(X)$ devuelve una variable que modeliza el número de ocurrencias del suceso X. Después, se cuenta la ocurrencia de los sinónimos de w en la lista de las palabras polarizadas en cada categoría. Esto se divide entre el total de palabras de la categoría. En otras palabras, la nueva palabra pertenecerá a la categoría que tenga una proporción de términos polarizados mayor. Este es un ejemplo de la utilidad de expandir un léxico polarizado, construir un clasificador.

Kim cita como problemas las distintas acepciones de las palabras y que su léxico polarizado sólo contempla unigramas, es necesario que el léxico polarizado contemple ngramas pues hay significantes que sólo denotan un significado si son expresados mediante ngramas.

Pero este no es el problema principal, pues los unigramas si son útiles para enfoques cuantitativos. Godbole [14] menciona que los recursos que utilizan Wordnet para ampliar los léxicos polarizados pueden sufrir incoherencias. Este problema aplica a aquellos métodos que simplemente utilizan un mecanismo recursivo para extender la polaridad de un término a otros. El problema radica en que la coherencia de los sets de sinónimos de Wordnet cae con la distancia. Por ejemplo, se puede llegar a obtener antónimos mediante la obtención de sucesivos sets de sinónimos. Por ejemplo, la siguiente figura detalla 4 vías de obtener *bad* desde *good* a base de recorrer recursivamente sets de sinónimos.

Godbole también extiende su léxico polarizado a través de sinónimos y antónimos en Wordnet, por lo cual puede incurrir en una pérdida de coherencia de su léxico

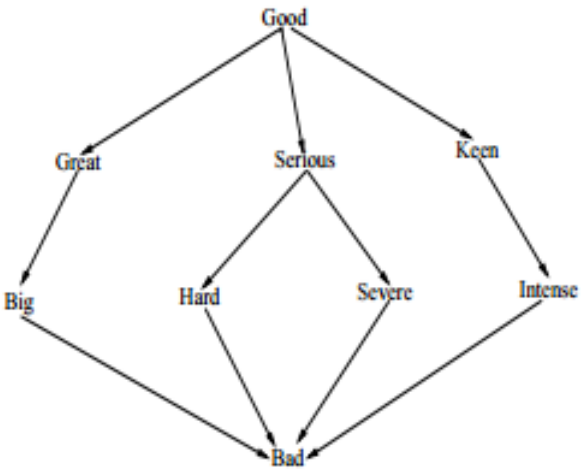

FIGURA 2.3: Cuatro formas de obtener *bad* desde *good* en tres saltos.

polarizado extendido. Para evitar esta pérdida de coherencia, Godbole propone un enfoque en el cual la polaridad de los términos disminuya según se expande la polaridad. Este enfoque se compone de los siguientes pasos: En primer lugar, se asocia una polaridad a priori a cada palabra perteneciente al léxico polarizado. Los sinónimos heredarán esta polaridad, los antónimos tendrán la polaridad opuesta.

Los autores asumen que la polaridad decrece en función a la longitud que separa al sinónimo o antónimo de la palabra semilla. Esta polaridad es calculada mediante la siguiente expresión: $score(W) = 1/c^d$. Donde c es una constante mayor que 1 y la variable d es la distancia de W con respecto a la palabra semilla.

La polaridad total de una palabra es el sumatorio de esta expresión para todas las palabras que estén relacionadas desde distintas profundidades con esta. Para encontrar todas las palabras relacionadas entre sí, se emplean iteraciones y se va redefiniendo la polaridad de cada término. Solamente se consideran los sentidos puestos en primer lugar por Wordnet y si hay términos con polaridades opuestas en el sumatorio se descartan. Pese a los problemas que se han comentado, Wordnet es un recurso que sirve como base para la construcción y expansión de léxicos polarizados como +/-EffectWordnet [38] entre otros. De todos modos, debido a las posibles incoherencias, es necesario ser crítico con estos enfoques y someterlos a las debidas evaluaciones que demuestren que el modelo de análisis de opiniones mejora con su inclusión. Por ello, se mencionan a continuación enfoques que no dependen de Wordnet para expandir un léxico polarizado.

2.4.2. Enfoques basados en reglas y métodos de aprendizaje no supervisado

Pero no sólo se han utilizado recursos lingüísticos para la expansión de léxico a partir de un léxico polarizado semilla. Existen otros enfoques que se basan en analizar textos y satisfacer reglas lingüísticas o ejecutar algoritmos para encontrar que términos se deben incluir en el léxico polarizado.

Huang [43] emplea este enfoque y al igual que Kim, después emplea el léxico para clasificar. En este caso, Huang primero detecta posibles candidatos a ser polarizados mediante diversas técnicas que va introduciendo en una lista de candidatos. Después, Huang mide la ocurrencia de estos candidatos junto a términos ya polarizados y en función a esta coocurrencia polariza.

Para introducir términos en la lista de candidatos Huang emplea reglas lingüísticas y técnicas de carácter no supervisado. Posteriormente hace una votación entre el resultado de estas técnicas para ver que candidatos se incluyen en la lista. Empieza con los más votados y acaba con los menos votados. Se comentan brevemente estos enfoques.

Huang, que trabaja con la lengua China, define expresiones regulares que simbolizan patrones léxicos que se repiten con frecuencia en un texto. Estos patrones están compuestos de combinaciones entre etiquetas de categorías gramaticales. Analiza textos y construye un listado con los patrones léxicos más comunes. Con estos patrones detecta adverbios que pueden ser modificadores y nuevas palabras candidatas a introducirse en el léxico polarizado semilla al ocurrir en el patrón junto a las ya polarizadas.

Aparte de reglas lingüísticas, Huang emplea algoritmos de carácter no supervisado. Estos algoritmos se comentarán en detalle en la sección Métodos estadísticos y basados en Aprendizaje No Supervisado de esta memoria. El objetivo de esta exposición es destacar que Huang emplea estas técnicas para introducir términos en la lista de candidatos, no describir las técnicas, aún así se incluye el siguiente ejemplo. Para detectar ngramas y que no sólo se polaricen unigramas en la lista de candidatos, Huang emplea la medida Enhanced Mutual Information (EMI). La idea de este algoritmo es detectar como de frecuentemente distintos términos ocurren simultáneamente frente a su ocurrencia en soledad. A mayor valor de esta medida, mayor probabilidad de que el término pertenezca a un ngrama. La expresión, cuyas variables se explican en el siguiente párrafo, es la siguiente:

$$EMI(w) = log_2 \frac{F/N}{\prod_{i=1}^{n} \frac{F_i - F}{N}}$$

F es el número de segmentos de texto en los cuáles una expresión de distintos términos ocurre, F_i es el número de segmentos en los cuales la palabra w_i ocurre y N es el número total de segmentos de textos. Esta medida obtiene la proporción de veces que la expresión ocurre frente a las palabras en separado. El logaritmo se emplea para que los valores extremos no tengan tanta importancia frente a los valores pequeños, puesto que la mayoría de veces el valor de esta medida es pequeño pero aún siendo pequeño es representativo.

Para polarizar, Huang calcula la coocurrencia de los ngramas que se han introducido en la lista de candidatos con respecto a palabras polarizadas previamente como positivas y palabras polarizadas previamente como negativas. Si la coocurrencia es mayor con respecto a las palabras positivas entonces se clasifica al ngrama como positivo y negativo en caso contrario. Para medir la coocurrencia emplea la Pointwise Mutual Information (PMI) y la votación por conteo. Con este enfoque, Huang expande léxicos polarizados empleando únicamente categorías gramaticales como recurso lingüístico para los patrones léxicos. Añadir más rasgos a las categorías gramaticales es recomendado para obtener mayor precisión.

Zhang, en el sistema propuesto en su artículo que se va a describir a continuación [1] también emplea reglas lingüísticas para polarizar las categorías gramaticales. A partir de una frase, Zhang emplea el generador del árbol de dependencias de Stanford para parsear la frase. Una vez está parseada, Zhang se queda sólo con los términos de aquellas relaciones que pertenecen a ciertos tipos como *amod* o *nsubj*. El resto de relaciones las denomina *nonsentiment triples*, y las ignora. A partir de un léxico base, busca añadir términos recogidos en estas tripletas para construir un léxico expandido.

Las tripletas resultantes son polarizadas mediante un modelo de aprendizaje automático, denominado Conditional Random Fields, que será explicado en la sección de Aprendizaje Automático. Los rasgos que emplea Zhang para categorizar los términos de estas tripletas son la categoría gramatical, si aparece junto a un término ya polarizado, la relación entre ambos, la aparición junto a términos clave, si está en una frase con algún signo de puntuación o entidad y la parte del texto donde aparece esta tripleta (principio, mitad y final del texto). Con este enfoque, Zhang consigue los siguientes resultados:

	ANEW			LIWC			ReNew		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Positive	0.59	**0.994**	0.741	0.606	0.975	0.747	**0.623**	0.947	**0.752**
Negative	0.294	0.011	0.021	**0.584**	0.145	0.232	0.497	**0.202**	**0.288**
Neutral	0	0	0	0	0	0	**0.395**	**0.04**	**0.073**
Weighted average	0.41	0.587	0.44	0.481	0.605	0.489	**0.551**	**0.608**	**0.518**

FIGURA 2.4: Resultados del procedimiento de expansión de léxico de Zhang [1].

Se aprecia cómo se mejora todas las cifras excepto la precisión de los términos negativos con respecto al LIWC. Este hecho se interpreta del siguiente modo: Dado que hay un mayor número de términos en el léxico polarizado, existe un mayor número de falsos negativos al contemplar más términos. Aún así, el procedimiento mejora la calidad de los textos puesto que la F-Measure es mejor para las tres categorías y dado que se contempla la categoría neutral, que no se contempla en los dos léxicos polarizados semilla. Se recomienda supervisar mediante métricas de evaluación el crecimiento de falsos negativos y no solo basarse en la precisión obtenida para emplear y juzgar acordemente este modelo. Pese a ello se considera que es un enfoque correcto que además hace uso de un recurso muy útil, el árbol de dependencias de Stanford. Por contra, el problema del multilingüismo es vigente en este enfoque, que tiene una dependencia fuerte hacia un recurso costoso de construir en otra lengua como es el árbol de dependencias de Stanford. Se debe procurar evitar la dependencia de un modelo de análisis de opiniones a recursos u otros modelos de difícil construcción.

Volkova propone un modelo que se puede consultar en su publicación [44] para obtener un léxico polarizado exclusivo para Twitter. El enfoque es tomar un léxico polarizado de propósito general, en este caso emplea el MPQA, y ampliarlo analizando textos no etiquetados. El proceso se guía mediante un corpus anotado.

El procedimiento seguido se divide en dos pasos, primero polariza y luego extiende el léxico considerado. Para polarizar, en cada iteración del algoritmo se anotan tweets mediante el léxico de la iteración anterior. Se anota un tweet en dos pasos. Primero, si el tweet contiene uno o más términos del léxico, entonces es categorizado como subjetivo, de lo contrario es categorizado como objetivo. Si es subjetivo y contiene más de un término positivo y ninguno negativo se categoriza como positivo, si contiene más de un término negativo y ninguno positivo se categoriza como negativo, si contiene términos positivo y negativo se categoriza como *ambos*. Se tiene en cuenta la negación. Una vez polarizado el tweet, se inicia el procedimiento

Para los términos del Tweet que no están en el léxico w y su frecuencia es mayor que una frecuencia umbral θ_{freq} se calcula primero la probabilidad de que el término sea subjetivo $p^{subj}(w)$. Esta probabilidad es calculada como una razón del número de veces que un término coocurre con uno que está en el léxico polarizado frente al número de veces que ocurre el término, es decir: $p^{subj}(w) = \frac{c(w,L_B(\vec{\theta}))}{c(w)}$. Donde $c(w, L_B(\vec{\theta})$ es un contador del número de veces que el término y uno que pertenece al léxico polarizado ocurre y $c(w)$ es el número de veces que el término a polarizar ocurre. El número de términos con mayor probabilidad de ser subjetivos se polarizan en cada iteración. Este número es un parámetro de configuración del modelo.

El procedimiento de polarización es similar al de categorización de término subjetivo u objetivo. Se calcula la probabilidad de que este término sea positivo mediante la siguiente expresión: $p^{pos}(w) = \frac{c(w,L_B^{pos}(\vec{\theta}))}{c(w,L_B(\vec{\theta}))}$. Esta expresión mide la proporción de veces que el término aparece junto a un término polarizado positivamente. Si es mayor a 0.5, el término se incluye en el léxico polarizado con categoría positiva y si es menor a 0.5 se incluye en el léxico polarizado con categoría negativa.

El algoritmo itera siguiendo este procedimiento hasta que se decide interrumpir las iteraciones. En la evaluación se observa que se incrementa el recall empleando este procedimiento pero se penaliza la precisión. Es lógico, puesto que la forma de expandir el léxico categoriza términos en una categoría pese a que también ocurran junto a términos de la otra categoría. Se considera un buen mecanismo para expandir el léxico polarizado en lo relativo al recall, pero debe estar más controlada la precisión. Además de los enfoques vistos, existe otro mecanismo para expandir los léxicos polarizados que se comenta a continuación.

2.4.3. Enfoques basados en modelos de optimización

Además de construir léxicos polarizados en categorías positiva, negativa y neutral existen otros enfoques que tratan de polarizar en un mayor número de categorías. Uno de los más novedosos es el que se puede consultar en el artículo de Patra [45], que utiliza un modelo para construir un léxico que anote sus términos en las categorías *anger*, *disgust*, *fear*, *happy*, *sad* y *surprise*.

Para ello, utiliza un léxico base que consta de una sola palabra anotada a mano que pertenece a cada uno de estas categorías. Para expandir este léxico base necesita construir un grafo cuyas uniones entre nodos estén ponderadas. Se construye esta red léxica uniendo dos palabras si una aparece en la glosa de la otra palabra. El recurso

empleado para consultar las glosas es un diccionario. Cada unión entre dos palabras representa la misma orientación, conjunto SL o distinta orientación, conjunto DL. Se considera la negación. Se termina con una red léxica de 88015 términos. A las uniones entre términos, w_{ij}, se les asigna un peso mediante la siguiente expresión:

$$w_{ij} = \begin{cases} \frac{1}{\sqrt{d(i)d(j)}}(e_{ij} \in SL) \\ -\frac{1}{\sqrt{d(i)d(j)}}(e_{ij} \in DL) \\ 0 \text{ otherwise} \end{cases}$$

Tal que e_{ij} representa la unión entre la palabra i y la j y *d(i)* es el número de palabras a las que está unida la palabra i. A medida que i y j están unidas con mas palabras, el peso de su unión es más bajo para simular la pérdida de coherencia ya comentada por Godbole [14]. Además de esta red, Patra, construye otra que denomina *Gloss Thesaurus Network (GT)* que une sinónimos, antónimos e hiperónimos además de las palabras ya unidas en el enfoque presentado.

Una vez construido el grafo, Patra expande la polaridad de las palabras semilla que pertenecen a las 6 categorías mediante el Modelo de Potts. Los términos no considerados en la red construida se clasificarán como palabras no vistas. Este modelo toma como entrada 6 palabras ya clasificadas con un estado, que se relacionan con otras palabras del grafo mediante las uniones ponderadas. Con la clase de las palabras ya clasificadas, estima la clase del resto de palabras en el grafo. Se comenta a continuación la adaptación del modelo de Potts a la red léxica.

Sea n el estado colectivo de los términos, w_{ij} el peso de la unión entre términos, β una constante llamada temperatura inversa, L el set de índices para las términos observadas, α_i el estado de cada término y α una constante positiva si el término n_i ya ha sido anotado o es anotado inicialmente. El modelo de Potts define la función de energía $H(n)$ a minimizar, que indica el estado global de entropía de la red, como:

$$H(n) = -\beta \sum_{ij} w_{ij}\delta(n_i, n_j) + \alpha \sum_{i \in L} -\delta(n_i, \alpha_i)$$

La función δ devuelve 1 si dos argumentos son iguales y 0 en caso contrario. Se busca minimizar la expresión expuesta *H(n)*. Para que esta expresión se minimice, los estados de nodos adyacentes con gran peso deberán ser los mismos $w_{ij}\delta(n_i, n_j)$, manteniendo el estado inicial de las variables ya anotadas $\alpha \sum_{i \in L} -\delta(n_i, \alpha_i)$ para que la energía global de la red sea la misma. Se consigue que términos que el grafo

muestra como relacionados pertenezcan a la misma categoría expandiendo las categorías iniciales sin que se corrompa o se pierda coherencia en la expansión. Se resuelve este problema de optimización mediante un mecanismo iterativo que puede ser consultado en [45].

Los resultados de este experimento son positivos pero los autores insisten en que es prioritario en su trabajo futuro borrosificar las categorías de cada palabra. De esta forma un término podrá pertenecer a distintas categorías. Para la expansión de dichas categorías habría que hacer que el modelo funcione si cada término puede pertenecer a distintas categorías con un grado borroso. En su actual estado, se recomienda esperar al modelo ampliado de los autores.

Un último enfoque que propone la construcción de léxicos polarizados con múltiples categorías es el propuesto por Feng en su publicación [46], cuyo léxico determina la connotación de cada palabra en función a otras palabras. Por ejemplo, considérese la frase: *Geothermal replaces oil-heating: it helps reducing greenhouse emissions.*. En esta frase, la palabra *emissions* tiene una connotación negativa en la opinión de los lectores con respecto a *Geothermal*. En este contexto, la polaridad de la palabra *reduces* condicionada a la aparición simultánea de *emissions* y *Geothermal* es positiva.

Este modelo emplea la medida PMI para discernir las principales coocurrencias entre palabras y así construir las categorías condicionadas entre distintos términos que se deben polarizar. Tomando como léxico inicial el MPQA, emplea Wordnet con sus relaciones de sinonimia y antonimia para construir un grafo que dictará la propagación de la polaridad hacia otros términos.

Una vez ha construido el grafo, expande la probabilidad mediante un problema de optimización entera lineal que maximiza múltiples funciones objetivo. Estas funciones objetivo toman en cuenta que se mantenga la coherencia entre los sets de sinónimos y antónimos. Por ejemplo, si dos palabras están directamente conectadas mediante un enlace en el grafo, entonces deben tener una polaridad similar.

Se definen con esta idea objetivos que intentan mantener coherencia en el grafo expandiendo la polaridad. Las variables que manejan estos objetivos son variables binarias que indican si cada término tiene polaridad positiva x_i, negativa y_i o neutral z_i. Por ejemplo, el objetivo coordinación, explicado en el siguiente párrafo, es el siguiente:

$$\Phi^{coord} = \sum_{i,j}^{R} w_{i,j}^{coord}(d_{i,j}^{++} + d_{i,j}^{--} + d_{i,j}^{00})$$

En este ejemplo se muestra que las palabras $w_{i,j}$ deben tener coordinación. Es decir, cuando se expande la polaridad entre dos términos, esta se debe respetar. Los términos w_i y w_j vecinos en el grafo deben tener la misma polaridad. Habrá tantas variables d como vecinos w_i y w_j hay en el grafo. La variable $d_{i,j}^{++}$ valdrá 1 si los términos i y j tienen polaridad positiva los dos y 0 en caso contrario. La variable $d_{i,j}^{--}$ valdrá 1 si los términos i y j tienen polaridad negativa los dos y 0 en caso contrario. La variable $d_{i,j}^{00}$ valdrá 1 si los términos i y j tienen polaridad neutral los dos y 0 en caso contrario. Por lo tanto, si se maximiza el sumatorio de todas estas variables, se construye un grafo en el cual la coherencia entre las polaridades de los vecinos es la máxima posible.

Dado que se trata de un problema multiobjetivo, esto no será siempre así, sino que se obtendrá el valor máximo de esta expresión sujeto a valores máximos de otros objetivos. Es un problema multicriterio y hay que encontrar una solución no dominada en todos los objetivos. El criterio de cuál será la mejor solución para todos los objetivos de entre las no dominadas podrá ser determinado por el usuario o mediante una función de utilidad, empleadas en análisis de decisiones multicriterio.

Los detalles del resto de funciones objetivo se presentan en [46]. Con este enfoque se aumenta la precisión, recall y F-Measure del MPQA. El concepto es muy parecido al de Patra [45] con su función de energía a minimizar en el grafo que construía. Ambos enfoques proponen la expansión del léxico polarizado como un problema de optimización y se considera que son una alternativa viable a los presentados en los otros apartados de la sección de expansión de léxicos polarizados.

Hasta este punto del estado del arte, se han comentado enfoques que son capaces de realizar análisis de opiniones y los recursos que estos utilizan para funcionar. Estos enfoques no son los únicos que se han empleado para el análisis de opiniones. A lo largo de la última década, se ha recurrido a enfoques estadísticos para resolver esta tarea con muy buenos resultados como alternativa a los modelos mencionados. Dado el alto volumen de publicaciones que emplean modelos de esta naturaleza, se hace imprescindible dedicar un apartado del estado del arte a estos modelos.

2.5. Métodos basados en Estadística y Aprendizaje Automático

2.5.1. Introducción

El lenguaje se trata de una herramienta rica, mediante la cual se puede expresar la misma idea usando distintas combinaciones de palabras. Del mismo modo, una palabra o frase puede expresar distintas ideas en distintos contextos. Este concepto se conoce como ambigüedad, y es inherente al lenguaje. Además, el empleo de una misma lengua varía entre culturas y con el paso del tiempo, lo que hace que en distintos contextos se entienda una frase de distinta forma.

Por ello, es muy difícil analizar el lenguaje siguiendo reglas o patrones definidos a priori. La validez de estas reglas está condicionada a factores como la cultura, la época o la acepción de las palabras analizadas. Aún sin considerar estos, elaborar modelos de análisis de opiniones en un dominio amplio o varios dominios considerando únicamente modelos cualitativos es una tarea ardua. Los modelos cualitativos, pese a contar con gran precisión y ser el mejor enfoque para dominios reducidos, cuentan con la problemática de escalar en ámbitos generales.

Se necesita un gran número de expertos y de tiempo para dividir el conjunto total de textos en categorías y tener un conjunto de reglas para cada una de estas categorías. En contraste, si se dispone de un corpus de textos anotados, incluso por modelos cualitativos, los modelos cuantitativos son capaces de crear un conjunto de reglas automáticamente que no alcanzan la precisión de los modelos cualitativos, pero que cubren todos los dominios. Otros modelos de aprendizaje automático clasifican los textos mediante otros mecanismos que no son reglas, como se estudiará más adelante. Por esto motivo, para clasificar los textos en categorías teniendo un conjunto de textos anotados del cual se extrae características, se emplea el aprendizaje automático.

En el procesamiento del lenguaje natural, los modelos de aprendizaje automático generan clasificadores capaces de generalizar comportamientos a partir de una lista de textos anotados. Gracias a esta capacidad de generalización, aportan un valor añadido a los modelos cualitativos, más precisos pero de menor alcance. Desde un punto de vista personal, el enfoque óptimo es combinar la precisión de los modelos cualitativos con la generalización de los cuantitativos.

En la última década, como alternativa a los modelos cualitativos, han emergido publicaciones que tratan de procesar el lenguaje natural empleando enfoques estadísticos, de aprendizaje automático y de clustering. Estos modelos encuentran patrones que se repiten en distintos textos con los que infieren el significado subyacente de significantes no considerados, al encontrar analogías de estos significantes en textos categorizados a priori. Ejemplos de problemas tratados por estos modelos han sido la categorización de textos, de la cual se puede consultar una útil aplicación para la detección de Spam en el artículo de Metsis [47], y el análisis de opiniones.

Los modelos de aprendizaje automático son aquellos que generalizan comportamientos a partir de información no estructurada. En el caso del análisis de opiniones, esta información no estructurada es suministrada en forma de un corpus de documentos. Se les denomina modelos de aprendizaje supervisado a aquellos modelos que necesitan que esta información esté clasificada en categorías para poder clasificar nuevas instancias. Los modelos de aprendizaje no supervisado o clustering y los enfoques estadísticos o de minería de patrones no necesitan que esta información esté categorizada.

Los modelos de aprendizaje supervisado y de clustering necesitan extraer características de los textos. Estas características se denominan rasgos. Por ejemplo, una variable cuyo valor es 1 si el texto contiene la palabra *buen* y 0 sino aparece es un rasgo. El resultado de usar una técnica estadística o de minería de patrones puede considerarse otro rasgo. La metodología seguida por los modelos es transformar cada texto en un vector de rasgos. Con este vector, los modelos etiquetan al texto en una de las categorías en función del valor de las variables pertenecientes a su vector de rasgos y de las instancias convertidas en vectores del corpus. A continuación se muestran los rasgos comentados en la literatura.

2.5.2. Características a extraer de los Conjuntos de Textos

En este apartado se exponen los distintos rasgos considerados en la literatura para los modelos de aprendizaje automático. Estos rasgos generan variables que pueden ser continuas, categóricas o binarias. Cada texto contenido en el corpus es representado por el vector que generan estas variables. Este vector puede ser anotado con una categoría. Los modelos utilizarán estos vectores y el vector generado por un nuevo texto para anotar el nuevo texto.

Cada uno de los rasgos comentados genera distintas variables en función de cómo se analizan. Por ejemplo, considérese el rasgo unigramas. Se genera: Una variable

continua que simboliza la frecuencia de aparición del unigrama, una variable binaria que simboliza la aparición o no del unigrama, una variable categórica cuyo valor es *Alta* si el unigrama se presenta más de 5 veces, *Medio* si el unigrama es detectado entre 2 y 5 veces *Detectado* si el unigrama sólo aparece una vez y *No Detectado* sino se detecta.

Esta lógica aplica a todos los rasgos comentados. Puede ser una buena estrategia considerar múltiples variables pertenecientes a un solo rasgo, depende del problema. Para elegir que rasgos son mas óptimos para modelizar cada problema, existen métodos conocidos como *Feature Subset Selection Methods* que, presentados todos los rasgos, se quedan con los representativos.

Es importante destacar que es muy posible que un rasgo en soledad no sea indicativo de polaridad, pero que condicionado al valor de otros si determine polaridad. Los modelos son los encargados de descubrir estas dependencias. Los rasgos que se emplean en el análisis de opiniones son los siguientes, para los cuáles, se expone una breve descripción del rasgo y se enumeran todas las referencias en las que se encuentran modelos de análisis de opiniones que emplean el uso de cada uno de los rasgos:

- **Unigramas**: El unigrama es un término compuesto por una única palabra. A efectos del analista de opiniones, se puede considerar un sinónimo de palabra. Depende del análisis de opiniones que se quiera efectuar, será útil construir un diccionario de unigramas que necesitan ser detectados en el texto. Las variables más mencionadas en la literatura con respecto a unigramas son: Una variable binaria que simbolice la aparición del unigrama en el texto o una variable real que simbolice la puntuación del algoritmo TF-IDF sobre el unigrama. Los modelos cuantitativos son capaces de establecer dependencias entre un subconjunto de estas variables que explican la clasificación de un texto en una categoría. Por ello, se recomienda desde un punto de vista personal el empleo de un gran conjunto de N-Gramas. Si resulta que no son explicativos de la clasificación de un texto en una categoría, no hay problema, puesto que la falta de valor añadido de información de uno de estos rasgos será detectada por un filtro. Un ejemplo es la detección del término *bueno*. En las siguientes referencias se pueden encontrar modelos que emplean unigramas: [48] [49] [50] [38] [51] [52] [26] [53] [54] [55] [56] [57].
- **N-Gramas**: Se persigue detectar una secuencia de términos en el texto que sigue un orden predeterminado. Esta secuencia de términos hace referencia a un concepto que sólo es expresado mediante la unión de términos en un orden

determinado. Se entiende como N-Grama a cualquier secuencia de palabras compuesta por 2 o más términos. Estos pueden o no referirse a un único significado. Depende del análisis a realizar será útil recuperar N-Gramas de distinto número de términos. El N-Grama es un término genérico que engloba a los bigramas, términos formados por 2 palabras; trigramas, términos formados por 3 palabras, etc... No hay una longitud máxima predefinida. Ejemplo, detección de la secuencia: *Estados Unidos*, que es un bigrama que hace referencia a un único significado. En las siguientes referencias se pueden encontrar modelos que emplean N-Gramas: [48] [49] [50] [38] [58] [52] [53] [54] [57]. Como ya se ha mencionado en el anterior apartado, los modelos cuantitativos encuentran dependencias entre varios unigramas. Por ello, considerar como rasgo a un N-Grama puede ser útil si se trata de una secuencia de términos que simbolice un único significado, como *Estados Unidos*, pero es más costoso elaborar un recurso que considere N-Gramas y el valor añadido de información es menor con los unigramas. Si la dependencia entre dos variables que modelan la aparición de dos unigramas es detectada, el valor añadido de incluir el bigrama que modela la aparición de las dos variables es altamente probable que sea muy pequeño con respecto al de las dos variables anteriores. En consecuencia, un filtro eliminará la variable que modela el bigrama.

- **Raíz de términos**: Se denomina raíz o lexema de un término a la parte básica de la palabra obtenida tras remover las unidades con significado gramatical que lo complementan añadiendo género, número, aumentativo, diminutivo, etc.. Por ejemplo, el lexema de *destornillador* es *tornillo*. Se emplean lemmatizadores para obtener estos lexemas. Se considera un rasgo útil, puesto que una sola variable modela lo que más variables representantes de unigramas modelan. En consecuencia, un filtro elimina las variables representantes de unigramas que pueden ser modelados con su raíz . En las siguientes referencias se describen modelos que usan raíces de términos: [59] [51] [52].

- **N-Gramas con frecuencia mínima**: Este rasgo es representado mediante una variable binaria que indica si el N-Grama es reconocido con una frecuencia igual o superior a una determinada. Por ejemplo: Detección de *Muy bien* al menos 3 veces. En el artículo de Pla se describe un modelo que hace uso de este rasgo [59]. En determinados problemas, se considera un rasgo útil que puede aportar información adicional.

- **N-Gramas con palabras intermedias**: Mediante una expresión regular, se reconoce un N-Grama que contiene términos entre él mismo. Considérese el N-Grama: *Verdaderamente bien*. Si se analiza la frase *Verdaderamente muy bien*,

este N-Grama sería detectado. El modelo de Elming cita este rasgo [48]. No se considera de utilidad puesto que un modelo cuantitativo detectaría la correlación entre varios unigramas y la clasificación de un texto en una categoría. Si no se usa un modelo cuantitativo, puede resultar útil.

- **Categorías gramaticales**: Se identifica la categoría gramatical de cada término en el texto. Por ejemplo, *bonito* es un adjetivo. Se busca la frecuencia o aparición de categorías en frases como distintos rasgos. Se encuentran los siguientes modelos en la literatura que emplean este rasgo: [49] [59] [60] [52] [61] [26] [53]. En el análisis de opiniones se ha considerado un enfoque común el polarizar un texto en base a analizar únicamente objetivos. Además, se puede combinar el rasgo de la detección de un unigrama con la condición de que este pertenezca a una categoría gramatical determinada. Esto puede ayudar a la desambigüación de una acepción de un término y a la detección de información útil por modelos cuantitativos. Por ello, se recomienda modelar un conjunto de variables que simbolicen categorías gramaticales.

- **Términos categorizados como ambiguos**: El enfoque presentado por Ali [61] utiliza un recurso que contiene términos que clasifica como no polares al ser ambiguos en lo respectivo a su categoría polar. Necesitan ser desambiguados en función al contexto. Por ejemplo: El término *grande* condicionado a la aparición del término *premio* tendrá polaridad positiva, pero si aparece junto al N-Grama *cola de espera* tendrá polaridad negativa. En principio es un rasgo de carácter cualitativo, pero que su inclusión en modelos cuantitativos garantiza una mayor calidad en el análisis de opiniones.

- **Patrones léxicos y posición de los mismos**: Se busca detectar combinaciones de categorías gramaticales en un orden predefinido. Por ejemplo: Sustantivo+Adjetivo+Verbo+Adverbio. Se ha incluido junto a la posición en las cuáles se incluyen estas categorías gramaticales en el texto. Esto es debido a que ambos rasgos suelen considerarse de forma conjunta. Por ejemplo: Se busca el anterior patrón léxico en la última frase del texto. En los siguientes artículos se encuentran modelos que emplean estos rasgos: [60] [52] [26] [53] [54]. También es posible considerar como rasgo separado la posición en la que se encuentra un término, N-Grama o categoría gramatical en el texto. Los modelos cuantitativos pueden concluir en estos patrones mediante la modelización de variables más sencillas. Aun así, si ya son nutridos por estos enfoques cualitativos, se garantizará que la calidad del modelo final será superior.

- **Presencia parcial de N-Gramas**: Una parte de los N-Gramas pueden contener significado sin incluir todos los términos que lo componen. Por ejemplo:

Verdaderamente muy bien se puede descomponer en dos bigramas que tienen polaridad positiva: *Muy bien* y *Verdaderamente bien*. Por ello se considera como rasgo la detección parcial de N-Gramas. Elming incorpora en su modelo estas variables [48]. Se considera un rasgo secundario puesto que los modelos de aprendizaje automático pueden crear estas dependencias sin la intervención de un experto.

- **Distribución de la Longitud de los Términos**: El enfoque de Abbasi [52] propone la distribución de la longitud de los términos de un texto como rasgo. Se analiza si emplear palabras con longitud superior está relacionado con la polaridad de un texto. En principio, es un rasgo secundario.

- **N-Gramas pertenecientes a léxico polarizado**: Se analiza la polaridad que propone un léxico polarizado de los términos que aparecen en el texto a analizar. Se trata de un rasgo muy popular entre la comunidad encontrado en las siguientes publicaciones.: [49] [38] [59] [52] [61] [26] [56] [57]. Desde un punto de vista personal, se considera uno de los rasgos más útiles pues combina la información de modelos cualitativos desarrollados por expertos con el potencial de los modelos cuantitativos capaces de encontrar dependencias entre estas variables y las categorías en las cuales se pretende clasificar un texto.

- **Clusters de Brown y otros mecanismos de Clustering**: Los clusters de Brown pertenecen al clustering jerárquico. La finalidad es agrupar términos basados en los contextos en los que aparecen y generar probabilidades de ocurrencia de términos.

  Se agrupan a los términos w_i en clusters que simbolizan categorías gramaticales que se anotan como c_i. Con esto se puede modelizar el lenguaje y predecir la aparición de un término w_i basado en el término reconocido w_{i-1} mediante la siguiente expresión: $P(w_i|w_{i-1}) = P(w_i|c_i) * P(c_i|c_i - 1)$. Mediante mecanismos de *Smoothing* se puede modelizar la probabilidad de ocurrencia de términos que no se encuentran en el corpus.

  Los rasgos extraíbles son las probabilidades de los términos y de los contextos o en función de los contextos. Este mecanismo se encuentra en los siguientes enfoques: [48] [55] Cualquier modelo de clustering del procesamiento del lenguaje natural puede aportar rasgos.

- **Presencia de números cardinales** [52] [26]: La presencia de ciertos números cardinales puede ser indicativo de polaridad. Por ejemplo, en críticas de cine, la presencia del cardinal *5* o *10* puede simbolizar el concepto de 5 estrellas o de valoración 10/10, conceptos que expresan una polaridad positiva. Es un rasgo

muy dependiente del dominio, pero que en el análisis de un dominio cerrado como las críticas de cine, puede aportar información crítica para clasificar el texto en una categoría.

- **Palabras con todas las letras mayúsculas**: En las redes sociales es común encontrar palabras como *TREMENDO* escritas en mayúsculas. Estas palabras pueden denotan subjetividad en un texto como citan los siguientes artículos: [48] [49]. Se trata de un rasgo que sin duda denota subjetividad pero no una polaridad concreta. Esta tendrá que ser resuelta mediante la dependencia con otros rasgos.

- **Signos de puntuación**: El uso de exclamaciones o interrogaciones es analizado y cuantificado [48] [52] [57]. Por ejemplo: Encontrar exclamaciones seguidas, *!!!!*, denota subjetividad. Se trata de un rasgo de naturaleza muy similar al anterior.

- **Uso de Hashtags**: En Twitter, red social analizada en con este rasgo en las siguientes publicaciones [49] [57], es muy normal usar hashtags antes de usar expresiones, a estas formaciones se les denomina simplemente hashtags. Por ejemplo, un Hashtag muy popular en 2013 fue *#FollowFriday*. Estas expresiones denotan polaridad, pero deben ser analizadas junto a otros rasgos.

- **Presencia de interrogante o exclamación en la última frase** [48] [54]: La última palabra o frase del texto revierte en ocasiones la polaridad del texto. Analizar sus características es útil para discernir si está pasando este fenómeno.

- **Estructura polar del texto con atención a la última frase**: Los modelos que analizan este rasgo estructuran un texto en función a la polaridad de sus frase. Por ejemplo: Polaridad de las frases Negativa,negativa,negativa,positiva. Este patrón puede ser asociado con polaridad positiva, pues es conocido que la última frase revierte la polaridad de los textos. Por ejemplo: *No me ha gustado(...), tampoco me ha gustado(....) pero la película en general me encanta.* Un empleo de este rasgo es anotar a los textos con etiquetas que simbolizan patrones de flujo creados a priori. Es un rasgo muy referenciado [51] [26] [29] [57] e implementado en modelos de análisis de opiniones.

- **Empleo de conectores entre segmentos de texto** : Conocida la polaridad de una frase, si está conectada por un conector como *pero* con otra frase, se puede inferir que la segunda frase tiene la polaridad contraria [51] [26] [29] [56]. Es un rasgo que aporta conocimiento cualitativo a los enfoques de aprendizaje automático. Se toman en cuenta estos factores.

- **Palabrotas, términos relacionados con sexo y violencia** [58]: Es muy común el uso de estos términos en redes sociales e incluso en críticas. En ocasiones estos términos están asociados a polaridades negativas, por eso se toman como rasgos.

- **Pertenencia a clases semánticas** [52] [54]: Considérese que se categorizan palabras en clases semánticas como por ejemplo la clase *Enfado*. Si esta clase está asociada a una polaridad, entonces reconocer términos asociadas a esa clase es un rasgo que denota polaridad.

- **Palabras con caracteres seguidos repetidos**: En redes sociales es frecuente encontrar términos con caracteres repetidos. Por ejemplo: *Fataaaaal*. En vez de normalizar estas palabras, se toma este rasgo [48] [57]. Este hecho se toma como rasgo pues suele ser indicativo de subjetividad. Debe ser acompañado por otros rasgos.

- **Intensificadores** [49] [26]: Los intensificadores son términos que incrementan la polaridad de un término. Por ejemplo, el intensificador *muy* incrementa la polaridad de *bien*. Su uso frecuente denota subjetividad.

- **Negación**: Un uso frecuente de la negación puede ser capturado como un rasgo [48] [51] [26] [56] [57]. Si el valor de la variable que denota negación es verdadero, por sentido común se concluye que la polaridad es revertida. Para más detalle, se dedica un apartado de este estado del arte a estudiar como distintos enfoques tratan la negación.

- **Abreviaciones**: Las abreviaciones de palabras son comúnmente encontradas en redes sociales. Kouloumpis [49] las toma como rasgo.

- **Mayor presencia de términos polares positivos, neutrales o negativos** [48] [26]: Una hipótesis puede consistir en que si se encuentra un mayor número de términos que denotan polaridad positiva en el texto entonces el texto tiene polaridad positiva. Dado que es una hipótesis que procede del sentido común, se incluye como factor.

- **Emoticonos**: La construcción de un léxico polarizado que incluya sólo emoticonos es de utilidad en el análisis de opiniones en redes sociales [49] [59] [57]. Dada la subjetividad de los emoticonos, se incluyen variables resultado del procesamiento de los emoticonos. Se debe tener precaución en la clasificación de la polaridad de los emoticonos dado a problemáticas como la ironía. Ejemplo: *Comprando a todo el mundo... ¿Qué bien le ha ido a Pedro en el concurso eh? :) :) :).*

- **Reconocimiento de una entidad y consulta a sus propiedades**: Si se tiene información sobre una entidad y sus propiedades se puede asociar a una polaridad si se reconoce su presencia en el texto a analizar [50] [29]. Es un rasgo muy dependiente del contexto.

- **Uso y forma de pronombres reconocidos**: Mulholland [58] asocia el pronombre en primera persona singular a una polaridad negativa. El conteo y la categorización de los pronombres es un factor usado en el análisis de opiniones. Se puede consultar un mayor detalle en las siguientes publicaciones: [61] [26] [54]. Es un rasgo que depende de otros para una clasificación de un texto en categorías.

- **Empleo de la construcción pasiva** : Wilson [26] y Mulholland [58] asocian la construcción pasiva a una polaridad negativa. Por si solo este rasgo no es suficiente para clasificar a un texto pero asociado a otros factores puede incorporar información útil.

- **Estilo de escritura** [52]: El estilo de escritura puede ser un indicativo de subjetividad en el texto. Se puede analizar el empleo reiterado de frases cortas, por ejemplo. Un rasgo que debe ser combinado con otros en enfoques cualitativos y cuantitativos.

- **Empleo de Referencias o Citaciones**: La citación a un gran número de fuentes es analizada como factor en el trabajo de Abbasi [52].

- **Riqueza de vocabulario**: Abbasi [52] analiza el número de términos utilizado en el texto. Los modelos cuantitativos pueden incorporar cualquier número de rasgos. Si no incorporan información adicional con respecto a otros rasgos serán eliminados por filtros. Por ello, una metodología recomendada desde el punto de vista personal es incorporar el mayor número de rasgos posibles aunque parezcan extraños, emplear filtros para deshacerse de los que no aportan información adicional y entrenar posteriormente un modelo de aprendizaje automático.

- **Presencia de caracteres especiales** [52]: La presencia consecutiva de caracteres como *@#!%* puede denotar polaridad negativa en textos de redes sociales.

- **Categoría del documento analizado**: En distintos contextos la aparición de rasgos puede denotar distintas categorías. El contexto de un texto es una variable que hace que el valor de otra dependa de su valor. Este rasgo, de utilidad en textos anotados, es estudiado por Wilson en su modelo [26].

- **Detección de expresiones que denotan risa** [54]: Detectar *jajaja* puede ser un indicativo de subjetividad en un texto.

Estas características son las que utilizan los modelos de aprendizaje automático y de clustering para clasificar o agrupar textos en categorías. Dado que las características son inútiles sino se dispone de un clasificador, se van a detallar los clasificadores más populares en el siguiente apartado de esta memoria.

2.5.3. Métodos basados en Aprendizaje Supervisado

2.5.3.1. Modelos empleados en Análisis de Opiniones

Los modelos descritos a continuación categorizan una nueva instancia en función del vector de sus rasgos y de un corpus anotado. En cada una de las referencias que se van a citar se utilizan los modelos que se describen a continuación con la diferencia de que se emplean distintos rasgos de los comentados en la sección anterior. Existe una gran variedad de clasificadores, pero sólo se van a citar los más utilizados en análisis de opiniones. Estos son los siguientes:

Support Vector Machines: Existen múltiples tipos de Máquinas de Soporte Vectorial (SVM). En este capítulo se describe las Máquinas de Soporte Vectorial originales, en adelante, L-SVM que viene de SVM lineal y la Soft-SVM, y su evolución para clasificar puntos linealmente no separables.

La L-SVM es un modelo que trabaja con un conjunto de instancias, cada una anotada en una de dos categorías $D = \{(x_i, y_i)|x_i \in \mathbb{R}, y_i \in \{-1, 1\}\}_{i=1}^{n}$. y_i es el valor de la categoría ,-1 o 1, y x_i de un total de n es el vector de rasgos llamado instancia, que pertenece a una categoría. El algoritmo de entrenamiento de la L-SVM produce un modelo que es capaz de asignar a una nueva instancia en una de las dos categorías, se trata de un clasificador binario no probabilístico.

Para la primera versión de la L-SVM se requiere que los puntos sean linealmente separables, es decir, que si los puntos x_i pertenecen a un espacio de n dimensiones, $\mathbb{R}^n$, existe un hiperplano, $w \cdot x - b = 0$, de n-1 dimensiones que es capaz de separar a todos correctamente en función de su categoría y_i. $w \cdot x$ simboliza el producto escalar entre un vector de pesos no necesariamente normalizado y el valor de cada uno de los rasgos del vector de la instancia x. Por tanto, si una nueva instancia se encuentra por debajo del hiperplano, $w \cdot x_i \leq 0$, se le asignará una categoría y en caso contrario, $w \cdot x_i \geq 0$, se le asignará otra.

Hasta aquí la L-SVM consigue lo mismo que el perceptrón, pero cuenta con más propiedades. La L-SVM consigue generar un hiperplano de separación óptimo entre las categorías de las instancias. Es decir, un hiperplano cuya distancia a las nubes de puntos a ambas categorías sea máxima. Con esto se maximiza la probabilidad de clasificar bien nuevas instancias.

Sea $||w||$ la norma del vector w, que simboliza su magnitud y dirección y viene dada por la expresión $||w|| = \sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}$. Si los puntos son linealmente separables, entonces, además del hiperplano construido óptimo, $w \cdot x - b = 0$, existen dos vectores paralelos más que son los que intersecan con las instancias de cada categoría más próximas al hiperplano separador: $w \cdot x - b = -1$ y $w \cdot x - b = 1$. La región que se encuentra entre estos 3 vectores se le denomina margen, y es la que se quiere maximizar. Se demuestra por geometría que esta distancia es igual a $2/||w||$. En la siguiente figura se muestran estos vectores:

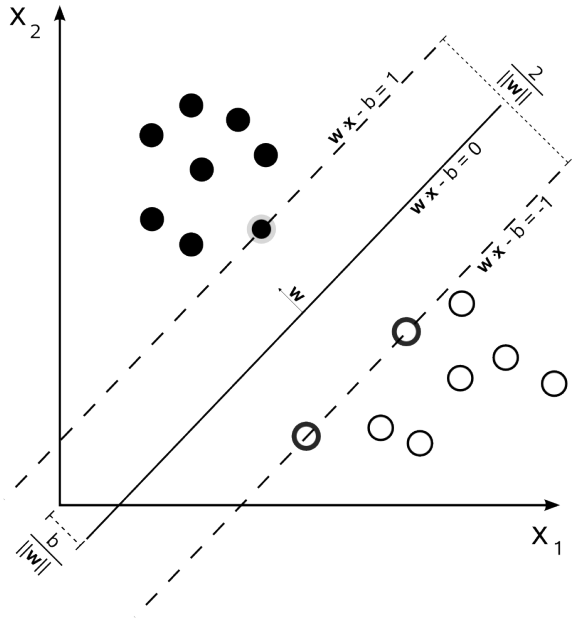

FIGURA 2.5: Hiperplano óptimo de separación creado por la L-SVM [2].

Por tanto, el objetivo del algoritmo de aprendizaje de la L-SVM es minimizar $||w||$, para maximizar el margen $\frac{2}{||w||}$ y con ello maximizar la probabilidad de clasificar un nuevo punto x_j en una categoría y_i. Esto se traduce en un problema de programación u optimización matemática. Se quiere minimizar la función objetivo $||w||$ sujeto a las restricciones de que ningún punto se encuentre en el margen, para lo que se usan los vectores paralelos. Estas restricciones se pueden mostrar como $w \cdot x_i - b \geq 1$

y $w \cdot x_i - b \leq 1$ o juntándolas en una en función de la categoría de los puntos $y_i(w \cdot x_i - b) \geq 1$.

Dado que la norma $||w||$ incluye una raíz cuadrada la complejidad del problema de optimización es grande, por lo que la función objetivo se transforma en la equivalente $argmin\frac{1}{2}||w||^2$, lo que convierte al problema en uno de optimización por programación cuadrática que puede ser resuelto mediante los Multiplicadores de Lagrange, α_i.

Mediante el uso de los multiplicadores y tomando en cuenta la restricción, el problema se queda en minimizar la expresión:

$$\arg \min_{w,b} \max_{\alpha \geq 0} \{\frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i [y_i(w \cdot x_i - b) - 1]\}$$

Por la condición de Karush-Kuhn-Tucker, se tiene que el vector de pesos es equivalente al producto de las categorías con la posición de las instancias que se encuentran en la frontera, o contenidas dentro de los vectores paralelos que se quieren encontrar: $w = \sum_{i=1}^{n} \alpha_i y_i x_i$.

El problema de optimización genera que α_i, los multiplicadores de Lagrange, sólo sea distinto de cero en las instancias contenidas en el vector paralelo al hiperplano óptimo de separación. Solucionando el problema de optimización, se genera el hiperplano que es capaz de separar y categorizar puntos linealmente separables.

Para solucionar el problema de que pocos puntos pueden hacer que se genere un margen muy pequeño, se introdujo en 1995 unas variables de suavidad ξ_i penalizando la función objetivo:

$$arg \min_{w,\xi,b} \{\frac{1}{2}||w||^2 + C \sum_{i=1}^{n} \xi_i\}$$

Estas variables también se introdujeron en la restricción del problema de optimización: $y_i(w \cdot x_i - b) \geq 1 - \xi_i$. Al estar sumando en la función objetivo, generalmente las variables ξ_i serán cero al resolver el problema de optimización, pero si, por ejemplo, es sólo un punto el que impide que se genere el margen máximo, este será admitido. A esta SVM se le denominó Soft-SVM, que genera un hiperplano que admite puntos en el margen con tal de que este sea lo más grande posible para la mayoría.

Para solucionar el problema de poder clasificar sólo puntos linealmente separables, se elaboró el siguiente desarrollo. Si se sustituye $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ en $arg \min_{w,b} \max_{\alpha \geq 0} \{\frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i [y_i(w \cdot x_i - b) - 1]\}$, teniendo que $||w||^2 = w \cdot w$, si se operan los términos multiplicando a todas las expresiones entre paréntesis se obtiene el problema de maximizar α_i en:

$$\frac{1}{2}(\underbrace{\sum_{1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j)}_{ExpresionA} - \underbrace{\sum_{1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j}_{ExpresionB} - \sum_{i=1}^{n} \alpha_i y_i b + \sum_{i=1}^{n} \alpha_i$$

Se traslada a la b como una restricción en función de los multiplicadores, $\sum_{i=1}^{n} \alpha_i y_i = 0$ por lo que se iguala a 0. Por otro lado se resta la Expresión A con la Expresión B, dejando al problema con la siguiente expresión:

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2}(\sum_{1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j)$$

Se conoce a esta expresión como la forma dual de la L-SVM, y es necesaria para que el hiperplano de n-1 dimensiones se convierta en una función polinómica que puede adquirir cualquier forma separando cualquier nube de puntos. Para ello se utiliza el conocido Truco del Kernel, que se expone a continuación.

Dentro de la expresión anterior, se denomina al kernel de esta expresión a $x_i^T x_j$, esta expresión kernel de una SVM se denotará $k(x_i, x_j)$ y mediante su modificación la SVM podrá clasificar todo tipo de puntos. En el caso de la L-SVM el Kernel es igual a $k(x_i, x_j) = x_i^T x_j$. La SVM con el Kernel modificado generará un hiperplano en un espacio de dimensión superior al de las instancias, incluso en un hiperplano de infinitas dimensiones denominado Espacio de Hilbert.

Mediante su proyección en el espacio de dimensiones de las instancias se genera una función polinómica que es capaz de separar cualquier conjunto de instancias en dos categorías. La función objetivo de la SVM con Kernel queda por tanto expresada con la siguiente expresión, en la cual hay que declarar una función Kernel y sustituirla por $k(x_i, x_j)$ para resolver el problema:

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2}(\sum_{1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j))$$

Un ejemplo de función Kernel que hace esto posible es la Gaussiana:

$$k(x_i, x_j) = exp(-\gamma||x_i - x_j||^2)$$

Otro ejemplo es la función Kernel Hiperbólica Tangencial:

$$k(x_i, x_j) = tanh(kx_i \cdot x_j + c)$$

El resultado final de emplear una Soft-SVM con una función Kernel Gaussiana genera un clasificador que puede tener cualquier forma en $\mathbb{R}^n$ y cuyo margen maximiza la distancia contra las dos categorías separables. La siguiente figura ilustra la función generada por este modelo:

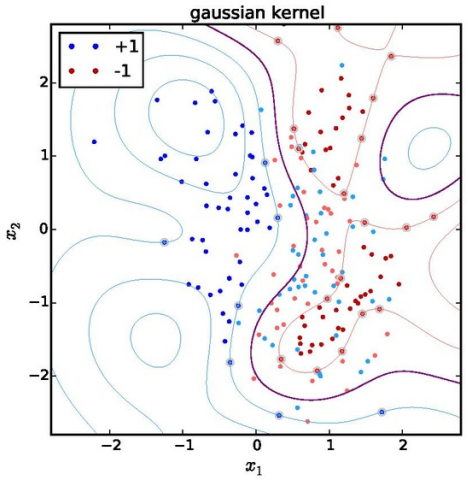

FIGURA 2.6: Función de separación creada creado por la Soft-SVM con Kernel Gaussiano [3].

Esta SVM es muy dinámica, pero sólo es capaz de separar instancias que pertenecen a dos categorías. En la actualidad se ha conseguido construir una SVM que es capaz de clasificar puntos que pertenecen a más de dos categorías, llamada SVM Multiclase.

La filosofía para la construcción de la SVM multiclase es la construcción iterativa de SVMs que clasifiquen a los puntos en dos categorías. Una de ellas será una de las categorías y_i y la otra será el conjunto del resto de categorías $\{y_j/i \neq j\}$. Una vez construida esta SVM se entrena otra SVM partiendo el espacio de las categorías $\{y_j/i \neq j\}$ hasta que se tengan n-1 funciones construidas por SVMs donde n es el número de categorías. Esta última SVM, evolución de la Soft-SVM con función Kernel gaussiana, es la más comúnmente empleada en análisis de opiniones para clasificar las instancias en las categorías positiva, negativa y neutral.

Los siguientes enfoques utilizan distintas versiones de la SVM como modelos de aprendizaje automático: [59] [51] [52] [61] [53] [54] [62] [63] [64] [65] [66] [57].

Logistic Regression: La regresión logística es muy citada y conocida también como Maximum Entropy Classifier. Al contrario de la SVM, se trata de un clasificador probabilístico, que da una probabilidad $P(c_i/x_i)$ a cada instancia x_i de pertenecer a las categorías consideradas c_i. Cada instancia x_i es un vector de rasgos, donde la variable que denota un rasgo es x_{ij}. Es un análisis de regresión empleado para predecir la clase de una variable categórica.

La regresión logística, al contrario que la regresión lineal que minimiza el error cuadrático de la recta que aproxima una nube de puntos, modeliza a las categorías de estos puntos como su media. Estas categorías, en principio dos en la regresión logística estándar $C = 0$ *y* $C = 1$, son modelizadas por una variable aleatoria que sigue una distribución de Bernoulli.

Dado que la esperanza de una distribución de Bernoulli condicionada a un vector de variables predictivas es la media de estas $E(C|x_j) = \mu_j$, se modeliza a las categorías por la media del valor de estos puntos. El modelo busca la relación entre la variable dependiente categoría C y las variables predictivas x_j, que son el vector de rasgos de una nueva instancia.

La relación de una categoría con sus variables predictivas debe ser un número $\mu_j = P(C = j|\{x_1, ..., x_n\}) \in [0, 1]$ y además se debe cumplir que $P(C = 0|\{x_1, ..., x_n\}) + P(C = 1|\{x_1, ..., x_n\}) = 1$ para que la relación simbolice una variable aleatoria de probabilidad.

Para garantizar estas propiedades, se utiliza una función logística, que toma cualquier valor y su salida es siempre un número entre 0 y 1. Para estimar el valor de esta relación, que simboliza la probabilidad de que una instancia sea clasificada en una categoría, se emplea por tanto la siguiente expresión:

$$\mu_j = P(C = 1/x_j) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{j1} + \cdots + \beta_k x_{jk})}}$$

El opuesto de esta probabilidad para la instancia x_j será la probabilidad de que la instancia x_j pertenezca a la otra categoría, $C = 0$. El modelo tiene que estimar los parámetros β_i con respecto al conjunto de datos que utiliza para el entrenamiento. Se estiman por máxima verosimilitud. No es posible encontrar una expresión analítica para resolver este problema por lo que se hace empleando el método iterativo de

Newton-Raphson hasta convergencia. Una vez encontrado el vector de valores β_i el modelo está listo para clasificar nuevas instancias.

Si se necesita discriminar entre más de dos clases se utiliza la regresión logística multinomial, que en vez de considerar una distribución de Bernoulli para modelizar las categorías considera una distribución multinomial. Una distribución multinomial es la extensión de la distribución de Bernoulli, en la que la variable aleatoria X tiene valor x_i con probabilidad p_i, pudiendo haber más de dos valores y tal que $\sum_{i=1}^{k} p_i = 1$. La regresión logística multinomial es un modelo muy utilizado en el análisis de opiniones.

Las siguientes publicaciones incorporan la regresión logística como modelo cuantitativo: [51] [61] [53] [63].

Näive Bayes Classifiers: Como en las SVMs, existen una gran variedad de clasificadores basados en el Naive Bayes original. De entre todos ellos, el más empleado en análisis de opiniones es el Naive Bayes Multinomial. Al igual que la regresión logística multinomial, este modelo es capaz de asignar la probabilidad de pertenecer a una categoría de un vector de rasgos o instancia. Al igual que la regresión logística, se trata de un clasificador probabilístico.

Este modelo calcula la probabilidad de que un vector de rasgos o instancia X pertenezca a una categoría como $P(C|X) = \frac{P(X|C)P(C)}{P(D)}$. Esta expresión es el Teorema de Bayes. La probabilidad a priori de una categoría $P(C)$ es calculada por máxima verosimilitud, es decir, contando cuántas instancias del conjunto de datos pertenecen a esa categoría y dividiendo por el total. La probabilidad del denominador es la misma para todas las instancias por lo que se elimina, quedando $P(C|X) = \underset{c \in C}{argmax} P(X|c)P(c)$. Como las variables $x_1, x_2, ..., x_n \in X$ del vector de rasgos son independientes entre sí la expresión final es: $P(C|X) = \underset{c \in C}{argmax} P(c_j) \prod_{x \in X} P(x|c)$.

Las probabilidades $P(x|c)$, tal que $c \in C$ puedes ser calculadas por máxima verosimilitud como la razón entre la instancias que pertenecen a c entre las instancias totales. Si se quiere generalizar a más valores para no incurrir en un posible *Overfitting* se puede emplear en vez de máxima verosimilitud una técnica de *Smoothing*, que asigne masa de probabilidad a instancias nunca vistas y quite una proporción similar a todas las vistas. Los resultados mejoran. Junto a las SVM y la regresión logística, es el clasificador mas empleado en análisis de opiniones, dada su sencillez y porque todos los rasgos son independientes entre sí.

Se encuentra este modelo en los siguientes artículos: [51] [52] [61] [53] [67] [64].

Metaclasificadores: Son modelos que utilizan el resultado de otros modelos para clasificar. Existen múltiples variantes de estos modelos pero la más popular es el AdaBoost. Se tiene una serie de modelos que clasifican las instancias en categorías por separado. Los metaclasificadores utilizan primero uno de ellos y comprueban que instancias se han clasificado correctamente, para las que no se han clasificado correctamente, se envían a los siguientes clasificadores o se incrementa su importancia para los siguientes clasificadores. Estos clasificadores intentarán clasificar estas instancias, que se denominan instancias frontera. Se intenta que las instancias que fallan los primeros clasificadores sean clasificadas por los siguientes. Se comprueba la mejoría de la clasificación empleando modelos distintos. Los siguientes autores emplean metaclasificadores: [51] [52] [53] [49]. Desde un punto de vista personal se recomienda combinar varios modelos de aprendizaje automático creando un metaclasificador ya que ningún modelo es mejor a cualquier otro en todos los problemas. Ante análisis de opiniones en dominios amplios o en varios dominios, se recomienda emplear varios modelos, tanto cuantitativos como cualitativos. Posteriormente, se recomienda crear un metaclasificador que clasifique el texto entrante en la categoría que cuente con el mayor número de votos entre los modelos.

Otros clasificadores vistos en la literatura son los Modelos Ocultos de Markov [60], Árboles de Decisión [58] [63] y diferentes tipos de Redes Neuronales [68] [69]. Un último enfoque emplea Deep Learning [70] en dichas redes neuronales.

En algunos problemas del análisis de opiniones, no solo basta con clasificar un texto en una categoría, sino que es necesario clasificarlos en varias categorías. Por ejemplo, un texto puede pertenecer a las categorías negativo y deportivo. A continuación se detalla información sobre estos problemas.

2.5.3.2. Modelos de aprendizaje supervisado multietiqueta

Los modelos vistos clasifican a una instancia en una categoría. Es posible, sin embargo, que lo más adecuado para un tipo de problema sea clasificar a una instancia X en distintas categorías $c \in C$, con un grado de pertenencia $w_j \in [0, 1]$ para todas ellas. En análisis de opiniones, por ejemplo, si se quiere analizar una crítica simultáneamente en las categorías *Enfado* sobre una característica y *Alegría* sobre otra característica, se necesita uno de estos modelos, denominados modelos de aprendizaje supervisado multietiqueta *Multilabel models*.

Marcheggiani [71] emplea modelos basados en Conditional Random Fields (CRFs) para categorizar frase a frase aspectos particulares de un producto. A cada segmento

de texto, t, le es asignado las siguientes variables que modelizan opinión: La opinión general expresada en el segmento $y_0 \in Y$ y las variables de opinión $y_t^a \in Y$ expresadas del aspecto a si estos son reconocidos. Las categorías son positivo, negativo y sin opinión. Al poder asignar a una instancia tantas variables como aspectos se reconozcan se trata de un problema multietiqueta.

El modelo de Potts empleado en [45] que se comentó en la sección de Enfoques basados en modelos de optimización tiene en cuenta categorizar a cada palabra en múltiples clases para el trabajo futuro. De momento los autores consiguen asignar a cada palabra una probabilidad de pertenecer a cada clase, pero finalmente la clasifican en una sola clase. No obstante, citan la importancia de que una palabra, por ejemplo *succumb* pertenezca a las categorías *sad* y *fear*. El enfoque es identificar estas palabras y representarlas en clases borrosas. Estas palabras se introducirían en el léxico polarizado semilla que se ampliaría mediante los grafos mencionados en el modelo de Potts.

Se trata de un problema no muy tratado en la literatura dada la novedad de los modelos de aprendizaje supervisado multietiqueta. Estos modelos podrían proporcionar aplicaciones muy útiles en lo relativo al análisis de opiniones. Además de poder clasificar una palabra categorías distintas o categorizar aspectos en una sola frase, la utilización de modelos multietiqueta podría etiquetar la opinión de múltiples usuarios, productos, distintas categorías afectivas y polares, asignar categorías asumiendo que el texto pertenece a distintos contextos y tomando en cuenta distintos factores para todos ellos, asignar categorías en función a factores, etc...

No siempre se dispone de un corpus anotado de textos ni se quiere clasificar un texto en una categoría. Un escenario muy común es disponer de textos sin procesar. Una metodología a emplear es agrupar estos textos en función de sus similitudes para poder estudiar cada grupo de textos por separado y que estos grupos sean clasificados en categorías por expertos. Para ello, son útiles los modelos de clustering y estadísticos, comentados a continuación.

2.5.4. Métodos estadísticos y basados en Aprendizaje No Supervisado

En esta sección se citan métodos de carácter estadístico que se emplean en el análisis de opiniones. El objetivo de estos métodos es agrupar N-Gramas en categorías sin basarse en ninguna supervisión previa. Se basan en encontrar patrones que se repiten en los mismos términos a lo largo de múltiples textos. Esta agrupación hace posible

que se clasifique un nuevo término en distintas categorías o un texto basado en las categorías de sus términos.

2.5.4.1. Métodos Estadísticos

Mutual Information: Prabowo [72] define un set de palabras que denotan sentimiento y quiere averiguar si otras expresiones también denotan sentimiento. Esta técnica mide como de frecuente es encontrar la expresión junto a una palabra que denota sentimiento frente a encontrarlas por separado. Si es más frecuente encontrarlas juntas que por separado entonces se asume que se extiende la polaridad de la palabra que denota sentimiento a la expresión que aparece junto a ella.

Para usar esta medida se necesitar definir otra medida de proximidad, es decir, cuando se dice que dos expresiones coocurren. Por ejemplo, se dice que dos expresiones coocurren si están en la misma frase, en el mismo párrafo, a una distancia de 3 palabras, etc... Prabowo define la Mutual Information de una expresión e con respecto a la palabra que denota sentimiento s como:

$$MI(e,s) = log_2 \frac{P(e,s)}{P(e) * P(s)}$$

Donde la probabilidad de coocurrencia se mide si dos palabras están presentes en un mismo documento. La probabilidad de ocurrencia se mide por máxima verosimilitud, es decir, el conteo de palabras que ocurren en el documento entre el total de documentos. Este valor deberá ser calculado para todas las expresiones que denoten sentimiento. Las expresiones cuyas $MI(e,s)$ superen un umbral predefinido estas ser categorizadas como polares y la polaridad que tenga la expresión será expandida a ellas.

Chi-Square χ^2 [72]: Si se asume que encontrar una expresión con una determinada frecuencia en textos que denotan opinión hace que esta expresión sea categorizada como polar, entonces se puede emplear esta medida de contraste. Esta medida compara la frecuencia esperada E_i con la frecuencia observada O_i, en cada texto i. La hipótesis que se desea rechazar es la de que no aparece la expresión con la frecuencia deseada y por lo tanto no se puede categorizar a la expresión como polar. Prabowo acepta valores de χ^2 por encima de 5 para rechazar la hipótesis. La expresión de esta medida es la siguiente:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Latent Semantic Analysis (LSA) [30]: Esta técnica establece una asociación entre palabras y los conceptos que ellas simbolizan teniendo una base de documentos. LSA categoriza un documento en una categoría y encuentra relaciones entre los distintos términos considerados.

Esta técnica construye un *espacio conceptual* en $\mathbb{R}^2$ que, dadas las coordenadas de dos palabras, simboliza lo cercanos que están los conceptos de las palabras consideradas entre sí. Si los conceptos a los que se quieren asociar las palabras son las categorías positiva, negativa y neutral esta técnica se puede emplear para el análisis de opiniones. Para ello se deben definir palabras representativas de estas categorías o conceptos y definir los conceptos.

En esta técnica, los documentos son representados mediante conjuntos de palabras sin orden determinado *bag of words*. Los conceptos deben ser asociados a palabras representativas de los mismos, por ejemplo: *Caballo, Torre, Peón y Alfil* representantes del concepto *Ajedrez*. Con esta representación, la técnica LSA construye una matriz que tiene por filas las palabras, por columnas los documentos y el elemento M_{ij} es un contador de la aparición de la palabra i en el documento j. Una variante también propone que el elemento M_{ij} sea el resultado de aplicar el algoritmo *TF-IDF* de la palabra i en los documentos.

A partir de esta matriz, se ejecuta el algoritmo Singular Value Decomposition (SVD) sobre la matriz. Los detalles de este algoritmo pueden ser consultados en la literatura. Este algoritmo encuentra una representación dimensional reducida de la matriz enfatizando las relaciones que encuentra, devolviendo tres matrices, de las cuales, es importante en análisis de opiniones la primera y tercera matriz.

La primera matriz tiene tantas filas como términos considerados y tres columnas, de las cuales, son importantes las dos últimas. Estas columnas representan las coordenadas de las palabras en el espacio conceptual de los conceptos considerados en $\mathbb{R}^2$. Si dos palabras están cercanas entre sí, entonces es que pertenecen a la misma categoría o concepto. La tercera matriz tiene tantas columnas como documentos considerados y tres filas. Las últimas dos filas son las coordenadas el espacio conceptual $\mathbb{R}^2$ de los textos, pueden ser dibujados junto a las términos para clasificar a los textos. En caso del análisis de opiniones, categorías positiva, negativa y neutral.

Esta técnica es útil para averiguar si una palabra no considerada antes está cercana a una serie de palabras categorizadas, si la distancia de la nueva palabra junto a las otras es pequeña, se puede categorizar a este término en una categoría. En análisis de opiniones sirve para extender un léxico polarizado y emplear posteriormente al término como rasgo en un modelo de aprendizaje supervisado. Se puede considerar la distancia de un documento a un centroide de términos como un rasgo para un modelo de aprendizaje que clasifica a un documento en distintas categorías.

Latent Dirichlet Allocation (LDA) [57]: Este método probabilístico, al igual que el LSA, categoriza a los documentos por categorías en función de las palabras que aparecen en este documento. Esta técnica representa a cada documento como una mixtura de categorías. El documento es clasificado en la categoría que maximiza su probabilidad condicionada a las palabras del documento. El LDA necesita conocer el número de categorías en las cuales se desea clasificar los documentos. A partir de este número, el modelo LDA aprende la representación de cada categoría y las palabras que la representan.

Este algoritmo asume que un documento es generado por una serie de distribuciones que modelizan su longitud y la proporción de categorías a la que pertenece. Cada categoría está representada por un conjunto de palabras clave. Las palabras del documento son generadas aleatoriamente en función a la distribución que modeliza cada categoría. Con esas asunciones sobre la generación de un documento, el LDA aprende las palabras que componen cada categoría.

El modelo aprende una probabilidad condicionada $p(w|C)$ para cada palabra w representante de C. Emplea un proceso iterativo en el cual se asigna a cada palabra w una categoría C, y se calcula las probabilidades $p(C|D)$ y $p(w|C)$. Las palabras cambian de categoría para maximizar $p(C|D) * p(w|C)$ en todas las categorías y documentos. Al final de las iteraciones, se asume que los tópicos representan a los documentos y que las palabras representan a las categorías [73].

Este algoritmo puede ser empleado como rasgo en la clasificación de un texto si estas categorías simbolizan a las utilizadas en el análisis de opiniones. Los términos que esta técnica extrae pueden ser usados como rasgos en un modelo de aprendizaje supervisado o como términos para expandir su polaridad.

2.5.4.2. Clustering

En esta sección se citan enfoques que han empleado Clustering en análisis de opiniones. Se entiende como Cluster a una colección de objetos similares entre sí y lo

más diferentes a los objetos que habitan en otros clusters. Los métodos de clustering son aquellos procedimientos que, dados una colección de puntos, construyen estos clusters. En análisis de opiniones, estos métodos sirven para asignar a cada texto o término en un cluster, por ejemplo: De textos positivos, negativos o neutrales.

Wachsmuth, en [29], analiza los textos en función de su estructura polar. Estos textos serán asociados a la categorización de la estructura polar más similar con respecto a las estructuras polares referencia. El centroide de cada cluster que se obtiene en el proceso, ω, se convierte en una estructura polar referencia. La estructura polar de los centroides es estudiada y anotada en una categoría posteriormente al proceso de clustering. Para determinar la amplitud de los clusters emplea un dendograma, construido mediante clustering jerárquico ascendente.

Este proceso agrupa a los objetos más cercanos entre sí en sucesivas iteraciones creando categorías, en la cual va agrupando objetos cada vez más alejados, hasta que en la última iteración todos quedan englobados en una categoría. Todos los mecanismos de clustering necesitan una función de semejanza entre objetos para conocer como de alejados están unos de otros y saber si se agrupan en un cluster o no. En coordenadas polares, esta puede ser la distancia euclídea o de Manhattan.

Wachsmuth emplea una función de semejanza de estructuras polares para comparar las estructuras de los textos. Los clusters finales quedan definidos mediante un umbral, que simboliza la iteración a partir de la cual las categorías que se definiesen ahí son las definitivas.

Una vez terminado el proceso de clustering jerárquico ascendente, en este tipo de clustering un nuevo texto será asociado al cluster que minimice la distancia con respecto con él. Para saber el punto que simboliza a un cluster existen criterios como el más cercano, el más lejano, etc... Wachsmuth emplea el centroide de cada cluster como el punto referencia. De esta forma categoriza nuevos textos en función a su estructura polar.

Formalmente, todo modelo de clustering jerárquico que asigna a un texto a una categoría, *Hard clustering*, puede seguir el siguiente enfoque, referenciado por Popat [55]. Este enfoque emplea la agrupación previa en distintos clusters de cada una de las palabras del texto a agrupar para clasificar al texto en una categoría. Sea C una función de proximidad genérica a definir en cada problema que clasifica o agrupa palabras que componen un texto, $w = [w_j]_{j=1}^{m}$, en un cluster K_i. Si $K_i = C(w_j)$, es decir, esta función agrupa a cada palabra en un cluster, entonces, un texto será asignado al cluster K_i que maximice la siguiente expresión, cuyas variables se explican en el siguiente párrafo:

$$L(w; C; F) = \prod_{j=1}^{m} p(w_j|C(w_j))p(C(w_j)|C(w_{j-1}))$$

Donde M es la longitud del texto. F es el método para calcular la probabilidad $P(w_j)$, este método es genéricamente el criterio de máxima verosimilitud, es decir el conteo de las palabras sobre el total de palabras del texto. La probabilidad $P(C(w_j))$ se calcula de la misma forma, contando el número de palabras de esta categoría que aparecen en los textos en función al total de palabras. También es empleado para calcular la probabilidad algún modelo de *Smoothing*, para una mejor generalización. Se recomienda utilizar mecanismos de *Smoothing* si el problema tiene un dominio amplio para que sea más generalista.

La expresión descrita arriba simboliza que la agrupación final del texto será en el cluster en el cual se agrupa la mayoría de sus palabras $\prod_{j=1}^{m} p(w_j|C(w_j))$ teniendo en cuenta lo factible de que distintas categorías sean presentadas en el orden de las palabras del texto $p(C(w_j)|C(w_{j-1})$. El texto será clasificado en la categoría que maximice esa expresión.

No todos los modelos de clustering asignan a un objeto una categoría o cluster en su totalidad. Los modelos denominados como modelos de clustering suave, *Soft Clustering*, asignan a un objeto una probabilidad o grado de pertenecer a uno o más clusters. De este modo un objeto puede ser asignado en categorías.

Xiang [57] emplea dos niveles de clustering suave para clasificar a un tweet x_i en una categoría polar c.

En el primer proceso de clustering pretende clasificar tweets en tópicos t_j. Previamente Xiang emplea la técnica LDA para generar los tópicos t_j de sus textos. Con estos k tópicos, emplea posteriormente el modelo K-Means de Soft Clustering para asignar a un tweet a un tópico. Este modelo necesita el número de clusters como parámetro y asigna una probabilidad a cada objeto de pertenecer a un cluster, por lo que se adapta perfectamente al problema. Dado que el modelo K-Means necesita una definición de distancia entre objetos, Xiang define la siguiente función de semejanza:

$$P_t(t_j|x_i) = max_k P_t(t_k|x_i)$$

Es decir, un tweet x_i es asignado a cada categoría t_j con probabilidad $P_t(t_k|x_i)$, probabilidad condicionada de la categoría k al tweet x_i, calculada por máxima verosimilitud. A partir de este paso el tweet es modelado como una distribución multinomial

de tópicos, puesto que $\sum_{i=1}^{k} P_t(t_k|x_i) = 1$. Esta modelización del tweet será empleada para clasificarlo en una categoría polar en el segundo modelo de clustering.

Previamente, Xiang [57] ha clasificado a los tópicos t_k en categorías polares c con una probabilidad $P(c|t_k)$. Un tweet x_i quedará clasificado en una de las T categorías c con probabilidad $P(c|x_i)$ mediante la siguiente expresión:

$$P(c|x_i) = \sum_{j=1}^{T} P_m(c|t_j, x_i) P_t(t_j|x_i)$$

El modelo K-Means es uno de los más populares en clustering suave, pues su aplicación es directa cuando se conoce el número de categorías en las que se quiere agrupar un conjunto de datos. Este modelo es también utilizado por Li [65] para detectar foros sobre un tópico que se va a comentar con asiduidad, *hotspot forums*.

En este estudio, Li comienza con un conjunto de foros que históricamente han sido muy concurridos en distintos tópicos. Li representa a cada foro como un vector de 5 elementos: Número de posts del tópico j en una franja de tiempo, número medio de respuestas a los posts de los tópicos considerados, el valor medio de la opinión de estas respuestas y la fracción de respuestas categorizadas positivamente y negativamente.

Estos vectores serán agrupados por el algoritmo de clustering K-Means en clusters. Li argumenta que el centroide de cada cluster representa las coordenadas de un foro en el cual comenta un gran número de gente, dado que los foros *semilla* han sido muy concurridos. Para cada nuevo foro, si su distancia a un centroide de cualquier cluster es muy pequeña entonces se predice que ese foro será concurrido.

Los recursos lingüísticos pueden ser usados en modelos de clustering. Popat [55] considera a los synsets de Wordnet como un cluster de palabras. Términos muy parecidos entre sí y además su significado es diferente a los del resto de clusters.

Popat no especifica función de distancia para incorporar una nueva palabra con el modelo de clustering, pero esta puede basarse en el conteo del número de palabras del texto que pertenecen a cada synset, máxima verosimilitud, y la polaridad asociada a estos synsets. Para esto se necesita un recurso como SentiWordnet y un proceso de desambigüación. Dado este último requisito, se recomienda ser especialmente crítico.

Se ha comentado información acerca de todos los enfoques que se pueden realizar en el análisis de opiniones. Vistos estos enfoques, es necesario comentar las principales problemáticas que hacen que el análisis de opiniones sea una tarea ardua. Estas problemáticas son listadas en el siguiente apartado del estado del arte.

2.6. Problemáticas específicas en Análisis de Opiniones

Todo problema relacionado con el proceso del lenguaje natural tiene que lidiar con problemáticas específicas que ocurren al procesar el lenguaje. En este apartado se citan problemáticas y se proponen enfoques existentes en la literatura para solucionarlas.

2.6.1. Multilingüismo

Existen métodos para gestionar el multilingüismo pero ninguno es aceptado por toda la comunidad científica como el más válido en todos los problemas. Consciente de este fenómeno, Chen [4], construye un grafo en el cual cada término está conectado el término traducido en distintas lenguas por un enlace que se puede representar como un vector de 6 valores. Cada valor de este vector representa el método de traducción entre el término y su traducción en otro lenguaje o una relación semántica entre dos términos en el mismo lenguaje, sinonimia y antonimia.

Por ejemplo, un valor del vector es 1 si dos términos están relacionados por su traducción en Google Translate y 0 en caso contrario. Chen elabora esta solución para poder propagar las categorías polares de un léxico semilla a un léxico multilingüe polarizado.

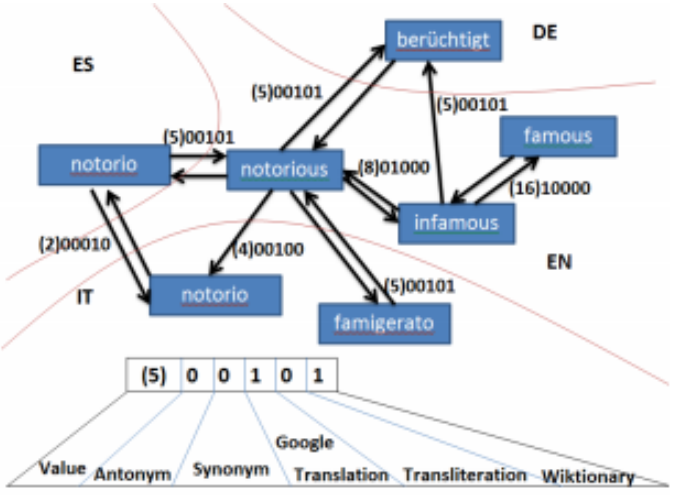

FIGURA 2.7: Grafo construido por Chen [4] para la expansión de un léxico polarizado a un léxico polarizado multilingüe.

Se trata de un enfoque que combina varios criterios entre sí, lo cual es positivo. Por contra, el problema del multilingüismo no es solucionado con alta precisión por algunos de los criterios que cita como Google Translate por lo que se pueden obtener resultados no deseados.

Boiy [51] propone un enfoque en el que emplea Aprendizaje Automático para analizar opiniones en múltiples lenguajes. Boiy se centra en que cada lenguaje tiene sus características que lo hacen distinto de otros y que no se puede tratar a todos los lenguajes del mismo modo. El enfoque propone utilizar distintos rasgos para el modelo de aprendizaje automático en cada lenguaje en función de sus peculiaridades.

Por ejemplo, el holandés es un idioma que junta palabras distintas en una sola palabra con asiduidad, por lo que necesita reglas adicionales para lidiar con las palabras y verbos compuestos. El francés necesita reglas específicas para tratar con la negación y con las abreviaturas. En definitiva, analizar opiniones en distintos lenguajes necesita de un experto en esa regla para diseñar los rasgos específicos de la misma.

Un enfoque muy común es la traducción automática de textos pertenecientes a cualquier lengua al inglés previo al análisis de opiniones. Mediante este procedimiento, se separa el problema en dos: Primero, lidiar con el multilingüismo para pasar un texto de cualquier lengua a inglés y luego diseñar el modelo de análisis de opiniones exclusivamente en lengua inglesa. Se recomienda este tipo de enfoques, puesto que reducen la complejidad del problema, de acuerdo a la popular metodología *Divide y vencerás*.

Bautin [74] emplea el IBM Websphere Translation Server para traducir 8 lenguajes al inglés. De nuevo, es un enfoque que no tiene una precisión total. Abbasi [52] considera distintos rasgos para el análisis de opiniones en lengua inglesa frente a lengua arábiga. Por ejemplo, sostiene que en inglés los n-gramas son más útiles que en árabe, donde son importantes las raíces de las palabras. Enfocar el análisis de opiniones por separado en varias lenguas hace que se obtenga una gran precisión, pero esta metodología es inabordable si se quiere obtener un análisis de opiniones en el multilingüismo masivo.

Las ventajas de esta metodología es que sólo se necesita diseñar un modelo de análisis de opiniones para una lengua y las desventajas es que dependiendo del mecanismo de traducción a lengua inglesa se puede perder el significado original del texto y por lo tanto hacer un análisis de opiniones de baja calidad.

El enfoque propuesto por Jadhav [27] empleando la interlingua UNL para crear reglas con las cuáles determinar términos que contienen polaridad podría ser ampliado para lidiar con multilingüismo. Empleando los codificadores UNL se puede traducir texto en cualquier lengua a su representación como documento UNL.

El análisis de opiniones se realiza con las reglas propuestas por Jadhav [27] o con otros enfoques que usen la representación UNL. El mecanismo de evaluación necesitaría *Gold Standards* en las lenguas traducidas a UNL para medir la efectividad del análisis de opiniones empleando UNL. Se recomienda el uso de una interlingua como UNL para lidiar con el multilingüismo, puesto que la complejidad del problema se hace menor que tener que traducir un idioma a cualquier idioma y todas esos idiomas a ese idioma. Mediante UNL, o cualquier otra interlingua, únicamente se necesitan expertos de cada idioma y de UNL para crear los codificadores y decodificadores de esa lengua a UNL. Se trata por tanto de una solución más escalable.

2.6.2. Negación

En análisis de opiniones la negación es un rasgo que se debe tener en cuenta. Por ejemplo, considérese que se tiene un léxico polarizado en el cual el término *bueno* tiene polaridad positiva. Si no se tiene en cuenta la negación, entonces la frase *Este filete no está bueno* es categorizada positivamente, cuando la polaridad de la frase es negativa.

Un enfoque muy utilizado en la literatura es el de negación local. Este enfoque divide un texto en frases que polariza por separado. Para cada frase, se detecta si existe en ella un término que denota negación, por ejemplo *not*. Si este término es detectado, entonces, la polaridad de la frase será la contraria de la polaridad en la que se le ha categorizado.

En el ejemplo anterior, *Este filete no está bueno*, la categorización de la frase es positiva pero como se encuentra un término que denota negación, *no*, la frase es categorizada como negativa.

Wilson [26] emplea una variable binaria como rasgo representante de la negación, con valor 1 si se detecta negación y 0 en caso contrario. De ser detectada se invierte la polaridad. Por razones de sencillez, se opina que pese a no ser el enfoque más preciso y probablemente no es apto para problemas muy específicos, se trata del más escalable. Kim [20] incluye palabras como *not* y *never* que invierten la polaridad, la idea es fundamentalmente la misma que Wilson. Zirn [56] elabora una lista de términos que

expresan negación e invierte la polaridad de las frases en las que se encuentran estos términos, personalmente es el enfoque recomendado dada su sencillez y a obtener una precisión relativamente alta.

La negación es también expresada mediante el uso de giros conjuntivos y conectores, como cita Vechtomova [22]. Por ejemplo, *The steak is good, although could be better*. Considérese un proceso que separa este texto en dos frases unidas por el giro conjuntivo *although*. Si se cuenta con un recurso efectivo para separar frases, es un enfoque recomendado.

Si se conoce que la primera frase tiene polaridad positiva, entonces se puede establecer que la segunda frase tiene polaridad negativa mediante una regla lingüística del tipo: Si hay dos frases unidas por un giro lingüístico, entonces la polaridad de la segunda frase es la inversa de la primera. Es necesario elaborar un listado de giros lingüísticos para poder ejecutar esta regla, por ejemplo: *but* y *although*.

Considerar que siempre que se encuentre un giro lingüístico o término que denota negación la polaridad de la frase es la inversa en un enfoque práctico dada la sencillez de su implementación pero no es totalmente preciso. Dependiendo de la frase, esta negación puede incurrir en una mayor o menor intensificación del cambio de polaridad. Si se desea una mayor precisión en los resultados del modelo de análisis de opiniones, es necesario seguir el siguiente enfoque.

Zhu [69], en su estudio de la negación en el análisis de opiniones, cita este hecho con ejemplos. Considérese la frase: *not very good* frente a *never has been good*. Con los enfoques anteriormente citados, la polaridad de las dos frases sería la misma: Negativa. En cambio, la polaridad de la primera frase puede ser incluso considerada como positiva, mientras que la segunda es claramente negativa.

Zhu define por tanto que la polaridad debe ser una variable continua en un intervalo cerrado para representar con exactitud el impacto de la negación dependiendo de la palabra negativa encontrada y la frase que se analiza. Zhu propone diversas heurísticas para lidiar con la negación basadas en una polaridad a priori de cada una de las palabras que denotan negación y una polaridad final calculada en función a la palabra que denota negación y la frase. Los enfoques de Zhu son más precisos que el enfoque clásico pero añaden complejidad al problema.

2.6.3. Independencia de Contexto

Los rasgos que denotan subjetividad en distintos tipos de texto son diferentes entre sí. Por ejemplo: En Twitter el análisis de emoticonos es esencial para clasificar al tweet en una categoría polar mientras que en artículos de opinión en periódicos este rasgo es inútil. Un modelo de análisis de opiniones robusto frente a cambios de contexto puede ser empleado para categorizar cualquier tipo de texto, evitando desarrollar un modelo diferente por cada contexto a considerar. Los enfoques de independencia de contexto o dominio, pretenden desarrollar soluciones para el análisis de opiniones más escalables.

Tan [67] emplea redes bayesianas con distintos rasgos para adaptarse al contexto de los textos que analiza. Tan divide a los rasgos que se pueden emplear en análisis de opiniones como específicos y no específicos de un dominio, se considera un enfoque muy acertado, puesto que hay rasgos que son determinantemente independientes de contexto y otros en cambio son fuertemente dependientes del dominio. El problema radica en identificar cuáles de estos rasgos se pueden aplicar independientemente del dominio para analizar opiniones. Por supuesto, un rasgo, como por ejemplo un unigrama, puede tener polaridad invertida si aparece junto a otros, pero los enfoques cuantitativos son capaces de determinar estas dependencias. Este enfoque pretende únicamente encontrar aquellos rasgos que se pueden emplear independientemente del dominio, no su polaridad final.

Tan únicamente considera N-Gramas para ver si son generalizables o no. Para encontrar N-Gramas generalizables en cualquier dominio emplea la métrica *Frequently Co-Occurring Entropy (FCE)*. Esta métrica incluye el criterio de quedarse con los rasgos que ocurren frecuentemente en dos dominios y que ocurren con probabilidad similar. Sea $P_0(w)$ la probabilidad de ocurrencia del N-Grama en el contexto original y $P_n(w)$ la probabilidad de ocurrencia en el nuevo dominio, entonces se busca maximizar:

$$f_w = log(\frac{(P_0(w)P_n(w))^{\pi}}{(|P_0(w) - P_n(w)| + \beta)^{1-\pi}})$$

Donde β es un parámetro con valor 0,00001 para evitar una indeterminación en el caso que $|P_0(w) - P_n(w)| = 0$ y el parámetro $\pi \in [0, 1]$ es parametrizable por el usuario y empleado para dar más importancia al criterio de quedarse con los rasgos que ocurren frecuentemente $P_0(w)P_n(w)$ o con los que ocurren con probabilidad similar $P_0(w) - P_n(w)$.

Este tipo de heurísticas son útiles para quedarse con los rasgos que mejor se adaptan a un cambio de contexto. Pueden, por ejemplo, servir para mejorar un léxico polarizado y adaptarlo a los dos contextos que se trabajan. Los autores sólo la aplican a N-Gramas pero esta métrica tendría mayor valor si fuese adaptada a cualquier tipo de rasgos en análisis de opiniones. También tendría un mayor valor si fuese empleada para un número mayor de dos contextos. Esta métrica puede ser embebida en un filtro para los sistemas de aprendizaje automático, para detectar los rasgos que mejoran la precisión.

Elming [48] propone un enfoque en el que en el proceso de aprendizaje automático no gobierne un rasgo sobre otros por ser más representativo en el contexto en el cual se ha entrenado el modelo. A este proceso le denomina *Data Corruption*. Consiste en que al entrenar el modelo de aprendizaje automático, si un rasgo gana peso para categorizar a textos entrantes, se le elimine del modelo. De esta forma, el peso de dicho rasgo se distribuye entre otros rasgos, logrando un modelo que en teoría es más generalista.

Este proceso de *Data Corruption* es aplicable en cambios de contexto, dado que cada uno de ellos es explicado por rasgos distintos. Al no centrar todo el peso en pocos rasgos, si se cambia el contexto y estos rasgos no aplican el modelo es robusto dado que ha distribuido el peso en otros rasgos. Se trata de un enfoque recomendado para modelos cuantitativos.

Una parte de los rasgos, como el de estructura polar definido por Wachsmuth [29], se consideran como independientes de contexto. Sea cual sea el texto a analizar, se han definido determinadas estructuras polares que denotan la polaridad del texto. Cada dominio puede albergar textos con una diferente frecuencia de estructuras. Pero dado que estas se agrupan mediante el proceso de clustering en las distintas categorías, es indiferente lo distintas que sean en función del contexto. Es por tanto importante, como ya introduce Tan para los N-Gramas, clasificar a los rasgos como dependientes o independientes del dominio, considerando cualquier rasgo y no sólo N-Gramas.

El problema de la independencia de dominio o contexto es poco tratado en la literatura pero de esencial uso práctico. Si el modelo no es robusto frente a cambios de contexto, hay que hacer un modelo distinto por cada dominio que se quiera analizar. Una metodología es tratar los contextos o dominios no estudiados por expertos con este tipo de modelos hasta que sea posible realizar su supervisión por expertos. Con esto se gana agilidad en la creación de modelos de análisis de opiniones. Es por tanto

necesario el diseño de heurísticas o modelos que sean robustos frente a cambios de contexto.

2.6.4. Extracción de Argumentos

Una variante de problemas en el análisis de opiniones trata de encontrar no sólo la opinión subyacente de un texto sino también los argumentos que soportan esta opinión. A partir de estos argumentos, el texto puede ser clasificado en distintas categorías, que explican las posibles posturas ante un problema o una decisión, expresadas a través de los argumentos detectados. Se trata de un problema extremadamente dependiente del dominio, dependiendo del dominio habrá categorías distintas, y de una información expresada de modos muy diferentes sin que ellos tengan nada que ver en común, por lo que, desde un punto de vista personal, se recomienda categorizar los textos entrantes en diferentes dominios y aplicar conocimiento experto en cada uno de ellos para extraer y clasificar estos argumentos. En consecuencia, los modelos cualitativos son una solución óptima.

Moreno [75] propone una metodología para lidiar con este tipo de problemas. En primer lugar, se propone que un experto elabore una clasificación manual de los argumentos de las opiniones para inferir la posición del sujeto expresada en el mensaje. A partir de esta clasificación, se propone que se elaboren reglas lingüísticas que expliquen esta clasificación basándose en los patrones lingüísticos que tienen en común los textos de cada categoría. Estas reglas lingüísticas son posteriormente codificadas en el sistema de minería de textos a implementar.

Además de las reglas lingüísticas codificadas, se necesita saber la polaridad de la opinión, para lo que se construye un recurso dependiente del dominio con la categoría gramatical de los términos seleccionados y su polaridad. Esta polaridad es aprovechada por las reglas lingüísticas para no sólo seleccionar los argumentos sino también la polaridad de cada texto. Con la polaridad de cada uno de los términos del recurso, las categorías gramaticales y un orden establecido entre ellas, se codifican reglas lingüísticas que si son reconocidas clasifican a un texto en una categoría. Por ejemplo, $DET?AP?NAP^*PP^* -> C1$ donde DET simboliza encontrar un determinante, AP una frase adjetival, N un sustantivo y PP una frase. Se codifica además que al menos uno de estos términos debe tener polaridad positiva. Si se reconoce esta regla en el texto, se clasifica en la categoría C1. En este artículo se detalla un caso de uso. Moreno comenta detalles adicionales de la aplicabilidad de esta metodología en un artículo posterior [75].

Estos modelos cualitativos adquieren gran precisión en problemas pertenecientes a dominios muy pequeños. Es difícil incluir modelos cuantitativos en este tipo de problemas puesto que se dispone de una base de textos muy pequeña. Una posible metodología para crear un modelo híbrido sería incluir un modelo cuantitativo paralelo al modelo cualitativo. Para obtener un mayor número de textos con el objetivo de que el modelo cuantitativo ofrezca buenos resultados, se emplea un mecanismo de *Bootstrapping* al modelo del lenguaje obtenido por el conjunto de los textos de los que se disponga para tener un mayor número de textos.

Para ello, se calcula la probabilidad de la aparición de unigramas a N-Gramas hasta un determinado N por máxima verosimilitud o mediante un mecanismo de Smoothing. Una vez obtenida la distribución de probabilidad del texto, se generan textos automáticamente acordes a estas probabilidades mediante una gramática generativa. Con este procedimiento, se aumenta el número de textos. Con un número de textos lo suficientemente grande, un modelo cuantitativo y el modelo cualitativo citados pueden combinarse para obtener un modelo híbrido que mejore los resultados de ambos por separado. La complejidad de este enfoque es grande y la presencia de ruido es relevante, por lo que, si el enfoque cualitativo da buenos resultados, se considera suficiente.

Del análisis de opiniones modelado para superar estas problemáticas se extrae información útil que explica todo tipo de tendencias. En el siguiente apartado de este estado del arte, se describen estas tendencias.

2.7. Tendencias explicadas por el Análisis de Opiniones

Los modelos de análisis de opiniones sirven como soporte para explicar tendencias. En esta sección se exponen los ejemplos más recientes sobre las tendencias que se pueden explicar mediante el análisis de opiniones.

En su artículo, Connor [19] categorizó a los tweets de Obama y McCain para las elecciones norteamericanas de 2008. Construye una serie temporal de la media de positividad leída en Twitter de cada uno de los candidatos a presidente durante todo 2008.

En estos estudios se debe tener en cuenta que la población sobre la que se hace el estudio está sesgada, por ejemplo por edad. El uso de Twitter es más común entre los jóvenes que entre los mayores. Aún así es un buen indicativo de la opinión de un

sector de la población. En la siguiente figura se puede observar como la tendencia es favorable a Obama, representado mediante una línea azul mientras que McCain está representado por una línea roja. Finalmente Obama fue elegido presidente.

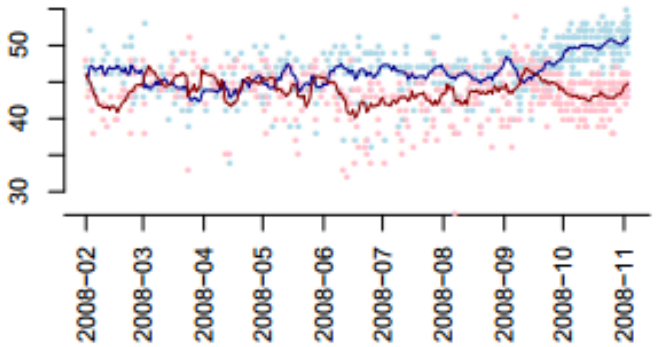

FIGURA 2.8: Serie temporal que refleja la positividad media de los dos candidatos a presidente en EEUU durante 2008.

Explicar las tendencias políticas a través de Twitter no sólo ha sido realizado en EEUU sino que este caso de estudio también ha sido efectuado en España por Pla [59]. En cambio, en este caso se predice la tendencia política de un usuario de Twitter en función a sus tweets. Para realizar esta tarea se emplea análisis de opiniones.

En primer lugar se reconoce si un usuario está hablando de un partido o movimiento político. Una vez queda reconocido el partido o movimiento y su tendencia política, se analiza si la opinión de ese usuario en lo relativo al movimiento político es positiva o negativa. Si es positiva se clasifica al usuario con esa tendencia política y de lo contrario con su opuesta.

Sea U_i el usuario que se quiere analizar y T_{ij} el tweet j del usuario i. Sea E_{ijk} la entidad k reconocida en el tweet j del usuario i. Cada entidad política es representada mediante un número comprendido en el intervalo cerrado $[-1, 1]$ donde -1 es izquierda y 1 es derecha. Si se obtiene la polaridad expresada en cada tweet que habla sobre una entidad del usuario i mediante otro número comprendido en $[-1, 1]$, entonces la tendencia política del usuario i se obtiene mediante la siguiente expresión:

$$Political\_Tendency(U_i) = \frac{\sum_{j=1\cdots|T_i|} \sum_{k=1\cdots|E_{ij}|} Polarity(E_{ijk})Tendency(E_{ijk})}{\sum_{j=1\cdots|T_i|} |E_{ij}|}$$

Es decir, la polaridad del usuario viene definida por la media del producto de sus opiniones sobre movimientos políticos y la tendencia política de estos partidos políticos. Hay que ser crítico desde el punto de vista de la rigurosidad del estudio, pues la muestra del espectro social que emplea Twitter no es uniforme según la edad de los usuarios.

Además de explicar la tendencia política, el análisis de opiniones en Twitter ha sido también empleado como factor para explicar los cambios en el valor de una acción en los mercados, los detalles de este estudio se encuentran en la publicación de Si [76]. Al igual que con las elecciones, en este estudio se construye una serie temporal diaria con la media de la positividad analizada sobre un valor. Al igual que en el anterior enfoque, se debe considerar solo como una muestra sesgada.

Construida esta serie temporal, se puede comparar con el comportamiento del valor en el mercado y hacer análisis predictivo. Mediante un estudio muy similar, Bollen [5] ha conseguido un 87.6 % de precisión prediciendo subidas y bajadas en los valores de cierre diario del Dow Jones. En los últimos años, el análisis de opiniones ha sido un factor clave en las decisiones de los inversores junto al análisis fundamental y técnico.

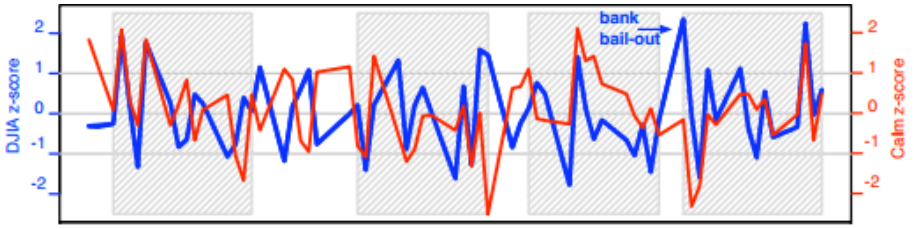

FIGURA 2.9: Series temporales del Dow Jones (azul) y de uno de los rasgos del sistema de Bollen [5] (rojo).

Aparentemente los resultados del caso de estudio de este experimento son muy positivos. Por contra, se sospecha que es un caso cerrado que no es capaz de generalizar. No se encuentra ningún modelo en la literatura basado en este estudio que contemple otro caso de uso. Es extraño, puesto que los resultados son muy positivos. Otra tendencia que se explica mediante el análisis de opiniones es la popularidad de un producto, en el artículo de Gao [50]. El primer paso es reconocer en el texto

el producto del cual se está expresando una opinión. Posteriormente se analizan las opiniones de los usuarios sobre ese producto. De esta forma es posible averiguar, por ejemplo, si una campaña propagandística ha mejorado la opinión de los usuarios sobre un producto.

Dada la eficacia del análisis de opiniones para explicar las tendencias anteriores han surgido estudios que han intentado explicar otras tendencias. Moniz [77] relaciona la satisfacción de los empleados de una empresa con sus ganancias. La hipótesis que se quiere contrastar es que si la opinión de los trabajadores es positiva entonces la empresa tiene más ganancias.

Ranganath [54] construye un modelo que averigua si una pareja que tiene su primera cita persigue tener una amistad o una relación romántica mediante una serie de factores entre los que se incluye análisis de opiniones. Li [65] predice que foros van a ser los más concurridos para un tema determinado. Nitta [78] detecta abuso cibernético entre adolescentes en las webs de colegios... Todos estos estudios, al tener sentido común sus implicaciones, se consideran acertados.

Se han citado las tendencias explicadas por análisis de opiniones más relevantes en la literatura actual pero existen otras. Se trata de un campo novedoso y de alta utilidad para la empresa por lo que se ha intentado explicar distintos fenómenos a través de las opiniones de los usuarios.

2.8. Conclusiones

Se ha elaborado un estado del arte de los modelos de análisis de opiniones. El análisis de opiniones es una tarea para la cual se han propuesto un gran número de modelos dada la complicación de la misma, al tratar con la ambigüedad del lenguaje natural y la subjetividad inherente de las opiniones. Cada uno de estos modelos cubre determinados dominios o facetas del problema con eficacia pero es raro encontrar algún modelo que escale bien en cualquier tipo de dominio.

En este estado del arte se ha comenzado con un repaso de los métodos tradicionales que se han empleado para esta tarea. Existen heurísticas que tratan de polarizar textos y opiniones y que han resultado efectivas. Pese a ser más básicas que modelos más avanzados, proporcionan buenos resultados, por lo que pueden ser contempladas como rasgos para un modelo de aprendizaje automático.

Se ha comentado también dentro de este apartado técnicas que hacen uso de la semántica de forma avanzada para analizar opiniones. Estas técnicas mejoran el

rendimiento del análisis de opiniones pero suelen ser muy complejas y requerir de recursos lingüísticos. De nuevo, sus resultados pueden ser también usados como rasgos.

Una vez comentados los métodos tradiciones se expusieron los recursos lingüísticos más populares en análisis de opiniones. Estos recursos incorporan términos polarizados que facilitan el análisis de opiniones para los modelos y aplicaciones. Pese a que ser muy extensos existen desventajas con su uso. Las principales desventajas son que la mayoría de ellos sólo están disponibles para la lengua inglesa, que el lenguaje es variante por época y los términos que expresan polaridad en un año no son los mismos que en 10 años después, las variedades geográficas del lenguaje tampoco son tomadas en cuenta, la mayoría de léxicos polarizados no toman en cuenta las diversas acepciones que puede tener un término y los términos que presentan son muy generalistas por lo que no se adaptan a ningún dominio particular.

Para solventar estos problemas se crean procedimientos de expansión de la polaridad del léxico polarizado a otros términos. Una parte de los recursos lingüísticos son creados mediante estos procedimientos. Estos modelos son útiles pues intentan crear un léxico que se adapte mejor a un problema. Por ejemplo se han creado procedimientos para solventar la barrera del multilingüismo y crear un léxico polarizado multilingüe. Aún así, es importante destacar que pocos son los modelos que han intentado crear un léxico polarizado que se adapte a cualquier dominio. Esto último es importante porque un modelo de análisis de opiniones debe ser lo más robusto e independiente del contexto o dominio como sea posible para garantizar un análisis de calidad sobre cualquier texto.

Los términos proporcionados por estos léxicos polarizados son comúnmente usados como rasgos por modelos de aprendizaje automático que categorizan textos en polaridades distintas. Se han revisado los modelos más frecuentemente empleados en análisis de opiniones con especial énfasis en la Máquina de Soporte Vectorial dada su popularidad en el análisis de opiniones. En este apartado se han expuesto los rasgos que se han tenido en cuenta en análisis de opiniones. Es importante destacar que una parte de estos rasgos son generalizables a todos los dominios mientras que otros son particulares del dominio. De nuevo es importante distinguir el dominio en el cuál se opera para poder analizar opiniones con calidad.

Se cierra el estado del arte con dos secciones que hablan sobre las problemáticas específicas del análisis de opiniones ya comentadas en estas conclusiones y las aplicaciones que se están elaborando que funcionan sobre modelos de análisis de opiniones.

El principal problema de los enfoques comentados como ya se ha hablado es el multilingüismo. Existen dos opciones viables. Una de las opciones es traducir los textos encontrados en cualquier idioma a inglés y a partir de ahí elaborar el análisis de opiniones con los modelos presentados. La otra opción es convertir recursos lingüísticos de la lengua inglesa a una interlingua como UNL y convertir el texto de cualquier lengua a esta representación que es en donde se ejecutaría el modelo de análisis de opiniones. Intentar hacer un modelo de análisis de opiniones que sea independiente del lenguaje es un enfoque muy complejo. Por eso se recomienda elaborar modelos de multilingüismo y de análisis de opiniones por separado.

Para saber qué dominio se está analizando es importante la frecuencia con la que se emplean los términos y el orden en el que se presentan. Se han presentado en el estado del arte métodos de carácter estadístico o no supervisado que analizan el texto extrayendo patrones y términos frecuentes. Esta extracción y categorización de términos puede ser empleada también como rasgo en un modelo de expansión de léxicos polarizados. La ventaja de este tipo de métodos es que son independientes del dominio en el que se está operando.

En cambio, dentro de un mismo lenguaje también existen distintos dominios que hay que analizar empleando distintos rasgos. No tiene sentido en este problema crear modelos independientes para cada dominio puesto que entonces habría que tener un dominio para operar en Twitter, otro para periódicos, otro para críticas de películas, etc... Dado que elaborar modelos independientes para cada dominio no es escalable, urge la necesidad de crear un modelo robusto que sea capaz de analizar opiniones con calidad en cualquier dominio.

Ninguno de los enfoques que se han presentado funciona mejor en todos los dominios. Esto se denomina en *Data Mining* como *No Free Lunch Theorem*. Es decir, ningún algoritmo es mejor en todos los dominios. Esta característica aplica también en el procesamiento del lenguaje natural. Por ello, un modelo que sea robusto debe incorporar múltiples de los enfoques que se han visto en el estado del arte y un modo o heurística inteligente de combinarlos para conseguir adaptarse a cada dominio perdiendo la menor generalización posible.

Capítulo 3

Modelo

3.1. Introducción y Motivación

Los enfoques actuales en análisis de opiniones son fuertemente dependientes del contexto y de los léxicos polarizados con los que operan, siendo incapaces de generalizar. Por ejemplo, considérense los siguientes dominios: Startups y Competencia. En el dominio Startups, se encuentran textos como el siguiente: *El beneficio de este mes es considerable*. Un léxico polarizado puede contener los términos *beneficio* y *considerable*. En el dominio de las startups, y en concreto en esa frase, la polaridad de esos términos es positiva. En cambio, en el dominio competencia, la frase *El beneficio de la competencia es considerable* hace que las polaridades de ambos términos sean negativas.

Se comprueba que la polaridad de los términos de léxicos polarizados es fuertemente dependiente del contexto. En consecuencia, el mismo procedimiento utilizado en distintos contextos ofrece muy buenos resultados en unos dominios pero resultados muy pobres en otros dominios. Esto aplica modelos de análisis de opiniones basados en léxicos polarizados y a procedimientos no supervisados de expansión de los mismos. Debido a esta problemática, equipos de lingüistas y expertos en cada dominio analizan cada contexto y elaboran reglas que, de ser detectadas por un modelo, infieren la polaridad de un texto, en función siempre al dominio analizado. Con ello se pretende obtener diferentes modelos de análisis de opiniones que obtengan resultados más precisos a tener un único modelo general para todas las categorías.

Esta metodología de trabajo, basada en categorizar los textos y analizar cada categoría de textos por separado, es aceptable, pero presenta problemas en la práctica.

Los expertos y lingüistas son un recurso muy costoso, tanto en tiempo como financieramente. Si el número de categorías es muy elevado, hecho muy común por ejemplo al analizar textos en redes sociales, esta metodología es difícilmente escalable. En casos extremos, el número de categorías puede llegar a ser equivalente al número de textos.

Un problema adicional es que, dentro de la misma categoría, una regla puede inferir una polaridad que no es correcta. Por ejemplo, en el dominio Startups, se definía que *beneficio* tiene polaridad positiva. Si únicamente se considera este rasgo para polarizar, en la frase *El beneficio fue elevado únicamente en los sueños del CEO* la polaridad inferida no es la correcta. Incluso, el término *elevado*, susceptible de aparecer en un léxico polarizado con polaridad positiva, continúa sin lograr que la polaridad predicha sea la correcta.

Este tipo de problemáticas son muy comunes en el lenguaje natural, en el que, desde un punto de vista cuantitativo, se da un gran número de combinaciones de términos en una frase que hacen revertir su polaridad. La única forma de lidiar con esta problemática es inferir combinaciones entre los términos y otros rasgos que juntas infieran polaridades. El uso de léxicos polarizados y reglas construidas por expertos logra resultados muy precisos si los dominios están muy bien definidos y se cuenta con un tamaño de textos razonable. Los problemas aparecen cuando el volumen de textos a analizar es muy grande y no está categorizado.

En esta circunstancia, se hace necesaria la ayuda de un modelo cuantitativo para, al menos, desarrollar un prototipo de cada una de las categorías no analizadas en profundidad por lingüistas. Los modelos cuantitativos se presentan como una alternativa de gran valor, puesto que estos modelos si cuentan con un conjunto de textos anotados lo suficientemente grande, son capaces de inferir distintos conjuntos de valores de variables dependientes altamente correlacionados con distintas polaridades. En el caso de esta tesis sólo se considera la polaridad positiva y negativa. Se asume, como más adelante se detalla, que todos los textos a procesar son subjetivos. El problema de discernir objetividad o subjetividad en un texto es independiente al de categorizar la polaridad de un texto.

No obstante, se recomienda que los prototipos creados por modelos cuantitativos como el que se va a presentar sean revisados posteriormente por lingüistas para incrementar su precisión. Los modelos cuantitativos en análisis de opiniones poseen la métrica de recuperación elevada pero peor precisión que los modelos cualitativos por norma general y viceversa. Un modelo híbrido, como el presentado en esta tesis tras ser revisado por expertos tal y como se recomienda, cuenta con las ventajas de

ambos. Si se le denomina *recall* a la alta capacidad de recuperación de los modelos cuantitativos, el siguiente gráfico comenta las características de cada tipo de modelo:

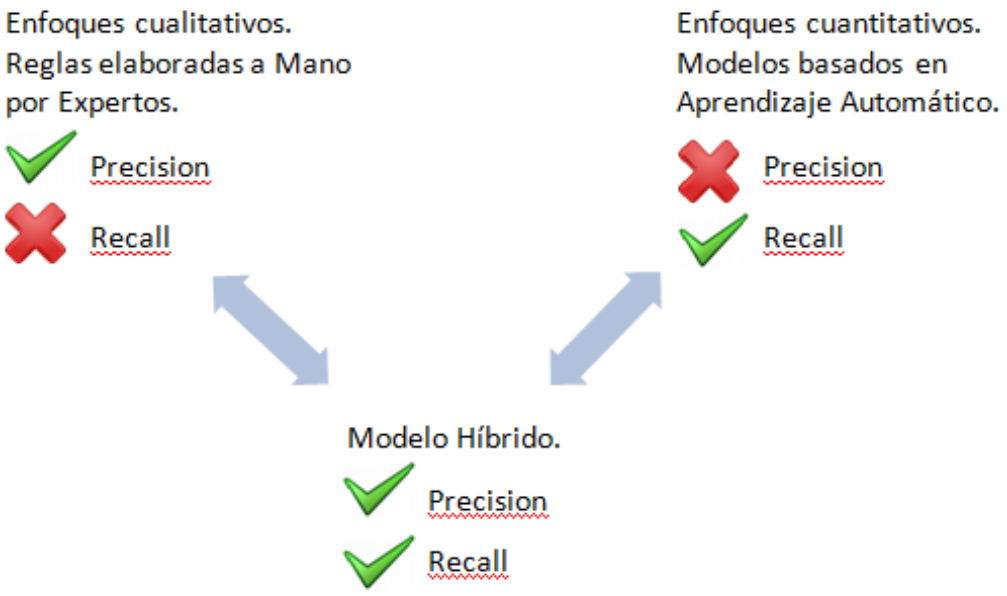

FIGURA 3.1: Características de los distintos tipos de modelos aplicables en el análisis de opiniones.

En esta tesis, se presenta un modelo cuantitativo que hace uso del conocimiento cualitativo contenido en léxicos polarizados y otros recursos para la construcción de prototipos de modelos de análisis de opiniones sobre un conjunto de dominios. Se recomienda su uso para situaciones en las cuáles se pretende analizar un gran volumen de textos anotados por un número muy extenso de categorías.

Dado este gran número de categorías, el tiempo y recursos necesario para elaborar modelos cualitativos sobre cada una de ellas es muy elevado. Hasta que se disponga de dichos recursos, este modelo ofrece una alternativa para tener un prototipo que funcione con eficacia independientemente de las categorías de los textos.

Como se verá a continuación, este modelo expande léxicos polarizados semilla de forma inteligente para obtener resultados de calidad en independientemente del dominio. Para ello, se va a emplear un modelo híbrido consistente en una mezcla de métodos tradicionales junto a aprendizaje automático. Se persigue que el modelo incorpore la precisión de los rasgos cualitativos con la alta capacidad de recuperación

de los modelos cuantitativos. De este modo el modelo incorpora las mejores características de ambos enfoques. Uno de los objetivos, que se describen más adelante, es que con este modelo, se efectúe un análisis de opiniones de calidad sin preocupación del dominio en el cual se opera.

Este modelo no sólo obtendrá prototipos eficientes en diferentes conjuntos de categorías sino que, de forma adicional, estos prototipos servirán como soporte a los lingüistas, los cuáles al crear modelos sobre dominios no estudiados tendrán ya rasgos preparados para estas categorías. Su trabajo consistirá no en crear diferentes modelos para cada una de las categorías desde cero, sino en modificar modelos ya existentes.

Se persigue que el modelo que se va a definir en esta Tesis obtenga un análisis de opiniones robusto frente a cambios de contexto y que solucione la problemática consistente en que el lenguaje nunca deja de cambiar. Con este modelo el usuario final podrá analizar las opiniones de cualquier medio o temática sin preocuparse de si el modelo tendrá malos resultados al no adaptarse a ese dominio particular.

El modelo presentado en esta Tesis pretende obtener un vector de características generalizable a un número indeterminado de diferentes dominios de los cuáles se quieren analizar opiniones. El modelo incluirá también parámetros para ponderar la importancia de cada uno de los dominios. Esto es debido a que se puede dar la circunstancia en la cual sea más prioritario obtener resultados más precisos en una de las categorías pero intentando mantener la mayor generalización posible.

El modelo propuesto se basa en la extensión de un trabajo propuesto con anterioridad, en el cuál se perseguía definir un modelo que obtuviese rasgos generalizables ante dos dominios diferentes. El modelo resultante pierde precisión en los dos dominios por separado, pero obtiene mejores resultados que aplicar el modelo de una categoría a otra. Por tanto, sirve como prototipo para analizar opiniones en el dominio que todavía no ha sido estudiado por expertos. A continuación se explican las bases teóricas sobre las que se sustenta este Modelo, para más tarde explicar el enfoque del modelo propuesto en esta Tesis.

3.2. Trabajo previo y bases teóricas sobre las que se sustenta el Modelo

Descrita la motivación para definir este modelo, se procede a sentar las bases teóricas del modelo a definir. El modelo que se va a presentar en esta tesis es una generalización del propuesto por Tan [67]. El objetivo de Tan es adquirir características generalizables a dos dominios en la tarea del Análisis de Opiniones.

El objetivo de este modelo es adquirir características generalizables a un número indeterminado de categorías. La generalización, por tanto, consiste en que mientras el modelo de Tan únicamente obtiene los rasgos generalizables a dos categorías, el modelo propuesto en esta Tesis obtiene rasgos generalizables a las categorías que se necesiten analizar.

El modelo de Tan ha sido ya expuesto en el estado del arte, pero dado que el modelo que se va proponer es una ampliación de este, es importante revisar el modelo de Tan. Una vez revisado el modelo, se propondrá la generalización y ampliación del mismo.

Para resolver el problema de la generalización de rasgos que representan polaridad en opiniones, Tan publicó el enfoque que se comenta a continuación. En primer lugar, es necesario diferenciar entre dos tipos de rasgos en el análisis de opiniones dentro de la misma categoría. Estos tipos de rasgos son los que son específicos al dominio y los que son generalizables a cualquier dominio de entre los considerados. En esta tesis se amplía esta definición a conjuntos de rasgos dependientes e independientes del dominio. Se piensa que esta definición es más adecuada para los modelos cuantitativos.

Por ejemplo, considérese la combinación de los valores de las siguientes variables en un texto: Valor de la variable 1, *Aparición del término beneficio* cierto; valor de la variable 2, *Aparición del término bajo* cierto. Si las dos variables poseen valor cierto, entonces se tiene un conjunto de rasgos que implica polaridad negativa en el dominio Startups. La variable artificial creada por la intersección del valor de ambas variables es un rasgo específico del dominio Startups, pero que, probablemente, sea generalizable a más categorías. Existen rasgos que combinan valores de más de dos variables.

En su artículo, Tan busca los rasgos generalizables para dos categorías. Los encuentra mediante la expresión que se va a definir en el siguiente párrafo. Una vez encontrados, los modelos de aprendizaje automático a los que recurre encuentran

las combinaciones de valores de las variables citadas en el anterior párrafo que clasifican los textos en categorías. Pero estos modelos no pueden inferir correctamente la polaridad de un texto si no cuentan con los rasgos adecuados. Es por lo tanto importante encontrar una expresión que obtenga los rasgos generalizables a, en el caso de Tan, dos dominios diferentes.

Para ello, Tan propone una heurística que, a partir de un léxico polarizado, se queda únicamente con aquellos rasgos que calcula que son generalizables a dos dominios. Tan denomina a esta métrica la *Frequently Co-Ocurring Entropy (FCE)*, que obtiene características generalizables en dos dominios diferentes. Estas características son, en su mayoría, unigramas. Pese a ello, es posible generalizar esta expresión para que no únicamente considere unigramas. La expresión es la siguiente:

$$f_w = log(\frac{P_0(w) * P_n(w)}{|P_0(w) - P_n(w)|})$$

Donde $P_0(w)$ es la probabilidad de encontrar la palabra w en el modelo del lenguaje del dominio 0 y $P_n(w)$ es la probabilidad de encontrar la palabra w en el modelo de lenguaje del dominio n. Esta expresión, denominada *Frequently Co-Ocurring Entropy (FCE)*, mide dos factores por separado. El numerador trata de extraer características que ocurran en ambos dominios. Para que una característica sea generalizable, debe ocurrir en ambos dominios con una probabilidad elevada. Lo ideal, es que en los dos dominios la palabra w ocurra con probabilidad similar, entonces se dice que es generalizable. El denominador trata de extraer características que ocurran con igual probabilidad, penalizando aquellas que ocurren más en un dominio que en otro a igual numerador.

La probabilidad $p_i(w)$ puede ser calculada por máxima verosimilitud o por cualquier otro método, como por ejemplo una medida basada en un método de Smoothing. Dado que Tan sólo considera dos dominios, es razonable emplear una probabilidad basada en máxima verosimilitud. Una de las generalizaciones que se propondrá para este modelo es no sólo considerar unigramas para esta expresión sino otro tipo de variables.

El analista de opiniones puede considerar, dadas dos categorías a analizar, que la generalización tiene una importancia diferente a que las palabras ocurren con probabilidad similar en dominios diferentes. Para modelizar esta circunstancia, Tan [67], otorga un peso a cada uno de estos dos factores, expresados por el numerador y el denominador. Con ello, Tan concluye en la siguiente expresión que encuentra características generalizables:

$$f_w = log(\frac{(P_0(w)P_n(w))^{\mu}}{(|P_0(w) - P_n(w)| + \beta)^{(1-\mu)}})$$

Donde el parámetro $\mu \in [0, 1]$ otorga más importancia a cada uno de los dos criterios presentados con anterioridad. El parámetro β se incluye para evitar una indeterminación cuando $|P_0(w) - P_n(w)| = 0$ y su valor es aproximadamente cero. Para evaluar el léxico polarizado resultante de aplicar esta expresión, Tan evalúa cómo se comporta esta medida mediante aprendizaje automático.

Una vez decididas las características que se van a extraer de cada texto, en función a las que más puntuación obtengan por la expresión propuesta, Tan construye un conjunto de datos con el valor de estas características. Cada instancia del conjunto de datos simboliza un texto. La instancia está compuesta por un conjunto de variables que simbolizan la aparición de los términos extraídos por la expresión. Estos valores están acompañados por la polaridad de cada texto del cual se extraen los valores.

Con este conjunto de datos, Tan evalúa la métrica en ambos dominios. Para ello se sirve de las medidas de evaluación extraídas de analizar opiniones con Multinomial Naive Bayes. Estas son la precisión, la recuperación y el F-Measure. Este modelo logra mejores resultados en ambos dominios con respecto a aplicar el modelo de análisis de opiniones de una categoría para otra categoría.

Se piensa que el modelo elaborado por Tan es generalizable por varias vías. En primer lugar, no únicamente considerando unigramas sino cualquier otro tipo de rasgo que pueda aportar un mayor valor cualitativo. En segundo lugar, generalizando el modelo para un número indeterminado de categorías. En tercer lugar, integrando la expresión en un modelo de aprendizaje automático para que elabore una expansión de léxicos polarizados a partir del modelo de lenguaje de cada una de las categorías. En cuarto lugar, definiendo una metodología para la construcción de prototipos de categorías a partir del modelo que se va a proponer. En último lugar, proponiendo variables para una mayor parametrización del analista de opiniones.

A continuación, se detalla una lista de hipótesis, asunciones, restricciones, limitaciones y objetivos que tendrá el modelo que cuenta con las características que se han citado. Una vez detallada esta lista, se expondrán los detalles del modelo que supone una generalización al presentado por Tan.

3.3. Hipótesis

Antes de describir los detalles del modelo, es preciso describir el alcance del mismo, que es definido mediante hipótesis. A continuación, se presenta un listado de hipótesis que se quieren evaluar del modelo que se propondrá posteriormente. En el apartado de conclusiones se revisarán estas hipótesis, con el objetivo de comprobar cuáles de ellas han podido ser evaluadas.

- El prototipo creado por el modelo de análisis de opiniones presentado en esta Tesis, analizando un volumen lo suficientemente grande de textos, presentará mejores resultados en las categorías que se analicen en su fase de entrenamiento que en categorías no consideradas. Se pretende demostrar que, pese a ser generalista, el modelo toma en cuenta las características de las categorías a analizar, construyendo un prototipo que presenta resultados adecuados en dichas categorías.

- El modelo propuesto es capaz de mejorar las medidas de precisión, recuperación y F-Measure con respecto a las medidas obtenidas por medio de un modelo de análisis de opiniones basado únicamente en un léxico polarizado semilla en cualquier categoría. Para que sea aplicable esta hipótesis, se debe entrenar el prototipo generado por el modelo de análisis de opiniones únicamente en esa categoría. Es decir, el prototipo construido mediante el procedimiento que se describe en el modelo es independiente del contexto sobre el que opere. Si son analizadas varias categorías, se requiere que se analicen todas las categorías conjuntamente en la fase de entrenamiento del modelo.

- El modelo propuesto es capaz de mejorar, en al menos la mitad de los casos y habiendo analizado un número lo suficientemente grande de textos, las medidas de precisión, recuperación y F-Measure con respecto a las medidas obtenidas por medio de un modelo de análisis de opiniones generalista en textos que pertenezcan a cualquier categoría. Se requiere que se analicen todas las categorías conjuntamente en la fase de entrenamiento del modelo.

- El modelo propuesto es capaz de mejorar las medidas de precisión, recuperación y F-Measure con respecto a las medidas de los léxicos polarizados semilla, en otros contextos distintos al que pertenece el conjunto de textos a analizar, una vez se ha expandido el léxico polarizado semilla que se utilice. Esta hipótesis únicamente aplica si el número de categorías analizadas y el volumen de textos analizados son suficientemente grandes. Se pretende demostrar que, con

el número suficiente de categorías, el modelo obtenido es generalista. Estas categorías deben ser analizadas conjuntamente por el modelo en su fase de entrenamiento.

- El modelo propuesto es capaz de expandir los léxicos polarizados semilla en cualquier contexto de entre los que se consideran en los experimentos, mostrando una mejoría en las medidas de precisión, recuperación y F-Measure con respecto a las medidas de los léxicos polarizados semilla. Es decir, con la parametrización correcta, el modelo siempre encuentra términos para todas las categorías que no están presentes en los léxicos polarizados semilla. Se requiere que se analicen todas las categorías conjuntamente en la fase de entrenamiento del modelo.
- El modelo propuesto es capaz de mejorar las medidas de precisión, recuperación y F-Measure con respecto a las medidas obtenidas por medio de un modelo de análisis de opiniones construido para una categoría distinta a la que pertenece el conjunto de textos que se analiza. Es decir, el prototipo construido por el procedimiento descrito en el modelo se comporta mejor en la categoría de textos a analizar que un modelo de análisis de opiniones diseñado para otra categoría. Se requiere que se analicen todas las categorías conjuntamente en la fase de entrenamiento del modelo.
- El modelo propuesto es capaz de construir un vector de características para cualquier número de categorías. Desde una categoría al número de categorías que se deseen analizar.
- El modelo debe obtener resultados similares independientemente del léxico polarizado semilla empleado con un número lo suficientemente grande de iteraciones y textos analizados para cualquier conjunto de categorías.

3.4. Asunciones

En el listado de hipótesis se comprueba que el alcance que cubre el dominio que se va a proponer es muy extenso. Dado que el alcance de este problema es muy grande con respecto al que se debe abordar en una Tesis Fin de Máster, se van a definir las siguientes asunciones, con el objetivo de restringir el dominio que se va a abarcar.

- Se asume que los textos que se van a analizar poseen coherencia y cohesión. En otras palabras, mantienen un mismo tema central alrededor del cual exponen

sus ideas y están bien formados por la relación entre sus oraciones. No se pretende analizar la coherencia y cohesión de los textos, sino su polaridad.

- Se asume que los textos contienen términos que no pertenecen a ningún diccionario o recurso empleado o están mal escritos, dado el carácter informal de algunos contextos. Por ejemplo, Twitter. Se asume que el modelo debe funcionar en estos casos, sin necesidad de corregir el texto.
- Se asume que las medidas de evaluación que se emplearán reflejan la adaptación del léxico polarizado expandido al contexto de los textos que se están analizando. No se pretende discutir la bondad de estas medidas.
- Se asume el correcto funcionamiento del software del que se va a hacer uso en el modelo con respecto a sus especificaciones.
- Se asume que los textos etiquetados en cada una de las categorías que se consideran pertenecen realmente a esas categorías. Este no es un problema de clasificación de textos en categorías, sino de clasificación de su polaridad.
- Se asume que los textos que se van a analizar poseen un contenido subjetivo. No se pretende clasificar un texto como objetivo o subjetivo.
- Se asume que el número de categorías que se van a analizar es lo suficientemente elevado como para asumir que el modelo es escalable a un número mayor de categorías.
- Se asume que las características de las categorías a analizar son lo suficientemente distintas como para asumir que el modelo se comportaría de forma similar en categorías no consideradas.
- Se asume que el modelo funcionaría en una lengua distinta de la inglesa.

3.5. Restricciones y Limitaciones

Debido a la limitación de tiempo existente para esta investigación, se deben de tener en cuenta una lista de limitaciones y restricciones. Estas restricciones y limitaciones podrán ser no tomadas en cuenta para futuros trabajos que se realicen sobre esta investigación. Se van a tener en cuenta las siguientes restricciones:

- El modelo únicamente se va a probar en lengua inglesa.

- Sólo se considerarán textos pertenecientes a los contextos sobre los que se van a trabajar. Estos contextos se describen en el apartado de experimentación.
- Únicamente se trabaja con textos que exponen opiniones. Este modelo no pretende discernir si un texto posee un carácter objetivo o expresa una opinión. No es objetivo de esta tesis fin de máster.
- Solamente se considerarán en el modelo las técnicas de aprendizaje automático descritas en el apartado de diseño. No se pretende averiguar que técnica de aprendizaje automático de entre todas las existentes es más efectiva. No es objetivo de esta tesis fin de máster.
- Solamente se considerarán en el modelo los léxicos polarizados que se describen en el apartado de experimentación. No es objetivo de esta tesis fin de máster averiguar que léxico polarizado se adapta mejor a cada texto, sino probar que el modelo mejora los resultados de analizar opiniones con los léxicos polarizados que se consideren.
- Únicamente se utilizarán en el modelo los recursos descritos en el apartado de experimentación. No es objetivo de esta Tesis averiguar qué recursos se adaptan mejor a cada categoría, sino comprobar cómo el modelo crea un prototipo que mejora el comportamiento de cada uno de estos recursos en cualquier categoría.
- Únicamente se contará con el número de textos anotados descritos en el apartado de experimentación. Por razones de tiempo y recursos, no es factible anotar un conjunto de textos adicional, aunque fuese conveniente.
- Únicamente se considerarán los rasgos para el vector de características de cada texto comentados en el apartado de experimentación. Como ya se ha listado en el estado del arte, existe un gran número de características a extraer de los textos para el análisis de opiniones pero por razones de tiempo y recursos es imposible analizar cada una de estas características. No obstante, se recomienda utilizar el mayor número de características posible.

3.6. Objetivos

La lista de objetivos propuestos en esta investigación es directamente extrapolable de las hipótesis que se quieren evaluar. En todo momento, los objetivos son corroborar o refutar las hipótesis que se han propuesto en el apartado de hipótesis. En consecuencia, se quieren verificar, o refutar, los siguientes hechos:

- Comprobar que el modelo obtiene mejores resultados en categorías consideradas que en categorías sin considerar.
- Comprobar si el modelo propuesto hace que los léxicos polarizados mejoren sus medidas de evaluación con respecto a las obtenidas cuando estos léxicos no están expandidos.
- Comprobar si el modelo obtiene prototipos de modelos que analizan opiniones en contextos o dominios que no han sido analizados por lingüistas o expertos en esa categoría.
- Comprobar el correcto funcionamiento del dominio independientemente de la naturaleza de cada categoría.
- Comprobar que el modelo obtiene mejores resultados en las categorías que ha considerado que un modelo diseñado para otra categoría adicional.
- Comprobar el correcto funcionamiento del modelo independientemente del número de categorías considerado.
- Comprobar que el prototipo generado posee independencia con respecto al léxico polarizado semilla empleado.

Una vez acotado el alcance del modelo por todas las hipótesis, restricciones, asunciones y objetivos descritos anteriormente, se deben explicar las características del modelo. De este modo, se podrá contrastar si se cumplen las hipótesis mencionadas.

3.7. Modelo Propuesto

Para dar cobertura a la lista de hipótesis y otras características mencionadas anteriormente, el modelo presentado en esta Tesis, pretende obtener un vector de características generalizable no sólo a dos, sino a un número indeterminado de diferentes dominios de los cuáles se quieren analizar opiniones. El prototipo resultante será válido para todas las categorías que se analicen.

El prototipo se podrá aplicar a cualquier número de categorías de entre las consideradas en su fase de entrenamiento, desde una sola categoría a cualquier número de ellas. Para obtener este prototipo, se extiende el enfoque propuesto por Tan con su Frequently Co-Ocurring Entropy (FCE), proponiendo en este apartado la medida

Generalized Entropy Appearance (GEA), que además de suponer una generalización frente a la FCE añade nuevas características.

$$Frequently\ Co-Ocurring\ Entropy\ (FCE) \xrightarrow{aplicable\ a} Dos\ Dominios$$
$$Generalized\ Entropy\ Appearance\ (GEA) \xrightarrow{aplicable\ a} De\ 1\ a\ N\ Dominios$$

Donde N es un número entero que puede tomar cualquier valor superior a uno. Esta medida, desde el punto de vista del aprendizaje automático, se trata de un modelo de Feature Subset Selection (FSS) especializado para unigramas. En vez de tomarse todos los posibles unigramas que pertenecen a todos los textos, se toma únicamente un conjunto de unigramas que pertenecen a un léxico polarizado más un conjunto de unigramas generales y uniformes en todas las categorías y que además denotan polaridad en estos textos.

El modelo que se va a proponer expande el conjunto de términos de un léxico polarizado a un conjunto de términos mayor. El espacio de soluciones es cualquier combinación entre los unigramas pertenecientes a los textos analizados que no se encuentran en el léxico polarizado empleado más los términos del léxico polarizado. El modelo encuentra la solución que maximiza una función fitness que representa la evaluación de los modelos de aprendizaje automáticos empleados para clasificar los textos en categorías. A continuación se analizará este modelo en mayor detalle.

A diferencia de los métodos de Feature Subset Selection de aprendizaje automático, esta medida, aplicada a textos, no sólo se queda con los rasgos más representativos de un léxico polarizado semilla, sino que averigua que rasgos de los no contenidos en el léxico polarizado son aplicables a todas las categorías consideradas. Para ello se sirve de distintos clasificadores de aprendizaje automático, como se verá en este apartado.

La GAE es la heurística encargada de elegir, iterativamente, que términos de los textos analizados son más generales, uniformes y representativos de polaridad en el conjunto de textos analizados. Estas características serán denominadas de aquí en adelante los tres pilares de la GAE: Generalización, Uniformidad y Polarización. El modelo planteado en esta memoria, en cada iteración que aplica la GAE, extrae el vector de características o rasgos para que sea utilizado por los modelos de aprendizaje automático.

Para este modelo se analizarán únicamente unigramas, debido al alcance de la tesis fin de máster. En la sección de trabajo futuro de esta memoria se propone un modelo que expande el espacio de soluciones de todas las posibles combinaciones

de unigramas presentes en el texto además de los términos del léxico polarizado a cualquier combinación de N-Gramas presentes en el texto más los términos del léxico polarizado. Este espacio de soluciones es mucho mayor, por lo que la convergencia de la heurística GAE será más lenta, pero se presume que se obtendrá un mejor rendimiento.

Como se ha mencionado previamente, la GAE extrae iterativamente los términos más generales, uniformes y representativos de polaridad. Esta heurística se basa en una combinación lineal de estas tres características para determinar qué términos deben ser añadidos a los del léxico polarizado. Serán añadidos los términos más valorados por la GAE. La puntuación $GAE(w) \in [0, 1]$ del unigrama w es calculada por la siguiente expresión:

$$GAE(w) = q_1 g(w) + q_2 u(w) + q_3 pol(w)$$

Donde $g(w) \in [0, 1]$ simboliza la generalidad del término w, $u(w) \in [0, 1]$ simboliza la uniformidad del término w y $pol(w)$ simboliza la polarización del término w. Los términos $q_1 \in [0, 1]$,$q_2 \in [0, 1]$ y $q_3 \in [0, 1]$ son pesos configurables por el analista. Estos tres términos deben sumar uno. Con estos términos el analista puede ponderar la importancia de que un término sea general, uniforme o representativo de polaridad en el texto.

Estos pesos también pueden ser representados de otra forma para el analista. Generalidad y uniformidad expresan lo común que es un término, mientras que polarización es un concepto independiente de los dos anteriores. Denomínese $c(w)$ a lo común que es un término, es decir lo general y uniforme que es en el conjunto de los textos. La heurística permite el cálculo de la GAE de la siguiente forma:

$$GAE(w) = i * pol(w) + (1 - i) * c(w)$$

Donde $i \in [0, 1]$ es un peso configurable por el analista que expresa lo importante que es que el término sea representativo de polaridad frente a que sea común. Del mismo modo, dentro de lo común que es un término se le propone al analista un peso gi que representa lo importante que es que el sea general frente a que sea uniforme. Las dos expresiones de la GAE son equivalentes, pero se proponen las dos para mayor facilidad de configuración de la heurística por parte del analista.

El modelo presentado realiza un ranking en base a la GAE de todos los unigramas pertenecientes a todos los textos analizados en su primer paso. Antes de seguir

adelante, es necesario explicar cómo se calculan los tres pilares de la heurística GAE: Generalización $g(w)$, uniformidad $u(w)$ y polarización $pol(w)$.

En primer lugar, antes de poder calcular estos tres pilares, hay que calcular la probabilidad de encontrar cada uno de los términos considerados en cada una de las categorías. En su artículo, Tan únicamente considera una forma de calcular la probabilidad. En cambio, en la literatura se encuentran muchas formas de calcular la probabilidad de un término en un texto. Por ello, la primera ampliación del modelo de Tan es considerar más de un único método para calcular esta probabilidad. Depende del problema, podrá ser más o menos adecuado emplear cada uno de los métodos.

Esta probabilidad se puede calcular bien por máxima verosimilitud o mediante un modelo de Smoothing. El cálculo de la probabilidad de ocurrencia de un término en un corpus de documentos por máxima verosimilitud es calculado como el número de veces que se encuentra el término en un texto dividido entre el número de términos totales.

En su modelo, Tan propone que se calcule la probabilidad de encontrar un término en una categoría como el número de textos que contienen el término entre el número de términos totales. Se incluye esta medida de la probabilidad también en este estudio. La expresión para calcular la probabilidad del término w en la categoría c es la siguiente:

$$P_c(w) = \frac{N_w + \alpha}{D + 2\alpha}$$

Donde D es el número de textos pertenecientes a la categoría c, $P_c(w)$ es la probabilidad de encontrar el término w en la categoría c, N_w es el número de textos que contienen el término w y α es un parámetro con valor 0.00001 para evitar indeterminaciones en la expresión.

El modelo de Smoothing propuesto en esta tesis para calcular la probabilidad de ocurrencia de un término en una categoría es el Generalized Laplace Smoothing. Este modelo consiste en la siguiente expresión, cuyos miembros se explican a continuación:

$$\theta_i = \frac{x_i + \alpha}{N + \alpha d}$$

Donde x_i es el número de veces que aparece el término i en el texto. α es una variable entera mayor de 1. d es tamaño del vocabulario empleado en el texto y N el número

de términos totales del texto. A mayor valor toma α, mayor valor tomarán aquellas palabras que aparecen muy pocas veces en el texto o no aparecen.

El total de masa de probabilidad eliminado de las palabras que si aparecen en el texto es introducido en el denominador como αd. Existen más modelos de Smoothing, pero este es de los más utilizados. El modelo permitirá al analista elegir que método de cálculo de probabilidad desea ejecutar, o si desea ejecutar todos los métodos de probabilidad ponderados. Si así lo desea, la probabilidad final será calculada mediante la siguiente expresión.

Sea $P_f^c(w)$ la probabilidad final de encontrar el término w en la categoría c, W un vector de variables reales cuyo valor está contenido entre 0 y 1 y $p^c(w)$ un vector de variables que simbolizan los distintos métodos de calcular la probabilidad del término w tales que $\sum_{i=1}^{n} w_i = 1$ y n el número de métodos de cálculo de probabilidad; entonces la probabilidad final es calculada por la siguiente expresión:

$$P_f^c(w) = \sum_{i=1}^{n} w_i * p_i^c(w)$$

El modelo, en su etapa inicial, calculará el modelo del lenguaje para todos los términos de los léxicos polarizados semilla empleados. Construirá una matriz $P(w)$ cuyas variables $P_i^c(w_i)$ simbolizan la probabilidad P_i^c de encontrar el término w_i en cada categoría c. La GAE se servirá de las probabilidades, bien sean calculadas por máxima verosimilitud o un modelo de smoothing, de cada término para estimar que términos son las más representativos en un conjunto de categorías.

A continuación se detalla la expresión de la GAE para que se observe como esta medida utiliza estas probabilidades para estimar los términos más representativos de un conjunto de categorías. Se explica cómo se calcula la generalidad $g(w)$, uniformidad $u(w)$ y polarización $p(w)$.

La generalización de la GAE trata de otorgar una puntuación mayor a una mayor ocurrencia del N-Grama w en todas las categorías que se consideren para la construcción del prototipo de análisis de opiniones. Es una generalización del modelo de Tan, que únicamente considera dos dominios. Se propone el siguiente nuevo numerador para calcular lo general que es un N-Grama en cada una de las n categorías a considerar. Se incluye a su vez a la izquierda la expresión de Tan para ilustrar como la nueva expresión es una generalización de la anterior.

$$p_0(w)p_n(w) \xrightarrow{\textit{Expresión Generalizada a:}} (\prod_{i=1}^{n} p_i(w))^{1/n}$$

Donde i simboliza el dominio i de un total de n dominios. El exponente $1/n$ se introduce con el propósito de evitar el problema del underflow y que el resultado sea similar independientemente del número de dominios. El problema del underflow se origina al multiplicar numerosas probabilidades obteniendo resultados muy bajos. Con esta expresión, aquellos términos w con altos valores de probabilidad en las n categorías obtendrán una mayor puntación y viceversa. La GAE por tanto favorece con este factor a aquellos términos altamente probables en las n categorías.

Un requisito a cumplir por el modelo es que el analista debe poder parametrizar los pesos que quiere otorgar a cada categoría. Si todas las categorías tuviesen el mismo peso, la expresión anterior es válida. Dado que esto no tiene porque ser así, pese a que lo es por defecto, se modifica la expresión anterior introduciendo un vector θ de pesos. Este vector θ de pesos debe cumplir que $\sum_{i=1}^{n} \theta_i = 1$ donde n es el número de categorías. La expresión ,por tanto, queda modificada de la siguiente forma:

$$(\prod_{i=1}^{n} p_i(w))^{1/n} \xrightarrow{\textit{Expresión Convertida a:}} \prod_{i=1}^{n} \theta_i p_i(w)$$

El exponente $1/n$ queda modificado por el vector de pesos. Si el analista no especifica ninguna parametrización y decide que todas las categorías tienen el mismo peso, el valor de cada peso θ_i será igual a $\theta_i = 1/n$, donde n es el número de categorías. Dado que esta expresión simboliza lo general que es el término en las categorías, se la define como la generalización del término w o $g(w)$:

$$g(w) = \prod_{i=1}^{n} \theta_i p_i(w)$$

Calculada la generalización de un término, se explica la uniformidad de un término. Tan intenta que se favorezcan características que ocurren frecuentemente en dos dominios. Para ello usaba la expresión $|P_0(w) - P_n(w)|$. Con esta expresión, si un término aparece muy frecuentemente en una categoría y apenas ocurre en la otra o no se detecta, el valor de esta expresión será muy elevado. En la FCE, este término aparece en el denominador, en la GAE no. En consecuencia, a mas similares sean los

valores $P_0(w)$ y $P_n(w)$, menos se puntuará la importancia del término w, al contrario que en la FCE, donde se puntuaba más a mas similares eran esos valores.

Para generalizar esta expresión a n dominios, se emplea la desviación típica. La desviación típica es un estadístico cuyo valor es menor a menor dispersión con respecto al valor promedio presente la variable aleatoria en la muestra analizada. Por tanto, a menor sea la desviación típica, menor será el resultado final. En consecuencia, la GAE promociona a características generalizables en las categorías consideradas. La expresión que se introduce, con respecto a la propuesta por Tan, es la siguiente:

$$|P_0(w) - P_n(w)| \xrightarrow{\textit{Expresión Generalizada a}} \frac{\sum_{i=1}^{n}(p_i(w) - \overline{p(w)})^2}{n-1}$$

Si el número de categorías es 1, entonces se asigna a n un valor igual a 2. Como en el caso del numerador, se pretende que el analista pueda asignar una importancia mayor a una categoría frente a otra distinta. En el caso de que el analista asigne más peso a una categoría, se debe hacer que la media $\overline{p(w)}$ tienda a ese valor. Por ello, sea ω el vector de pesos parametrizado por el analista para cada categoría tal que $\omega_i \in [0,1]$ y $\sum_{i=1}^{n} \omega_i = 1$, entonces:

$$\overline{p(w)} \xrightarrow{\textit{Incluyendo Pesos}} \sum_{i=1}^{n} \omega_i p_i$$

A esta nueva expresión se le define como $\overline{p_\omega(w)}$. Donde n es el número de categorías. Ponderada la media en la desviación típica, hay que ponderar también la importancia de cada sumando, que simboliza a la probabilidad de cada categoría con respecto a su media. De no hacerlo, el estadístico no modelizaría la dispersión ponderada de cada una de las variables.

Cada sumando corresponde a la expresión: $(p_i(w) - \overline{p_\omega(w)})^2$. Para lograr la ponderación, simplemente se multiplica a cada sumando por su peso asignado. Dado que el valor sería tantas veces más pequeño como categorías existen, se le multiplica al sumatorio por el número de categorías para eliminar ese factor. Con ello, se obtiene expresión que se va a definir a continuación para la desviación típica ponderada. Dado que esta expresión modeliza la uniformidad del término en las categorías, se le define como $u(w)$:

$$u(w) = \frac{n \sum_{i=1}^{n} \omega_i (p_i(w) - \overline{p_\omega(w)})^2}{n-1}$$

Donde n es el número de categorías, si sólo se considera una categoría entonces al denominador se le asigna un valor igual a 1. Si el analista no especifica ninguna parametrización y decide que todas las categorías tienen el mismo peso, el valor de cada peso ω_i será igual a $\omega_i = 1/n$, donde n es el número de categorías.

Enfatizar que el vector de pesos de uniformidad ω y el vector de pesos de generalización θ no tienen que ser iguales. Se le proporciona de este modo al analista una mayor flexibilidad en el modelo. Con la definición de esta expresión ya se tienen dos de los tres pilares de los que se compone la GAE, a continuación se explica el cálculo de la polarización de un término.

La polarización de un término en la GAE expresa la correlación de la aparición de cada término en el texto con respecto a la polaridad del texto. Para poder calcular la polarización, es necesario calcular en primer lugar la proporción positiva y negativa de cada uno de los términos en el conjunto de textos a analizar.

La variable real proporción positiva $PP(w,T) \in [0,1]$ de un término w en un conjunto de textos T viene dada por la siguiente expresión, donde T_{POS} es una variable entera que simboliza el subconjunto con respecto a T de textos categorizados positivamente y $count(w, T_{POS})$ es una variable entera que simboliza el número de textos categorizados positivamente tal que en los cuáles aparece el término w.

$$PP(w,T) = \frac{count(w, T_{POS})}{T_{POS}}$$

La variable real proporción negativa $PN(w,T) \in [0,1]$ de un término w en un conjunto de textos T viene dada por la siguiente expresión, donde T_{NEG} es una variable entera que simboliza el subconjunto con respecto a T de textos categorizados negativamente y $count(w, T_{NEG})$ es una variable entera que simboliza el número de textos categorizados negativamente tal que en los cuáles aparece el término w.

$$PN(w,T) = \frac{count(w, T_{NEG})}{T_{NEG}}$$

Calculadas estas proporciones, se introduce el concepto positividad $PS \in [0,1]$ y negatividad $NG \in [0,1]$ del término w. Las proporciones positivas y negativas no nos

indican realmente como de polarizado está el término w con respecto al conjunto de textos T, indican únicamente una proporción. Por ejemplo, el término no está polarizado simplemente por tener una proporción positiva igual a 0.95, pues puede tener una negativa de 0.98. Por ello, se introduce la variable positividad $PS \in [0, 1]$ que obedece a la siguiente expresión, donde se omite el conjunto de textos T para simplificar, pero del que todas las expresiones son dependientes:

$$PS(w) = \frac{PP(w)}{PP(w) + PN(w)}$$

Si la expresión $PP(w) + PN(w)$ es igual a cero, es decir, el término es neutral dado que no aparece en ningún texto, entonces $PS(w)$ es igual a cero. Del mismo modo, se introduce la variable negatividad $NG \in [0, 1]$ que obedece a la siguiente expresión, donde se omite el conjunto de textos T para simplificar, pero del que todas las expresiones son dependientes:

$$NG(w) = \frac{PN(w)}{PP(w) + PN(w)}$$

Dada la positividad y negatividad de un término, se deben introducir dos nuevos conceptos que son los dos pilares de los que se compone la polarización de un término w. Estos conceptos son la Pureza $PZ(w) \in [0, 1]$ y la Intensidad Polar $IP(w) \in [0, 1]$ del término w. Son explicados a continuación.

La Pureza $PZ(w) \in [0, 1]$ del término w es una variable real que modela lo polar que es un término. Un término puede ser polar sin estar presente en muchos textos. Por ejemplo, si un término tiene una positividad igual a 0.2 y una negatividad igual a 0, el término será polar. Un término polar no implica que sea negativo o positivo. Conocer que un término es polar es de utilidad para la GAE, pues estos términos son de mayor interés para ser tomados en cuenta con respecto a los que no son polares. La pureza del término viene dada por la siguiente expresión:

$$PZ(w) = |PS(w) - NG(w)|$$

El valor absoluto de la resta entre positividad y negatividad obtiene la pureza del término w. Sin embargo, la polarización de un término puede ser tomada como distinta con purezas similares. Una pureza igual a 0.4 puede obedecer a que un término tenga positividad 0.4 y negatividad 0 o positividad 0.8 y negatividad 0.4.

En este segundo caso, el término aparecería en un número mucho mayor de textos. Esta aparición polar en más textos del término puede interesar al analista.

Por ello, se define el concepto Intensidad Polar $IP(w) \in [0, 1]$ como:

$$IP(w) = PP(w, T)$$

si $PP(w, T) \geq PN(w)$ o

$$IP(w) = PN(w)$$

en caso contrario. De esta forma si un término tiene mayores proporciones polares entonces será mayor puntuado.

Definidos los dos pilares de la polarización del término w, únicamente resta explicar cómo se combinan estos dos pilares, para la creación de la expresión que simboliza la polarización del término w. Estos dos términos son combinados mediante un peso, $k \in [0, 1]$, para que el analista otorgue a la intensidad y pureza la importancia que considere necesaria. La polarización $pol(w) \in [0, 1]$ de un término se define por la siguiente expresión:

$$pol(w) = k * PZ(w) + (1 - k) * IP(w)$$

Mediante la polarización del término, el analista añade a la GAE no únicamente el hecho de que los rasgos, o unigramas en este modelo, sean generales y uniformes, sino que añade un mecanismo específico para comprobar como de polares son. Se trata de un enfoque que se especializa más para la tarea de análisis de opiniones y se aleja algo de la categorización de textos, problema más genérico.

Una vez explicados los tres pilares de la GAE, únicamente resta explicar cómo se combinan estos pilares para el cálculo de la GAE. Sea $g(w)$ la generalización del término w en las n categorías, $u(w)$ la uniformidad del término w en las n categorías y $pol(w)$ la polarización del término. La expresión final de la GAE, $GAE(w)$, queda definida como:

$$GAE(w) = q_1 g(w) + q_2 u(w) + q_3 pol(w)$$

Los tres pilares explicados son levemente modificados para que los tres rankings creados por $g(w)$, $u(w)$ y $pol(w)$ obtengan resultados uniformes. De este modo, ninguno de los pilares domina a los otros. Por ejemplo, si un ranking tiene el término más alto puntuado con 0.5 y otro lo tiene con 0.99, entonces el segundo domina al primero. Se desea evitar este sesgo.

Por otro lado, también son levemente modificados dado que la diferenciación en importancia entre términos con valores muy bajos con respecto a valores medios es superior que la diferenciación en importancia entre términos con valores medios con respecto a altos. Esto es debido a que, como más tarde se explica, los términos de puntuaciones más altas serán directamente elegidos por la expresión pero aquellos filtros de puntuaciones más bajas tendrán que competir. Por ello se quiere diferenciar de la mejor forma posible su importancia.

Para evitar el sesgo y elevar la importancia de la diferencia entre valores bajos, se utilizan raíces de grados elevados. En cada uno de los pilares se utiliza un grado. Este grado ha sido extraído en base a numerosos experimentos, hasta que se han observado que los rankings son uniformes y la puntuación de los términos es elevada en cada uno de ellos. En la sección de líneas futuras, se propone un modelo para optimizar estos grados de tal modo que la GAE devuelva resultados óptimos. Se emplea la raíz cuadrada para la generalización, la raíz quinta para la uniformidad y la polarización es elevada a 0.35.

Mediante la GAE, se extraen términos aplicables a cualquier dominio independientemente del mismo. Estos términos, son denominados como independientes del dominio en la investigación de Tan por medio de su FCE, pero únicamente considerando dos dominios. Como se ha visto, la GAE puede considerar todo tipo de dominios, especializándose en análisis de opiniones.

Una vez formulada la GAE, se detalla cómo utilizar esta expresión para construir un prototipo de análisis de opiniones. En esta tesis, se usará la GAE para filtrar, en primer lugar, léxicos polarizados semilla, eliminando aquellos términos no generalizables o no aplicables en los dominios considerados y posteriormente expandir los léxicos polarizados con nuevos términos.

Para el filtrado, una iteración de la expresión se quedará solo con los k términos cuya f_w sea superior a un umbral u configurable por el analista. Del mismo modo el analista puede elegir también emplear todos los términos o quedarse sólo con los k términos que él considere.

Para estos k términos se extrae una variable booleana que simboliza su apariencia en cada uno de los textos considerados para las n categorías. Para cada texto se extrae una variable nominal que indica la categoría de ese texto. Al mismo tiempo para cada texto se tiene como variable nominal que simboliza la clase la polaridad del texto: Positiva o negativa. Estas variables constituyen cada una de las instancias representativas de cada texto que forman el conjunto de datos a procesar.

A partir de este vector característico, se efectuará una clasificación de los textos de los N dominios usando modelos de aprendizaje automático supervisado. Los clasificadores elegidos para esta investigación serán el Discriminative Multinomial Naive Bayes, la Regresión Logística y la linear SVM. Como se ha visto en el estado del arte, son los modelos más empleados en el análisis de opiniones, aunque no los únicos, puesto que se puede realizar análisis de opiniones también con árboles de decisión, random forests o con el algoritmo K-NN, entre otros.

De nuevo, el analista tendrá un vector de pesos ϕ con la misma naturaleza que los anteriores vectores de pesos para calibrar la importancia que quiere dar al resultado de cada clasificador. Mediante este vector de pesos el modelo es más flexible.

De forma adicional, el modelo permitirá al analista configurar si, para cada término, desea incluir variables que simbolicen la frecuencia de aparición del término en el texto y la puntuación del término en los textos empleando Smoothing Add one. También podrá añadir una variable que simboliza si en el texto se han encontrado más términos categorizados positivamente o negativamente por los léxicos polarizados semilla.

Mediante el léxico polarizado semilla filtrado para analizar estas categorías, se obtendrán medidas de evaluación del modelo tras la fase de entrenamiento: La precisión, recall y F-Measure medias. Este será el punto de partida de un algoritmo iterativo, que pretende maximizar el valor de esta expresión mediante la GAE. La expresión a maximizar, fit, donde τ es de nuevo un vector de pesos con la importancia que el analista da a la precisión, recall y f-measure, será la siguiente:

$$fit = max(\tau_1 Precision(prot) + \tau_2 Recall(prot) + \tau_3 F - Measure(prot))$$

Donde *prot* es el modelo obtenido tras ser entrenado por los modelos de aprendizaje automático. Se entra en un algoritmo iterativo en el que se quiere maximizar la expresión definida, que se le denominará fitness. El fitness obtenido en la primera iteración es el rendimiento inicial del prototipo, resultado de los términos relevantes del léxico semilla que se desee emplear. La precisión, recall y F-Measure obtenidas serán las de base, cuyo valor deberá incrementar según se añadan nuevos términos mediante la GAE en una serie de iteraciones.

El objetivo es maximizar la función fitness. Para ello, se expande el léxico polarizado mediante la GAE. Para lograr este objetivo, en primer lugar, se necesita un nuevo recurso. Este recurso es una lista de Stop Words del lenguaje en el que se están analizando los textos. Se añade esta lista puesto que estas stop words simbolizan

términos que en ningún caso, o en casos muy raros, polarizan un texto o son variables condicionadas para que otra variable defina la polaridad de un texto.

Con este recurso, se puede comenzar la expansión del léxico polarizado con nuevos términos que se encuentren en el texto. Se calcula la GAE para todos los términos que aparezcan en los textos excepto para los considerados en el léxico polarizado y los que se encuentren en la lista de Stop Words. No se conoce la polaridad de los nuevos términos, simplemente se conoce que su aparición es general en todas las categorías y que se encuentran uniformemente en todas ellas.

Por ello, pese a que los términos por sí solos, probablemente en la mayoría de los casos pero no siempre, no tengan semántica polar si se sabe que son generales y uniformes. Una vez modelizados todos los términos, se hace un ranking con la puntuación que obtienen mediante la GAE. En cada iteración, se introducen los términos más valorados por la GAE al léxico polarizado.

Se podrán introducir uno a uno o en bloque según parámetros de configuración habilitan en el modelo. El analista configurará cuantos términos se querrán añadir en cada iteración al léxico polarizado de la iteración anterior. Esto se define mediante la variable entera E, de valor mayor o igual a 1. A menor E, mayor precisión tendrá el prototipo final pero el prototipo necesitará un mayor tiempo de computación y viceversa.

Se extraen las variables de aparición en el texto, frecuencia o las consideradas por el analista para cada uno de los términos extraídos por la GAE para cada texto. El procedimiento es similar al que se elabora en la primera iteración, para los términos que se incluyen en el léxico polarizado.

Se añaden estas variables al vector característico representante de cada texto. Las variables que simbolizarán a estos términos en el vector característico condicionadas al valor de las variables del léxico polarizado podrán aportar un mejor comportamiento a los modelos de aprendizaje automático.

Si la función fitness ha mejorado con los nuevos términos, se incorporan al léxico polarizado, sin polaridad, pero representativos del vector característico de cada instancia representante de un texto.

Si los nuevos términos decrementan las medidas de evaluación obtenidas en la iteración anterior, no se incluirán los nuevos términos. En la siguiente iteración, se desechan los E términos más valorados por la GAE y se extraen los siguientes E términos. Para estos términos, se vuelven a extraer sus variables para cada texto

y se vuelve a evaluar el conjunto de datos mediante los modelos de aprendizaje automático, y así sucesivamente.

En este punto, es necesario definir un criterio de convergencia o de parada, puesto que de lo contrario el algoritmo únicamente terminaría cuando se analicen todos los términos. Esto es muy costoso. Por ello, se definen criterios de parada para el algoritmo iterativo de maximización.

El criterio de convergencia será también configurable por el analista. Se proponen varios criterios en este modelo:

- Obtener una variación v, en porcentaje, menor de las medidas de evaluación de una cantidad α a parametrizar por el analista.
- Permitir un máximo de i iteraciones.
- No haber incluido γ términos en x iteraciones.
- Mejorar un p porcentaje con respecto al rendimiento inicial.

Una vez alcanzado el criterio de convergencia, se detiene el algoritmo alcanzando las medidas de evaluación logradas por la última iteración. Con esto, el prototipo de análisis de opiniones construido para analizar las categorías consideradas por el analista es construido. A continuación se incluye un gráfico en el que se visualiza toda la lógica explicada en estos párrafos.

El prototipo obtenido por el modelo es general para las categorías consideradas por el analista. Debido a la flexibilidad del modelo, si el número de categorías es superior a dos y se dispone de tiempo y recursos para obtener prototipos, se recomienda seguir la siguiente metodología para obtener prototipos de análisis de opiniones en todos los subconjuntos las categorías que se consideren. Mediante la siguiente metodología, se obtendrán prototipos de más generales a más específicos para todas las categorías a analizar.

Supóngase que se deben analizar N categorías. Se recomienda emplear el modelo en las N categorías para obtener un prototipo general para todas ellas. Posteriormente, se recomienda obtener todos los subconjuntos de categorías de las N categorías y generar prototipos para todas ellas. Se recomienda también generar un prototipo para cada una de las categorías por separado.

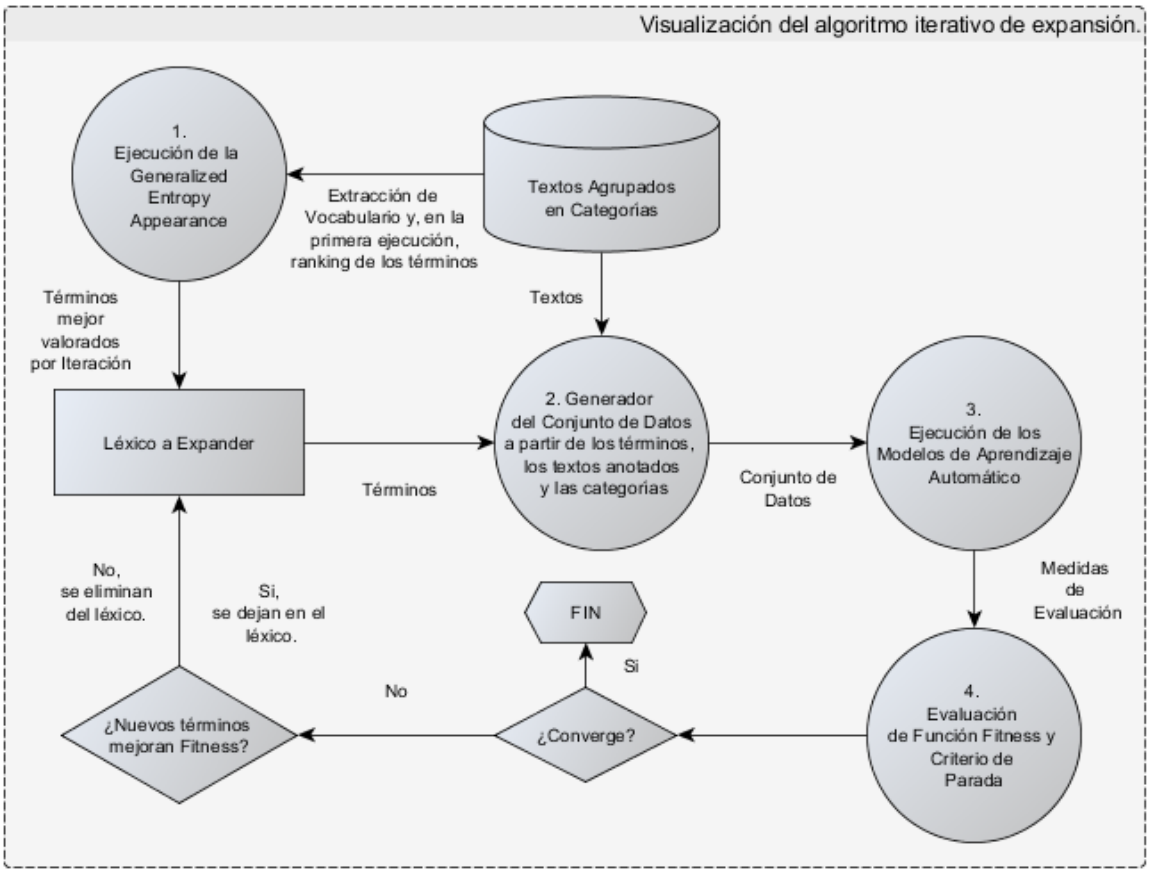

FIGURA 3.2: Algoritmo de expansión del léxico y construcción de prototipos de modelos de análisis de opiniones.

Por ejemplo, si se tiene las categorías de textos CA,CB Y CC se recomienda generar un prototipo P que las considere todas: $P(CA, CB, CC)$. Una vez obtenido, se recomienda generar los prototipos de todos los subconjuntos de las categorías: $P(CA, CB)$,$P(CA, CC)$,$P(CB, CC)$,$P(CA)$,$P(CB)$ y $P(CC)$.

De este modo, si el analista de opiniones necesita obtener un modelo para una situación en la que sólo se encuentren textos de la categoría CC ya tiene el prototipo. Si por el contrario se encuentra en la casuística de que en un dominio sólo existen textos pertenecientes a las categorías CA y CB ya tiene el prototipo. En resumen, tiene prototipo para cualquier casuística que se pueda encontrar.

Además de generar el prototipo, el modelo presentado ofrece la expansión del léxico polarizado semilla presente en el algoritmo. Los términos que la GAE va extrayendo

para mejorar el análisis de un texto por parte de los algoritmos de aprendizaje automático son añadidos en este léxico polarizado.

Para, además de añadirlos, descubrir su polaridad, se emplea la pureza y la positividad y negatividad, descritas anteriormente. El analista dispone de un umbral de polarización, $UP \in [0, 1]$. Este umbral decidirá si se desconoce la polaridad del término que se inserta o se polariza.

Si la pureza es inferior al umbral, se añade el término al léxico polarizado sin polaridad. Si es superior al umbral, se polariza. Si la positividad es superior a la negatividad, se categoriza positivamente. En caso contrario, se categoriza negativamente.

De esta forma, el modelo presentado en este apartado obtiene, a partir de un conjunto de textos ordenados por categorías y un léxico polarizado semilla, un modelo para analizar futuras opiniones de estas categorías y un léxico polarizado expandido especializado en las categorías analizadas.

Una observación importante es que en este modelo se prefiere no lemmatizar ni eliminar algunos caracteres que no son propios del unigrama con el tokenizador. Por ejemplo: Great!!!!. Esto hace que el término Great!!!! sea tratado de forma distinta a Great!. Esto se hace así dado que el modelo soporta grandes volúmenes de datos y dado que la semántica subyacente a los dos términos mostrados puede ser distinta.

Una vez definido el modelo, el siguiente apartado definirá los experimentos que se van a efectuar para observar los resultados obtenidos por este modelo. Esto se detalla en el siguiente apartado, el apartado de experimentación.

Capítulo 4

Experimentación

Como se ha visto en el capítulo anterior, en el apartado en el cuál se propone el modelo, este modelo se presta a la realización de un gran número de experimentos. Esto es debido a las características del modelo, que contiene parámetros ajustables a las necesidades de cada circunstancia y es capaz de analizar un número indeterminado de categorías.

Para poder evaluar las hipótesis listadas en el anterior apartado y cumplir con los objetivos propuestos es necesario elaborar múltiples experimentos. En este capítulo, se proponen experimentos para evaluar las hipótesis y cubrir los objetivos propuestos. Estos experimentos están sujetos a las asunciones y las restricciones listadas en el capítulo anterior. A estas restricciones se les añaden otras aquí listadas con respecto al modelo propuesto debido a la limitación de tiempo existente para la elaboración de la Tesis.

Este capítulo se divide en dos apartados. En primer lugar, se expondrán los recursos de los que se dispone en esta investigación para elaborar los experimentos, adjuntando para cada uno de ellos una explicación sobre el porqué se han elegido.

Posteriormente, se incluye un apartado en el cuál se muestra el diseño de cada uno de los diferentes experimentos. Para cada experimento, se añaden los objetivos que se quieren cumplir mediante su realización, las hipótesis que se desean evaluar y el listado de asunciones y restricciones que se toman en cuenta en el experimento. Los resultados de cada experimento, sobre las hipótesis y objetivos planteados para cada uno de ellos, se comentan en el capítulo de Evaluación.

4.1. Recursos Utilizados en los Experimentos

El modelo descrito en el anterior capítulo necesita de múltiples recursos para poder ser implementado, realizar experimentos con el mismo y ser posteriormente evaluado. Algunos de los recursos que, de forma ideal, necesita este modelo para ser evaluado en toda posible casuística, son muy costosos, en parámetros de tiempo y económicos. Por ello, se incluye en el capítulo anterior como restricciones que sólo se utilizará para la experimentación y evaluación del modelo los recursos que se mencionan en este apartado.

A continuación, se exponen los distintos tipos de recursos de los que se dispone para efectuar experimentos:

4.1.1. Corpus de textos clasificados por categoría y anotados por polaridad

El modelo se encarga de construir prototipos que analizan opiniones en cualquier número de categorías distintas. Por ello, en primer lugar, se necesita disponer de distintos textos clasificados por categorías. De forma adicional, estos textos necesitan estar clasificados en función de la polaridad del escritor. Es importante recordar que el modelo emplea la polaridad como variable clase para los modelos de aprendizaje automático.

Como conjunto de textos clasificados por categoría y polarizados, se cuenta con un corpus categorizado por opinión de 5 GB de críticas sobre distintos productos descargado de Amazon [79]. Las críticas de cada tipo distinto de producto simbolizan las distintas categorías que se consideran. Algunos ejemplos de estas categorías son críticas de usuarios efectuadas sobre música o vivienda.

Como categoría adicional, se cuenta con un corpus anotado por polaridad de Tweets de Stanford que simboliza otra categoría. Esta categoría es especialmente diferente a las anteriores dado que la gente escribe distinto en Twitter. Esto será útil para diseñar experimentos. Se cuenta, por último, de un corpus de texto anotado de críticas de películas, que simboliza una nueva categoría con respecto a las comentadas previamente.

Debido a la restricción de tiempo y a la gran utilidad del corpus de Amazon, se utiliza este para los experimentos.

Como comentario sobre los textos de los que se dispone, el volumen de textos no es todo lo grande de lo que a priori es deseable. Este es un modelo cuyo funcionamiento, gracias al aprendizaje automático y la estadística, es óptimo en entornos en los que se deben analizar volúmenes de texto del orden de Terabytes. Pese a ello, el volumen de textos del que se dispone es aceptable para realizar un prototipo.

Adicionalmente, todos los textos están anotados por polaridad y clasificados por categoría y han sido empleados en otros experimentos dado que tienen licencia para uso en la investigación. Esto supone una garantía de calidad, ya que han sido revisados por otros investigadores.

4.1.2. Léxicos Polarizados Semilla

Para su correcto funcionamiento, el modelo necesita ser inicializado mediante un léxico polarizado semilla. Por ello, adquiere especial importancia la elección de un léxico polarizado semilla que se adapte a este modelo. Este debe contener términos y su polaridad asociada.

En el estado del arte, se elaboró una lista con los léxicos polarizados semilla más representativos en el análisis de opiniones. Los léxicos polarizados que se han elegido para la experimentación son algunos de los que se describen en el estado del arte. Para mayor información sobre ellos se referencia al lector a consultar el apartado del estado del arte en el cual se describen.

Debido a que contienen licencia para su uso en investigación y cuentan con polaridades asociadas a cada uno de los términos que contienen, los léxicos polarizados semilla que se van a considerar en los experimentos son el General Inquirer y el Bing Liu Lexicon.

En los experimentos, se comprobará si el análisis de opiniones realizado, una vez converge el algoritmo, varía con respecto a las medidas de evaluación obtenidas, si se inicializa el modelo mediante distintos léxicos polarizados semilla.

A priori, a menor adaptación del léxico polarizado semilla a las categorías que se consideren, el algoritmo necesitará un mayor número de iteraciones para converger y, en consecuencia, obtener las medidas de evaluación finales. Se verificará si el resultado final es independiente de un distinto número de iteraciones, entre otros factores.

4.1.3. Modelos de Aprendizaje Automático

Para comprobar si los términos añadidos al léxico polarizado en cada iteración, determinados a partir de que sean los mejor valorados por la GAE, mejoran el análisis de opiniones realizado por el léxico sin estos términos, se emplearán modelos de aprendizaje automático.

Estos modelos de aprendizaje automático se entrenarán a partir del conjunto de datos formado por un vector característico por texto. La variable clase es la polaridad del texto. Las variables predictivas son las representativas de cada término del léxico en función del texto. Por tanto, los modelos de aprendizaje automático a emplear en la expansión del léxico son muy importantes para lograr medidas de evaluación que superen las del léxico polarizado semilla y se adapten a las categorías consideradas.

Para los experimentos, se utilizarán el siguiente modelo de aprendizaje automático: El Discriminative Multinomial Naive Bayes [80]. La razón por la cual se emplea este método es debido a que es el más rápido computacionalmente y es muy efectivo en categorización de texto, lo que lo hace ideal para la construcción de un prototipo.

Para los experimentos, se utilizará la implementación de este clasificador proporcionada por la plataforma Weka [81]. Esta plataforma, que tiene licencia de código abierto, cuenta con la ventaja de ser utilizada y modificada por un gran número de investigadores. Por ello, se garantiza su buen funcionamiento, dado que cualquier posible mal función en los clasificadores utilizados habrá sido corregida por la comunidad. Esto es debido a que los clasificadores citados son empleados en un gran número de experimentos de distinta naturaleza.

4.2. Diseño y Descripción de los Experimentos

Tras describir los recursos que se van a utilizar los experimentos, se necesita exponer los detalles de los distintos experimentos que se van a contemplar. Todos los experimentos se efectúan sobre una implementación escrita en Java SE 7 del modelo que se ha propuesto.

A continuación, se muestra el detalle de cada experimento. Los experimentos están ordenados por importancia. Para cada experimento, se adjunta información sobre los recursos que se emplean en él, las hipótesis que se desean evaluar y los objetivos que se desean alcanzar.

Todos los experimentos descritos a continuación están sujetos a las restricciones y asunciones descritas en el capítulo anterior y en este capítulo. Los resultados de cada experimento, junto a si consiguen validar o refutar las hipótesis que se plantean en cada uno de ellos, se incluyen en el capítulo de evaluación.

4.2.1. Experimento I

Descripción: Este experimento desea corroborar si el modelo es capaz de expandir el léxico polarizado. Se quiere comprobar si el modelo que toma el léxico expandido mejora el comportamiento del modelo que toma el léxico polarizado semilla. Se experimenta en una casuística en la que se analizan textos pertenecientes a dos categorías distintas. El experimento tendrá éxito si las medidas de evaluación, función fitness del modelo, del prototipo generado, son mejores que las del léxico polarizado semilla.

Parámetros del modelo en el experimento:

- Importancia de la generalización, q_1, en la GAE: 0.2.
- Importancia de la uniformidad, q_2, en la GAE: 0.2.
- Importancia de la polarización, q_3, en la GAE: 0.6.
- Importancia de la Pureza , k, con respecto a la Intensidad Polar en la Polarización: 0.8.
- Umbral de polarización, UP: 0.3.
- Número de términos, E, a insertar en el léxico por iteración: 5.
- Umbral de términos a preservar del léxico polarizado: 100
- Vector de pesos de generalización de las categorías, θ: Uniforme.
- Vector de pesos de generalización de las categorías, ω: Uniforme.
- Vector de pesos para la función fitness ,τ: Precisión: 0.3. Recall: 0.2. F-Measure: 0.5.
- Criterio de convergencia: Permitir un máximo de iteraciones.
- Número de iteraciones ,i, para la convergencia: 30.

- Número de categorías consideradas, N: 2.
- Categorías consideradas: Automotive, Beauty.

Recursos empleados: Bing Liu Lexicon. Corpus de Amazon. Discriminative Multinomial Naive Bayes.

Hipótesis que se quieren evaluar: El modelo propuesto es capaz de mejorar las medidas de precisión, recuperación y F-Measure con respecto a las medidas obtenidas por medio de un modelo de análisis de opiniones basado únicamente en un léxico polarizado semilla en cualquier categoría.

El modelo propuesto es capaz de mejorar, en al menos la mitad de los casos y habiendo analizado un número lo suficientemente grande de textos, las medidas de precisión, recuperación y F-Measure con respecto a las medidas obtenidas por medio de un modelo de análisis de opiniones generalista en textos que pertenezcan a cualquier categoría.

Objetivos que se quieren alcanzar: Comprobar si el modelo propuesto hace que los léxicos polarizados mejoren sus medidas de evaluación con respecto a las obtenidas cuando estos léxicos no están expandidos.

Comprobar si el modelo obtiene prototipos de modelos que analizan opiniones en contextos o dominios que no han sido analizados por lingüistas o expertos en esa categoría.

4.2.2. Experimento II. Descripción

Descripción: Este experimento desea corroborar si el modelo obtiene resultados similares independientemente del léxico polarizado semilla empleado. Para ello, se repiten los parámetros del anterior experimento a excepción de las categorías para seguir probando nuevas combinaciones, pero esta vez empleando el resto de léxicos polarizados. También a excepción del número de iteraciones, para demostrar que aún empleando un menor número de iteraciones, se obtiene una mejoría.

A priori, los resultados de crecimiento de la función objetivo deberían ser similares, excepto el número de iteraciones empleado mediante el cual se llega a medidas estables. Se prueba que el modelo es independiente de la categoría a analizar y de los recursos previos. Se piensa a priori que con 15 iteraciones añadiendo 5 términos por iteración debería ser suficiente.

Parámetros del modelo en el experimento:

- Importancia de la generalización, q_1, en la GAE: 0.2.
- Importancia de la uniformidad, q_2, en la GAE: 0.2.
- Importancia de la polarización, q_3, en la GAE: 0.6.
- Importancia de la Pureza , k, con respecto a la Intensidad Polar en la Polarización: 0.8.
- Umbral de polarización, UP: 0.3.
- Número de términos, E, a insertar en el léxico por iteración: 5.
- Umbral de términos a preservar del léxico polarizado: 100
- Vector de pesos de generalización de las categorías, θ: Uniforme.
- Vector de pesos de generalización de las categorías, ω: Uniforme.
- Vector de pesos para la función fitness ,τ: Precisión: 0.3. Recall: 0.2. F-Measure: 0.5.
- Criterio de convergencia: Permitir un máximo de iteraciones.
- Número de iteraciones ,i, para la convergencia: 10.
- Número de categorías consideradas, N: 2.
- Categorías consideradas: Toys and games, tools and hardware.

Recursos empleados: General Inquirer. Corpus de Amazon. Discriminative Multinomial Naive Bayes.

Hipótesis que se quieren evaluar: El modelo debe obtener resultados similares independientemente del léxico polarizado semilla empleado con un número lo suficientemente grande de iteraciones y textos analizados para cualquier conjunto de categorías.

Objetivos que se quieren alcanzar: Comprobar que el prototipo generado posee independencia con respecto al léxico polarizado semilla empleado.

4.2.3. Experimento III. Descripción

Descripción: En este experimento se desea comprobar que el prototipo generado mediante el propuesto generado es capaz de analizar opiniones de forma más efectiva en categorías que se han considerado frente a categorías que no se han considerado. Para ello, se generará un prototipo considerando 2 categorías. Se anotará el rendimiento que el prototipo posee analizando dichas categorías. Por último, se anotará el rendimiento del prototipo en una nueva categoría. El experimento tendrá éxito si esta última cifra es inferior al rendimiento presentado en las categorías consideradas.

Parámetros del modelo en el experimento:

- Importancia de la generalización, q_1, en la GAE: 0.2.
- Importancia de la uniformidad, q_2, en la GAE: 0.2.
- Importancia de la polarización, q_3, en la GAE: 0.6.
- Importancia de la Pureza , k, con respecto a la Intensidad Polar en la Polarización: 0.8.
- Umbral de polarización, UP: 0.3.
- Número de términos, E, a insertar en el léxico por iteración: 5.
- Umbral de términos a preservar del léxico polarizado: 100
- Vector de pesos de generalización de las categorías, θ: Uniforme.
- Vector de pesos de generalización de las categorías, ω: Uniforme.
- Vector de pesos para la función fitness ,τ: Precisión: 0.3. Recall: 0.2. F-Measure: 0.5.
- Criterio de convergencia: Permitir un máximo de iteraciones.
- Número de iteraciones ,i, para la convergencia: 30.
- Número de categorías consideradas, N: 2 más una de test.
- Categorías consideradas: Books, Tools and Hardware, Beauty.

Recursos empleados: Bing Liu Lexicon. Corpus de Amazon. Discriminative Multinomial Naive Bayes.

Hipótesis que se quieren evaluar: El prototipo creado por el modelo de análisis de opiniones presentado en esta Tesis, analizando un volumen lo suficientemente grande de textos, presentará mejores resultados en las categorías que se analicen en su fase de entrenamiento que en categorías no consideradas.

Objetivos que se quieren alcanzar: Comprobar que el modelo obtiene mejores resultados en categorías consideradas que en categorías sin considerar.

Comprobar que el modelo obtiene mejores resultados en las categorías que ha considerado que un modelo diseñado para otra categoría adicional.

4.2.4. Experimento IV. Descripción

Descripción: En este experimento se desea comprobar si el modelo propuesto es capaz de mejorar el rendimiento del análisis de opiniones que efectúa el prototipo que genera en cualquier categoría. Para ello, e emplean un total de 5 categorías.

Parámetros del modelo en el experimento:

- Importancia de la generalización, q_1, en la GAE: 0.2.
- Importancia de la uniformidad, q_2, en la GAE: 0.2.
- Importancia de la polarización, q_3, en la GAE: 0.6.
- Importancia de la Pureza , k, con respecto a la Intensidad Polar en la Polarización: 0.8.
- Umbral de polarización, UP: 0.3.
- Número de términos, E, a insertar en el léxico por iteración: 5.
- Umbral de términos a preservar del léxico polarizado: 100
- Vector de pesos de generalización de las categorías, θ: Uniforme.
- Vector de pesos de generalización de las categorías, ω: Uniforme.
- Vector de pesos para la función fitness ,τ: Precisión: 0.3. Recall: 0.2. F-Measure: 0.5.

- Criterio de convergencia: Permitir un máximo de iteraciones.
- Número de iteraciones ,i, para la convergencia: 50.
- Número de categorías consideradas, N: 5.
- Categorías consideradas: Camera and Photo, Electronics, Music, Magazines, Jewelry and Watches

Recursos empleados: Bing Liu Lexicon. Corpus de Amazon. Discriminative Multinomial Naive Bayes.

Hipótesis que se quieren evaluar: El modelo propuesto es capaz de construir un vector de características para cualquier número de categorías. Desde una categoría al número de categorías que se deseen analizar.

El modelo propuesto es capaz de expandir los léxicos polarizados semilla en cualquier contexto de entre los que se consideran en los experimentos, mostrando una mejoría en las medidas de precisión, recuperación y F-Measure con respecto a las medidas de los léxicos polarizados semilla.

Objetivos que se quieren alcanzar: Comprobar el correcto funcionamiento del modelo independientemente del número de categorías considerado.

Comprobar el correcto funcionamiento del dominio independientemente de la naturaleza de cada categoría.

4.2.5. Experimento V. Descripción

Descripción: Cubiertos todos los objetivos planteados en el documento de diseño, se desean efectuar pruebas adicionales del modelo para comparar su comportamiento en función a los parámetros introducidos. Estos experimentos, de menor importancia, resultan interesantes desde el punto de vista computacional, para observar el rendimiento del modelo en distintas circunstancias.

En este experimento, se desea comparar el rendimiento del modelo variando su criterio de convergencia. Se efectuará una primera ejecución del modelo añadiendo un único término por iteración y con 50 iteraciones. Se efectuará una segunda ejecución del modelo añadiendo 5 términos por iteración y con 10 iteraciones. Se espera un mejor rendimiento en el primer caso frente al segundo. El primer caso tiene la penalización del factor tiempo frente al segundo.

Parámetros del modelo en el experimento. Ejecución 1:

- Importancia de la generalización, q_1, en la GAE: 0.2.
- Importancia de la uniformidad, q_2, en la GAE: 0.2.
- Importancia de la polarización, q_3, en la GAE: 0.6.
- Importancia de la Pureza , k, con respecto a la Intensidad Polar en la Polarización: 0.8.
- Umbral de polarización, UP: 0.3.
- Número de términos, E, a insertar en el léxico por iteración: 1.
- Umbral de términos a preservar del léxico polarizado: 100
- Vector de pesos de generalización de las categorías, θ: Uniforme.
- Vector de pesos de generalización de las categorías, ω: Uniforme.
- Vector de pesos para la función fitness ,τ: Precisión: 0.3. Recall: 0.2. F-Measure: 0.5.
- Criterio de convergencia: Permitir un máximo de iteraciones.
- Número de iteraciones ,i, para la convergencia: 50.
- Número de categorías consideradas, N: 2.
- Categorías consideradas: Baby, Video.

Parámetros del modelo en el experimento. Ejecución 2:

- Importancia de la generalización, q_1, en la GAE: 0.2.
- Importancia de la uniformidad, q_2, en la GAE: 0.2.
- Importancia de la polarización, q_3, en la GAE: 0.6.
- Importancia de la Pureza , k, con respecto a la Intensidad Polar en la Polarización: 0.8.
- Umbral de polarización, UP: 0.3.

- Número de términos, E, a insertar en el léxico por iteración: 5.
- Umbral de términos a preservar del léxico polarizado: 100
- Vector de pesos de generalización de las categorías, θ: Uniforme.
- Vector de pesos de generalización de las categorías, ω: Uniforme.
- Vector de pesos para la función fitness ,τ: Precisión: 0.3. Recall: 0.2. F-Measure: 0.5.
- Criterio de convergencia: Permitir un máximo de iteraciones.
- Número de iteraciones ,i, para la convergencia: 10.
- Número de categorías consideradas, N: 2.
- Categorías consideradas: Baby, Video.

Recursos empleados: Bing Liu Lexicon. Corpus de Amazon. Discriminative Multinomial Naive Bayes.

4.2.6. Experimento VI. Descripción

Descripción: En este experimento, se desea comprobar que factor es más relevante, si la generalización o la uniformidad, si se cuenta con un número bajo de iteraciones. Se proponen dos ejecuciones del modelo, una con mayor factor de generalización y otra con mayor factor de uniformidad.

Parámetros del modelo en el experimento. Ejecución 1:

- Importancia de la generalización, q_1, en la GAE: 0.3.
- Importancia de la uniformidad, q_2, en la GAE: 0.1.
- Importancia de la polarización, q_3, en la GAE: 0.6.
- Importancia de la Pureza , k, con respecto a la Intensidad Polar en la Polarización: 0.8.
- Umbral de polarización, UP: 0.3.

- Número de términos, E, a insertar en el léxico por iteración: 5.
- Umbral de términos a preservar del léxico polarizado: 100
- Vector de pesos de generalización de las categorías, θ: Uniforme.
- Vector de pesos de generalización de las categorías, ω: Uniforme.
- Vector de pesos para la función fitness ,τ: Precisión: 0.3. Recall: 0.2. F-Measure: 0.5.
- Criterio de convergencia: Permitir un máximo de iteraciones.
- Número de iteraciones ,i, para la convergencia: 10.
- Número de categorías consideradas, N: 2.
- Categorías consideradas: Toys and games, tools and hardware.

Parámetros del modelo en el experimento. Ejecución 2:

- Importancia de la generalización, q_1, en la GAE: 0.1.
- Importancia de la uniformidad, q_2, en la GAE: 0.3.
- Importancia de la polarización, q_3, en la GAE: 0.6.
- Importancia de la Pureza , k, con respecto a la Intensidad Polar en la Polarización: 0.8.
- Umbral de polarización, UP: 0.3.
- Número de términos, E, a insertar en el léxico por iteración: 5.
- Umbral de términos a preservar del léxico polarizado: 100
- Vector de pesos de generalización de las categorías, θ: Uniforme.
- Vector de pesos de generalización de las categorías, ω: Uniforme.
- Vector de pesos para la función fitness ,τ: Precisión: 0.3. Recall: 0.2. F-Measure: 0.5.
- Criterio de convergencia: Permitir un máximo de iteraciones.

- Número de iteraciones ,i, para la convergencia: 10.
- Número de categorías consideradas, N: 2.
- Categorías consideradas: Toys and games, tools and hardware.

Recursos empleados: Bing Liu Lexicon. Corpus de Amazon. Discriminative Multinomial Naive Bayes.

El resultado de los experimentos planteados se exponen en el siguiente apartado de esta Tesis: El capítulo de Evaluación. Tras este capítulo se expondrán las conclusiones extraídas con respecto al modelo propuesto y el trabajo futuro a realizar en esta investigación.

Capítulo 5

Evaluación

5.1. Introducción

En esta sección del documento se muestran los resultados planteados en la sección anterior, la de experimentación. Cada experimento evalúa si se han alcanzado los objetivos y se han verificado las hipótesis asignadas a cada experimento. Cada experimento es descrito mediante su descripción, si se quiere conocer un detalle mayor de los parámetros del modelo en ese experimento, se tiene en la sección anterior, la de experimentación.

En la sección de Modelo se encuentra un listado completo de estos objetivos y asunciones. Para cada experimento se muestra un gráfico que muestra la evolución de la función fitness del modelo a lo largo de las iteraciones. Se muestra el léxico polarizado expandido resultante en el primero de los experimentos para que muestre que el modelo polariza. Se recuerda que la evaluación de los modelos de aprendizaje automático se efectúa con 10-Fold Cross-Validation.

5.2. Experimentos

5.2.1. Experimento I

Descripción: Este experimento desea corroborar si el modelo es capaz de expandir el léxico polarizado. Se quiere comprobar si el modelo que toma el léxico expandido mejora el comportamiento del modelo que toma el léxico polarizado semilla. Se experimenta en una casuística en la que se analizan textos pertenecientes a dos categorías distintas. El experimento tendrá éxito si las medidas de evaluación, función

fitness del modelo, del prototipo generado, son mejores que las del léxico polarizado semilla.

Evaluación del prototipo generado por el modelo: A continuación se muestra una gráfica que ilustra la evolución de la evaluación del modelo con 10-Fold Cross-Validation a lo largo de las iteraciones:

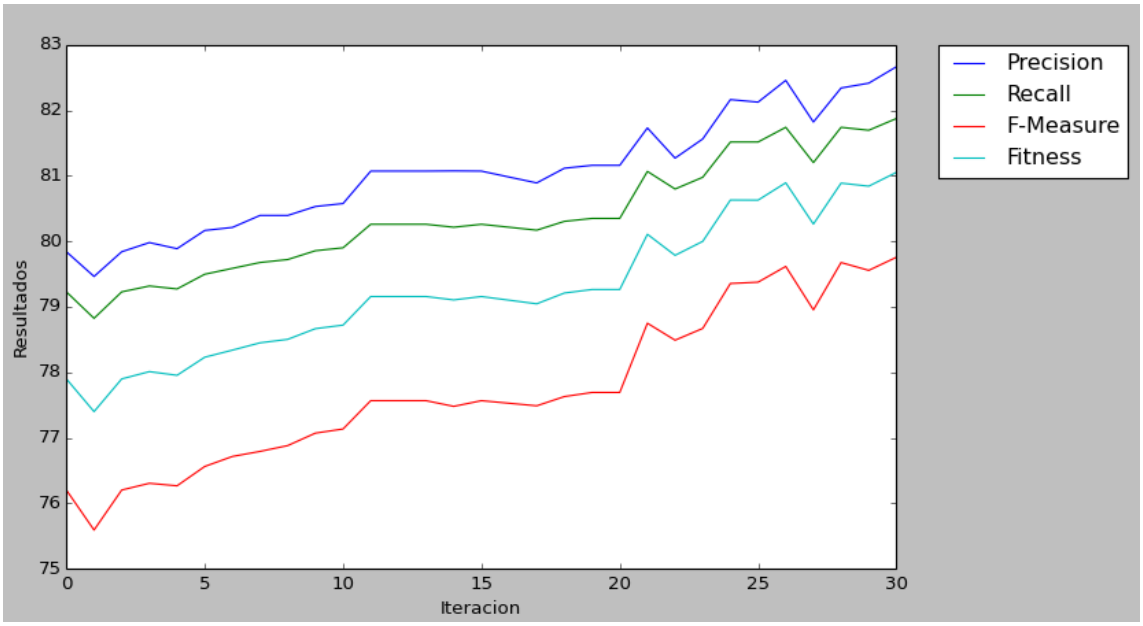

FIGURA 5.1: Evolución de los resultados.

Se puede observar que el modelo propuesto en esta tesis incrementa el rendimiento del prototipo con respecto a uno que únicamente funciona con un léxico polarizado semilla en un 3 % aproximadamente. Es importante destacar que añadiendo sólo 26 términos al léxico polarizado, que contiene un total de 6784 términos. Este valor añadido tan significativo en la evaluación añadiendo un número tan reducido de términos pese a contar con un volumen muy reducido de textos es un resultado muy positivo.

Léxico polarizado resultante: El léxico polarizado obtenido es el existente mas los siguientes términos:

- Redo: Positive.
- Photos: Positive.

- Dissapointed: Negative.
- Natural: Negative.
- Most: Negative.
- Contents: Negative.
- Up: Positive.
- Prevents: Positive.
- Coated: Negative.
- Anybody: Negative.
- Hopes: Negative.
- Yum!: Positive.
- Pony: Positive.
- w/o: Not polarized.
- Batch: Negative.
- Listen: Negative.
- Built-in: Positive.
- Sheen: Positive.
- Vents: Positive.
- This: Not polarized.
- Advertising: Negative.
- Flip: Positive.
- Heats: Positive.
- Speeds: Positive.
- Adds: Positive.
- Street: Positive.

- What: Not polarized.

Objetivos, hipótesis y comentario: Se verifica la hipótesis: El modelo propuesto es capaz de mejorar las medidas de precisión, recuperación y F-Measure con respecto a las medidas obtenidas por medio de un modelo de análisis de opiniones basado únicamente en un léxico polarizado semilla en cualquier categoría.

Se alcanzan los objetivos: Comprobar si el modelo propuesto hace que los léxicos polarizados mejoren sus medidas de evaluación con respecto a las obtenidas cuando estos léxicos no están expandidos.

Comprobar si el modelo obtiene prototipos de modelos que analizan opiniones en contextos o dominios que no han sido analizados por lingüistas o expertos en esa categoría.

Este experimento demuestra empíricamente que el modelo mejora las medidas de evaluación obtenidas por los modelos de aprendizaje automático a lo largo de las iteraciones.

El modelo presentado en este documento es capaz de mejorar el rendimiento de un modelo cuantitativo, incluso pese a que este está basado únicamente en rasgos que denotan la presencia o ausencia de unigramas en cada texto.

5.2.2. Experimento II

Descripción: Este experimento, desea corroborar si el modelo obtiene resultados similares independientemente del léxico polarizado semilla empleado.

Para afirmar esta hipótesis, se repiten los parámetros del anterior experimento, con ligeros cambios, pero esta vez empleando un léxico polarizado que no se empleaba en el experimento anterior.

A priori, los resultados deberían ser similares, excepto el número de iteraciones empleado, mediante el cual se llega a medidas estables. Se piensa que con 15 iteraciones añadiendo 5 términos por iteración debería ser suficiente.

Evaluación del prototipo generado por el modelo: A continuación se muestra una gráfica que ilustra la evolución de la evaluación del modelo con 10-Fold Cross-Validation a lo largo de las iteraciones:

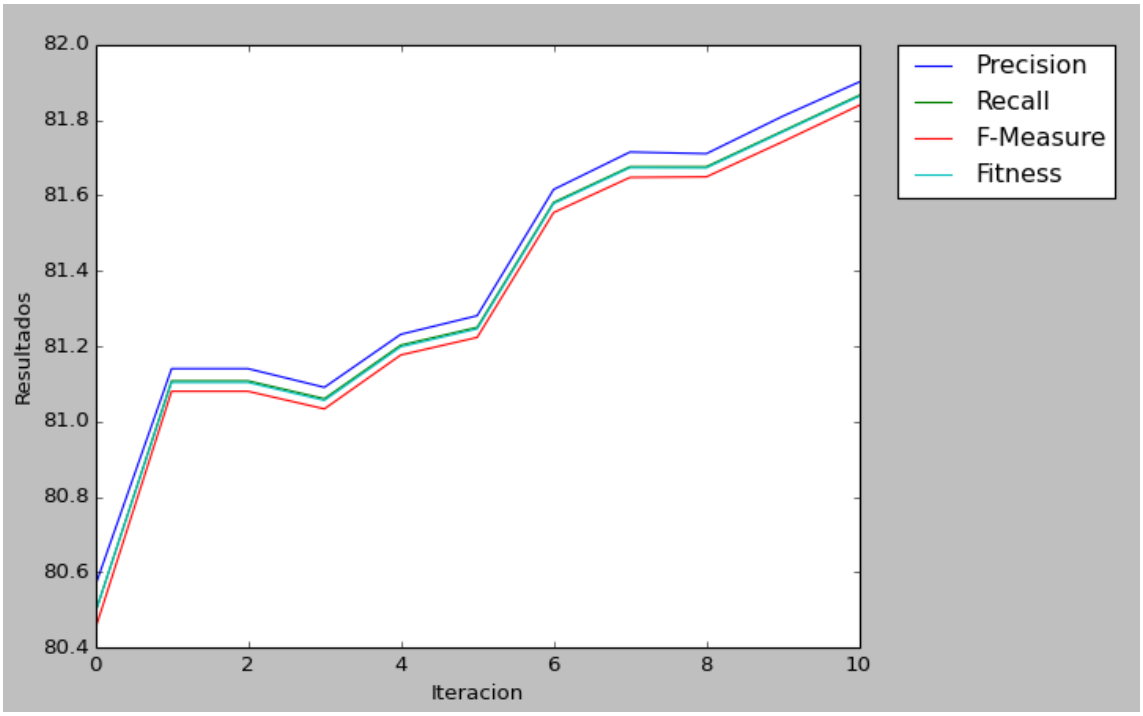

FIGURA 5.2: Evolución de los resultados.

Se puede observar que el modelo propuesto en esta tesis incrementa el rendimiento del prototipo con respecto a uno que únicamente funciona con un léxico polarizado semilla en aproximadamente un 1.3 %.

Objetivos, hipótesis y comentario: Se verifica la hipótesis: El modelo debe obtener resultados similares independientemente del léxico polarizado semilla empleado con un número lo suficientemente grande de iteraciones y textos analizados para cualquier conjunto de categorías.

Se alcanza el objetivo: Comprobar que el prototipo generado posee independencia con respecto al léxico polarizado semilla empleado.

Este experimento demuestra empíricamente que el modelo presentado en este documento es independiente no sólo de la categoría, sino de los recursos léxicos que se emplean para su funcionamiento. Para funcionar con resultados aceptables y en un tiempo de ejecución razonable, el modelo necesita un léxico polarizado, pero no tiene ninguna dependencia con ninguno en particular. El experimento II demuestra que, independientemente del léxico polarizado, el modelo es capaz de incrementar

la evaluación obtenida por los modelos de aprendizaje automático a lo largo de las iteraciones, obteniendo un prototipo de calidad superior a la inicial.

5.2.3. Experimento III

Descripción: En este experimento se desea comprobar que el prototipo generado mediante el propuesto generado es capaz de analizar opiniones de forma más efectiva en categorías que se han considerado frente a categorías que no se han considerado. Para ello, se generará un prototipo considerando 2 categorías. Se anotará el rendimiento que el prototipo posee analizando dichas categorías. Por último, se anotará el rendimiento del prototipo en una nueva categoría. El experimento tendrá éxito si esta última cifra es inferior al rendimiento presentado en las categorías consideradas.

Evaluación del prototipo generado por el modelo: A continuación se muestra una gráfica que ilustra la evolución de la evaluación del modelo con 10-Fold Cross-Validation a lo largo de las iteraciones:

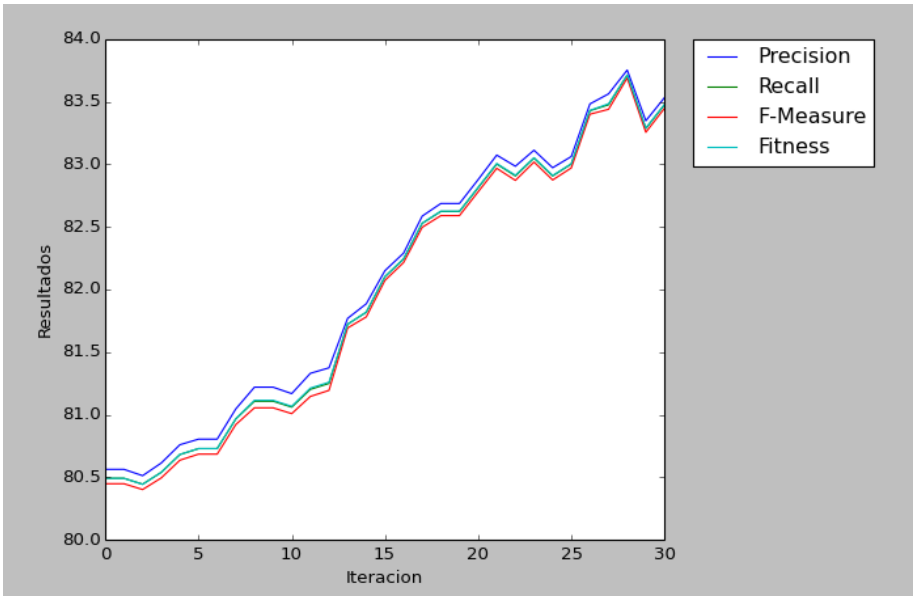

FIGURA 5.3: Evolución de los resultados.

Se puede observar que el modelo propuesto en esta tesis incrementa el rendimiento del prototipo con respecto a uno que únicamente funciona con un léxico polarizado semilla en un 3 %.

En la categoría de test en la que el modelo no se ha entrenado, este ha obtenido una función fitness del 78 %, inferior al rendimiento obtenido en las categorías en las cuáles el modelo ha sido entrenado.

Objetivos, hipótesis y comentario: Se verifica la hipótesis: El prototipo creado por el modelo de análisis de opiniones presentado en esta Tesis, analizando un volumen lo suficientemente grande de textos, presentará mejores resultados en las categorías que se analicen en su fase de entrenamiento que en categorías no consideradas.

Se alcanza el objetivo: Comprobar que el modelo obtiene mejores resultados en categorías consideradas que en categorías sin considerar.

5.2.4. Experimento IV

Descripción: En este experimento se desea comprobar si el modelo propuesto es capaz de mejorar el rendimiento del análisis de opiniones que efectúa el prototipo que genera en cualquier categoría. Para ello, e emplean un total de 5 categorías.

Evaluación del prototipo generado por el modelo: A continuación se muestra una gráfica que ilustra la evolución de la evaluación del modelo con 10-Fold Cross-Validation a lo largo de las iteraciones:

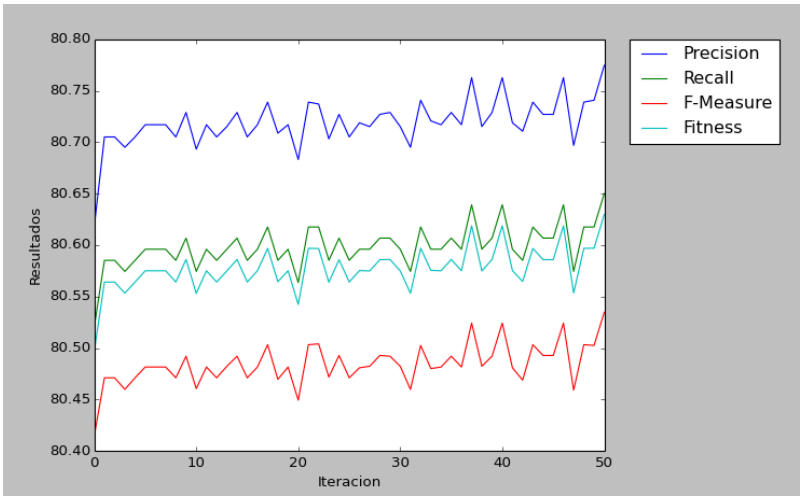

Figura 5.4: Evolución de los resultados.

Se puede observar que el modelo propuesto en esta tesis incrementa el rendimiento del prototipo con respecto a uno que únicamente funciona con un léxico polarizado semilla en un 0.15 %.

Objetivos, hipótesis y comentario: Se verifican las hipótesis: El modelo propuesto es capaz de construir un vector de características para cualquier número de categorías. Desde una categoría al número de categorías que se deseen analizar.

El modelo propuesto es capaz de expandir los léxicos polarizados semilla en cualquier contexto de entre los que se consideran en los experimentos, mostrando una mejoría en las medidas de precisión, recuperación y F-Measure con respecto a las medidas de los léxicos polarizados semilla.

Se alcanzan los objetivos: Comprobar el correcto funcionamiento del modelo independientemente del número de categorías considerado.

Comprobar el correcto funcionamiento del dominio independientemente de la naturaleza de cada categoría.

Las hipótesis se corroboran y los objetivos se alcanzan, pese a que el incremento del rendimiento en este experimento es menor, únicamente un 0.15 %. La causa del menor rendimiento es interpretable. Esto es debido a que al contemplar un número mayor de categorías, existe un menor número de factores que son generalizables a todas ellas. Debido a ello, el incremento del rendimiento del modelo es menor, aunque logra incrementar algo el rendimiento del léxico polarizado inicial.

5.2.5. Experimento V. Descripción

Descripción: Cubiertos todos los objetivos planteados en el documento de diseño, se desean efectuar pruebas adicionales del modelo para comparar su comportamiento en función a los parámetros introducidos.

Estos experimentos, de menor importancia, resultan interesantes desde el punto de vista computacional, para observar el rendimiento del modelo en distintas circunstancias.

Evaluación del prototipo generado por el modelo: A continuación se muestra una gráfica que ilustra la evolución de la evaluación del modelo con 10-Fold Cross-Validation a lo largo de las iteraciones en la primera ejecución:

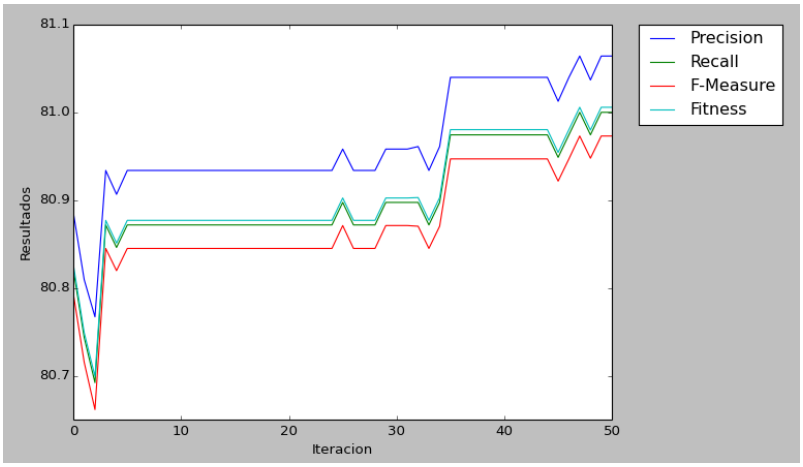

FIGURA 5.5: Evolución de los resultados.

Se puede observar que el modelo propuesto en esta tesis incrementa el rendimiento del prototipo con respecto a uno que únicamente funciona con un léxico polarizado semilla en más de un 1 %.

A continuación se muestra una gráfica que ilustra la evolución de la evaluación del modelo con 10-Fold Cross-Validation a lo largo de las iteraciones en la segunda ejecución:

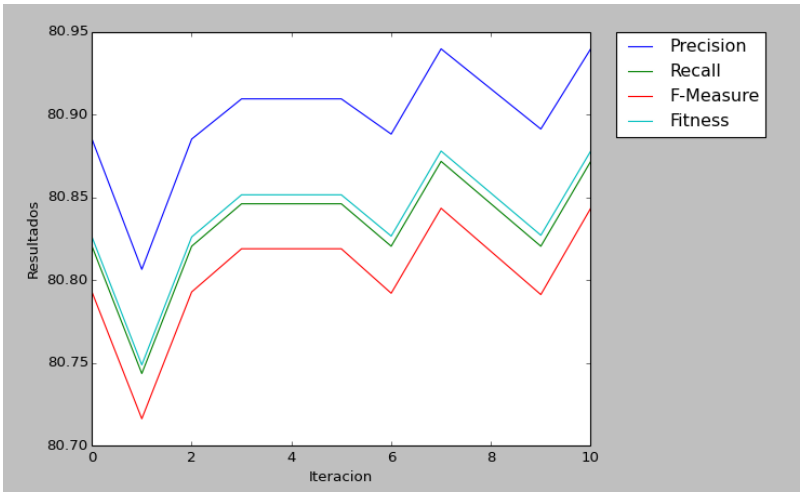

FIGURA 5.6: Evolución de los resultados.

Se puede observar que el modelo propuesto en esta tesis incrementa el rendimiento del prototipo con respecto a uno que únicamente funciona con un léxico polarizado semilla en aproximadamente un 0.1 %.

Comentario: El experimento es una muestra clara de que si se proponen varios términos por iteración el rendimiento global del modelo disminuye, pese a que el tiempo de ejecución es, lógicamente, menor. Esto es debido a que el incremento en el rendimiento de la función fitness de añadir un único elemento por iteración puede ser inferior al deterioro en la función fitness de añadir otro elemento. De este modo, si se añaden los dos elementos simultáneamente, los dos elementos son rechazados al no incrementar la función fitness. Por el contrario, si se añaden en dos iteraciones independientes, se rechazará uno y el otro no. Si se hace de este último modo, la función fitness obtiene una evaluación mayor después de las dos iteraciones.

5.2.6. Experimento VI. Descripción

Descripción: En este experimento, se desea comprobar que factor es más relevante, si la generalización o la uniformidad, si se cuenta con un número bajo de iteraciones. Se proponen dos ejecuciones del modelo, una con mayor factor de generalización y otra con mayor factor de uniformidad.

Evaluación del prototipo generado por el modelo: A continuación se muestra una gráfica que ilustra la evolución de la evaluación del modelo con 10-Fold Cross-Validation a lo largo de las iteraciones en la primera ejecución:

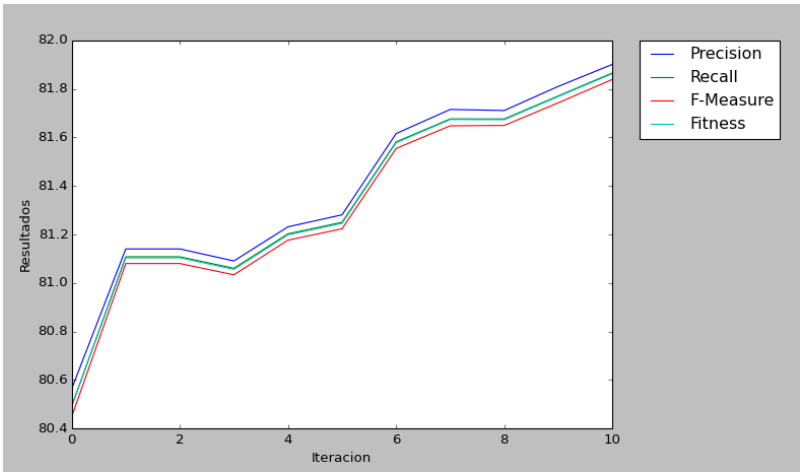

FIGURA 5.7: Evolución de los resultados.

Se puede observar que el modelo propuesto en esta tesis incrementa el rendimiento del prototipo con respecto a uno que únicamente funciona con un léxico polarizado semilla en un 1.4 %.

A continuación se muestra una gráfica que ilustra la evolución de la evaluación del modelo con 10-Fold Cross-Validation a lo largo de las iteraciones en la segunda ejecución:

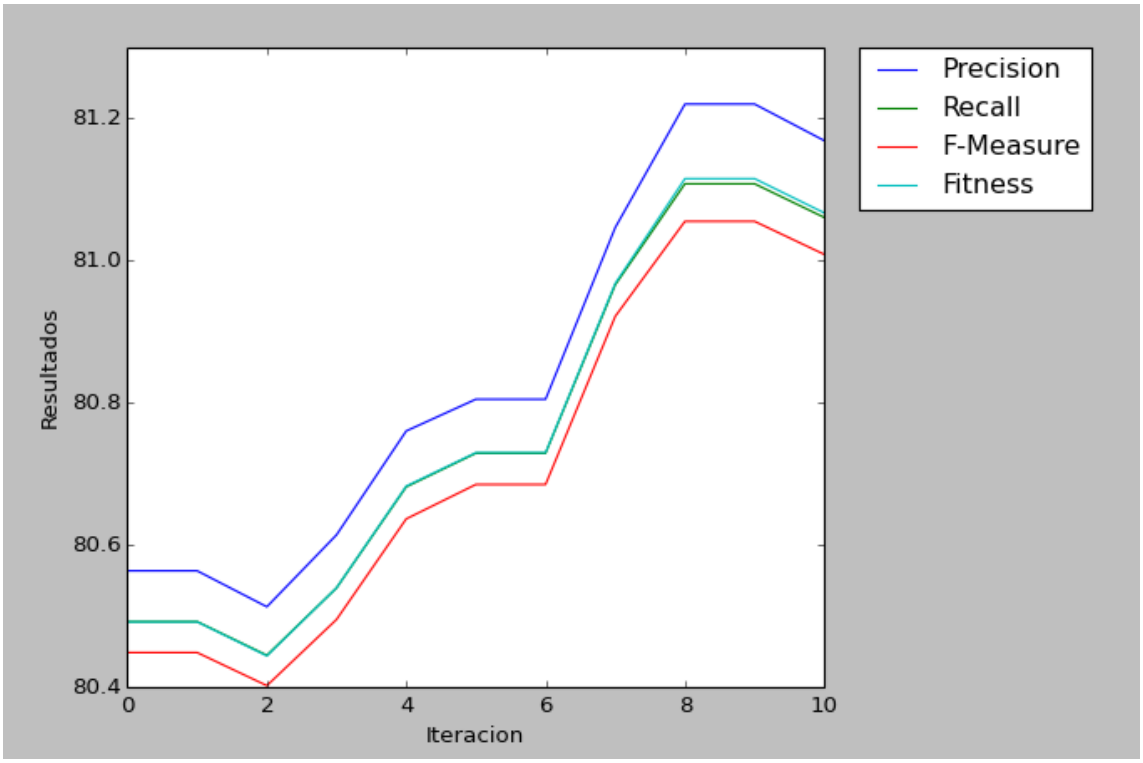

FIGURA 5.8: Evolución de los resultados.

Se puede observar que el modelo propuesto en esta tesis incrementa el rendimiento del prototipo con respecto a uno que únicamente funciona con un léxico polarizado semilla en aproximadamente un 0.6 %.

Comentario: El resultado de este experimento indica que la generalización es un factor más relevante que la uniformidad, en el conjunto de textos planteado. Los resultados de este experimento no son generalizables a cualquier conjunto de categorías que se quieran analizar. Se recomienda que el analista pruebe distintas combinaciones de pesos para comprobar cuál logra un mayor resultado de la evaluación del prototipo a lo largo de las iteraciones.

Una vez mostrado y comentado el resultado de todos los experimentos se obtienen conclusiones de carácter general sobre el trabajo presentado en este documento. Estas conclusiones, que preceden a un capítulo en el cuál se exponen líneas futuras

para mejorar las características del modelo presentado, se muestran en el siguiente capítulo.

Capítulo 6

Conclusiones

6.1. Conclusiones

El modelo presentado en esta tesis, de carácter cuantitativo, ofrece una solución alternativa a los modelos cualitativos de procesamiento del lenguaje natural, y de forma más concreta, al análisis de opiniones. El análisis de opiniones es un campo en el cuál se han presentado en los últimos años una amplia variedad de modelos. Una amplia parte de estos modelos cuenta con el inconveniente de que son extremadamente dependientes de la categoría o categorías de los textos a analizar.

La problemática de la dependencia de contexto en el análisis de opiniones ralentiza considerablemente el tiempo de desarrollo de soluciones que deben operar con textos de los cuáles no se conoce su categoría. Estos textos son genéricos, pertenecientes a cualquier categoría, pero deben de ser polarizados. Un área donde ocurre esta problemática es, por ejemplo, las redes sociales.

Existen dos metodologías para lidiar con esta problemática. La primera de ellas consiste en el estudio de las diferentes categorías que presentan los textos que se analizan y en elaborar un modelo cualitativo para cada una de ellas. Los principales inconvenientes de esta metodología son los siguientes. En un gran número de escenarios, existe un número muy grande de estas categorías, por lo que hacer un modelo para cada categoría conlleva un tiempo muy elevado. En segundo lugar, el tiempo de estudio de estas categorías, si el volumen de textos es muy elevado, es considerable.

La segunda metodología por la cual enfrentarse a esta situación es mediante un modelo cuantitativo. Las principales virtudes de los modelos cuantitativos son la gran cobertura que ofrecen todos sus resultados y su habilidad para analizar grandes

volúmenes de texto en tiempos de ejecución razonables. Los modelos cuantitativos carecen de la precisión de los cualitativos en categorías acotadas, pero representan una alternativa viable si el problema a tratar es el de la independencia de contexto en el análisis de opiniones.

El modelo propuesto, lidia con la problemática de la dependencia de contexto acorde a la segunda metodología de las propuestas en el párrafo anterior. Este modelo ofrece una solución que construye automáticamente prototipos especialistas en las categorías en las cuáles han sido entrenados. A través de los experimentos evaluados en el capítulo anterior, se observa que la dependencia de contexto o categoría es un problema que se puede tratar mediante enfoques cuantitativos. En todos los experimentos propuestos, la evaluación de los modelos ofrece mejores resultados a lo largo de las iteraciones.

No obstante, los enfoques cuantitativos, pese a su gran flexibilidad, también cuentan con desventajas. El modelo presentado en esta tesis, para ser realmente efectivo, necesita una amplia gama de textos etiquetados. Los resultados de la evaluación del modelo sobre los textos propuestos con 10-Folds Cross-Validation son mejores a lo largo de las iteraciones, pero al carecer de un corpus de texto mayor, no es posible determinar con seguridad si un número lo suficientemente grande de nuevos textos de las categorías en las que ha sido entrenado el prototipo serán correctamente analizados.

Por ello, se recomienda que el prototipo generado sea sometido a la supervisión de un especialista en las categorías analizadas. El especialista proporcionará la precisión necesaria para no obtener falsos positivos, restringiendo los resultados del modelo mediante un conjunto de reglas. El prototipo generado cuenta con la cobertura que el especialista, debido a las limitaciones humanas, jamás podrá obtener. Mediante la base de reglas proporcionada por el especialista, el modelo presentado será híbrido y contará con las virtudes de los modelos cualitativos y cuantitativos.

La heurística presentada en esta tesis, al igual que los modelos cuantitativos, resulta especialmente efectiva si el volumen de textos es muy elevado. Esta tesis de fin de máster cuenta con recursos limitados, pero contando con un mayor corpus de texto anotado por polarización, los resultados del léxico expandido semilla del experimento 1 serían más adecuados para las categorías analizadas.

En otras palabras, de nada sirve elaborar un estudio tan detallado de la generalización y la uniformidad si, por ejemplo, la mayoría de los términos únicamente aparece 1 vez en el conjunto global de textos. Para volúmenes de texto reducidos y

categorías perfectamente acotadas, se recomienda usar un modelo cualitativo en vez del modelo cuantitativo presentado.

Un problema adicional es caer en el problema del overfitting mediante el modelo presentado. Es decir, que si se configuran un número muy grande de iteraciones y términos a añadir por iteración, entonces el prototipo presentado puede tener el riesgo de especializarse demasiado únicamente en los textos presentados, si su volumen no es muy grande. Para evitar este riesgo se emplea 10-Folds Cross-Validation, pero se sospecha que si el volumen de textos no es muy grande y se añaden muchas iteraciones, entonces el rendimiento puede no ser el esperado en nuevos textos pertenecientes a las categorías analizadas.

Dado el volumen reducido de textos del corpus anotado de textos que se ha empleado para esta investigación, únicamente se han considerado unigramas para el modelo presentado en este documento. No obstante, en el capítulo siguiente de líneas futuras, se presenta una generalización de este modelo que emplea como rasgos N-Gramas de longitud personalizable por el analista.

Una conclusión importante que se extrae, es que se necesita un volumen más elevado de textos anotados por polaridad del que se ha empleado para realizar esta Tesis para seguir extrayendo más conclusiones a partir de experimentos. Con volúmenes de textos pequeños se recomienda elaborar enfoques cualitativos.

En el capítulo posterior se presentará como, además de la generalización de unigramas a N-Gramas, otros problemas, como el multilingüismo, son abordables con ligeras modificaciones del modelo presentado. Esta es una de las principales ventajas de los modelos cuantitativos, que su gran flexibilidad e independencia casi total de recursos lingüísticos hace que se analice un texto independientemente de sus características, incluso del lenguaje en el que está escrito.

A continuación, se cierra este documento mediante el capítulo de líneas futuras. En este capítulo se exponen generalizaciones de este modelo para cubrir casuísticas que no se han contemplado en el modelo presentado y se proponen nuevos trabajos que aportan un valor añadido al presentado en este documento.

Capítulo 7

Líneas Futuras

7.1. Líneas Futuras

En esta memoria, se ha presentado un modelo que, a partir de textos que pertenecen a categorías independientes, es capaz de expandir un léxico polarizado semilla. Este léxico polarizado semilla tiene la particularidad de que se adapta únicamente a las categorías que se quieren analizar.

Adicionalmente, el modelo presentado en esta memoria es capaz de generar un clasificador para clasificar nuevos textos pertenecientes a esas categorías en función a su polaridad. Tanto el clasificador como el léxico polarizado expandido pueden ser generados a partir de cualquier número de categorías y de textos, independientemente de sus características particulares.

El modelo presentado es una generalización de un modelo expuesto en el estado del arte, al que se le han añadido nuevas ideas. Del mismo modo que se ha generalizado el modelo expuesto en el estado del arte, el modelo que se ha expuesto en esta memoria puede ser también ampliado, como se verá a continuación.

En este apartado, se enumeran y explican las líneas futuras que se deben tomar para seguir trabajando en este modelo.

- **Emplear un mayor número de textos y categorías:** Dado el carácter cuantitativo de este modelo, se garantiza que su funcionamiento será más efectivo si se emplea un mayor número de textos del que se dispone para esta tesis fin de máster. Se recomienda construir un corpus anotado por polaridad con tantas instancias o textos como atributos se generen para los clasificadores de

aprendizaje automático. De esta forma, el resultado obtenido en la expansión del léxico polarizado será más preciso.

- **Ampliación de la funcionalidad del modelo:** Este modelo ha sido diseñado en última instancia para construir clasificadores de análisis de opinión automáticos independientes de la categoría que se quiera analizar. No obstante, el modelo puede ser empleado para otras labores. Por ejemplo, considérese que se quiere analizar no la polaridad, sino la tendencia política del que escribe una opinión.

  Esta opinión puede ser escrita en diferentes contextos: Artículos de opinión, Twitter, Blogs, etc... Se quiere obtener la tendencia política del autor: Izquierda o Derecha. Tal y como está construido el modelo, dado que izquierda y derecha son tendencias opuestas, el modelo puede extrapolarse con ligeros cambios en la parte de polaridad de la GAE para resolver este nuevo problema. Del mismo modo, habría que construir un léxico polarizado por ideología. El modelo obtendría como salida un léxico expandido y un clasificador automático independiente de la categoría.

  Toda clasificación entre categorías polarmente opuestas puede ser resuelta mediante esta metodología. Otros ejemplos pueden ser: Nivel de satisfacción de los empleados de una empresa, adaptación de los alumnos a un colegio nuevo, detección de patologías psicológicas en escritos, etc...

- **Expansión de N-Gramas en el léxico polarizado:** El modelo actualmente únicamente incorpora unigramas en el léxico polarizado semilla. En consecuencia, el modelo de aprendizaje automático empleado únicamente incorpora variables que simbolizan la existencia o no existencia de unigramas en el texto para polarizarlo. Tal y como se ha expuesto en el estado del arte, existen muchos rasgos más que puede tomar en cuenta un modelo de aprendizaje automático para clasificar un texto en categorías.

  Por tanto, una mejora a realizar en el modelo consiste en la búsqueda de N-Gramas por parte de la GAE. Como candidatos, no sólo se considerarían unigramas sino términos de una longitud de palabras inferior a un umbral personalizable por el usuario.

  Esta mejora otorgaría flexibilidad al analista de opiniones, que podría configurar hasta que longitud de palabras quiere que la GAE extraiga candidatos. La GAE, por tanto, aumenta el espacio de búsqueda para buscar candidatos, y puede ser configurada para que trabaje en el mismo espacio de búsqueda en el cuál lo hace en el modelo propuesto en esta memoria.

Aumentando el espacio de búsqueda puede ralentizarse la convergencia hacia una solución pero también se puede evitar un mínimo local. Los nuevos candidatos formados por un número configurable de palabras pueden mejorar la evaluación del clasificador obtenido por la GAE. Es una medida por tanto recomendada para el modelo presentado en esta memoria.

- **Construcción de una metaheurística para la optimización de la GAE:** Como se ha presentado en el apartado de modelo de esta memoria, la GAE utiliza para sus tres pilares expresiones que se han aproximado manualmente para la obtención de mejores resultados. En estos tres pilares: uniformidad, generalidad y polarización, se han tenido que emplear expresiones como raíces de alto grado para que los rankings de los términos calificados por estas categorías tengan los mismos valores. De esta forma, ninguno de los tres pilares domina sobre el resto.

  Esta mejora propone emplear una metaheurística que optimice el valor de estas raíces para obtener los resultados óptimos. Se propone emplear la GAE en un número concreto de experimentos. Se emplea como función fitness de la metaheurística el resultado de la evaluación del clasificador de aprendizaje automático obtenido cuando el modelo converge. El espacio de búsqueda está compuesto por los números reales que serán los que se irán asignando a las raíces de los tres pilares de la GAE. La solución final consistirá en los tres números reales bajo los cuales la GAE construya clasificadores automáticos que optimicen la media de las funciones fitness en el número concreto de experimentos que se han propuesto.

  Para realizar esta medida, se debe asumir que los experimentos que se diseñen para esta mejora del modelo son representativos tanto en naturaleza como en número del conjunto total posible de categorías que se puedan analizar. Dado que se puede analizar cualquier número de categorías, siempre pueden acontecer casos bajo los cuales la optimización de la GAE obtenga resultados menos precisos. Por ello, si se desea acometer esta mejora, se deben analizar el mayor número de categorías posibles y que sean lo más diferentes entre sí.

- **Construcción de un sistema de aprendizaje reforzado para la optimización de la GAE**: Del mismo modo que los valores de las raíces se pueden optimizar para que la GAE obtenga resultados más precisos, se pueden optimizar los pesos que se otorgan a cada uno de los tres pilares de la GAE para que esta obtenga resultados más precisos.

  Esto puede hacerse con una metaheurística como se ha propuesto en el anterior apartado o bien mediante un sistema de aprendizaje reforzado en el usuario. Se

debe comenzar con un vector aleatorio de pesos para los tres pilares de la GAE. A partir de esta configuración, se genera el clasificador para las categorías. Se ejecuta este clasificador en un conjunto de textos de prueba elaborados por el usuario. El usuario introduciría en el sistema su grado de aceptación de los resultados. Este grado de aceptación de los resultados sería la función fitness de la metaheurística, que, iterativamente, cambia el valor de las variables del vector de pesos de los pilares mediante algún comportamiento de naturaleza similar al recocido simulado y genera nuevos clasificadores, que deberán ser juzgados por el usuario.

El criterio de convergencia deberá ser elegido por el analista. La solución final sería el clasificador que mayor grado de aceptación tenga por el usuario y el léxico polarizado asociado a este clasificador.

- **Empleo de un mayor número de rasgos para los clasificadores**: Del mismo modo del que se pueden añadir N-Gramas a este modelo, el modelo puede incorporar más rasgos para que el clasificador los tenga en cuenta. Estos rasgos son los comentados en el estado del arte, como por ejemplo, reglas gramaticales del tipo: Que la frase contenga un sustantivo seguido por un adjetivo. Otro rasgo que se puede tener en cuenta es la presencia o ausencia de deícticos pertenecientes a una categoría determinada.

  Para implementar esta mejora en el modelo, se propone construir un set de rasgos a ser tomados en cuenta en el clasificador. En el algoritmo iterativo del modelo, además de los términos considerados por la GAE, cada número de iteraciones personalizable por el usuario, se propondrá añadir al clasificador variables que simbolicen uno o más rasgos pertenecientes a este set. Si la evaluación del modelo mejora, se añaden estos rasgos al clasificador, de no hacerlo, se desechan.

  Esta mejora aumenta aún más el espacio de búsqueda de la GAE, que ya no sólo consideraría N-Gramas sino el conjunto formado por N-Gramas y los rasgos que los analistas consideren que hay que estudiar. Si los analistas no quieren considerar ningún rasgo dejarían este set en blanco. Por lo tanto se puede generar también el modelo presentado en esta memoria. Esta mejora es, por tanto, otra generalización adicional del modelo propuesto en esta memoria.

- **Adaptación al Multilingüismo**: En este modelo, la única dependencia existente con el lenguaje se halla en el léxico polarizado semilla. Gracias a funcionar mediante estadística y aprendizaje automático, la barrera del cambio del lenguaje es menor. Aun así, por utilizar un recurso, si se quiere emplear este

modelo en otro lenguaje se debe usar un léxico polarizado semilla perteneciente al lenguaje con el que se quiera trabajar. También se puede optar por la traducción automática de los términos contenidos en el léxico polarizado semilla. Del mismo modo, se necesitarán tokenizadores en el lenguaje destino para obtener los unigramas. Si se cuentan con estos dos recursos, entonces el modelo generará prototipos y expandirá léxicos independientemente del idioma.

Las mejoras propuestas en este apartado podrían mejorar la evaluación de los clasificadores obtenidos por este modelo, por ello, se recomienda experimentar sobre ellas. Del mismo modo, ninguna de estas mejoras impide que se genere el mismo modelo del presentado en esta memoria, son únicamente generalizaciones. La principal virtud del modelo presentado en esta memoria es la gran cobertura que presenta, a costa de una pérdida de precisión de los resultados obtenidos en el clasificador.

Este es el principal inconveniente de los modelos cuantitativos como el presentado en esta memoria. Por tanto, se recomienda combinar los resultados obtenidos por este modelo para analizar opiniones con los resultados que obtenga un modelo cualitativo construido por expertos sobre las categorías a considerar. Para averiguar qué pesos aplicar a cada uno de los dos modelos se recomienda construir un sistema de ayuda a la toma de decisiones que, a partir de los resultados de los dos clasificadores y en función al texto entrante, asigne un peso al modelo cualitativo y cuantitativo para la decisión final de la polarización del texto entrante. Un modelo híbrido como el resultante de integrar un sistema de análisis de decisiones al modelo cualitativo y cuantitativo gozaría de las ventajas de ambos: Precisión y Cobertura.

Bibliografía

[1] Singh Zhang. Renew: A semi-supervised framework for generating domain-specific lexicons and sentiment analysis. 2014.

[2] cyc, 2014. URL `http://commons.wikimedia.org/wiki/File:Svm_max_sep_hyperplane_with_margin.png#mediaviewer/File:Svm_max_sep_hyperplane_with_margin.png`.

[3] Kkddkkdd, 2014. URL `http://upload.wikimedia.org/wikipedia/commons/4/47/SVM_with_soft_margin.pdf`.

[4] Yanqing Chen and Steven Skiena. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 383–389, 2014.

[5] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[6] Erik Cambria, Bjorn Schuller, Yunqing Xia, and Catherine Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, page 1, 2013.

[7] Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *AAAI*, volume 4, pages 755–760, 2004.

[8] Nitin Indurkhya and Fred J Damerau. *Handbook of natural language processing*, volume 2. CRC Press, 2012.

[9] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[10] J Schafer. The application of data-mining to recommender systems. *Encyclopedia of data warehousing and mining*, 1:44–48, 2009.

[11] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[12] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[13] Christopher Potts. Sentiment analysis tutorial, 2011.

[14] Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. Large-scale sentiment analysis for news and blogs. *ICWSM*, 7, 2007.

[15] Prem Melville, Wojciech Gryc, and Richard D Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1275–1284. ACM, 2009.

[16] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[17] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[18] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[19] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010.

[20] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.

[21] Justin Martineau and Tim Finin. Delta tfidf: An improved feature space for sentiment analysis. In *ICWSM*, 2009.

[22] Olga Vechtomova, Kaheer Suleman, and Jack Thomas. An information retrieval-based approach to determining contextual opinion polarity of words. In *Advances in Information Retrieval*, pages 553–559. Springer, 2014.

[23] Xiaowen Ding, Bing Liu, and Philip S Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 231–240. ACM, 2008.

[24] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[25] Eduard C Dragut and Christiane Fellbaum. The role of adverbs in sentiment analysis. *ACL 2014*, 1929:38–41, 2014.

[26] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

[27] Nikhikumar Jadhav and Pushpak Bhattacharyya. Dive deeper: Deep semantics for sentiment analysis. *ACL 2014*, page 113, 2014.

[28] Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[29] Henning Wachsmuth, Martin Trenkmann, Benno Stein, and Gregor Engels. Modeling review argumentation for robust sentiment analysis. In *Proceedings of the 25th International Conference on Computational Linguistics COLING*, 2014.

[30] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

[31] Mitra Mohtarami, Man Lan, and Chew Lim Tan. Probabilistic sense sentiment similarity through hidden emotions. In *ACL (1)*, pages 983–992, 2013.

[32] Alena Neviarouskaya and Masaki Aono. Extracting causes of emotions from text. 2013.

[33] Voula Giouli and Aggeliki Fotopoulou. Linguistically motivated language resources for sentiment analysis. In *Workshop on Lexical and Grammatical Resources for Language Processing*, page 39, 2014.

[34] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. The general inquirer: A computer approach to content analysis. 1966.

[35] Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Twenty-eighth AAAI conference on artificial intelligence*, 2014.

[36] Kerstin Denecke. Using sentiwordnet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, pages 507–512. IEEE, 2008.

[37] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[38] Wiebe Choi. +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. *EMNLP*, 2014.

[39] Subhabrata Mukherjee and Sachindra Joshi. Sentiment aggregation using conceptnet ontology. IJCNLP, 2013.

[40] Carlo Strapparava and Alessandro Valitutti. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086, 2004.

[41] Alexandra Balahur, Jesús M Hermida, Andrés Montoyo, and Rafael Muñoz. Emotinet: A knowledge base for emotion detection in text built on the appraisal theories. In *Natural Language Processing and Information Systems*, pages 27–39. Springer, 2011.

[42] Jesús Ibáñez, Oscar Serrano, and David García. Emotinet: A framework for the development of social awareness systems. In *Awareness Systems*, pages 291–311. Springer, 2009.

[43] Minlie Huang, Borui Ye, Yichen Wang, Haiqiang Chen, Junjun Cheng, and Xiaoyan Zhu. New word detection for sentiment analysis.

[44] Svitlana Volkova, Theresa Wilson, and David Yarowsky. Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. In *ACL (2)*, pages 505–510, 2013.

[45] Braja Gopal Patra, Hiroya Takamura, Dipankar Das, Manabu Okumura, and Sivaji Bandyopadhyay. Construction of emotional lexicon using potts model. *Construction of Emotional Lexicon Using Potts Model*, 2014.

[46] Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *ACL (1)*, pages 1774–1784, 2013.

[47] Vangelis Metsis, Ion Androutsopoulos, and Georgios Paliouras. Spam filtering with naive bayes-which naive bayes? In *CEAS*, pages 27–28, 2006.

[48] Jakob Elming, Dirk Hovy, and Barbara Plank. Robust cross-domain sentiment analysis for low-resource languages.

[49] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! *ICWSM*, 11:538–541, 2011.

[50] Wenliang Gao, Naoki Yoshinaga, Nobuhiro Kaji, and Masaru Kitsuregawa. Modeling user leniency and product popularity for sentiment classification. *Proceedings of IJCNLP, Nagoya, Japan. to appear*, 2013.

[51] Erik Boiy and Marie-Francine Moens. A machine learning approach to sentiment analysis in multilingual web texts. *Information retrieval*, 12(5):526–558, 2009.

[52] Ahmed Abbasi, Hsinchun Chen, and Arab Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12, 2008.

[53] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[54] Rajesh Ranganath, Dan Jurafsky, and Daniel A McFarland. Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech & Language*, 27(1):89–115, 2013.

[55] Kashyap Popat2 Balamurali AR, Pushpak Bhattacharyya, and Gholamreza Haffari. The haves and the have-nots: Leveraging unlabelled corpora for sentiment analysis. 2013.

[56] Cäcilia Zirn, Mathias Niepert, Heiner Stuckenschmidt, and Michael Strube. Fine-grained sentiment analysis with structural features. In *IJCNLP*, pages 336–344, 2011.

[57] Bing Xiang and Liang Zhou. Improving twitter sentiment analysis with topic-based mixture modeling and semi-supervised training. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 434–439, 2014.

[58] Matthew Mulholland and Joanne Quinn. Suicidal tendencies: The automatic classification of suicidal and non-suicidal lyricists using nlp.

[59] Ferran Pla and Lluis-F Hurtado. Political tendency identification in twitter using sentiment analysis techniques. In *Proceedings of COLING*, 2014.

[60] Wei Jin, Hung Hay Ho, and Rohini K Srihari. Opinionminer: a novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1195–1204. ACM, 2009.

[61] Tanveer Ali, David Schramm, Marina Sokolova, and Diana Inkpen. Can i hear you? sentiment analysis on medical forums.

[62] Wang Jiang Gemulla Weikum Qu, Zhang. Senti-lssvm: Sentiment-oriented multi-relation extraction with latent structural svm. In *ACL (2)*, pages 155–168, 2014.

[63] David Martens, Liesbeth Bruynseels, Bart Baesens, Marleen Willekens, and Jan Vanthienen. Predicting going concern opinion with data mining. *Decision Support Systems*, 45(4):765–777, 2008.

[64] Qiang Ye, Ziqiong Zhang, and Rob Law. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3):6527–6535, 2009.

[65] Nan Li and Desheng Dash Wu. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, 48(2): 354–368, 2010.

[66] Andrea Vanzo, Danilo Croce, and Roberto Basili. A context-based model for sentiment analysis in twitter. In *Proceedings of COLING*, pages 2345–2354, 2014.

[67] Songbo Tan, Xueqi Cheng, Yuefen Wang, and Hongbo Xu. Adapting naive bayes to domain adaptation for sentiment analysis. In *Advances in Information Retrieval*, pages 337–349. Springer, 2009.

[68] Ozan Irsoy and Claire Cardie. Opinion mining with deep recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 720–728, 2014.

[69] Xiaodan Zhu, Hongyu Guo, Saif Mohammad, and Svetlana Kiritchenko. An empirical study on the effect of negation words on sentiment. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL*, volume 14, 2014.

[70] Cıcero Nogueira dos Santos and Maıra Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING), Dublin, Ireland*, 2014.

[71] Diego Marcheggiani, Oscar Täckström, Andrea Esuli, and Fabrizio Sebastiani. Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In *Advances in Information Retrieval*, pages 273–285. Springer, 2014.

[72] Rudy Prabowo and Mike Thelwall. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157, 2009.

[73] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[74] Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. International sentiment analysis for news and blogs. In *ICWSM*, 2008.

[75] José María Moreno-Jiménez, Jesús Cardeñosa, and Carolina Gallardo. *Arguments that support decisions in e-cognocracy: A qualitative approach based on text mining techniques*. Springer, 2009.

[76] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. Exploiting topic based twitter sentiment for stock prediction. In *ACL (2)*, pages 24–29, 2013.

[77] Andy Moniz and Franciska de Jong. Sentiment analysis and the impact of employee satisfaction on firm earnings. In *Advances in Information Retrieval*, pages 519–527. Springer, 2014.

[78] Taisei Nitta, Fumito Masui, Michal Ptaszynski, Yasutomo Kimura, Rafal Rzepka, and Kenji Araki. Detecting cyberbullying entries on informal school websites based on category relevance maximization. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, 2013.

[79] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 219–230. ACM, 2008.

[80] Jiang Su, Harry Zhang, Charles X Ling, and Stan Matwin. Discriminative parameter learning for bayesian networks. In *Proceedings of the 25th international conference on Machine learning*, pages 1016–1023. ACM, 2008.

[81] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.