Método probabilista para clasificación de polaridad: Negación e Intensificación en Análisis de Sentimientos

por:

Samara Gretel Villalba Osornio

Tesis sometida como requisito parcial para obtener el grado de

MAESTRÍA EN CIENCIAS DE LA ESPECIALIDAD DE CIENCIAS COMPUTACIONALES

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica
Tonantzintla, Puebla

Septiembre 2016

Supervisada por:

Dr. Luis Villaseñor Pineda
Investigador titular del INAOE
Dr.. Manuel Montes y Gómez
Investigador titular del INAOE

A mi familia
por hacerme ver siempre
que seguir adelante es mi obligación.
En especial, a mi nana yaya y mi tata chepe... hasta donde estén.

RESUMEN

El Análisis de Sentimientos (AS) es un área que utiliza técnicas de procesamiento de lenguaje natural y de aprendizaje automático para extraer información subjetiva de los textos. En el AS aún quedan muchos problemas abiertos, uno de ellos es el tratamiento de la negación. La Negación es un fenómeno lingüístico presente en todos los idiomas humanos. En documentos, la negación está dada por la presencia de señales o partículas negativas. Las partículas negativas invierten el valor de verdad de una frase. Para lograr un correcto entendimiento del significado de un texto es necesario identificar y tratar estos fenómenos lingüísticos. La finalidad de este trabajo es considerar los fenómenos lingüísticos de negación e intensificación para mejorar la clasificación por polaridad en textos de opinión. Para ello se utilizará un enfoque de tipo probabilista, proponiendo algunas modificaciones al método de Naive Bayes Multinomial (NBM), las cuales permiten añadir información lingüística a los textos mejorando con ello su clasificación. El método propuesto es poco dependiente del lenguaje y la temática de los textos. Se realizaron experimentos en Español e Inglés y en varios dominios tales como cine, hoteles, libros, electrónicos, etc. Los resultados de los experimentos fueron comparados con métodos del estado del arte.

Palabras Clave: Minería de Opiniones, Análisis de Sentimientos, Transferencia de Información, Tratamiento de la Negación.

ABSTRACT

Sentiment Analysis (SA) is an area that uses Natural Language processing and Machine Learning techniques to extract subjective information from texts. In SA area, several problems are still open, one of them is negation handling. Negation is a linguistic phenomenon presented in all human languages. In written documents, negation is presented as marks or negative particles. Negative particles invert the true value of a sentence. In traditional text classification, semantic information is lost and with that, the capacity to recognize some linguistic phenomena like negation and intensification is lost too. To correctly understand the meaning of a text it is necessary to identify and to treat these linguistic phenomena. The aim of this work is to consider the negation and intensification to improve polarity classification in opinion texts. A probabilistic approach that suggests some modifications to the Multinomial Naive Bayes (MNB) that allows the handling of negation and intensification in the texts improving their classification is proposed. The work proposes a method that is little dependent of language and kind of text. Experiments in English and Spanish texts and in some domains like movies, hotels, books, electronics, etc. were performed. The results were compared with the ones published in related works.

keywords: Opinion Mining, Sentiment Analysis, Information Transfer , Negation Handling.

AGRADECIMIENTOS

Mi agradecimiento, en primer lugar al pueblo mexicano y al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo otorgado a través de la beca no. 627295.

A mis asesores, los Dres. Luis Villaseñor y Manuel Montes gracias por su apoyo y por todas sus enseñanzas que han ido desde mostrarme cómo hacer procesamiento de lenguaje hasta dónde comer si algún día visito Francia o España. Les agradezco cada hora que se tomaron para contarme sus experiencias y compartirme un poco de su amor por la ciencia y la investigación.

A mis sinodales, los Dres. Aurelio López, Hugo Escalante y Saúl Pomares por sus comentarios y correcciones.

A mis papás, gracias por su apoyo y por dejarme luchar siempre por lo que considero mejor aunque probablemente no lo sea. Mami, papi, saben de sobra que siempre estaré agradecida por ser los mejores padres y por haberme dado todo, mi compromiso siempre será darles lo mejor de mi y seguir luchando por estar bien.

A mis hermanos, Sebastian, Gexabel y Eliud gracias por cada llamada, cada mensaje, cada favor, gracias por ser incondicionalmente mis hermanos y aplicarme el bullying que eso conlleva.

A Andrés, gracias cochito por haberme acompañado en este camino y por haberme apoyado en cada momento y sobre todo, gracias por seguir estando para mí.

A mis amigos, Víctor, Mike, Josue, Viviana y Pastor gracias por cada comida y cada fiesta compartida, pero sobre todo gracias por cada plática, aprendí mucho de ustedes no sólo en cuanto a ciencia, también de la vida.

A mis amigos Sinaloenses, Cinthia, Muri, Fer, Raquel, Zurrita, Cacharpitas, Pedro, Wapo, Ñurro y Fredy, Hemos pasado por muchas alegrías y también por tristezas y no hemos podido compartirlas como quisiera. Gracias por no olvidarse de mí y estar en todos mis días con algún mensaje dándome ánimos o enviándome memes.

Gracias a todos los que se tomen el tiempo de leer este documento de tesis.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ACRÓNIMOS

AS	Análisis de Sentimientos
BOW	Bag of words
CMR	Corpus of Movie Reviews
COAH	Corpus of Analucia's Hotels
fp	*false positives*
GI	Ganancia de Información
IMDb	Internet Movie Database
JST	Join Sentiment Topic Model
LDA	Latent Dirichelet Allocation
LSM	Latent Sentiment Model
MO	Minería de Opinión
NB	Naive Bayes
NBM	Naive Bayes Multinomial
POS	Parts of Speech
SFU	Simon Fraser University Corpus
SVM	Support Vector Machines
tp	*true positives*

CAPÍTULO 1

INTRODUCCIÓN

En este capítulo se describe el problema abordado, así como su importancia social y computacional. Además se especifican los objetivos y la hipótesis sobre la cuál se desarrolló esta tesis así como las contribuciones derivadas de la misma. Por último se da una breve explicación de los capítulos siguientes.

1.1. Descripción del problema

Hoy en día el internet se ha vuelto parte fundamental de nuestras vidas. La mayoría de los seres humanos hacemos muchas cosas de nuestra vida diaria en línea: hablamos con nuestros amigos y familiares, trabajamos, hacemos movimientos bancarios y por supuesto, compramos toda clase de productos. Esto ha provocado un incremento desmedido en el número de documentos que hay en internet.

En años pasados cuando queríamos comprar cualquier producto que no conocíamos preguntábamos a familiares, amigos o al vendedor. Hoy en día usamos internet. Hay cientos de sitios en línea donde podemos ver y añadir comentarios sobre qué nos parece un producto, un servicio, una persona, etc., es por ello que el análisis automático de críticas se ha vuelto un área con creciente interés en el círculo científico y en el círculo mercantil. Cada día más empresas públicas y privadas están inclinadas en automatizar el tratamiento y clasificación de esos documentos de opinión, esto se debe a que esos

documentos son cada vez más utilizados tanto por usuarios para tomar decisiones de compra como por los negocios para mejorar sus productos y su mercadotecnia.

El analizar y tratar los textos de opinión ha sido un área creciente dentro de la investigación y ha sido nombrada Minería de opinión. La minería de opinión tiene subáreas, entre las que se encuentra el análisis de sentimientos. El Análisis de Sentimientos busca encontrar el sentimiento del usuario que escribió una determinada opinión. Hay también varias tareas en el análisis de sentimientos, siendo una de las más comunes la definición de polaridad. Es decir, determinar si una opinión refleja una impresión positiva o negativa sobre el objeto de crítica.

La clasificación de polaridad ha sido mayormente estudiada con enfoques de clasificación apoyados en lexicones o diccionarios [Taboada et al., 2011, Jiménez et al., 2015, Zhu et al., 2014]. Sin embargo, si comparamos los resultados de enfoques de aprendizaje computacional con los que usan lexicones resultan mejores los de aprendizaje. Es necesario investigar cómo mejorar los enfoques de aprendizaje para esta tarea. Elementos como la negación y la intensificación son de gran ayuda para mejorar el desempeño de los clasificadores que utilizan diccionarios, por lo que resulta natural pensar que también serían de ayuda en los métodos con base en aprendizaje. En esta tesis se utilizan estos elementos que son fácilmente aplicados en enfoques de diccionarios pero que no son tan sencillos de aplicar en enfoques de aprendizaje.

Es importante realizar las modificaciones necesarias a métodos de aprendizaje para tratar la clasificación de polaridad. La polaridad de un documento se ve realmente afectada y modificada por fenómenos lingüísticos tales como la negación y la intensificación.Un fenómeno lingüístico es es un elemento que rompe el uso normal del lenguaje. Estos fenómenos lingüísticos son poco tomados en cuenta en clasificación de documentos con enfoque de aprendizaje tradicional. Por lo anterior, la motivación de este trabajo de tesis es añadir al estado del arte un método de clasificación de documentos con enfoque de aprendizaje que incluya tratamiento de fenómenos lingüísticos que afectan directamente en la clasificación de polaridad.

Según estudios, el Inglés es unos de los tres idiomas más utilizados en el mundo [Kloumann et al., 2012]. Poco a poco se ha convertido en el idioma oficial mundial,

sin embargo, la existencia de documentos en otros idiomas está creciendo desmedidamente. Idiomas como Español, Francés, Chino, etc., son cada vez más utilizados en internet. La mayoría de los trabajos, no sólo de clasificación de polaridad, sino de cualquier tarea automática son principalmente pensados y realizados para el idioma Inglés, pero el uso del resto de los idiomas hace que sea necesario trabajar en lenguajes distintos al Inglés. Enfocándonos solamente en el trabajo realizado en clasificación de polaridad encontramos que hay pocos trabajos diseñados para Español, es por ello que este trabajo de tesis tiene como meta principal desarrollar un método de clasificación de opiniones que sea funcional tanto para Inglés como para Español, y que sea además adaptable con facilidad para el resto de los idiomas que compartan las características composicionales o de conjugación de estos dos idiomas.

1.2. Justificación y Motivación

Para entender la relevancia del problema tratado es necesario primero ubicarlo en el área de investigación. El área es el Análisis de Sentimientos es explicada en la siguiente sección.

1.2.1. Minería de Opinión VS Análisis de Sentimientos

La Minería de Opinión (MO) y el Análisis de Sentimientos(AS) normalmente son utilizados como sinónimos, sin embargo, aunque se trata de cosas que están muy relacionadas hacen referencia a diferentes tareas. La Minería de Opinión es un área extensa que busca encontrar información sobre entidades nombradas en textos de opinión. Una entidad nombrada puede ser cualquier cosa sobre la cual se exprese una opinión, por ejemplo: un producto, un servicio, un servidor público, un evento, etc. [Chinchor y Robinson, 1997].

Según [Liu y Zhang, 2012] el objetivo de la MO es, dado un grupo de documentos de opinión, realizar las tareas de:

1. Extracción y agrupamiento de entidades nombradas.- reconocer sobre qué o quién hablan el grupo de opiniones,

2. Extracción y agrupamiento de aspectos.- definir los aspectos de la entidad que son criticados en la opinión,

3. Extracción de tiempo y de titular de la opinión.- ordenar las opiniones dependiendo de cuando fueron escritas o publicadas y encontrar datos del autor,

4. Clasificación de sentimientos.- separar las opiniones por el sentimiento o la polaridad que reflejan, y

5. Combinación de las tareas anteriores.

Con la definición de estas tareas podemos encontrar que el AS es una área dentro de la Minería de Opiniones. El Análisis de Sentimientos tiene varias subtareas entre las que podemos mencionar, análisis de subjetividad, de emoción, de actitud, y una de las más comúnes, el análisis de polaridad. La polaridad puede tomar los valores de positivo, negativo y neutral. En la figura 1.1 se muestran los distintos niveles de clasificación de polaridad.

Figura 1.1: Niveles de clasificación de polaridad.

La clasificación de polaridad puede realizarse a distintos niveles dependiendo del tipo de textos y dominios sobre los cuales se esté trabajando. Hay tres niveles principales [Vinodhini y Chandrasekaran, 2012]:

1. Nivel de documento.- El objetivo en este nivel es clasificar el documento como positivo, negativo o neutral.

2. Nivel de sentencia.- La finalidad a este nivel es reconocer cuáles sentencias del documento tienen una carga de polaridad positiva, cuáles negativa y cuáles neutral. Este nivel está muy relacionado a la clasificación de subjetividad. La clasificación de subjetividad trata de distinguir entre textos objetivos y subjetivos.

3. Nivel de atributo.- Es también llamado nivel se entidad o nivel de característica. Este nivel surgió debido a que en los niveles de documento y de sentencia no se define cuál es el objeto de crítica. Una opinión puede hablar sobre más de un objeto, por ejemplo "Me encantó la comida, pero el servicio dejó mucho que desear". En el ejemplo podemos ver que hay una opinión positiva, sin embargo también hay una opinión negativa. El nivel de atributo nos permite identificar que hay dos polaridades en esa crítica y que se está dando la opinión de dos cosas diferentes.

Algunos autores incluyen un cuarto nivel llamado nivel de grupo. Lo que se busca en el nivel de grupo es encontrar la opinión general de un grupo de textos. Esta tarea se denomina cuantificación [Barranquero et al., 2015, Gao y Sebastiani, 2015].

Como se mencionó anteriormente uno de los problemas más importantes dentro del AS es la clasificación de polaridad. Este trabajo de tesis se enfoca en este problema a nivel de documento y en las problemáticas que pueden estar alrededor de la asignación de una etiqueta de polaridad. En la siguiente subsección se detalla más a fondo el problema y los puntos que giran alrededor de él y que fueron tratados en este trabajo.

1.2.2. El problema de clasificación de polaridad.

Podemos resumir el problema a resolver como: *Dado un documento de opinión asignarle una etiqueta de polaridad.* La polaridad puede tomar los valores de positivo, negativo o neutral. La polaridad representa la sensación o sentimiento que tiene el autor respecto al servicio o producto sobre el cuál está emitiendo una opinión. La clasificación por polaridad tiene muchas problemáticas alrededor, entre las que destacan la polaridad según el dominio y los fenómenos lingüísticos de negación e intensificación.

Polaridad según el dominio:

Uno de los enfoques más utilizado para hacer clasificación de polaridad es el que se soporta en diccionarios. En este caso los diccionarios son listas de palabras positivas y negativas. La clasificación bajo este enfoque se realiza mediante el conteo de palabras de cada uno de los diccionarios (positivo y negativo), el conteo mayor es la etiqueta que se asigna. Aunque es un enfoque muy usado y con resultados aceptables su principal

problema es su dependencia al dominio. Cada palabra puede tomar un valor positivo o negativo dependiendo de la temática sobre la cuál se esté emitiendo la opinión. Por ejemplo: la palabra "tranquila" en la oración *"La historia es muy tranquila, incluso aburrida"* tiene un grado de polaridad negativa, en cambio, en la oración *"El hotel está ubicado en una zona muy tranquila"* tiene carga de polaridad positiva.

El hecho de que cada palabra pueda variar su valor de polaridad según lo que se esté hablando hace que una sola lista de palabras positivas y negativas no sea funcional para todos los dominios.

La negación.

La negación es un fenómeno lingüístico presente en la mayoría de los idiomas humanos. Su presencia está dada por palabras de negación o partículas negativas tales como: no, ni, sin, nunca, nada, etc., o bien por prefijos y sufijos por ejemplo, *"in**feliz**"* o *"dis**funcional**"* [Abu-Jbara y Radev, 2012]. La negación tiene la capacidad de cambiar por completo el valor de verdad de una oración y también de modificar fuertemente su polaridad. Hay varias problemáticas dentro del tratamiento de la negación de forma automática [Zou et al., 2014], entre los que se enlistan los siguientes:

a) **Afectación:** Definir el tipo de afectación que tiene una negación en una oración en términos de polaridad es algo complicado, ya que puede depender de la perspectiva de quien esté leyendo o escribiendo la opinión. Invertir la polaridad es un enfoque muy común en el tratamiento de negación, sin embargo, es el menos intuitivo: Al invertir la polaridad estamos asumiendo que tratamos de decir lo contrario, por ejemplo, la oración *"La película no es aburrida"* podemos asumir que es igual a la oración *"La película es divertida"*. Las palabras aburrida y divertida son antónimos y su carga polar es contraria. Sin embargo, al hablar en un Español coloquial podemos entender que la película fue no "divertida", pero estuvo "bien" o "normal". Al ver este tipo de ejemplos surgió la idea de no invertir la polaridad sino sólo modificarla al encontrar una negación. Este enfoque también tiene sus problemas ya que no existe alguna regla o documento que diga cuánto hay que modificar el valor de polaridad debido a la negación.

b) **Alcance:** El alcance es la parte de la oración que resulta afectada por una negación. Una de las formas más simples de definir el alcance es con uso de ventanas. La definición de la ventana es mediante la asignación de un número de palabras despúes y/o antes de la negación que se verán afectadas por la partícula negativa [Hogenboom et al., 2011, Narayanan et al., 2013]. Otro enfoque utilizado es el uso de árboles de dependencia y la definición de reglas sobre ellos [Jiménez et al., 2015]. El enfoque de árboles de dependencia es menos utilizado que el enfoque de ventana por ser más complicado y porque genera ruido con errores que surgen en el análisis sintáctico, debido a estos errores no pueden hacer una gran mejora en la clasificación de polaridad que tiende, por el contrario, a empeorar los resultados.

c) **Enfoque:** En enfoque es la parte del alcance que está más altamente negada [Abu-Jbara y Radev, 2012]. El significado de una oración puede ser diferente según cuál sea el enfoque de la negación. Para identificar el enfoque de la negación se utiliza la información de contexto. En el siguiente ejemplo tomado de [Zou et al., 2014] podemos ver los diferentes enfoques que podemos tener en una oración:

Hellen no permite tocar el violín a su hijo menor...

- *... pero su esposo sí...* Centra la negación en Hellen.
- *... porque ella piensa que él no tiene el talento artístico de su hijo mayor...* Centra la negación en el hijo menor.
- *... porque sus vecinos se molestan...* Centra la negación en el violín.

d) **Representación:** Es muy importante tener en cuenta que la negación es un fenómeno difícil de tratar en clasificación tradicional de textos. La bolsa de palabras es una de las maneras más comunes de representación de textos para tareas de clasificación y recuperación. El principal problema de la representación de bolsa de palabras es que no mantiene el orden en que los términos aparecieron en los documentos. Es necesario buscar una representación o una modificación a la bolsa de palabras que permita incluir información de contexto para poder rescatar el fenómeno de la negación.

Figura 1.2: Problemáticas de la negación.

Intensificación.

La intensificación es un fenómeno lingüístico muy frecuente en el idioma. A diferencia de la negación este fenómeno no puede invertir el valor de polaridad de una expresión pero sí hacerla más fuerte o más débil. La intensificación se da por palabras de intensificación o atenuación. Algunos ejemplos de intensificadores son: muy, tan, más, etc. y de atenuantes son: menos, poco, entre otras. Hay listas predefinidas de las palabras que actúan como intensificadores y atenuantes y la proporción de modificación que tendrán las palabras afectadas. La lista utilizada en este trabajo es la utilizada en el sistema SOCAL [Taboada et al., 2011].

La intensificación tiene un nivel de alcance. El nivel de alcance de la intensificación se define por el uso de ventanas o bien con enfoques apoyados en etiquetas POS (Parts of Speech). En los enfoques de etiquetas POS se afectan la siguientes palabras después de la aparición de un intensificador o atenuante que tenga la etiqueta de adjetivo, adverbio o verbo.

1.3. Propuesta

En este trabajo de tesis se propone desarrollar un método de clasificación supervisada de tipo probabilista para la clasificación de polaridad. Se eligió atacar el problema desde un enfoque probabilista por considerarse un método de los más utilizados en tareas de clasificación de textos. Además normalmente requiere un menor tiempo de ejecución y de entrenamiento que otros clasificadores como el caso de SVM. Se pretende que el método (1) sea lo más independiente del idioma y del dominio posible y que (2) incluya un tratamiento a los fenómenos lingüísticos de negación e intensificación. Se busca que se mejore la clasificación de polaridad gracias al tratamiento de la

negación y de la intensificación.

1.3.1. Hipótesis

Como se muestra en los siguientes capítulos, la clasificación de polaridad, el análisis de la negación y otros elementos lingüísticos como la intensificación, ya han sido abordados en diversos trabajos [Jiménez et al., 2015, Taboada et al., 2011]. Sin embargo la unión de estas tres problemáticas no han sido lo suficientemente abordadas. No se han encontrado reportes de sistemas de clasificación de polaridad que puedan ser utilizados y tener resultados competitivos en varios idiomas, dominios o tipos de documentos.

La hipótesis sobre la cuál se realizó esta tesis es que agregar un tratamiento de la negación y de la intensificación a un sistema de clasificación pueden mejorar sus niveles de exactitud en la clasificación de polaridad. Esto aplicado a cualquier idioma y dominio que presente estos dos tipos de fenómenos lingüísticos.

1.3.2. Preguntas de investigación

Tras el análisis de la problemática de la negación surgen las siguientes preguntas:

1. ¿Dar tratamiento a la negación y a la intensificación en textos de opinión mejora la clasificación de polaridad usando cualquiera de los dos tipos de enfoques de clasificación propuestos?

2. ¿Tratar la negación e intensificación en la clasificación de polaridad con enfoque probabilista afecta de la misma manera en distintos dominios?

3. ¿Tratar la negación e intensificación en la clasificación de polaridad con enfoque probabilista tiene el mismo efecto en textos escritos en Español o Inglés?

4. ¿Qué enfoque de clasificación de opiniones se comporta mejor en combinación con el tratamiento de la negación?¿El enfoque de lexicones o el enfoque de aprendizaje computacional?

5. ¿Cómo se puede añadir un método de tratamiento de negación y de la intensificación a los dos tipos de métodos?

1.3.3. Objetivos

Con la finalidad de responder a las preguntas anteriores se plantean los siguientes objetivos:

Objetivo general

- El objetivo general de este trabajo es desarrollar métodos de clasificación de polaridad que incluyan un tratamiento de la negación y de la intensificación en Español e Inglés, con la finalidad de demostrar que con un apropiado tratamiento de estos fenómenos puede mejorar los niveles de exactitud en la clasificación de polaridad.

Objetivos específicos

- Desarrollar un método de clasificación de polaridad con enfoque híbrido que incluya el tratamiento de la negación e intensificación. Con enfoque híbrido nos referimos a un método de clasificación de incluya elementos de clasificación basada en diccionarios y elementos de clasificación basada en aprendizaje computacional.
- Desarrollar un método de clasificación de polaridad con enfoque de aprendizaje computacional que incluya el tratamiento de la negación e intensificación.
- Determinar si un tratamiento de la negación y de la intensificación afecta la exactitud de clasificación de polaridad de manera similar en documentos escritos en Inglés y en Español.

1.3.4. Contribuciones

Entre las contribuciones de este trabajo de tesis se encuentran las siguientes:

- Se presentan modificaciones a la representación de Naive Bayes Multinomial que permiten incluir información de negación e intensificación en la clasificación de documentos.
- Se desarrollan algoritmos de definición del alcance de la negación y la intensificación.

- Se propone representaciones para los documentos que permitan darle mayor importancia a las palabras afectadas por una negación o una intensificación.
- Se modifica la forma de calcular las probabilidades de las palabras a cada clase para incrementar su afectación en la clasificación de documentos.
- Se propone un nuevo esquema de pesado de palabras para enfoques basados en diccionarios que incluye datos sobre el vocabulario de cada una de las clases.
- Otra de las contribuciones es el análisis de las colecciones y el conteo de negaciones, intensificaciones y atenuaciones que ocurren en esos documentos. Además se incluye un análisis comparativo entre las clases.

1.3.5. Estructura de la tesis

Este trabajo de tesis se divide en los siguientes capítulos:

Capítulo 2: Marco teórico. En este capítulo de detallan las definiciones y conceptos preliminares necesarios para comprender la tesis.

Capítulo 3: Estado del arte. Se analiza algunos de los trabajos más sobresalientes en las distintas temáticas abordadas en esta tesis.

Capítulo 4: Clasificación de polaridad. Se exponen las variaciones propuestas para los métodos basados en lexicones y a los métodos basados en aprendizaje. Las modificaciones propuestas fueron pensadas para mejorar la clasificación de opiniones según su polaridad mediante la inclusión de información de negación e intensificación.

Capítulo 5: Experimentos. Las colecciones y corpus son analizados y detallados en este capítulo. También se detallan las configuraciones de los distintos experimentos realizados y los recursos externos utilizados.

Capítulo 6: Resultados y análisis. Se exponen los resultados obtenidos en este trabajo y se analizan dichos resultados.

CAPÍTULO 2

MARCO TEÓRICO

En este capítulo se detallan los conceptos preliminares necesarios para la comprensión de la tesis. La definición más importante en esta tesis es la clasificación, específicamente, la clasificación de textos. A continuación se describen las problemáticas de la clasificación de textos y de la clasificación de textos según su polaridad.

2.1. Clasificación de textos

La tarea de clasificación es uno de los problemas más estudiados actualmente. La clasificación es el ordenamiento o la disposición por clases [Michalski et al., 2013]. La meta de la clasificación automática es asignar una etiqueta a un elemento desconocido. Las etiquetas son predefinidas. En la figura 2.1 se muestra el comportamiento de un clasificador de dos clases. Los círculos indican los elementos más representativos de una clase, mientras los cuadrados son los más representativos de la otra. La estrella representa a el elemento al que debe asignarse una clase dependiendo de con cuál de las dos clases existentes tenga más similitudes.

En algunos problemas de clasificación, puede darse el caso de que un elemento pertenezca a más de una clase o bien que no pertenezca a ninguna. Un ejemplo de estos casos podría ser si queremos distinguir animales que pueden respirar dentro o

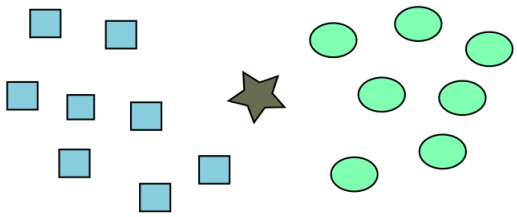


Figura 2.1: Problema de clasificación de dos clases.

fuera del agua y debemos clasificar a un anfibio. Este tipo de problemas se denominan de clasificación multi-clase o multi-etiqueta. Un ejemplo de problema multietiqueta de documentos y específicamente de clasificación por polaridad puede ser un documento que contenga tanto críticas con polaridad positiva como negativa. En el recuadro se encuntra un ejemplo tomado del corpus CMR. Podemos ver que tiene elementos de la clase positiva al decir "¡¡Me ha gustado muchísimo!!" y también de la clase negativa al mencionar "la película es mala, muy mala.".

> *Ahora es cuando quedo como un friki al decir. ¡¡Me ha gustado muchísimo!! Sobre todo, claro, cuando la cámara se pone en primer plano tal y como era el juego. Ainss, qué recuerdos. Lo único que he echado de menos es que se pusiera en modo inmortal y avanzase a puñetazos. Aunque en parte lo hace. Vale, antes de nada, la película es mala, muy mala. Pero la nostalgia me puede. ...*

Existen dos tipos principales de clasificación [Liu y Zhang, 2012]:

Clasificación supervisada.

La clasificación supervisada se distingue porque tiene datos ya etiquetados. Los datos pueden ser etiquetados con la ayuda de expertos en el área de interés. El funcionamiento de la clasificación supervisada se basa en crear dos grupos de datos etiquetados. El primer grupo son los datos de entrenamiento, los cuáles sirven para aprender qué atributos son discriminantes de cada clase. El segundo grupo de datos es llamado de prueba, en esta partición se realizan los experimentos [Liu y Zhang, 2012, Michalski et al., 2013].

Este trabajo se realizó con técnicas de clasificación supervisada, por lo que las definiciones anteriores y siguientes se encuentran dentro del marco de este tipo de

clasificación. El funcionamiento de un sistema de clasificación se muestra en la figura 2.2.

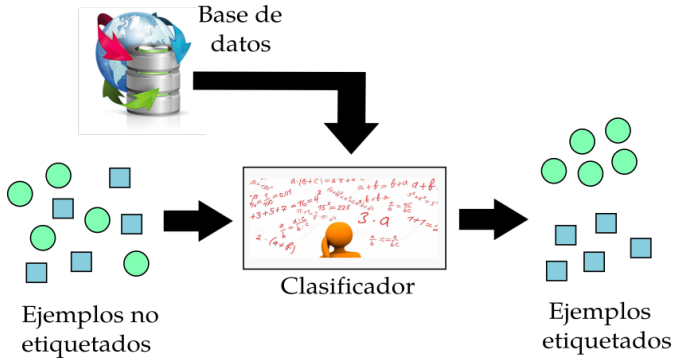


Figura 2.2: Proceso de clasificación.

Hay muchas aplicaciones para la clasificación en diferentes áreas, algunos ejemplos son la identificación de imágenes con de personas con cancer, la identificación de mensajes escritos por depredadores sexuales, o la definición de las razones del llanto de un bebé según su sonido. Dentro de la clasificación de textos existen muchas tareas altamente estudiadas. Algunos ejemplos es la clasificación temática, atribución de autoría, perfilado de autores, clasificación de polaridad etc.. Las categorías entre las cuáles quieran ordenarse un conjunto de documentos es dependiente de la aplicación y el problema que quiera resolverse.

Clasificación no supervisada.

A diferencia de la clasificación supervisada, en este caso no se cuentan con elementos ya etiquetados. Los métodos de clasificación no supervisada funcionan con la ayuda de recursos externos ya definidos [Liu y Zhang, 2012]. Un recurso externo puede ser una lista de palabras distintivas de cada clase.

2.1.1. Representación de documentos

Una de las formas más comunes de representar documentos es con la *bolsa de palabras*. La bolsa de palabras es una representación muy simple que muestra cuáles palabras están y cuáles no en el documento. Esta representación no toma en cuenta

el orden en que las palabras aparecieron en el documento ni la frecuencia de cada palabra [Sebastiani, 2002].

Otra representación de documentos es la vectorial. Un documento puede ser representado como un vector $\overrightarrow{d} =< x_1, x_2, ...x_j >$ dónde dicho vector puede ser binario o un vector de enteros. En caso de ser un vector binario representa si la palabra apareció o no en el documento. Si es un vector de enteros representa la frecuencia con que la palabra apareció en el documento.

2.1.2. Selección de características

La selección de características o atributos es un paso importante dentro de la clasificación de documentos y de la clasificación en general. La selección de características es el proceso de seleccionar un conjunto de característica que sean más relevantes o que resulten de mayor utilidad para diferenciar entre clases [Peng et al., 2005, Tripathy et al., 2016].

Hay muchas técnicas utilizadas para reducir el número de características. Entre las técnicas más simples se encuentra el filtrado de texto, que consiste en eliminar dígitos, emoticones o incluso palabras vacías. Las palabras vacías son altamente usadas en cualquier tipo de documento, es por ello que son poco relevantes para hacer cualquier distinción entre clases. Existen listas ya definidas de palabras vacías para muchos idiomas. Algunos ejemplos de esas palabras son conjunciones, preposiciones o verbos auxiliares. Existen también otros procesos de selección de atributos que incluyen el ranking de palabras como medida de decisión sobre qué atributos serán utilizados. Se utiliza un criterio de ranking para ordenar las palabras. Entre los criterios más utilizados se encuentran los siguientes.

Ganancia de Información

Una de las técnicas más utilizadas de selección de atributos es la Ganancia de Información (GI). La GI se utiliza para seleccionar los atributos que sean más impor-

tantes para cada una de las clases. La importancia de los atributos se mide mediante la impureza que tengan. La impureza se puede definir como qué tan dudoso es que el atributo pertenezca a una clase o a otra [Aggarwal y Zhai, 2012]. En Teoría de Información se llama entropía a esa impureza. Formalmente, la entropía se define como el grado de incertidumbre asociado a una distribución de probabilidad. Entre más bajo sea el valor de la entropía de una característica es más claro a qué clase pertenece.

$$\text{Ganancia de Información}(y|x) = \text{Entropía}(y) - \text{Entropía}(y|x) \tag{2.1}$$

Podemos resumir a la Ganancia de Información como una medida de cuánto nos ayuda conocer el valor de una variable x para conocer el valor de otra variable y. Supongamos que x es un atributo cualquiera y y es la clase, en la expresión 2.1 se calcula su GI.

La entropía de y condicionada a x se calcula con la expresión 2.2 donde v_j representa los posibles valores que puede tomar la variable x.

$$\text{Entropía}(y|x) = \sum_j Prob(x = v_j)E(y|x = v_j) \tag{2.2}$$

Información Mutua

La información mutua también llamada transinformación es una escala que mide la dependencia mutua entre dos variables, es decir, mide la reducción de incertidumbre de una variable a debido al conocimiento de otra variable b [Aggarwal y Zhai, 2012, Liu y Zhang, 2012]. En la figura 2.3 podemos ver un ejemplo de información mutua. Supongamos que a es el conjunto de documentos que tiene escrita la palabra "bonita" y b es el conjunto de documentos que son de la clase positiva. El conjunto a, b son aquellos documentos que tienen escrita la palabra bonita y son de la clase positiva.

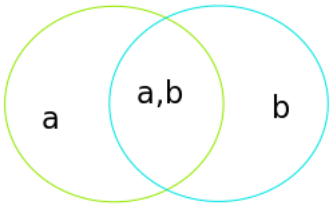


Figura 2.3: Información Mutua.

La información mutua puede calcularse con la expresión 2.3. El numerador indica la probabilidad conjunta de a y b y el denominador es la multiplicación de las probabilidades de las dos variables independientes.

$$\text{Información Mutua}(a,b) = \frac{Pr(a,b)}{Pr(a)Pr(b)} \tag{2.3}$$

Mínima Redundancia, Máxima Relevancia

Esta medida busca encontrar los atributos que tengan la más alta dependencia a las clases y que a su vez, tenga la mínima dependencia a otros atributos [Peng et al., 2005, Ding y Peng, 2005]. Para seleccionar las mejores características, se utiliza Información Mutua entre característica contra clase (para obtener los atributos más relevantes) y característica contra característica (para obtener los atributos más redundantes).

Con este criterio de evaluación se eligen las características que compartan más información con las clases pero que a su vez estén menos correlacionadas con el resto de las características.

Las dos principales razones para utilizar este tipo de técnicas es la reducción de dimensionalidad o bien para simplificar el problema de clasificación quitando atributos que pueden estar introduciendo ruido.

2.1.3. Clasificadores

Existen un sin número de métodos de clasificación que son utilizados en tareas automáticas de documentos. Entre los que destacan los siguientes:

Clasificadores basados en lógica: Son métodos que tratan de encontrar relaciones entre las características. Entre los más utilizados se encuentran los clasificadores basados en reglas [Aggarwal y Zhai, 2012].

Clasificadores geométricos: Este tipo de métodos tratan de mapear los documentos a un hiperplano y encontrar la línea que los separa según su clase. Algunos de los métodos dentro de esta categoría son las máquinas de soporte vectorial (SVM) y los métodos de K vecinos mas cercanos [Michalski et al., 2013, Sebastiani, 2002].

Redes Neuronales: Un clasificador de textos de red neuronal es una red de unidades. La figura 2.4 representa una red neuronal. Las entradas son los términos del documento. Las unidades de salida simbolizan las categorías de interés y los pesos de los ejes corresponden las relaciones de dependencia entre los términos y las categorías [Aggarwal y Zhai, 2012, Sebastiani, 2002]. Las redes neuronales también están dentro de los clasificadores geométricos.

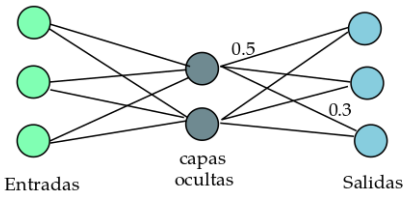


Figura 2.4: Clasificador de Redes Neuronales.

Clasificadores probabilistas: Un clasificador probabilista es un sistema que realiza desambiguación mediante la asignación, de un conjunto de clases, aquella que es más probable de acuerdo a un modelo probabilístico [McCallum y Nigam, 1998]. El modelo expresa las relaciones entre la variable de clasificación y las variables que corresponden a propiedades de un objeto ambiguo y el contexto dónde ocurre [Vapnik,

1999]. El clasificador Naive Bayes es uno de los más comunes entre los clasificadores probabilistas.

En este trabajo se tomó Naive Bayes como base para las modificaciones propuestas. Es por ello que este clasificador es visto más a fondo a continuación.

Naive Bayes

El clasificador Naive Bayes surge con la regla de Bayes y tiene un fuerte supuesto de que todos los atributos son independientes entre sí [Narayanan et al., 2013]. De acuerdo a la regla de Bayes la probabilidad de que un documento d pertenezca a la clase c_i está dada en la expresión 2.4.

$$P(c_i|d) = \frac{P(d|c_i) * P(c_i)}{P(d)} \tag{2.4}$$

Asumiendo que las palabras son independientes unas de otras, podemos simplificar la expresión 2.4 a la expresión 2.5. w_j son las palabras del documento. Además se elimina $P(d)$ pues todos los documentos son igualmente probables.

$$P(c_i|d) = \prod P(w_j|c_i) * P(c_i) \tag{2.5}$$

La salida del clasificador está representada en la expresión 2.6. El resultado será la clase que haya obtenido mayor probabilidad.

$$C = \arg\max_{c_i} P(c_i|d) \tag{2.6}$$

Las probabilidades de las palabras de pertenecer a cada una de las clases puede calcularse de varias maneras. Una de las más comunes es distribución de probabilidad de Bernoulli que se detalla en la expresión 2.7. Dicha distribución se calcula con la presencia de las palabras dentro de los documentos de cada clase.

$$P(w_j|c_i) = \frac{\text{documentos de } c_i \text{ que contienen el término } w_j}{\#\text{documentos de } c_i} \tag{2.7}$$

La probabilidad a priori $P(c_i)$ se calcula con la expresión 2.8

$$P(c_i) = \frac{\#\text{documentos en } c_i}{\text{total de documentos}} \tag{2.8}$$

El problema de probabilidad cero.

Algunas palabras que aparecen en la etapa de prueba pudieron no existir en el conjunto de entrenamiento, debido a esto su probabilidad de pertenecer a cualquiera de las clases es 0. Esto es un problema porque la fórmula de Bayes incluye una multiplicación de probabilidades. Para solucionar este problema podemos utilizar alguna técnica de suavizado. Uno de los suavizados más simples es el Laplaciano. Se muestra en la expresión 2.9.

$$P(w_j|c_i) = \frac{K + \text{ documentos de } c_i \text{ que contienen el término } w_j}{K + 1 * \#\text{documentos de } c_i} \tag{2.9}$$

Normalmente K es igual a 1. De esta manera, podemos decir que una palabra que no apareció en la etapa de entrenamiento tiene una aparición en la clase c_i.

Variaciones de Naive Bayes

La representación de *bolsa de palabras* es una de las más utilizadas en el método de clasificación de NB. Cómo se mencionó en la subsección "Representación de documentos" hay algunos tipos de vectores dentro de esta representación. debido a esta diferencia de vectores surgen algunas variaciones en el algoritmo NB. El utilizado en este trabajo es Naive Bayes Multinomial que funciona con la frecuencia de las palabras en los documentos [Aggarwal y Zhai, 2012, Fach, 2012, Vapnik, 1999]. En este trabajo de tesis se hicieron algunos modificaciones a Naive Bayes Multinonial. Se eligió esta variación bajo la premisa de que una palabra es más importante si aparece repetidas

veces en el documento. La expresión 2.10 es la modificación del Bayes Multinomial. El exponente f es la frecuencia de la palabra en el documento.

$$P(c_i|d) = \prod P(w_j|c_i)^f * P(c_i) \tag{2.10}$$

El calculo de las probabilidades de pertenencia de las palabras a cada clase se realiza con la expresión 2.11. V_{c_i} representa el vocabulario de la clase i.

$$P(w_j|c_i) = \frac{f \text{ de } w_j \text{ en } c_i + 1}{|V_{c_i}| + (\text{total de frecuencias de palabras en d de } c_i)} \tag{2.11}$$

2.2. Evaluación

Comúnmente, para evaluar el desempeño de un sistema de clasificación se utilizan las medidas de precisión y recuerdo, ambas son medidas estándar en Recuperación de Información y poco a poco han sido adoptadas también en tareas de clasificación. La precisión es la proporción de documentos etiquetados con clase c_i que realmente pertenecen a esa clase. El recuerdo es la proporción de documentos de la clase c_i que fueron etiquetados con esa clase.

Para calcular estas medidas se toma en cuenta los aciertos y errores de clasificación. La tabla 2.1 resume el comportamiento de la evaluación de un sistema de clasificación.

	Predicción positiva	**Predicción negativa**	**Total de predicciones**
Clase positiva	a	b	a+b
Clase negativa	c	d	c+d
	a+c	b+d	a+b+c+d=n

Tabla 2.1: Evaluación: aciertos y errores.

La tabla 2.1 representa el número de predicciones positivas y negativas. Así, $a + d$ son los verdaderos positivos (tp) o las predicciones correctas del sistema y $c + b$ son los falsos positivos (fp) o los casos en qué el sistema se equivocó. La suma de todos $a + b + c + d$ es el total de predicciones hechas por el sistema. Con la información de la tabla pueden estimarse las medidas de precisión y recuerdo para cada una de las clases.

La precisión indica en qué medida el clasificador ubicó a los elementos en la clase que les correspondía. La precisión se calcula con la expresión 2.12:

$$P_i = \frac{tp_i}{tp_i + fp_i} \tag{2.12}$$

Suponiendo que estemos calculando de precisión de la clase positiva tp_i es igual a a en la tabla y fp_i es igual a c.

El recuerdo expresa cuantos de los documentos de una clase son clasificados en ella.

$$P_i = \frac{tp_i}{tp_i + fn_i} \tag{2.13}$$

En la fórmula 2.13 fn_i son los elementos de la clase i que se clasificaron en la clase incorrecta. En este caso, si calculamos el recuerdo de la clase positiva fn_i tomará el valor de b.

Para medir el desempeño de un sistema de clasificación a veces no es conveniente tener dos medidas, por lo que también se utiliza la medida F. La medida F es la media armónica entre la precisión y el recuerdo. Esta métrica se calcula con la expresión 2.14

$$F_\beta = \frac{(1 + \beta^2) \text{ precisión * recuerdo}}{\beta^2 * \text{precisión} + \text{recuerdo}} \tag{2.14}$$

β es un parámetro que controla la importancia relativa de las dos medidas (precisión y recuerdo). Usualmente se asigna a β el valor de 1 para dar la misma importancia a ambas medidas.

Otra medida utilizada en este trabajo es la exactitud. La exactitud es el porcentaje de predicciones correctas que realizó el sistema. Tomando en cuenta la tabla 2.1 la exactitud se calcula con la expresión 2.15

$$E = \frac{a+d}{a+b+c+d} \tag{2.15}$$

La precisión y el recuerdo pueden ser calculados para cada clase c_i o mediante un promedio de todas las clases. Cuando se busca evaluar el comportamiento del sistema en todas las clases y no por cada clase las medidas de evaluación se establecen de la siguiente manera:

- Micro-Promedio, el cuál se usa para dar una importancia proporcional al número de ejemplos que tiene cada clase.
- Macro-promedio, el cuál asume que todas las categorías tienen la misma importancia.

Dicho de otra manera, el Micro-promedio calcula la precisión y el recuerdo considerando todas las predicciones como un sólo grupo. Por otro lado el Macro-promedio consiste en calcular las medidas de manera individual para cada clase y después promediar los resultados.

2.3. Clasificación de polaridad

Como se mencionó anteriormente hay un sin número de tareas relacionadas a clasificación de documentos. La abordada en este trabajo es la clasificación por polaridad.

La clasificación de polaridad es parte del Análisis de Sentimientos. Desde el punto de vista del AS un documento puede ser etiquetado según el sentimiento o la emoción

que refleja, o bien por la carga de polaridad que contiene [Tripathy et al., 2016, Taboada et al., 2011]. Aunque la polaridad puede tomar tres valores (positivo, negativo o neutral) es muy común reducir el problema a solamente dos clases (positivo y negativo). Hacer la reducción de clasificación de polaridad a un problema binario ha sido adoptado por varios investigadores, la razón principal es que se considera que las opiniones de tipo neutral contienen elementos de la clase positiva y de la clase negativa, por lo que hacen que tomar la decisión de etiqueta para el clasificador sea muy confuso [Narayanan et al., 2013, Jiménez et al., 2015].

La clasificación de polaridad se realiza principalmente con dos métodos: el primero es la clasificación apoyada en lexicones o diccionarios y el segundo es la clasificación que usa aprendizaje computacional. Ambos métodos son detallados en las siguientes secciones.

2.3.1. Clasificación basada en lexicones

La clasificación basada en lexicones o en diccionarios funciona con listas de palabras que son distintivas de cada una de las clases. Para el caso de la clasificación de polaridad se utilizan listas de palabras positivas y negativas [Taboada et al., 2011, Jiménez et al., 2015, Liu y Zhang, 2012]. Las listas tienen valores de positividad y negatividad de cada una de las palabras, en algunos casos también se tienen valores de objetividad.

El método de clasificación basada en diccionarios se muestra en la figura 2.5. La decisión de etiquetas se define con el conteo de las palabras que contiene el documento. Suponiendo que tenemos que asignar una clase a una crítica debemos contar cuántas palabras positivas y cuántas negativas hay en esa opinión. Se asigna la etiqueta del conteo que resulte mayor. Existen variaciones en este método dónde pueden sumar valores de positivo y negativo de cada palabra, o hacer multiplicaciones con las probabilidades de las palabras de pertenecer a cualquiera de los dos grupos. Sin embargo, el conteo sigue siendo la forma más popular de hacer este tipo de clasificación.

Algunas palabras tienen carga de polaridad positiva o negativa dependiente del tema sobre el cuál se esté dado una crítica. La palabra "bonita" tiene un valor positivo

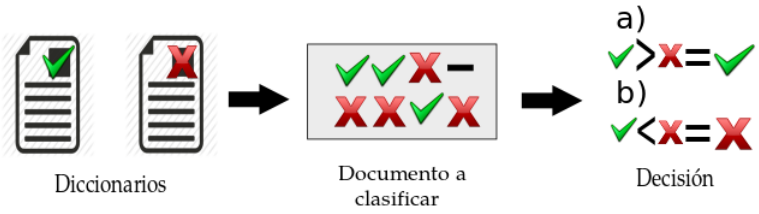


Figura 2.5: Clasificación basada en lexicones.

si se está hablando de una dama, sin embargo pierde esa polaridad si está hablándose de una película de terror. Con la finalidad de atender las diferentes cargas que puede tener una palabra en distintos dominios se han creado varias listas de palabras con polaridad para distintos dominios, especialmente para Español, algunos ejemplos son eSOL para el dominio del cine [Jiménez et al., 2015], eSOLHotel para el dominio turístico [González et al., 2015].

Otro punto importante en el uso de listas es el idioma. La mayoría de los recursos lingüísticos, no sólo en relación a tareas de análisis de sentimientos, están escritos, pensados y son utilizados para el idioma Inglés. La principal alternativa a esta problemática ha sido la traducción de listas a otros idiomas, tal es el caso de SentiWordNet que tiene su versión en Español y en Inglés [Esuli y Sebastiani, 2006]. Otra forma de solucionar este problema ha sido aprender listas de palabras en otros idiomas partiendo del Inglés [Pérez et al., 2012]

2.3.2. Clasificación basada en aprendizaje computacional

Como se mencionó anteriormente, el enfoque de aprendizaje computacional hace uso de un proceso inductivo que construye automáticamente un clasificador para la clase c mediante la observación de las características de un conjunto de documentos clasificados manualmente por un experto. En el caso de clasificación de polaridad, se tienen dos conjuntos de elementos previamente clasificados como positivos y negativos y se aprenden las características que tiene cada clase. El primer conjunto de documentos previamente clasificados se denomina conjunto de entrenamiento. Con base en las características aprendidas se clasifican los nuevos documentos. El segundo conjunto de documentos es el conjunto que se utilizará para realizar pruebas [Witten y Frank,

2005, Michalski et al., 2013]. En la figura 2.6 se muestra su funcionamiento.

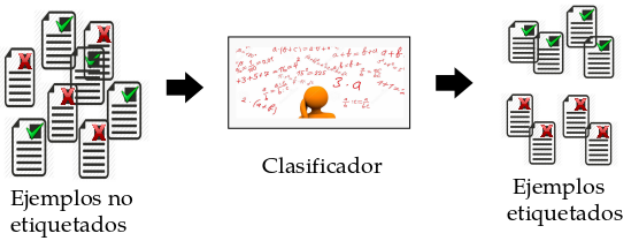


Figura 2.6: Clasificación basada en aprendizaje computacional.

Con este tipo de clasificación se busca encontrar que palabras o términos se consideran positivos o negativos y hacer uso de ellos para clasificar los nuevos documentos según su carga polar.

Existen varios enfoques entre los clasificadores basados en aprendizaje computacional [Sebastiani, 2002]. En este trabajo de tesis se trabajó más cercanamente con el enfoque probabilista.

CAPÍTULO 3

ESTADO DEL ARTE

El análisis de sentimientos y la clasificación de polaridad ya ha sido anteriormente estudiada. En este capítulo se muestran algunos de los trabajos más sobresalientes en las distintas problemáticas que influyen en la clasificación de polaridad y el análisis de sentimientos.

3.1. Clasificación de polaridad

En el capítulo anterior se describe la tarea de clasificación de polaridad y las principales formas utilizadas para atacar el problema. Para recordar: el problema consiste en dado un documento de opinión asignarle una etiqueta de polaridad y ha sido abordado desde dos principales enfoques: el primero es la clasificación con base en diccionarios [Brooke et al., 2009, Ohana et al., 2011, González et al., 2015] y el segundo es la clasificación que utiliza técnicas de aprendizaje computacional [Narayanan et al., 2013] La mayoría de los trabajos de clasificación de opiniones son para textos escritos en Inglés [Narayanan et al., 2013, Ohana et al., 2011, Taboada et al., 2011], aunque también se han hecho aproximaciones para otros idiomas como Chino [Zhang et al., 2009], Francés [Ghorbel y Jacot, 2011], Árabe [Shoukry y Rafea, 2012] y Español [González et al., 2015, Jiménez et al., 2015]. Otro aporte para idiomas distintos al Inglés han sido adaptaciones o traducciones de sistemas diseñados para el Inglés a otros idiomas [Brooke et al., 2009].

3.1.1. Problemáticas en la clasificación de polaridad : Enfoques de diccionarios

Existen muchas problemáticas que complican la tarea de clasificación de documentos según su polaridad. La mayoría de los problemas en torno a la categorización por polaridad son expuestas bajo el entorno de clasificación basada en diccionarios, esto se debe a que esta tarea se ha tratado en mayor medida con este tipo de enfoques y no con aprendizaje computacional. Algunas de las problemáticas destacadas se enlistan en los siguientes puntos:

La importancia de la temática.

El tópico sobre el cual se exprese una opinión es importante. Como se mencionó anteriormente una palabra puede ser positiva o negativa dependiendo el tema del que se esté hablando. Esto no ha sido ignorado por los estudiosos de la clasificación de polaridad. Para atacar esta problemática se ha optado por realizar listas de orientación semántica dependientes del dominio. Una de las listas más importantes e utilizadas en la clasificación de polaridad basada en lexicones es SentiWordNet [Esuli y Sebastiani, 2006]. SentiWordNet es una lista de palabras en Inglés y sus valores de positividad, negatividad y objetividad. Esta lista no está orientada a ningún dominio, y es por ello que se ha tomado como base para el desarrollo de otras listas que sí incluyan información de dominio o incluso, para hacer traducciones a otros idiomas. Un ejemplo de estas listas traducidas y modificadas es CriSOL [González et al., 2015]. CriSOL es una lista de palabras en Español y ha sido a su vez modificada en repetidas ocasiones. Por ejemplo, para ser adaptada para el dominio de cine [Jiménez et al., 2015] y al dominio de hoteles [González et al., 2015].

Se ha buscado superar los problemas de dominio haciendo combinaciones de listas de varios dominios. Un ejemplo de como unir varias listas de polaridad es mediante la suma de los valores de una palabra determinada en todas las listas [Ohana et al., 2011].

La importancia del idioma.

Los diccionarios tienen un importante problema en cuanto a idioma. Utilizar lexicones hace que un sistema de clasificación sea totalmente dependiente de en qué idioma están escritos estos lexicones. Como se mencionó en el punto anterior, esta problemática se ha abordado haciendo listas de palabras en distintos idiomas o bien traduciendo los diccionarios existentes al idioma necesario [González et al., 2015, Jiménez et al., 2015, Molina et al., 2013]. Otro acercamiento a la resolución de este problema ha sido utilizar traducciones automáticas de los textos a clasificar [Brooke et al., 2009]. La opción de traducir los textos de opinión no resulta tan buena debido a que utilizar máquinas de traducción automática arrastra errores que se llevan hasta el proceso de clasificación. Otra forma de traducir los textos podría ser usar expertos, sin embargo esto resultaría tardado. Otra alternativa es aprender lexicones en distintos idiomas tomando como base diccionarios escritos en Inglés [Pérez et al., 2012]. Este tipo de enfoques hacen uso de diccionarios bilingües y de anotaciones manuales para definir las diferencias de sentimiento o carga subjetiva de cada palabras según el lenguaje.

¿De qué opinamos?

Una opinión sobre un producto o un servicio tiene, por lo general, más información subjetiva que objetiva. Quienes brindamos una opinión buscamos dejar claro la impresión y el sentimiento que nos dejó el objeto de nuestra crítica. La información subjetiva suele contener datos de muchos objetos dentro de un mismo documento. Si hablamos del dominio de las películas, raramente estaremos criticando una película en un texto. Se hablará de la sala de cine, de la comodidad de las sillas, del actor principal y sus otras películas, del director de la película y su trayectoria. Al final en un texto de opinión sobre un producto o servicio se critican varios productos y varios servicios. Al hablar de distintas temáticas en una sola opinión se complica el encontrar la polaridad real de la opinión. Este problema suele atarse con el reconocimiento de entidades nombradas o bien la clasificación de los documentos a nivel frase o sentencia [Wilson et al., 2005]. En la clasificación por sentencia se divide el documento de opinión en las frases que lo componen y se clasifican éstas de manera independiente. Para conocer el valor general del documento se hace la unión de los valores de todas

las frases. Puede hacerse la unión de las polaridades por votos, por sumas o incluso por la exclusión de aquellas expresiones que tengan la polaridad menos frecuente [Llorente et al., 2015].

En el siguiente recuadro está un ejemplo de una opinión que menciona varios temas. Fue tomado del corpus CMR. Habla sobre un escritor, el director de la película, los actores y la temática, haciendo la opinión de la película una unión de todos estos elementos que, aunque están muy relacionados, no son la película en sí.

> *"El jardinero fiel" se basa en el libro del mismo nombre, escrito por John le Carré, famoso escritor inglés quién ya ha llevado a la pantalla grande otras novelas de espionaje, como "El sastre de Panamá" (2001), protagonizada por Pierce Brosnan.Hay tres buenos motivos para ver esta película:1. Su director: Fernando Meirelles, quién estuviera nominado al Oscar por su magnífica "Ciudad de Dios" (2003) (película que al igual que ésta, se basó en un libro).2. Los protagonistas: los excelentes actores británicos Ralph Fiennes (El paciente inglés, 1996) y Rachel Weisz (La momia, 1999).3. El tema: la grave denuncia contra el mundo de la industria farmaceútica, que realiza sus pruebas usando a los humanos más vulnerables del planeta y por lo tanto los más "deshechables": los pobres del continente africano, los perfectos humanillos de indias. En este caso, niños, lo cual hace que la denuncia sea más dura.Hace tiempo, cuando trabajaba en el área de comunicación de la Cruz Roja Internacional, me sorprendió una nota de prensa emitida en Ginebra que se titulaba algo así como "No a las donaciones internacionales de medicinas para el tercer mundo"...*

3.1.2. Representación de documentos

Generalmente los documentos en tareas de clasificación y de recuperación de información son representados como una bolsa de palabras. La bolsa de palabras es una representación simple, consiste en un arreglo con el vocabulario completo del problema dónde cada documento es representado como un vector del tamaño del vocabulario. Los números del vector dependen de si el documento contiene o no la palabra del vocabulario [Sebastiani, 2002].

Aunque en clasificación de polaridad no se han hecho muchas variaciones a la repre-

sentación de documentos si se ha buscado aplicar técnicas y representaciones cercanas a *Latent Dirichlet Allocation* (LDA) [Blei et al., 2003]. Una de las representaciones abordadas es *Latent Sentiment Model (LSM)* [Lin et al., 2010]. La representación LSM busca mapear los documentos en una distribución de sentimientos. En la distribución de sentimientos cada palabra es representada por el sentimiento que refleja (positivo, negativo o neutral). Otra representación es *Join Sentiment Topic Model (JST)* [Ramage et al., 2010]. La JST busca representar cada palabra con la carga de sentimiento que contiene y el tópico al que hace referencia. En esta representación la definición del tópico depende del sentimiento de la palabra. Es decir, primero se mapean todas las palabras en la distribución de sentimientos y después se mapea esa representación a una distribución de tópicos. También se ha trabajado de manera contraria mediante la representación de *Reversed Join Sentiment Topic Model* [Lin et al., 2010].

Una representación que actualmente está siendo bastante investigada es la representación vectorial de palabras, de párrafos y de documentos [Mikolov et al., 2013,Le y Mikolov, 2014]. En el caso de la representación vectorial de palabras (o Word2Vec [Mikolov et al., 2013]), cada palabra es mapeada en un vector único, representado por una columna en una matriz W. La columna es indexada por la posición de la palabra en el vocabulario. La suma o unión de varios vectores es utilizada para predecir la siguiente palabra en una sentencia. La idea principal es que aquellas palabras que estén relacionadas tengan vectores muy parecidos. Para aprender los vectores de cada palabra se utilizan grandes bases de datos como los documentos de Wikipedia [Le y Mikolov, 2014]. En el caso de análisis de sentimientos, se han utilizado la base de datos de Stanford (*Stanford sentiment treebank dataset* [Socher et al., 2013]) y la base de datos IMDB [Maas et al., 2011].

La representación vectorial de párrafos sigue la misma idea de representar al párrafo completo en un sólo vector [Le y Mikolov, 2014]. Estas representaciones se utilizan en clasificación de polaridad bajo la suposición de que aquéllos párrafos que tengan carga positiva serán parecidos entre ellos y lo mismo sucederá para los párrafos con carga negativa.

Otros enfoques ampliamente utilizados en tareas de análisis de sentimientos se basan en modificar las representaciones añadiendo información de etiquetas, frecuencia

o medidas que reflejen que tan discriminante es un término. Definir que tan relevante es un término para una clase ha sido bastante estudiado [Aggarwal y Zhai, 2012,Ding y Peng, 2005]. Añadir medidas que brinden la información discriminante de cada término a la representación de documentos es una tendencia en la tarea de clasificación por polaridad [Amir et al., 2014].

Este enfoque es muy relevante en tareas de clasificación de polaridad debido a que se considera que algunos tipos de palabras como verbos, adjetivos o adverbios tienen mayor carga polar que el resto de los términos.

3.1.3. Selección de atributos para clasificación de polaridad

La selección de atributos es un tema importante desde cualquier enfoque de clasificación. Se han utilizado un sin fin de técnicas y métodos de selección de atributos. Los métodos más simples aplican modelos de n-gramas de letras y de palabras [Peng y Schuurmans, 2003]. Los trabajos que incluyen n-gramas de palabras tratan de rescatar atributos estilísticos para determinar la clase de un documento. Usar atributos de estilo es de gran ayuda en tareas de atribución de autoría o perfilado de autores. En el caso de clasificación por polaridad no se han hecho muchos estudios sobre que tan útil puede ser la información de estilo.

Métodos más complejos en cuanto a caracterización de documentos de opinión incluyen el uso de lógica [Nadali et al., 2010]. Los trabajos que utilizan lógica tratan de convertir el documento a una expresión lógica que pueda ser manipulada bajo ese contexto. La utilización de árboles sintácticos de dependencia también tienen gran afluencia en la clasificación de textos, especialmente en textos de opinión. Los documentos son representados mediante las entidades de las oraciones y sus relaciones directas con el resto de las palabras dentro de las sentencias [Jiménez et al., 2015]. Existen varios recursos externos con árboles de dependencia de oraciones muy comunes en textos de opinión [Socher et al., 2013,Kramer y Gordon, 2014]. Otra forma de caracterización muy popular para documentos de opinión es la inclusión de características de etiquetado POS (Parts Of Speech) [Gamallo y Garcia, 2014].

Como se mencionó en puntos anteriores, la selección de características o atributos

es fundamental en las tareas de clasificación de textos. En categorización por polaridad hay dos tipos de selección de atributos fundamentales. La primera es selección de atributos según el tópico o el tema sobre el cuál hable el documento. Recuperar las palabras claves de cada tópico es importante en clasificación de polaridad porque conociendo el tema es posible asignar la orientación de polaridad a aquellas palabras que pueden resultar con sentimientos encontrados. Es decir, palabras que son positivas en una temática pero resultan negativas en otra [Lee et al., 2004].

El segundo punto importante para hacer selección de atributos en tareas de clasificación de polaridad es la tarea misma. La información subjetiva se encuentra comúnmente en palabras que contienen carga de positividad o negatividad. Las palabras con carga de polaridad generalmente tienen las etiquetas de adjetivos, verbos y adverbios. Por esta razón es común hacer una selección de atributos basándose en estas etiquetas POS [Gamallo y Garcia, 2014].

Sin importar que se traten de documentos de opinión, la selección de características también puede ser métodos más tradicionales como Ganancia de Información, Información Mutua y Mínima Redundancia Máxima Relevancia [Agarwal y Mittal, 2013].

3.1.4. Enfoques de clasificación

Algunos trabajos relevantes de los dos tipos de enfoques de clasificación de polaridad más utilizados se exponen en las siguientes secciones.

Clasificación basada en diccionarios

La clasificación basada en lexicones o diccionarios funciona con listas de palabras con cargas de polaridad. Se cuenta con dos listas de palabras: una de palabras positivas y otra de palabras negativas. Estas listas de palabras pueden ser sólo palabras y su orientación semántica o bien contener los grados de positividad o negatividad. La clasificación de polaridad depende de si hay más presencia de palabras positivas o más negativas en el texto. Bajo esta idea se han hecho muchas aproximaciones a la clasificación de textos de opinión [Taboada et al., 2011, Jiménez et al., 2015, González et al., 2015].

Las variaciones dentro de este tipo de enfoque de clasificación han seguido la vertiente de hacer modificaciones a los diccionarios más que modificar el sistema de clasificación. Inicialmente, se tomaron listas en Inglés para hacer los conteos de positividad o negatividad de los documentos, una de las listas más utilizadas es SentiWordNet [Esuli y Sebastiani, 2006]. SentiWordNet ha sido tomada como base para realizar listas en otros idiomas, especialmente en Español. Haciendo uso de traducciones automáticas y correcciones manuales han surgido listas como Crisol [González et al., 2015]. Tanto en Inglés cómo en Español se han desarrollado listas que son dependientes de cada dominio. Este tipo de listas dependientes de dominio ha sido más común en Español, por ejemplo la lista eSol [Jiménez et al., 2015] creada para el dominio de cine o la lista eSOLHotel para el dominio turístico [González et al., 2015]. Además de realizar traducciones se han realizado trabajos que tratan de aprender las listas en nuevos idiomas partiendo de listas en Inglés y diccionarios bilingües [Pérez et al., 2012].

En cuando al método de clasificación basada en diccionarios no hay muchas variaciones además del desarrollo de listas dependientes de cada dominio o de traducciones a distintos idiomas, la base de todos estos métodos es el conteo de palabras de cada una de las listas de terminos distintivos de cada clase. Los posibles cambios en la decisión de etiqueta de los documentos depende de la inclusión de eventos como la negación o la intensificación. Ambos eventos son explicados con detalle en las siguientes subsecciones.

Clasificación basada en aprendizaje automático

La clasificación basada en aprendizaje automático ha sido bastante estudiada en muchas aplicaciones de la clasificación automática [Dobre et al., 2007]. En el caso de clasificación de textos, uno de los métodos más utilizados son los enfoques probabilistas o estadísticos, específicamente, enfoques y variaciones de Naive Bayes [Peng y Schuurmans, 2003, Kibriya et al., 2004, Lin et al., 2010]. La razón para utilizar Naive Bayes es porque es un método sencillo de implementar y de modificar para ser adaptado a distintas particularidades del lenguaje. Sin embargo, el problema de clasificación de textos también se ha atacado con Máquinas de Soporte Vectorial, redes neuronales, etc. [Pang et al., 2002].

A continuación se exponen algunas de las aplicaciones y variaciones de Naive Bayes en el contexto de clasificación de documentos según su polaridad.

Modificaciones a Naive Bayes

El clasificador Naive Bayes es uno de los más utilizados para categorización de textos. Han sido muchas las problemáticas atacadas desde este enfoque probabilista. Entre las variaciones más relevantes para este trabajo se encuentran el Naive Bayes Bernoulli y el Naive Bayes Multinomial. La diferencia sobresaliente entre estas dos variaciones es que en el Multinomial se toma en cuenta la frecuencia de las palabras en el cálculo de las probabilidades del vocabulario y en la toma de decisión de a qué clase pertenece un documento. En el caso del Bernoulli solamente se toma en cuenta la presencia de una palabra o su ausencia, es decir que el valor de una palabra en la representación sólo puede tomar los valores de 0 o 1. En el caso de clasificación de textos de opinión no hay un estándar sobre qué variación de Bayes utilizar. Algunos trabajos defienden que es importante saber el número de apariciones de una palabra con carga polar [Kibriya et al., 2004], mientras otros alegan que saber cuántas veces apareció una palabra no brinda información relevante [Narayanan et al., 2013].

El método de Naive Bayes que ha recibido mayor atención es el Naive Bayes Multinomial. Uno de los principales problemas de este variación es que al tomar en cuenta la frecuencia de las palabras se pierde el valor de aquellos términos que aunque son muy distintivos de una clase son poco frecuentes. Para manejar esta problemática se han propuesto varias formas de modificar el elemento de frecuencia en la expresión de clasificación de NBM. Una de las principales ideas es la inclusión de suavizado a las frecuencias de las palabras. Por ejemplo en la expresión $f' = log(1 + f)$ se propone la inclusión de una funsión logaritmica para disminuír los valores de la frecuencia de las palabras [Schneider, 2005]. La idea de utilizar logaritmos se ha usado tambien por otros autores, otra idea fue la expresión $f' = 1 + logf$ [Mendoza et al., 2011]. Otra forma de suavizado de frecuencias fue el utilizado por [Qiang, 2010]. En ese trabajo se añadió además de la funsion logaritmica una decisión, la expresión utilizada fue $f' = min\{1 + log_2^f, f\}$, con esto el método de clasificación quedó definido como se muestra en la expresión 3.1.

$$P(c_i|d) = \frac{\prod_{w_j \in d}^{|d|} P(w_j|C_i)^{min\{1+log_2^{f_{w_j}}, f_{w_j}\}} * P(c_i)}{P(d)} \tag{3.1}$$

Otras variaciones dentro del cálculo de probabilidades de palabras a cada clase son la inclusión de TF-IDF [Kibriya et al., 2004], información mutua [Narayanan et al., 2013], y modificaciones en el suavizado [Rennie et al., 2003]. En cuanto a modificaciones de suavizado se propone usar la expresión 3.2 para calcular la frecuencia de las palabras y con ello ingresar un efecto de suavizado. f_{w_j} es la frecuencia de la palabra j en la clase y f el total de frecuencias de las palabras dentro de la clase. El elemento α_i es un valor a priori para la frecuencia de la palabra. Este valor puede ser distinto para cada una de las palabras. α es la suma de todos los valores a priori.

$$f'_{w_j} = \frac{f_{w_j} + \alpha_i}{f + \alpha} \tag{3.2}$$

En el mismo trabajo se emplea la regla de clasificación de mínimo error y se llega a representar a NBM con la expresión 3.3. f_{w_n} Es la frecuencia de la palabra n dentro del documento.

$$P(C_i|d) = log(P(C_i)) + \Sigma_n(f_{w_n} log \frac{f_{w_j} + \alpha_i}{f + \alpha}) \tag{3.3}$$

Otro tipo de variaciones a NB se ha realizado sobre el vocabulario de trabajo. Se ha propuesto hacer una separación de los documentos en varios subconjuntos, teniendo un documento X se puede dividir en varios componentes X^1, X^2, X^C. C podría ser = 1. Los componentes o subconjuntos del documento pueden tener distintos tamaños $N^1, N^2 ... N^c$ [Shen y Jiang, 2003]. Con esa separación se puede remplazar a NB tradicional por la expresión 3.4.

$$P(c_i|d) = \prod_{c=1}^{C} (\prod_{j=1}^{N_c} P(w_j^c|c_i)^{\beta_c/N_c}) \tag{3.4}$$

La idea de separar los documentos y vocabularios en varios subconjuntos se propone para dar cierto pesado a cada conjunto, esto con la idea de que algunas partes del documento pueden estar más relacionadas a la etiqueta. Por otro lado, se defiende y hace más fuerte la suposión de independencia entre los términos que tiene NB [Shen y Jiang, 2003]. En la expresión 3.4, β_c permite dar la ponderación a el subconjunto y el denominador N_c funciona como una normalización del exponente.

3.2. Tratamiento de la negación

El estudio de la negación en textos de manera automática surgió en tareas con documentos médicos [Zou et al., 2014,Abu-Jbara y Radev, 2012]. Los estudios sobre la negación se basan en reglas lingüísticas [Zou et al., 2014]. El tratamiento de la negación tiene tres puntos fundamentales: 1)Detección de las señales o banderas, 2)Definición del alcance y 3)Identificación del enfoque. Las señales o banderas son las palabras o letras que nos permiten identificar cuándo hay una negación dentro de una expresión. El alcance es la parte de la oración que es afectada por la negación. El enfoque es la parte del alcance que es más fuertemente negado. Identificar los tres puntos del tratamiento de la negación han sido atacadas desde varias perspectivas.

Encontrar las señales de negación es el punto más importante del tratamiento de la negación. Se han utilizado dos métodos de identificación de la negación [Abu-Jbara y Radev, 2012,Morante y Blanco, 2012,Read et al., 2012]. El primer método es el más simple. Se trata de utilizar diccionarios de palabras negativas. Las palabras negativas están definidas por lingüistas o por desarrolladores. Algunos ejemplos de palabras negativas en Español son: no, ni, sin, etc.. El segundo método es la identificación de prefijos y sufijos en las palabras, por ejemplo, in y dis. Identificar la negación con prefijos y sufijos es complicado porque existen palabras que inician o terminan con esta combinación de letras pero no contienen ninguna negación. Algunos ejemplos de palabras con prefijos negativos son: infeliz y disfuncional. Por otro lado, algunos ejemplos de palabras que inician con estas letras pero no contienen negación son: interno y discurso.

Los siguientes puntos son la identificación del alcance y del enfoque. Ambas tareas son comúnmente atacadas bajo las mismas líneas. Los métodos más utilizados se

enlistan a continuación.

- Ventanas.- Consiste en definir un número de palabras que serán afectadas después de encontrar la bandera de negación [Narayanan et al., 2013, Abu-Jbara y Radev, 2012]. La ventana puede ser interrumpida si aparece un signo de puntuación ya que índica que la oración termina. Otra variación de las ventanas es que no se tomen en cuenta todas las palabras de la oración. La palabras vacías suelen ser ignoradas. En la figura 3.1 Se encuentra un ejemplo de afectación por ventana y se detiene el alcance por final de la oración.

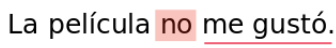


Figura 3.1: Definición del alcance mediante ventanas.

- Semántica proposicional.- En este enfoque las oraciones que contienen una negación son transformadas a expresiones lógicas y operadas sobre reglas lógicas. La definición del alcance depende de los resultados de las operaciones lógicas aplicadas al enunciado [Packard et al., 2014].

- Reglas.- La definición de reglas para conocer el alcance y el enfoque de la negación es comúnmente utilizado cuando se tienen oraciones representadas en árboles de dependencia. Las reglas definen qué ramas del árbol serán afectadas por la señal de negación [Read et al., 2012, Jiménez et al., 2015]. Por ejemplo: utilizando las reglas usadas en [Jiménez et al., 2015], un no afecta al nodo padre y al árbol formado por el hermano derecho del nodo de negación. El árbol quedaría afectado como se muestra en la figura 3.2 dónde las palabras subrayadas con rojo son las afectadas.

- Patrones sintácticos.- Otro método es convertir las oraciones a representaciones sintácticas. Es decir, las oraciones no son representadas por las palabras si no por las etiquetas sintácticas o etiquetas POS de cada palabra [Blanco y Moldovan, 2011]. Se definen reglas sobre los patrones sintácticos para definir qué etiquetas serán afectadas por la negación.Lo más común es afectar a las palabras que tengan etiqueta de adjetivo, verbo o adverbio. En la figura 3.3 se encuentra el ejemplo del alcance con esta regla, solamente el verbo es afectado.

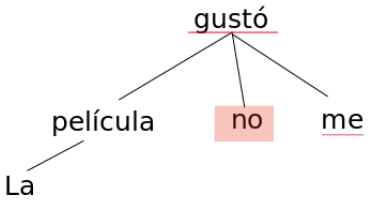


Figura 3.2: Definición del alcance mediante reglas y árboles de dependencia.

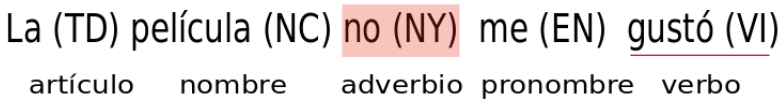


Figura 3.3: Definición del alcance con etiquetas POS.

- Modelo de grafos.- Otra representación para sentencias es el modelo de grafos. Cada expresión es un grafo. Los nodos del grafo representan las palabras de contenido de la expresión (se omiten las palabras vacías) y las ramas que conectan los nodos representan la correlación entre las palabras. Los pesos de las ramas o correlación entre palabras es la probabilidad de transición de un nodo a otro [Zou et al., 2014]. Bajo este método se asume que toda la oración es afectada cuando se encuentra una negación y que el enfoque de la negación es aquél nodo que tenga más correlación con el resto de los nodos.

El manejo de la negación en tareas de análisis de sentimientos toma las mismas ideas del tratamiento de la negación en general [Hogenboom et al., 2011]. La negación tiene especial atención en documentos de opinión por su capacidad de modificar completamente los valores de verdad o el significado de una expresión. Tomando en cuenta que la mayor parte de la investigación sobre clasificación de polaridad se basa en las palabras positivas y negativas que se encuentren en un documento, el correcto tratamiento de la negación resulta un punto bastante importante.

Varias investigaciones han tratado de medir que tan importante es el tratamiento de la negación en las tareas de análisis de sentimientos [Zhu et al., 2014, Wiegand et al., 2010, Lapponi et al., 2012]. Todos ellos concluyen que es de suma importancia tomar en cuenta los negadores o partículas negativas debido a que tienen la capacidad

de invertir o modificar completamente el valor de polaridad de una frase de opinión. Entre los problemas más importantes de manejar la negación se encuentran la inclusión de información de negación en la representación de los documentos [Taboada et al., 2011, Lapponi et al., 2012] y el efecto que puede causar una negación dentro de la decisión de etiqueta de un documento [Councill et al., 2010, Zhu et al., 2014]. Una de las representaciones más comunes es la bolsa de palabras y en esta representación no se mantiene el orden de las palabras por lo que resulta muy difícil reconocer qué palabras fueron afectadas por alguna partícula negativa. Por otro lado, el efecto que pueden causar las negaciones también es un problema debido a qué no se tiene claro qué grado de influencia tienen las negaciones en una oración. Podría asumirse que cuándo negamos algo estamos tratando de decir lo contrario o bien que sólo moderamos un poco el valor de verdad en una oración. Para que quede más claro podemos recordar el ejemplo de la sección 1: *"La película no es aburrida"* si invertimos el significado entenderíamos la oración como *"La película es divertida"*, sin embargo, la mayoría de nosotros entenderíamos que la *"La película es pasable o normal"*. En el último ejemplo moderamos el cambió al significado de la oración.

La negación ha sido atacada dentro del marco de clasificación de polaridad con enfoques de lexicones y de aprendizaje.

3.2.1. La negación y el enfoque de lexicones

El primer problema al tratar la negación es cómo representar que hay una negación presente en un documento de texto. Sin embargo bajo el enfoque de listas no es necesario una modificación al documento puesto que generalmente sólo se realizan conteos de palabras [Taboada et al., 2008, Jiménez et al., 2015, González et al., 2015]. Por ejemplo: bajo este enfoque se tienen listas de palabras positivas y negativas, se buscan estas palabras dentro del documento a clasificar y se cuentan cuántas hay de cada clase. En caso de que la palabra esté en un contexto negado o sea afectada por una negación se cuenta en la clase contraria a la que pertenece inicialmente.

Con ese método de clasificación se resuelve también el problema de afectación por negación. Definir qué tipo de efecto tendrá la negación sobre las palabras que resulten marcadas. Aunque la inversión del significado y con ello de la polaridad de la expresión

puede ser lo menos intuitivo en una conversación real, es el enfoque más utilizado. Invertir la polaridad de una palabra cuando ésta sea alcanzada por una negación ha sido utilizado por la mayoría de los trabajos de clasificación de opiniones basados en listas [Taboada et al., 2011, Jiménez et al., 2015]. La idea resulta sencilla de aplicar, se asume que las palabras positivas tienen un valor de $+1$ y las palabras negativas un valor de -1. Al encontrarse una negación se invierten los signos de los valores.

Además de invertir la polaridad, también se han desarrollado enfoques que modifiquen los valores de polaridad de las palabras sin llegar a invertirla. Una de las formas de modificar la polaridad es definiendo valores de afectación [Wiegand et al., 2010]. Es decir, elegir cuánto modificará la polaridad un *no*, cuánto un *ni*, y cuánto cada una de las partículas negativas definidas. Este tipo de métodos dependen mucho de las ideas y consideraciones de quienes lo desarrollen ya que no está definida ninguna lista oficial de valores de modificación.

3.2.2. La negación y el enfoque de aprendizaje

En el caso de métodos basados en aprendizaje si resulta fundamental hacer una modificación a los documentos para ingresar información sobre la afectación por una negación. Esto se debe a que las representaciones comunes suelen perder el orden de las palabras y es necesario poder identificar qué palabras han sido afectadas por una negación para poder tratarlas.

Agregar información de la negación a la representación del documento ha sido atacada principalmente agregando marcaciones a las palabras que tienen alguna influencia negativa [Narayanan et al., 2013, Hogenboom et al., 2011, Zhu et al., 2014]. Dicho de otra manera se agrega una marca o bandera a aquellas palabras que resulten afectadas por una partícula negativa. Por ejemplo, en la oración *"La película no me gustó"*, si asumimos que las palabras *"me"* y *"gustó"* son las afectadas por la partícula negativa modificaríamos el documento de la siguiente forma: *"La película no no_ me no_ gustó"*. Con esta modificación tendríamos el conocimiento de cuáles palabras están negadas. Para hacer esta modificación de documentos se asume que ya se conoce cuál fue el alcance de cada negación y se modifican aquellas palabras que sean parte del alcance. Para definir el alcance se utiliza alguno de los métodos mencionados en

la subsección anterior *Tratamiento de la negación.*

En el caso de métodos de clasificación de polaridad con enfoque de aprendizaje no se tienen listas de palabras positivas y negativas ni valores únicos de cada palabra, por esta razón se ha utilizado el efecto espejo para hacer el pesado de aquellas palabras que se encuentren en contextos negados. Se asume que si se encuentra una palabra negada en una clase, esta palabra debería pertenecer a la clase contraria, por lo que, se le asigna los valores de polaridad de la clase contraria a la que está siendo clasificada [Narayanan et al., 2013]. La asignación de pesos o valores de las palabras marcadas con negación se calculan como se muestra en el siguiente ejemplo:

- **Palabra: Bonita**

  Valor positivo: 0.7

  Valor negativo: 0.3

- **Palabra: No_Bonita**

  Valor positivo: 0.3

  Valor negativo: 0.7

Se hacen los cálculos de los valores de las palabras normales para después hacer la asignación a las palabras negadas. Llamo palabras normales a aquellas que no tienen ninguna afectación por negación y negadas a las que están marcadas con *"No_"*.

3.3. Intensificadores y Atenuantes

Los intensificadores y atenuantes son palabras que pueden aumentar o disminuir el valor de polaridad de una palabra o expresión. Por ejemplo, es claro que *"Es muy bonita"* es más positivo que *"Es bonita"*. Los intensificadores son palabras que aumentan la fuerza positiva o negativa de una expresión. Por otro lado, los atenuantes son las palabras que los disminuyen. Por ejemplo, *"Es un poco agresivo"* tiene menor grado de polaridad negativa que *"Es agresivo"*.

Los intensificadores y atenuantes han sido poco estudiados en áreas distintas a la clasificación de polaridad. Sin embargo, también han sido tomadas en cuenta en otras problemáticas como detección de sarcasmo e ironía [Liebrecht et al., 2013].

Algunas investigaciones proponen no tomar la clasificación de polaridad como un problema binario si no otorgar valores de positividad y negatividad de un documento.

Este enfoque sólo se ha propuesto pero no ha sido desarrollado hasta el momento. Sin embargo, con esta idea se comenzaron a utilizar los intensificadores y atenuantes en la clasificación de opiniones. Aunque no se han reportado trabajos que muestren niveles de polaridad si se ha agregado la utilización de intensificadores y atenuantes a la decisión de etiqueta de documentos de opinión.

Los intensificadores y atenuantes tienen su nivel de afectación tal como la negación. Este nivel o alcance de afectación se define por medio de ventanas. Una ventana es el número de palabras que será afectada después de la aparición de un intensificador o atenuante [Taboada et al., 2011, Kennedy y Inkpen, 2006]. Una vez definido el alcance de la intensificación o atenuación se define el grado de afectación que se tendrá en las palabras alteradas.

3.3.1. Intensificadores y atenuantes y el enfoque de lexicones

Los intensificadores y atenuantes representan los mismos problemas que la negación: representación y afectación. En enfoques de diccionarios no es necesaria una representación de la aparición de intensificadores o atenuantes debido a que sólo se toman en cuenta las palabras de las listas. Se identifica si hay una palabra afectada en el documento a clasificar al encontrar una palabra del lexicón y ver su contexto [Taboada et al., 2008].

En cuanto a la afectación que puede tener una palabra alcanzada por un intensificador o un atenuante hay dos principales tipos de afectación estudiadas hasta el momento, ambas basadas en el incremento y decremento de los valores de las palabras afectadas. El primer tipo de afectación es aumentar o disminuir un número definido por el investigador, se han utilizado números de 1 y 2 [Polanyi y Zaenen, 2006, Kennedy y Inkpen, 2006]. Es decir, si la palabra *"bonita"* tiene un valor de 2 tendría un valor de 3 si se afectara por un intensificador, o un valor de 1 si se viera alcanzada por un atenuante. La otra idea de afectación es muy similar, sólo que se tienen números específicos para cada intensificador y atenuante [Taboada et al., 2011]. Estos números son también definidos por el desarrollador, esto quiere decir que no hay una lista general sobre cuáles son los intensificadores y los atenuantes y cuáles son sus grados de afectación.

El hacer un tratamiento de los intensificadores en este tipo de enfoque genera una mejora significativa en los resultados de clasificación, aunque en general se utilizan en conjunto con otros elementos de clasificación como selección de atributos [Taboada et al., 2008].

3.3.2. Intensificadores y atenuantes y el enfoque de aprendizaje

En el caso de tratamiento de intensificadores y atenuantes para clasificación de polaridad desde enfoques de aprendizaje computacional hay pocos trabajos reportados. Se han encontrado algunos trabajos que hacen modificación del documento marcando las palabras que han sido alcanzadas por un intensificador o un atenuante [Kennedy y Inkpen, 2006]. Sin embargo, el tratamiento es pobre ya que después de hacer la modificación sólo son utilizados como una característica extra dentro de la representación del documento.

3.4. Discusión

En la tabla 3.1 se resumen algunos de los trabajos más importantes y que tienen mayor relación con este trabajo de tesis. Los trabajos presentados utilizan negación o intensificación e incluso ambas. Hay algunos trabajos con enfoques híbridos. Los trabajos con enfoque híbrido usan diccionarios pero esos diccionarios son calculados con los documentos de entrenamiento.

Entre las diferencias y ventajas más destacadas de la propuesta podemos mencionar las siguientes: 1) Este trabajo propone métodos de clasificación con enfoque de aprendizaje computacional, la tarea de clasificación de polaridad es poco estudiada desde este enfoque y es importante generar aportes distintos a los enfoques de diccionarios. 2) Se hicieron experimentos en varias colecciones que tienen diferencias de temática, de tipo de documentos, tamaño, idioma, etc. Esto nos permite probar la robustez de los métodos propuestos e identificar algunas particularidades de cada corpus. 3) Se plantea un tratamiento de negación simple lo que hace que el método sea más sencillo, además se propone una nueva forma de tratar las palabras afectadas

por negación. 4) Se hace un tratamiento a la intensificación lo cual ha sido mayormente estudiado desde enfoques de diccionarios. Y 6) Se plantea utilizar todas las palabras de los documentos por dos razones: a) para probar el método manteniendo los documentos en su forma natural y b) para evitar hacer un método pesado con transformaciones de documentos.

En general, la propuesta de este trabajo es más completa que varias de las hechas en el estado del arte, por las pruebas realizadas en varias colecciones y la inclusión del tratamiento de la intensificación y la negación..

Autor	Enfoque	Base de datos	Negación (alcance y tratamiento)	Intensificación	Atributos
[Kennedy y Inkpen, 2006]	Diccionarios	IMB v2..0	Árboles de dependencia Cambio de signo	Cambio de valores +1 0 -1	Lemas y POS de las palabras del diccionario
[Councill et al., 2010]	Diccionarios	Bd propia	Árboles de dependencia Cambio de signo	NO	Palabras del diccionario
[Taboada et al., 2011]	Diccionarios	SFU (inglés)	Resto de la oración Cambio de signo	Cambio con lista de valores	Adjetivos+Verbos+Adverbios
[Hogenboom et al., 2011]	Diccionarios	IMB v2..0	Ventana y resto de la oración Cambio de signo	NO	Lemas y POS
[Lapponi et al., 2012]	Diccionarios	IMB v2..0	Reglas Cambio de signo	NO	Palabras del diccionario
[Narayanan et al., 2013]	Aprendizaje	IMB	Ventana Efecto espejo	NO	Selección de atributos por información mutua
[Zhu et al., 2014]	Híbrido	Stanford sentiment treebank	Árboles de dependencia (reglas por cada negador) Cambio de signo	NO	Todas las palabras
[Jiménez et al., 2015]	Diccionarios	CMR	Árboles de dependencia (reglas por cada negador) Cambio de signo	NO	Palabras del diccionario
Este trabajo	**Aprendizaje**	CMR, COAH IMB v1.0, SFU	Ventana Cálculo de valores	Ventana Cálculo de valores	Todas las palabras del vocabulario

Tabla 3.1: Negación e intensificación en el estado del arte.

Las colecciones utilizadas son en su mayoría en Inglés, excepto por la de CMR que es un grupo de críticas de películas escrito en Español [Cruz et al., 2008]. Las versiones de IMB son críticas de cine en Inglés y la versión 2.0 es un subconjunto del total de críticas en IMB [Pang et al., 2002].

En este trabajo se analizan los diferentes métodos de clasificación y se añaden tratamientos de la negación y la intensificación para mejorar los grados de exactitud. El problema de clasificación de polaridad y de tratamiento de la negación ha sido anteriormente atacados pero en su mayoría como dos problemas diferentes. Además, ambos problemas han sido abordados con especificaciones de idioma y de dominio. En el caso del idioma, la mayoría de las investigaciones que abordan alguna de las problemáticas que fueron expuestas en este capítulo fueron desarrolladas para el idioma Inglés. Por otro lado, los trabajos que abordan los problemas desde otros lenguajes han hecho sistemas o utilizado enfoques que hacen necesario modelar el problema para un tópico en específico. El presente trabajo de tesis busca sobrepasar las fronteras de idioma y de dominio desarrollando un método de clasificación que sea competitivo para distintos idiomas y diferentes dominios. Se pretende tener una mínima dependencia de idioma, esa dependencia sería las listas de partículas negativas e intensificadores de cada idioma.

La relación principal de este trabajo con el estado del arte se encuentra con él trabajo de [Narayanan et al., 2013] ya que utiliza un enfoque de aprendizaje probabilista para la clasificación de polaridad y el tratamiento de la negación. El tratamiento realizado en este trabajo es el descrito en la sección "La negación y el enfoque de aprendizaje". La mayor diferencia con este trabajo es que los valores asignados a las palabras que aparecen en contextos negados son calculados según su presencia en la clase y no con el efecto espejo. Se buscó no usar el efecto espejo debido a que basa su funcionamiento en negar los términos existentes y agregarlos al vocabulario, y asignar valores a las palabras dependiendo de su aparición en otra clase. Con esto se generan dos problemas, el primero es que e crean muchos términos negados que realmente no aparecen en ninguna etapa de la clasificación y segundo, la asignación de valores a las palabras puede ser muy distante a la relación que esas palabras negadas pueden tener con las clases. Además aquí se trabajó con el NBM y no con la versión Bernoulli. Por último otra diferencia es que también se hace un tratamiento a los intensificadores y

atenuantes.

La principal y más importante diferencia con el resto del estado del arte, independientemente de si se trata de enfoques basados en diccionarios o en aprendizaje computacional, es que se busca una independencia de dominio y de idioma, y con ello se busca que el método desarrollado sea competitivo con cualquier tipo de texto. Que el sistema sea más general es una ventaja en realización de tareas automáticas ya que disminuye el tiempo de desarrollo y los esfuerzos innecesarios para expertos de cada temática. Sin embargo, lograr una cobertura de un gran número de idiomas y de tópicos no es una tarea sencilla de realizar debido a las diferencias léxicas y sintácticas de los diferentes idiomas. Así como los distintos vocabularios y expresiones utilizadas para hablar de todos los diferentes temas sobre los cuáles se puede emitir una opinión.

Otros puntos más específicos sobre los cuales se desarrolla este trabajo de tesis es el modelado de la negación y de la intensificación en la representación de los documentos y en la toma de decisión de clasificación. Se proponen algunas variaciones al algoritmo Naive Bayes y a las fórmulas de cálculos de probabilidades de las palabras para permitir que la información de negación e intensificación sea de mayor ayuda en la clasificación de polaridad.

CAPÍTULO 4

MÉTODOS PROPUESTOS PARA LA CLASIFICACIÓN DE POLARIDAD

En este capítulo se describen dos métodos propuestos para realizar clasificación de polaridad: uno con enfoque híbrido y el segundo con enfoque probabilista. Los métodos descritos incluyen un tratamiento de la negación y la intensificación.

El pre procesamiento de los datos, así como el tratamiento de la negación y la intensificación son pasos que se pueden añadir a los dos métodos de clasificación propuestos de manera similar, por lo que dichos pasos serán detallados previo a definir las particularidades de cada método.

4.1. Pre procesamiento de los datos

La etapa de pre procesamiento de los documentos es realmente necesaria. La finalidad de hacer una limpieza en los datos es ponerlos en un formato uniforme que permita que sea más fácil manipularlos y evitar posibles errores o ruido en la clasificación.

El pre procesamiento de los datos constó de la eliminación de palabras vacías, números y algunos espacios. La lista de palabras vacías es descrita en el capítulo 5 en la sección de recursos externos. El último paso del pre procesamiento fue convertir todo

el texto a minúsculas para eliminar palabras que fueran diferentes en el vocabulario debido a las diferencias de fuentes.

4.1.1. Representación de documentos

La representación de bolsa de palabras o Bag of words (BOW) es una de las más utilizadas en clasificación de documentos. Consta de un vector que representa a todas las palabras que existan en el vocabulario. Esas palabras pueden ser representadas por valores de ocurrencia o frecuencia. Los valores de ocurrencia indican si la palabra apareció o no en el documento, mientras que el valor de frecuencia indica cuántas veces ocurrió la palabra en el documento. En este trabajo se usó la representación de frecuencia.

Al referirnos a las palabras del vocabulario podemos referirnos a distintas cosas. El vocabulario está compuesto de palabras o términos. Los términos son las características que vamos a utilizar para nuestra representación. Comúnmente, se utilizan solamente las palabras tal cuál aparecen en los textos, sin embargo se pueden añadir más elementos al vocabulario como puede ser etiquetas POS, características representativas o incluso representaciones de n-gramas. Las características representativas pueden ser elementos que nos indiquen si aparece algún fenómeno lingüístico en el texto, si contiene negaciones, si contiene más de una etiqueta, etc. Los n-gramas son términos compuestos por secuencias de palabras o de letras. Como base en este trabajo se usaron unigramas de palabras, es decir, las palabras tal cual aparecieron en los documentos, a excepción de palabras vacías o números y demás elementos eliminados en la etapa anterior. Además se usaron n-gramas de palabras con n de tamaño 2 con la finalidad de marcar los términos o palabras que se vieron afectadas por una negación o bien, por una partícula de intensificación o atenuación.

4.2. Tratamiento de la negación

En este trabajo el "tratamiento de la negación" se refiere a la identificación de la presencia de partículas negativas y a la definición del alcance que cada partícula tiene

dentro de una oración.

El primer paso, la identificación, se realizó con la ayuda de diccionarios de partículas negativas. Dicho diccionario fue obtenido de la RAE en caso de Español y en caso del Inglés se obtuvo una lista de negaciones de trabajos anteriores. Ambas listas son descritas en el capítulo 5. El segundo paso es la definición del alcance que las palabras negativas causan en las oraciones. Para definir el alcance se utilizaron las siguientes reglas.

- Las palabras son consideradas dentro del alcance de una negación hasta que:
  - Se encuentre un signo de puntuación.
  - Se encuentre un nexo adversativo.
  - Se encuentre otra negación.
  - Se cumpla una ventana. La ventana es un número de palabras después o antes de la negación.

Aunque todos esos puntos han sido aplicados para definir el alcance de una partícula negativa, no se han aplicado de manera simultánea. En este trabajo se propone un algoritmo que une las reglas usadas para definir el alcance de la negación. El algoritmo 1 muestra como se unieron esos puntos.

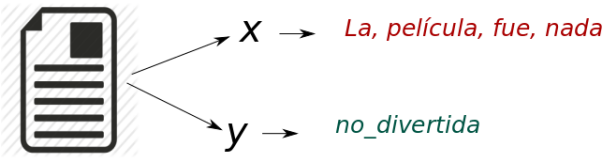


Figura 4.1: Representación de los documentos incluyendo negación.

Algoritmo 1 Algoritmo de definición del alcance de la negación.

```
1: Leer el documento
2: Definición de listas de palabras: Negativas y Normales
3: for Cada palabra en el documento do
4:    if Es una partícula negativa then
5:       for ventana=0 hasta 4 do
6:          if La siguiente palabra no es una negación ni es un signo de puntuación ni
           es un nexo adversativo then
7:             Marcar la palabra como afectada por negación
8:             Agregar la palabra a la lista Negativas
9:          else
10:            Agregar la palabra a lista Normales
11:            Terminar ciclo y continuar con la siguiente palabra en el documento
12:         end if
13:      end for
14:   end if
15: end for
```

Uno de los puntos de mayor importancia dentro del algoritmo es la definición de dos listas de palabras. Las palabras "Negativas" son palabras que se encontraron dentro del alcance de una negación, mientras las palabras "Normales" son aquellas que no han sido afectadas por ningún elemento. La finalidad de estas dos listas es poder representar al documento mediante la unión de dos listas de palabras. La representación se ilustra en la figura 4.1

Se eligió una ventana de tamaño 4 para realizar todos los experimentos de negación debido a que después de realizar algunas pruebas con distintos tamaños de ventana en una partición de las colecciones en Español resultó mejor la de ese tamaño. Estos experimentos se encuentran en el apéndice A.

Los nexos adversativos utilizados en el caso del español fueron: pero, aunque, sin embargo, no obstante,si no y a pesar de. Esta lista de nexos se obtuvo haciendo un conteo de los adversativos más comunes en el corpus CMR. En el caso del Inglés se utilizaron las traducciones.

4.3. Tratamiento de la intensificación

Los intensificadores y atenuantes son también un fenómeno lingüístico ampliamente utilizado en todos los idiomas. En el caso de documentos de opinión la intensificación y la atenuación son importantes para definir la polaridad o incluso para graduar la polaridad de una expresión.

Algoritmo 2 Algoritmo de definición del alcance de la intensificación y atenuación.

```
1: Leer el documento
2: Definición de listas de palabras: Intensificadas, Atenuadas y Normales
3: for Cada palabra en el documento do
4:   if Es un intensificador o un atenuante then
5:     for ventana=0 hasta 2 do
6:       if Es un intensificador y la siguiente palabra no es signo de puntuación
         then
7:         Marcar la palabra como afectada por intensificación
8:         Agregar la palabra a la lista Intensificadas
9:       else if Es un atenuador y la siguiente palabra no es signo de puntuación
         then
10:        Marcar la palabra como afectada por atenuación
11:        Agregar la palabra a la lista Atenuadas
12:      else
13:        Agregar la palabra a la lista Normales
14:        Terminar ciclo y continuar con la siguiente palabra en el documento
15:      end if
16:    end for
17:  end if
18: end for
```

Se tomaron en cuenta los intensificadores y atenuantes descritos en el capitulo 5 en la sección de recursos externos. Para poder tener información de la afectación que tienen los intensificadores se utilizó el algoritmo 2. El algoritmo funciona de manera muy parecida al algoritmo 1 que incluye información de la negación. Se busca modificar el documento ingresando palabras marcadas con intensificación o atenuación. Se

afectan las palabras que se encuentren dentro de una ventana definida o palabras que se encuentren antes de finalizar la oración. En este trabajo se usó una ventana de tamaño 2. La definición de la ventana se realizó mediante experimentación con distintos tamaños de ventana donde la de tamaño 2 resultó brindar mejores resultados.

Una vez aplicado el algoritmo a los documentos de entrenamiento y prueba se obtienen nuevos documentos con etiquetas de intensificación y atenuación. Los documentos modificados quedan como el siguiente ejemplo:

- Frase original: *" La película es muy buena, aunque el actor principal es poco interesante"*
- Frase modificada: *" La película es muy INT_buena, aunque el actor principal es poco ATE_interesante"*

Una vez modificado el documento con la información de intensificación podemos representar al documento como la unión de varios conjuntos de palabras. Los conjuntos de palabras posibles en esta representación son 3: 1) Las palabras que no fueron afectadas por ninguna intensificación ni atenuación, 2) Las palabras que están marcadas con una intensificación y 3) Las palabras que están marcadas con una atenuación. En la imagen 5.3 se muestra la representación de un documento que contiene los tres grupos de palabras.

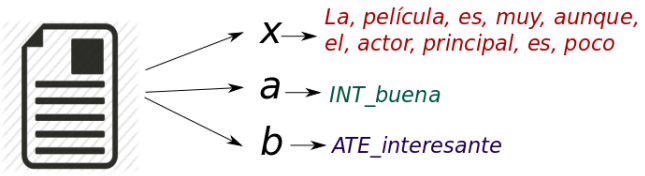


Figura 4.2: Representación de los documentos incluyendo intensificación y atenuación.

4.4. Clasificación

Como se mencionó en los capítulos anteriores, la clasificación de la polaridad tiene dos vertientes importantes: los métodos basados en lexicones o diccionarios y los métodos basados en aprendizaje. En este trabajo se abordaron ambas vertientes con la finalidad de hacer una comparación entre ambas y definir cuál da mejores resultados, además de encontrar cuál era el comportamiento del tratamiento de la negación en cada uno de los métodos. Los cambios propuestos a los métodos son descritos en las siguientes secciones. Los cambios realizados no solamente se enfocan en el tratamiento de la negación, hay también algunas aportaciones en esquemas de pesados y en representaciones de los documentos.

Los métodos basados en diccionarios fueron abordados con un enfoque híbrido mezclando técnicas de aprendizaje y de diccionarios. Mientras que se utilizaron variaciones del algoritmo de Naive Bayes Multinomial para atacar el problema desde el enfoque de aprendizaje. Ambas aproximaciones tienen dos pasos importantes: el cálculo de los valores de pertenencia de las palabras a cada clase y la clasificación. Ambos métodos son detalladas en las siguientes secciones.

4.5. Método híbrido para clasificación de opiniones

Los métodos basados en diccionarios utilizan listas de palabras positivas y negativas. Los valores de polaridad de cada una de las palabras están predefinidas en estas listas. La mayoría de esas listas han sido desarrolladas por expertos que etiquetan las palabras de manera manual.

Otros enfoques combinan aprendizaje con diccionarios. Este tipo de enfoques híbridos no tienen un lexicón como recurso externo así que lo calculan con los conjuntos de entrenamiento. Se obtienen la listas de términos positivos y negativos según las frecuencias de las palabras en las clases. Si una palabra es más frecuente en la clase positiva es considerada como positiva y si es más frecuente en documentos de la clase negativa se considera negativa.

En este trabajo se utilizó un método híbrido con base en métodos de diccionario. Es decir se calcularon las listas de palabras positivas y negativas desde una partición de los datos de prueba. Sin embargo a las listas de palabras se le asignaron valores de ambas polaridades. En la figura 4.3 se muestra el entorno general del método propuesto. Se tienen como entrada los documentos a clasificar, el primer paso es la segmentación del corpus en partición de entrenamiento y prueba. El siguiente paso es, a partir de los documentos de entrenamiento calcular los valores de pertenencia de las palabras a cada una de las clases, después se definen las listas de palabras positivas y negativas. Por último se clasifica el documento reuniendo pruebas de su pertenencia a cada clase dependiendo de si tiene mayor número de palabras positivas o negativas. La salida son los documentos con una etiqueta de polaridad.

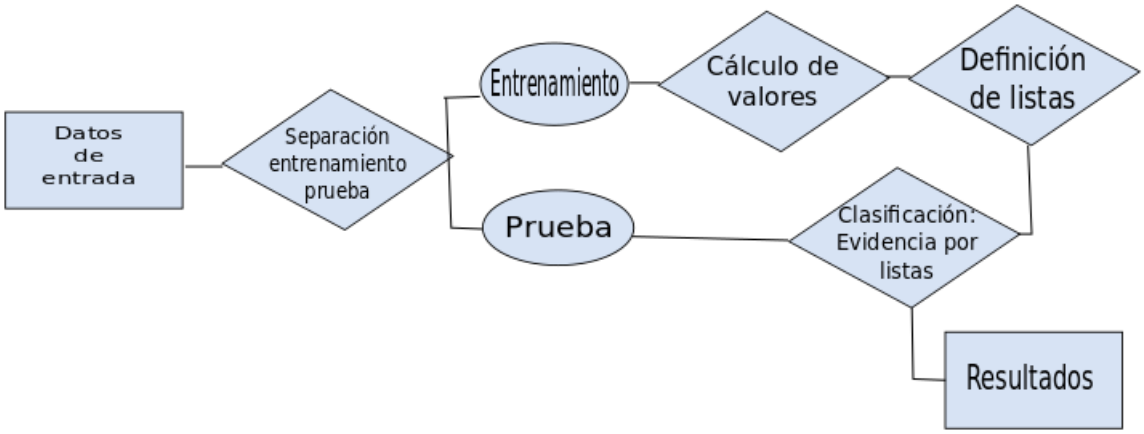


Figura 4.3: Método de clasificación híbrido.

4.5.1. Cálculo de valores de pertenencia: pesos

Los valores de pertenencia a cada una de las clases lo denominamos peso. El peso de cada palabra puede ser calculado de muchas maneras. Una de las maneras más comunes es la de frecuencia relativa de la palabra a cada una de las clases como se

muestra en la fórmula 4.1. Con esta expresión se busca encontrar cuántas veces ocurrió la palabra dentro de los documentos de la clase i y normalizar esa frecuencia entre el total de todas las frecuencias de las palabras en esa clase.

$$P_{cj}(x_j) = \frac{\text{frecuencia de } x_j \text{ en la clase } c_i}{\text{total de frecuencias de las palabras en documentos en } c_i} \tag{4.1}$$

En relación al tamaño del documento se encontró en los datos de prueba que los documentos de la clase positiva tenían un mayor número de palabras pero que las palabras tenían menores frecuencias que en los documentos negativos. Según algunos estudios el vocabulario con el que expresamos opiniones positivas es mayor al que utilizamos para decir cosas negativas [Kloumann et al., 2012]. Es decir, existen muchas más palabras para decir que algo nos gustó o nos pareció bien que para expresar que algo no nos gustó. Con la finalidad de tomar en cuenta esta información se sugiere una nueva forma de pesar las palabras tomando en cuenta el vocabulario de cada una de las clases.

$$P_{cj}(x_j) = \frac{\#\text{ de documentos de la clase } i \text{ en los que aparece la palabra } j}{\text{Total de documentos de la clase } i} * \frac{1}{|Voc_i|} \tag{4.2}$$

En la expresión 4.2 se encuentra una variación que incluye información del vocabulario. El primer término indica la frecuencia relativa del número de documentos dónde aparece una palabra entre el total de documentos y el segundo término funciona como una normalización entre el tamaño del vocabulario. Con esto se busca que las palabras no se normalicen por el total de frecuencia ya que en el caso de los positivos los valores eran más castigados porque aunque las palabras eran menos frecuentes los documentos eran mayores debido a que había mayor vocabulario. Esto provoca que las palabras positivas tengan valores menores. Al normalizar por el vocabulario se disminuye el problema de los valores menores.

4.5.2. Clasificación

La definición de etiqueta de cada documento está dada por la suma de los pesos que haya resultado mayor. Es decir, se realizan dos sumas de valores por cada documento, una de los valores positivos de cada palabra y otra con los valores negativos. La que haya resultado mayor será la clase del documento. En la ecuación 4.3 se representa matemáticamente la toma de decisión de asignación de etiqueta.

$$\begin{aligned} C &= \arg\max_{c_j} P_{c_j}(d) \\ P_{c_j}(d) &= \sum_{x_i \in d}^{|d|} P_{c_j}(x_i) \end{aligned} \tag{4.3}$$

4.5.3. Inclusión del tratamiento de la negación

En enfoques de clasificación basados en diccionarios o lexicones, la negación es tratada con enfoques simples. Una vez obtenido el alcance de una negación las palabras que hayan sido afectadas son clasificadas como clase contraria a la que tienen por defecto.

El tratamiento de la negación en este trabajo se realizó con efecto espejo. De esta manera, la palabra no tendría un valor totalmente contrario después de ser negada si no que se permitiría cierto rango de valores. Para aclarar: en el enfoque tradicional de listas al tener la frase " La película no es divertida" la pondríamos en la clase negativa por tener una negación en divertida, esto sería como decir que " La película es aburrida". En el enfoque que utiliza el efecto espejo podría variar el significado y la carga de positividad o negatividad dependiendo de los valores que tuviera la palabra divertida en cada una de las clases. Podríamos entonces entender que "La película es regular" o que "La película no es tan divertida pero tampoco está tan mal".

En el siguiente ejemplo podemos ver el comportamiento del efecto espejo. Los

valores de las palabras fueron calculadas como se mencionó anteriormente en la sección de "cálculo de valores de pertenencia: pesos".

- **Palabra: Divertida**

  Valor positivo: 0.7

  Valor negativo: 0.3

- **Palabra: No_Divertida**

  Valor positivo: 0.3

  Valor negativo: 0.7

En este trabajo el tratamiento de la negación tiene dos puntos importantes: Primero la definición del alcance de la negación, lo cuál se hizo con el algoritmo 1. Y segundo, la generación del vocabulario negado y la asignación de sus pesos. El vocabulario de la negación se realizó agregando la partícula "no_" a todas las palabras en el vocabulario y se les asignó pesos con el efecto espejo. En los experimentos con el efecto espejo se duplica el tamaño del vocabulario.

Clasificación:

La clasificación se realiza con sumas de los valores de cada clase como se muestra en la expresión 4.3, sin embargo al incluir la negación es necesaria una modificación en la ecuación debido a que ahora se cuentan con dos tipos de palabras. La formula 4.4 muestra la modificación al método de clasificación.

$$P_{c_j}(d) = \sum_{x_i \in d_1}^{|d_1|} P_{c_j}(x_i) + \sum_{y_i \in d_2}^{|d_2|} P_{c_j}(y_i) \tag{4.4}$$

Dónde:

X son las palabras normales, y

Y son las palabras afectadas por una negación

La diferencia de esta fórmula 4.4 con la expresión 4.3 es que se realizan dos sumatorias y los términos de las sumatorias provienen de dos tipos de conjuntos de palabras distintos: las palabras normales y las palabras negadas. Son vistas como dos listas de palabras distintas debido a que sus valores de pertenencia a cada clase son calculados de manera distinta.

4.6. Método basado en aprendizaje computacional

Se seleccionó un enfoque probabilista para realizar las pruebas con enfoque de aprendizaje computacional. Uno de los algoritmos probabilistas más populares es Naive Bayes. Se eligió trabajar con un enfoque probabilista por ser de sencillo desarrollo y porque brinda facilidades para incluir información de fenómenos lingüísticos de los textos en sus diferentes términos. Los términos son las probabilidades de las clases, su cálculo y su utilización para la toma de decisión de etiquetas. En la figura 4.4 se muestra el entorno del método basado en aprendizaje computacional. El método tiene como entrada la colección la cual se particiona en entrenamiento y prueba. De los documentos de entrenamiento se cálcula el vocabulario y su probabilidad de pertenecer a cada una de las clases. El método de clasificación contiene procedimientos que permiten el tratamiento de la negación y la intensificación.

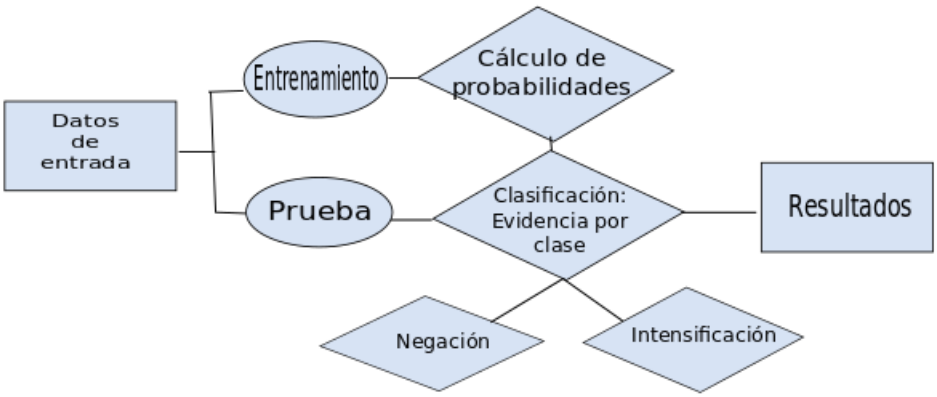


Figura 4.4: Método basado en aprendizaje.

Naive Bayes es muy utilizado en tareas de clasificación de documentos. Entre las variaciones de Naive Bayes la utilizada en este trabajo es la Multinomial. La principal característica del Naive Bayes Multinomial es que toma en cuenta la frecuencia que tengan las palabras para la clasificación. Esto es una característica importante en la clasificación de polaridad puesto que se busca que un documento sea más fácilmente clasificado con cierta polaridad entre más palabras con dicha polaridad contenga.

Visto con el siguiente ejemplo: " El inicio fue feo, pero la historia es muy bonita, y más bonita aún la forma en que la presentaron. Y ni que decir de la representación de escena. El jardín y la casa entera les quedó realmente bonita". Si tomamos en cuenta solamente las palabras que tienen una clara carga de polaridad positiva o negativa, la decisión se tomaría con la palabras fea y bonita. En Bayes tradicional solamente se toma en cuenta la presencia o ausencia de una palabra, por lo tanto bonita y fea tendrían el mismo valor. En el caso de Bayes Multinomial se toma en cuenta la frecuencia de una palabra, por lo que bonita tendría un valor mayor a fea.

4.6.1. Cálculo de valores de pertenencia: probabilidades

El cálculo de los valores de pertenencia de las palabras a cada una de las clases se hace con una frecuencia relativa a cada una de las clases y se agrega un suavizado Laplaceano para eliminar el error 0. El suavizado Laplaceano agrega 1 aparición extra a cada palabra del vocabulario para que aquellas palabras que no aparezcan en alguna de las clases tengan el valor mínimo y no causen problemas en la multiplicación de probabilidades en la clasificación. La fórmula 4.5 es con la que se calculan las probabilidades de las palabras de pertenecer a cada clase.

$$P(w_j|c_i) = \frac{\text{frecuencia de } w_j \text{ en la clase } c_i + K}{|V_{c_i}| + (\text{total de palabras en documentos en } c_i)} \tag{4.5}$$

K es una forma de eliminar el problema de que una palabra no aparezca en una de las clases. Normalmente, se da un valor de 1 a K para que las palabras tengan al menos una aparición en cada clase. $|V_{c_i}|$ es el tamaño del vocabulario de la clase i.

4.6.2. Clasificación

El método de clasificación con Naive Bayes Multinomial se realiza con la expresión 4.6. El exponente f_{w_j} es la frecuencia de la palabra en el documento.

$$P(c_i|d) = \prod_{w_j \in d}^{|d|} P(w_j|C_i)^{f_{w_j}} * P(c_i) \tag{4.6}$$

Esta versión de Naive Bayes Multinomial será tomada como referencia o baseline para todos las variaciones propuestas a los métodos basados en aprendizaje y específicamente, para los cambios a Bayes.

4.6.3. Inclusión del tratamiento de la negación

En el método basado en aprendizaje se utilizó el algoritmo 1 para definir el alcance de la negación y para modificar los documentos incluyendo la información de partículas negativas. La modificación de los documentos se muestran en el siguiente ejemplo: Teniendo la frase "La película no está divertida".

- Frase original: *"La película no está divertida"*
- Variación: *"La película no no_ está no_ divertida"*

A las secuencias de palabras obtenidas con el algoritmo de definición del alcance de la negación las nombramos palabras marcadas con negación o palabras negadas. Después de haber agregado los algoritmos de negación se hicieron algunos experimentos utilizando el NBM y usando la misma forma de cálculo de los pesos, sin embargo, las palabras negadas son menos frecuentes que las palabras que no han sido negadas. El hecho de que las palabras marcadas con negación sean menos frecuentes provoca que aún después de haber tenido cambios en los documentos y haber agregado la información de la negación la clasificación no marque gran diferencia. En el recuadro 4.1 se muestran ejemplos de las palabras más importantes del vocabulario para realizar la clasificación usando y sin usar la negación. La primera columna es el vocabulario de los documentos originales y están ordenados según su peso, dicho peso fue calculado con la expresión 4.5 descrito en la sección anterior. La segunda columna es el vocabulario de los documentos modificados con marcas de negación y el peso con el que fueron ordenados se obtuvo con la misma expresión 4.5. La última columna es

Vocabulario		**Vocabulario con negación**		**Vocabulario normalizado**	
no	nada	no	menos	no_duda	no_da
ni	mundo	ni	cinta	no	poco
vida	cinta	vida	mas	no_hay	no_visto
mas	menos	poco	gran	no_mas	no_mal
poco	peor	obra	no_mas	no_siquiera	no_gracia
parece	gran	parece	peor	ni	no_mucho
mal	mejores	film	mejores	no_deja	no_facil
obra	ver	mal	no_hay	vida	obra
hay	haber	nada	pelicula	no_llega	no_dejar
film	minutos	mundo	cine	no_sino	no_resulta

Tabla 4.1: Vocabulario con y sin palabras negadas.

el vocabulario de los documentos modificados con marcas de negación pero los pesos fueron calculados con la expresiones 4.7 y 4.8.

Para aumentar un poco la importancia de las palabras marcadas con negación dentro de la clasificación de los documentos se optó por representar a cada documento como la unión de dos tipos de vocabularios: el vocabulario normal y el vocabulario de las palabras negadas. La imagen 4.1 muestra la representación de los documentos. Formalmente, un documento es representado por $D =< x_1, x_2, ..x_n, y_1, y_2, ...y_n >$ dónde x_n; Son el conjunto de palabras dentro del documento que no fueron afectadas por ninguna negación y y_n; Son el conjunto de palabras negadas que ocurren en el documento.

Con la nueva representación se modificó también la forma en que los valores de pertenencia a cada una de las clases eran calculados. Al tenerse dos listas de palabras o dos tipos de vocabularios distintos se hicieron dos cálculos de frecuencias relativas a las clases por cada tipo de vocabulario. En las fórmulas 4.7 y 4.8 se muestran las fórmulas utilizadas para hacer el cálculo de las probabilidades de cada palabra.

$$P(x_j|c_i) = \frac{\text{frecuencia de } x_j \text{ en la clase } c_i + 1}{|Vx_{c_i}| + (\text{total de palabras normales en documentos en } c_i)} \tag{4.7}$$

En la ecuación 4.7 $x \in X$ dónde X es el conjunto de palabras que no fueron afectadas por una negación, Vx es el vocabulario de este conjunto. En la ecuación 4.8 $y \in Y$ dónde Y es el conjunto de palabras marcadas con negación y Vy es su vocabulario.

$$P(y_j|c_i) = \frac{\text{frecuencia de } y_j \text{ en la clase } c_i + 1}{|Vy_{c_i}| + (\text{total de palabras con negación en documentos en } c_i)} \tag{4.8}$$

Al tenerse dos tipos de vocabularios fue necesario también modificar el clasificador. En la fórmula 4.9 se muestra la modificación realizada al método para incluir los dos tipos de características. El primer factor representa a las palabras que pertenecen al vocabulario normal, es decir las palabras que no fueron afectadas por ninguna negación dentro del documento. El segundo factor son las palabras negadas, es decir las palabras que fueron alcanzadas por una negación, por ejemplo : "no_bonita", "no_gustar" , "no_linda".

$$P(c_i|d) = \prod_{x_j \in d_1}^{|d_1|} P(x_j|C_i)^{f_{x_j}} * \prod_{y_j \in d_2}^{|d_2|} P(y_j|C_i)^{f_{y_j}} * P(c_i) \tag{4.9}$$

4.6.4. Inclusión de la intensificación

El tener tres grupos de palabras hace necesario modificar nuevamente la expresión del clasificador para la categorización de los documentos. En este trabajo se propone la fórmula 4.10. Esta expresión sigue la idea descrita en la sección anterior en la ecuación 4.9. Se busca que cada tipo de característica sea igualmente importante en la clasificación y que no se pierda la información de intensificación por ser menos frecuente

que las palabras que no han sido afectadas. El primer término son las multiplicaciones de las probabilidades de pertenecer a la clase i de todas aquellas palabras que no han sido afectadas. El segundo factor son las probabilidades de las palabras intensificadas y la tercera las probabilidades de las palabras atenuadas.

$$P(c_i|d) = \prod_{x_j \in d_1}^{|d_1|} P(x_j|C_i)^{f_{x_j}} * \prod_{a_j \in d_2}^{|d_2|} P(a_j|C_i)^{f_{a_j}} * \prod_{b_j \in d_3}^{|d_3|} P(b_j|C_i)^{f_{b_j}} * P(c_i) \tag{4.10}$$

Las probabilidades de pertenecer a cada una de las clases de las palabras marcadas con intensificación y con atenuación se calculan con frecuencias relativas sobre cada tipo de palabras como se muestra en las fórmulas 4.11 y 4.12

$$P(a_j|c_i) = \frac{\text{frecuencia de } a_j \text{ en la clase } c_i + 1}{|Va_{c_i}| + (\text{total de palabras con intensificación en documentos en } c_i)} \tag{4.11}$$

$$P(b_j|c_i) = \frac{\text{frecuencia de } b_j \text{ en la clase } c_i + 1}{|Vb_{c_i}| + (\text{total de palabras con atenuación en documentos en } c_i)} \tag{4.12}$$

Una de las técnicas más utilizadas para el tratamiento de la intensificación es modificar los valores de las probabilidades de las palabras que hayan sido afectadas. Por ejemplo, si la palabra "bonita" tuviera un valor de 2 en positivo al encontrarse en el texto un " muy bonita" el valor de 2 sería modificado a 3. Las listas utilizadas para el tratamiento de intensificación y atenuación comúnmente tienen también los valores de aumento o disminución que provocan, sin embargo esos valores son definidos por estudiantes o personas hablantes naturales del idioma de la lista. En este trabajo se optó por modificar las frecuencias de las palabras que fueron afectadas por un intensificador o un atenuante. Se eligió agregar una presencia extra a las palabras que estuvieran marcadas con un intensificador y disminuir la mitad de una presencia a las palabras que estuvieran afectadas por un atenuador. Las modificaciones quedan más claras en el siguiente ejemplo.

- Frase 1: *"La película es muy buena"*
- Frase 2: *" La película es muy INT_ buena"*
- Frase 3: *"La película es poco interesante"*
- Frase 4: *"La película es poco ATE_ interesante"*

Las frecuencias se contabilizan en la tabla 4.2. Vemos las modificaciones en las palabras buena e interesante en las frases 2 y 4 que son las que están marcadas con información de intensificación y atenuación respectivamente.

Palabra	**La**	**película**	**es**	**muy**	**poco**	**buena**	**interesante**
Frase 1	1	1	1	1	0	1	0
Frase 2	1	1	1	1	0	2	0
Frase 3	1	1	1	0	1	0	1
Frase 4	1	1	1	0	1	0	0.5

Tabla 4.2: Ejemplo de frecuencias.

La modificación a la expresión de clasificación se llevó a cabo en el exponente que indica la frecuencia de cada palabra. La modificación se presenta en la expresión 4.13, seguido del desarrollo del exponente.

$$P(c_i|d) = \prod_{w_j \in d}^{|d|} P(w_j|C_i)^{f_{w_j}} * P(c_i) \tag{4.13}$$

El elemento f_{w_j} es el que se modifica en esta variación al método. En la expresión 4.14 se muestra que se añaden dos elementos en el cálculo de la frecuencia de una palabra y se modifica el de frecuencia. La frecuencia es el número de ocasiones en que apareció una palabra en el documento sin ser alcanzada por una intensificación o por

una atenuación. En este trabajo α representa a la intensificación y tiene un valor de 2. Es decir, a las palabras que fueron afectadas con una intensificación se les aumentó una aparición extra. En el caso de la atenuación es representada por β y tiene un valor de 0.5, las palabras que fueron afectadas por un atenuador se les disminuye sus apariciones a la mitad.

$$f_{w_j} = f(w_j, c_i) + \alpha f(INT\_w_j, c_i) + \beta f(ATE\_w_j, c_i) \tag{4.14}$$

La modificación de la frecuencia se da como en el siguiente ejemplo: supongamos que una palabra tuvo una frecuencia de 9 en un documento pero de esas 9 ocasiones 4 fueron mientras estaba en el alcance de afectación de una intensificación y 2 en el alcance de una atenuación. La frecuencia sería modificada con la fórmula 4.15 de la siguiente manera: $f_{w_j} = 3 + \alpha(4) + \beta(2)$ con $\alpha = 2$ y $\beta = 0{,}5$, el resultado sería:

$$\begin{aligned} f_{w_j} &= 3 + 2 * (4) + 0{,}5 * (2) \\ &= 3 + 8 + 1 \\ &= 12 \end{aligned} \tag{4.15}$$

4.6.5. Inclusión de la negación e intensificación

La negación y la intensificación son dos fenómenos muy utilizados en el lenguaje. Comúnmente estos dos fenómenos ocurren juntos como se muestra en el ejemplo. Al notar que en muchas de las opiniones de los corpus aparecían ambos fenómenos se buscó unir los tratamientos a estas dos problemáticas. El tratamiento de la negación y la intensificación en análisis de opiniones no es una temática nueva dentro de los enfoques de lexicones, sin embargo, en el caso de métodos basados en aprendizaje han sido poco abordados por separado y menos abordados juntos. En este trabajo se propone dos modificaciones a Bayes que permitan tratar con ambos fenómenos.

En el siguiente recuadro se muestra una parte de una opinión del corpus COAH. Podemos ver que la negación y la intensificación son fenómenos muy frecuentes incluso en textos de pocas palabras.

> *...“ Francamente **nada** recomendable a **no** ser por su precio, está situado en una calle casi imposible de localizar y mucho menos de llegar hasta allí, la calle es sucia y mal oliente el trato la verdad afable pero el apartamento esta anticuado y desfasado y algo sucio, la cama como para **no** dormir **ni** moverte un desastre **no** lo recomiendo a nadie y las toallas cada día faltaban y bastantes desgastadas una muy mala impresión del establecimiento y **nada** recomendable”...*

Para unir el tratamiento a las dos problemáticas se eligió extender el tratamiento a la negación que da igual importancia a los dos tipos de vocabulario (el normal y los bigramas de negación). A la variación de negación se agregaron los dos tipos de tratamiento a la intensificación que son la de tomar el documento como un conjunto de vocabularios distintos y el de modificar la frecuencia de las palabras según hayan sido afectadas por un intensificador o un atenuante.

Para preservar la información de negación e intensificación en los documentos se marcaron las palabras que fueron afectadas por alguno de esos elementos con los algoritmos 1 y 2. Primero se modificaron los documentos agregando la información de negación y después la información de intensificación. De esta manera los documentos quedan representados por la unión de seis listas como se muestra en la figura 4.5.

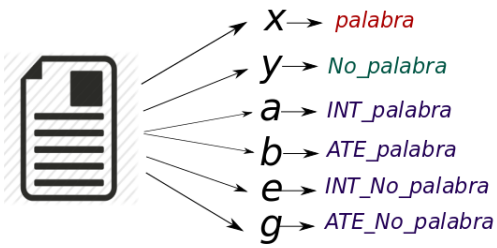


Figura 4.5: Representación de los documentos incluyendo negación, intensificación y atenuación.

Una vez modificados los documentos se calcularon las probabilidades de cada elemento con las fórmulas 4.7, 4.8, 4.11 y 4.12. Además se añadieron dos expresiones mas por los dos tipos de vocabulario agregados. Las palabras que fueron afectadas por una negación y una intensificación fueron convertidas en secuencias de tres palabras como se mostró en la figura 4.5 y sus pesos se calcularon con la expresión 4.16. En el caso de que una palabra fuera afectada por una negación y una intensificación también era convertida a una secuencia marcada por ambos eventos y sus probabilidades se calcularon con la fórmula 4.17.

$$P(e_j|c_i) = \frac{\text{frecuencia de } e_j \text{ en la clase } c_i + 1}{|Ve_{c_i}| + (\text{total de marcas de negación + intensificación en docs en } c_i)} \tag{4.16}$$

Dónde $e \in E$ y E es el conjunto de palabras afectadas por una negación y una intensificación. En los documentos este tipo de palabras está representado como "INT_No_palabra". En el caso de las atenuaciones se representan con "ATE_No_palabra" y su conjunto está nombrado con la letra G.

$$P(g_j|c_i) = \frac{\text{frecuencia de } g_j \text{ en la clase } c_i + 1}{|Vg_{c_i}| + (\text{total de marcas de negación + atenuación en docs en } c_i)} \tag{4.17}$$

En la tabla 4.3 se encuentra un ejemplo de las 15 características más frecuentes de cada uno de los vocabularios en una de las particiones de prueba del corpus CRM. Cada una de las columnas se refiere a las palabras más utilizadas, de cada uno de los tipos de palabras descritos en las secciones anteriores, en el corpus al haber hecho la modificación de los documentos para marcar las palabras afectadas por negación, intensificación o atenuación. El corpus CMR es descrito en el capítulo 5 en la sección de colecciones.

La modificación al clasificador se muestra en la expresión 4.18. Se añadieron más elementos al clasificador para incluir los seis distintos vocabularios.

Vocabulario normal	**Palabras negadas**	**Palabras intensificadas**	**Palabras atenuadas**	**Palabras negadas e intensificadas**	**Palabras negadas y atenuadas**
embucha	no_estanteria	INT_sabido	ATE_gusto	INT_no_deja	ATE_no_siempre
oigan	no_queda	INT_hubiera	ATE_guion	INT_no_ningun	ATE_no_venir
ramos	no_cabeza	INT_conocido	ATE_cuanto	INT_no_habla	ATE_no_apenas
golosina	no_interesante	INT_amor	ATE_gracias	INT_no_primero	ATE_no_he
cambiaria	no_deja	INT_propia	ATE_tarde	INT_no_primera	ATE_no_decir
distincion	no_le	INT_intensidad	ATE_otras	INT_no_darle	ATE_no_director
esperada	no_embargo	INT_dura	ATE_totalidad	INT_no_registra	ATE_no_sabria
fisk	no_ningun	INT_papeles	ATE_accion	INT_no_historia	ATE_no_queda
comprendia	no_realidad	INT_mostrar	ATE_evidente	INT_no_tema	ATE_no_volver
enmarcados	no_mismo	INT_merece	ATE_pantalla	INT_no_senalo	ATE_no_error
villanos	no_original	INT_ritmo	ATE_hecho	INT_no_pinta	ATE_no_reconozco
omitidos	no_claro	INT_comedia	ATE_leer	INT_no_tipo	ATE_no_cuela
desencadenara	no_ritmo	INT_visto	ATE_libro	INT_no_tenia	ATE_no_otra
innumerables	no_menos	INT_profundidad	ATE_dvd	INT_no_estando	ATE_no_quieren
susodicha	no_siempre	INT_mujeres	ATE_haber	INT_no_basada	ATE_no_entendia

Tabla 4.3: Vocabulario más frecuente.

$$P(c_i|d) = \prod_{x_j \in d_1}^{|d_1|} P(x_j|C_i)^{f_{x_j}} * \prod_{y_j \in d_2}^{|d_2|} P(y_j|C_i)^{f_{y_j}} * \prod_{a_j \in d_3}^{|d_3|} P(a_j|C_i)^{f_{a_j}} * \prod_{b_j \in d_4}^{|d_4|} P(b_j|C_i)^{f_{b_j}} * \prod_{e_j \in d_5}^{|d_5|} P(e_j|C_i)^{f_{e_j}} * \prod_{g_j \in d_6}^{|d_6|} P(g_j|C_i)^{f_{g_j}} * P(c_i) \tag{4.18}$$

La última variación al método de clasificación para agregar la información de intensificación y negación se hizo sobre la fórmula 4.19 modificando los exponentes de ambas listas con $\alpha = 2$ y $\beta = 0{,}5$.

$$\begin{aligned} P(c_i|d) &= \prod_{x_j \in d_1}^{|d_1|} P(x_j|C_i)^{f_{x_j}} * \prod_{y_j \in d_2}^{|d_2|} P(y_j|C_i)^{f_{y_j}} * P(c_i) \\ f_{x_j} &= f(x_j, c_i) + \alpha f(INT\_x_j, c_i) + \beta f(ATE\_x_j, c_i) \\ f_{y_j} &= f(y_j, c_i) + \alpha f(INT\_y_j, c_i) + \beta f(ATE\_y_j, c_i) \end{aligned} \tag{4.19}$$

4.7. Recapitulación

En este capítulo se exponen los principales aportes hechos en este trabajo de tesis, los cuales se enlistan resumidamente en los siguientes puntos.

- Se realizan aproximaciones a los dos enfoques de clasificación de opiniones y se propusieron modificaciones a cada uno de ellos con la finalidad de mejorar la exactitud de clasificación.
  - Métodos basados en lexicones o diccionarios.
    - Se realiza un enfoque híbrido, es decir, se combinaron elementos de aprendizaje computacional para mejorar el funcionamiento del método basado en diccionarios.
    - Se propone un nuevo esquema de pesado de palabras.
  - Métodos basados en aprendizaje computacional: Tomando como base las ideas probabilistas de MNB.
    - Se proponen cinco variaciones al método para incluir información de distintos fenómenos lingüísticos. Se busca que con la inclusión de esta información se mejore la clasificación. Las variaciones son:
      1. Negación,
      2. Dos variaciones de intensificación y
      3. Dos variaciones de negación e intensificación.
    - Se sugieren nuevos paradigmas de cálculo de probabilidades para las palabras que permiten dar mayor importancia a los términos que sean parte de los fenómenos de negación e intensificación.
    - Se exponen modificaciones a la representación de los documentos para incluir la información de los fenómenos lingüísticos expuestos.
- Se formulan algoritmos de definición de alcance de la negación e intensificación que tienen como salida documentos modificados con la inclusión de la información de estos dos fenómenos.

CAPÍTULO 5

MARCO EXPERIMENTAL

En este capítulo se detallan las colecciones utilizadas en la experimentación. También se muestran la listas de palabras o diccionarios de negaciones, intensificaciones y palabras vacías utilizadas. Por último se detallan los experimentos realizados y se definen nombres para la presentación de resultados en el siguiente capítulo.

5.1. Colecciones

La fase de experimentación hizo uso de colecciones de textos de opinión en los idiomas Inglés y Español.

Para todos los experimentos se consideraron solamente las opiniones positivas y negativas. Se descartaron las opiniones neutrales por contener tanto polaridad positiva como negativa. Acotar el problema a dos clases es muy común en la clasificación de polaridad. Los resultados publicados en el estado del arte para las colecciones utilizadas fueron realizados presentando el problema como clasificación entre dos clases.

Todas las opiniones de las colecciones fueron escritas por los usuarios de la red. Esto aumenta la dificultad de la tarea por los errores y fenómenos que pueden estar incluidos en los textos. Entre los errores más comunes están las faltas ortográficas y

los errores de escritura. Además, estas opiniones presentan el uso de emoticones, la presencia de expresiones fijas y el sarcasmo e ironía.

A continuación, se describen detalladamente las colecciones utilizadas.

CMR: Corpus of Movie Reviews.

El corpus de opiniones de películas o CMR por sus siglas es Inglés es un corpus de críticas en español. El corpus está formado por 3878 críticas de cine recolectadas del sitio MuchoCine[1] [Cruz et al., 2008]. Los textos están organizados en tres etiquetas: positivos (1351 textos), negativos (1274 textos) y neutrales (1253 textos). La etiqueta de cada crítica está definida por el número de estrellas con que fue puntuada. Las críticas con 1 o 2 estrellas son consideradas negativas y las que tienen 4 o 5 estrellas se consideran positivas. Las opiniones con 3 estrellas son las etiquetadas como neutrales. En el siguiente recuadro se encuentra un ejemplo de una opinión de este corpus.

<review author=Victor Trujillo title= El internado rank=2 maxRank=5 source=muchocine>
<summary>Bodrio de terror a la francesa</summary>
<body>"Joder que bodrio de internado. La peli no empieza mal y la presentación del personaje principal, que sin duda es el orfanato, un caserón impresionante en medio del campo, es muy preciosista, plagada de matices y con demasiada luz para una película de este género. La casa acojona a primera vista y el patio que la rodea tampoco se queda corto; tenían el ingrediente más importante para hacer una buena peli de terror, pero lo han desaprovechado. Sorprende la buena producción francesa y sorprende la belleza de la protagonista ...«</body>
</review>

Para los experimentos se utilizaron 1270 opiniones de la clase positiva y 1270 de la clase negativa. Los textos empleados fueron seleccionados de manera aleatoria.

COAH: Corpus of Andalucía's Hotels.

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

El corpus COAH[2] está compuesto por 1816 opiniones sobre hoteles ubicados en Andalucía. Las opiniones fueron recogidas del sitio web de TripAdvisor[3]. El etiquetado de este corpus se hizo según las estrellas relacionadas a cada documento. Se cuentan con 1020 opiniones positivas y 511 negativas. En el siguiente recuadro se encuentra un ejemplo de una opinión del corpus COAH.

<coah:hotel_review xmlns:coah="http://sinai.ujaen.es/coah»
<coah:id>1</coah:id>
<coah:rank>5</coah:rank>
<coah:abstract>Un hotel digno de mención!</coah:abstract>
<coah:review>
Çomo bien les comenté a los propietarios a la hora de abandonar el hotel, no dudaré un momento en recomendar una y otra vez el Hotel Albero de Granada. Su situación respecto del centro de Granada no es la mejor, pero para nuestros propósitos era perfecto (escapada de fin de semana con visita a la Alhambra). Se encuentra en la carretera de paso a Sierra Nevada y muy cercano a la Alhambra. Por la zona se puede encontrar aparcamiento y este se encuentra en una zona segura y tranquila. Los parkings del centro de Granada que nos recomendaron en el hotel fueron lo que nos dijeron (nada caros) y pudimos movernos por el centro perfectamente desde allí. Las habitaciones muy limpias y las camas confortables. El desayuno fue espectacular..."
</coah:review>
</coah:hotel_review>

Para evitar problemas por el desbalance entre las clases se tomaron en cuenta solamente 510 opiniones de cada clase. Las opiniones utilizadas fueron elegidas al azar.

Blitzer: Corpus multidominio de Blitzer.

Este es un corpus de opinión escrito en Inglés. Contiene opiniones de cuatro dominios distintos y generalmente es utilizado para experimentación del tema de dominios cruzados [Blitzer et al., 2007]. Los dominios contenidos en este corpus son: libros, películas, electrónicos y cocina. Por cada temática hay 2000 documentos de opinión

[2]http://sinai.ujaen.es/coah
[3]http://www.tripadvisor.es

de los cuales 1000 son positivos y 1000 son negativos.

Para los experimentos realizados con esta base de datos se tomaron en cuenta todas las opiniones disponibles.

IBM v1.0.

Esta base de datos contiene 5331 opiniones de cada clase de polaridad. Cada opinión es una frase o sentencia obtenida del corpus IBM. Dicho corpus se forma de opiniones del dominio de cine y están escritas en Inglés. IBM v1.0 fue publicada en 2005, su nombre oficial es *sentence polarity dataset* v1.0 [Pang y Lee, 2005]. En el siguiente recuadro se encuentran ejemplos de las opiniones contenidas en esta colección.

Negativas
**simplistic , silly and tedious .*
**unfortunately the story and the actors are served with a hack script .*
**too slow for a younger crowd , too shallow for an older one .*
Positivos
**a masterpiece four years in the making .*
**scores a few points for doing what it does with a dedicated and good-hearted professionalism .*
**occasionally melodramatic , it's also extremely effective*

En los experimentos con este corpus se utilizaron todas las opiniones disponibles.

Otros corpus

Se utilizó también la base de datos de SFU en su versión Español para realizar algunos experimentos. SFU es un corpus de textos de opinión de ocho dominios diferentes [Taboada et al., 2008]. Los dominios abordados son libros, autos, computadoras, máquinas de cocina, hoteles, películas, música y teléfonos. El corpus está compuesto por 400 opiniones, 50 de cada temática.

Con la finalidad de demostrar la funcionalidad del sistema propuesto y del impacto que tiene la negación y la intensificación en los documentos utilizados se realizó un

análisis comparativo de los corpora. En la tabla 5.1 se muestran las estadísticas de las colecciones utilizadas. Se contabilizaron el número de palabras por clase, el vocabulario, el promedio de palabras por documento, así como las razones de aparición de la negación, la intensificación y las palabras vacías.

Todos los conteos fueron realizados sobre el total de opiniones de cada corpus.

Colecciones								
		Palabras	**Vocabulario**		**Prom**	**Neg.**	**Int.**	**Stopwords**
CMR	**POS**	733698	41935	55844	543	0.0168	0.0296	0.4846
	NEG	558736	37281		438	0.0213	0.0317	0.4824
COAH	**POS**	109348	8614	13302	107	0.0165	0.0471	0.4885
	NEG	84534	8625		165	0.0358	0.0314	0.4835
Books	**POS**	146358	14147	21263	146	0.0109	0.0319	0.4968
	NEG	160284	14695		160	0.0182	0.0331	0.5134
Dvds	**POS**	169917	15507	21754	169	0.01244	0.0326	0.4915
	NEG	169036	14311		169	0.0189	0.0333	0.5089
Electronics	**POS**	98541	7721	10902	98	0.0162	0.0408	0.5243
	NEG	105341	7352		105	0.0236	0.0304	0.536
Kitchen	**POS**	86090	6863	9657	86	0.0148	0.0424	0.5312
	NEG	88558	6587		88	0.0224	0.0339	0.5404
CMR títulos	**POS**	39357	7425	11489	29	0.0173	0.0344	0.4719
	NEG	37437	7205		29	0.0268	0.0365	0.4682
IBM v1.0	**POS**	103022	12508	18244	19	0.01013	0.02944	0.433
	NEG	102735	12828		19	0.01671	0.03463	0.4416

Tabla 5.1: Comparativa de los corpus.

Entre los datos más interesantes que surgen de esta tabla podemos mencionar las diferencias de vocabularios y del tamaño de las opiniones según el idioma. El corpus de CMR contiene las críticas con mayor número de palabras y es además el corpus que tiene el vocabulario de mayor tamaño. Podemos ver también que incluso entre dominios los vocabularios y el tamaño de las opiniones es sumamente diferente. Por ejemplo, en el caso del corpus COAH sus opiniones tienen apenas la quinta parte de las palabras que contienen en promedio las críticas del corpus CMR. En el caso de los corpus en Inglés las diferencias son menos notorias, a excepción del dominio de

cocina cuyas opiniones tienen aproximadamente la mitad de palabras que el resto de dominios en Inglés. Las últimas dos colecciones de la tabla son opiniones pequeñas, sin embargo, conservan la relación de mayor número de palabras en las críticas escritas en Español. En el caso de vocabulario es mayor en el corpus escrito en Inglés. Por otro lado, es posible comparar los vocabularios según la clase, en la mayoría de los corpora analizados el vocabulario positivo resulta mayor que el vocabulario negativo. Con estas diferencias en el tamaño y vocabularios de las colecciones podremos saber si los métodos propuestos para la clasificación de polaridad serán afectados por este tipo de elementos.

Las columnas Neg, Int y Stopwords representan la razón de frecuencia de negaciones, intensificadores y palabras vacías en cada colección por clase. En cuanto a palabras negativas vemos que son más frecuentes en la clase negativa en todas las colecciones. En el caso de los intensificadores no parecen tener una inclinación hacía una clase puesto que se dan de manera similar en ambas clases. Por último, las palabras vacías tienen frecuencias muy parecidad en ambas clases. En cuanto al análisis por idioma, las palabras vacías son más frecuentes o más utilizadas en documentos en Inglés, dónde en 4 de los 5 corpora observados más del 50 % de las palabras son stopwords.

5.2. Recursos externos

La limpieza de los documentos para la experimentación y algunos de los experimentos requieren del uso de recursos externos. Los recursos externos utilizados en este trabajo son descritos en los siguientes puntos:

Lista de palabras vacías.

En muchas tareas de clasificación de textos es conveniente eliminar algunos elementos que pueden causar ruido o errores en la asignación de etiquetas. Los elementos eliminados suelen ser números, emoticones, algunos signos de puntuación y palabras que sean tan frecuentes en las clases que no ayude realmente para diferenciarlas entre sí. Las palabras vacías son palabras que no tienen elementos de contenido ni de sentimiento, algunos ejemplos de estas palabras pueden ser conectores, preposiciones, y algunos verbos auxiliares.

Existen muchas listas públicas de palabras vacías, para este trabajo se utilizaron las listas publicadas en la página RankNL[4] para Español y para Inglés.

Debido a que en este trabajo se proponen métodos de tratamiento para la negación y la intensificación se eliminaron de la lista de palabras vacías las negaciones e intensificaciones que se encontraban en ella.

Lista de negaciones

Uno de los puntos más importantes dentro de la experimentación realizada fue la lista de negaciones utilizada.

En Español, la negación es comúnmente expuesta mediante palabras negativas. Se utilizó una lista de seis palabras definidas por la RAE como negaciones. Éstas son: *no, ni, sin, nunca, nada* y *tampoco* [Española, 2009].

En el caso del Inglés, la negación no sólo se identifica por la presencia de palabras negativas si no también por contracciones añadidas a otras palabras como verbos auxiliares. Para el tratamiento de la negación en Inglés se tomaron en cuenta los mismos signos de negación utilizados en trabajos anteriores [Councill et al., 2010]. La lista de palabras negativas y algunos ejemplos de palabras que contienen la negación con la contracción "n't" se encuentran en la tabla 5.2.

neither	nobody	not	cannot
didnt	havent	neednt	wasnt
nor	none	n't	darent
hadnt	isnt	oughtnt	wouldnt
never	nothing	aint	dont
hasnt	mightnt	no	nowhere
cant	doesnt	mustnt	shouldnt

Tabla 5.2: Negaciones en el idioma Inglés.

[4]http://www.ranks.nl/stopwords

Español		Inglés	
Intensificadores	**Atenuantes**	**Intensificadores**	**Atenuantes**
bastante	menos	great	rather
mayor	apenas	huge	smaller
bien	solo	massive	some
altamente	algún	collosal	a bit
puro	un poco	biggest	minor
muy	ligeramente	super	a few
extra	moderado	extra	only
gran	casi	big	less
tan	pequeño	clear	out of
más	parcialmente	very	moderate

Tabla 5.3: Ejemplos de intensificadores y atenuantes.

Lista de intensificadores.

La lista de intensificadores y atenuantes utilizados en las pruebas fue la de SO-dictionaries [Taboada et al., 2011]. SO-dictionaries son un conjunto de listas de distintos tipos de términos como intensificadores, atenuantes, adjetivos, verbos y adverbios. La versión original de estos lexicones fue desarrollada para el idioma Inglés y después fue traducida manualmente por hablantes de Español.

En el caso del Inglés la lista de intensificadores y atenuantes contiene 217 palabras y los valores de afectación que provocan. En el caso del Español solamente hay 167 términos. En la tabla 5.3 se encuentran ejemplos de intensificadores y atenuantes de ambos idiomas.

5.3. Especificaciones

La limpieza de los datos y documentos, el cálculo de los vocabularios y pesos, así como el desarrollo de los clasificadores se realizó en Python[5]. Python es un lenguaje de

[5]https://www.python.org/downloads/

		Método	
Nombre	**Descripción**	**Híbrido**	**Aprendizaje**
UNI (Baseline) (U)	Unigramas (Pesado con frecuencia relativa)	X	X
UNI+Neg (UN)	Unigramas (Pesado con frecuencia relativa) + Palabras negadas	X	X
PP	Unigramas (Pesado propuesto)	X	
PP+Neg (PPN)	Unigramas (pesado propuesto) + Palabras negadas	X	
UNI+Neg+N (UNN)	Unigramas + Palabras negadas + normalizado por tipo de vocabulario		X
UNI+Int+N (UIN)	Unigramas + Palabras intensificadas + normalizado por tipo de vocabulario		X
UNI+Int+AD (UIAD)	Unigramas + Palabras intensificadas + Aumento y disminución de valores		X
UNI+Neg+Int+N (UNIN)	Unigramas + Palabras negadas + Palabras intensificadas + Normalización por tipo de voc		X

Tabla 5.4: Experimentos realizados.

programación que busca que el código sea legible. Es multiparadigma, esto quiere decir que soporta orientación a objetos, programación imperativa y también programación funcional.

La evaluación de todos los métodos de clasificación presentados se realizó mediante validación cruzada en k particiones (k-folds cross validation) [Arlot y Celisse, 2010]. En el caso de los métodos basados en lexicones se usó una k=10. En los métodos basados en aprendizaje se utilizaron 5 particiones. Según algunos estudios el nivel de confianza es similar usando k= 5-10 pliegues [Arlot y Celisse, 2010]. Se tiene como trabajo futuro realizar los experimentos con otras cinco particiones y conseguir así una doble validación cruzada de cinco particiones. En todas las colecciones las particiones fueron seleccionadas de forma aleatoria.

Los experimentos realizados fueron diseñados para probar las métodos y variaciones descritas en el capítulo 4. En la tabla 5.4 se resumen los experimentos realizados con base en Naive Bayes Multinomial y se asignan los nombres de cada uno.

CAPÍTULO 6

RESULTADOS

A continuación se muestran las tablas de resultados de los experimentos realizados en las distintas colecciones. También se hace un análisis sobre los resultados de la clasificación con cada una de las variaciones y métodos propuestos.

Las tablas y análisis serán presentados en el orden del capítulo 4. Es decir, primero se detallarán los resultados del método híbrido, seguido por los resultados del método Bayesiano. Los experimentos se realizaron primero en las colecciones en Español y después en las colecciones en Inglés para probar la robustez de los métodos propuestos. El detalle de todos los resultados se presentan en el apéndice A.

6.1. Resultados: Método híbrido

Los experimentos del método con enfoque híbrido se realizaron sobre el corpus CMR y el corpus COAH. Ambas colecciones están escritas en Español. Los resultados del estado del arte reportados en las tablas 6.1 y 6.2 son los que han obtenido mejores resultados sobre estas colecciones con enfoques de clasificación de diccionarios.

En la tabla 6.1 se muestran los resultados del corpus CMR. Podemos ver que los resultados con el pesado propuesto son mejores que los resultados con el pesado de

Método	Precisión	Recuerdo	Medida F	Exactitud
U	0.7471	0.5115	0.6072	0.5116
UN	0.7348	0.6315	0.6792	0.6470
PP	0.7645	0.6727	0.7156	0.6727
PPN	0.7487	0.7080	0.7277	**0.7373**
Jiménez,2015	0.6519	0.6430	0.6474	0.6465

Tabla 6.1: Resultados: Enfoque híbrido en CMR.
U.- Baseline, UN.- Unigramas + Negación, PP.- Pesado propuesto, PPN.- PP + Negación

frecuencia relativa. En ambos esquemas de pesado utilizados en estas pruebas hacer tratamiento de la negación provoca una mejora en la clasificación de polaridad. Es importante notar que incluso los resultados del método con el pesado propuesto son mejores que las variaciones de los métodos que usan pesado de palabras con frecuencia relativa. Con esto podemos pensar que tomar en cuenta el vocabulario presente en cada una de las clases es importante al menos en esta colección.

En el estado del arte el trabajo más cercanamente comparable es el de [Jiménez et al., 2015] ya que utiliza la misma base de datos para su experimentación y también hace un tratamiento a la negación. Las principales diferencias entre ese trabajo y el nuestro es que utiliza un enfoque de diccionarios y que el tratamiento de la negación lo hace mediante análisis de árboles de dependencia y cambio de signo. Además la configuración de experimentación es distinta, ya que en el enfoque de diccionarios no es necesario hacer particiones de entrenamiento y prueba por lo que utilizan el diccionario para clasificar la colección completa, mientras que en el enfoque propuesto en este trabajo si se hacen particiones de entrenamiento y prueba. El resultado más alto obtenido en esa investigación se logra con tratamiento de la negación y utilizando una lista de palabras que fue creada especialmente para el dominio del cine. En el artículo de [Jiménez et al., 2015] se reporta una exactitud de clasificación de 0.6465 y una medida F1 de 0.6474.

En la tabla 6.2 Se encuentran los resultados obtenidos en la colección COAH.

Dicha colección contiene opiniones de hoteles en Español. La principal diferencia de este corpus con el de CMR es el tamaño de las opiniones, ya que las críticas del CMR son cinco veces más grandes en cuánto a número de palabras que las opiniones del COAH.

Método	**Precisión**	**Recuerdo**	**Medida F**	**Exactitud**
U	0.7843	0.6950	0.7359	0.7019
UN	0.8075	0.7753	0.7918	**0.7831**
PP	0.7646	0.6313	0.6914	0.6375
PPN	0.7962	0.6999	0.7449	0.7069
González,2015	0.9026	0.8713	0.8866	0.9007

Tabla 6.2: Resultados: Enfoque híbrido en COAH.
U.- Baseline, UN.- Unigramas + Negación, PP.- Pesado propuesto, PPN.- PP + Negación

En la tabla 6.2 vemos que se conservan las mejoras en los métodos al agregar el tratamiento de la negación. Sin embargo, los mejores resultados no se obtienen con el pesado propuesto, sino con el pesado de frecuencia relativa normal. Esto puede deberse a que en el caso del dominio de hoteles, o en esta colección, los vocabularios entre documentos positivos y negativos no es tan diferente en cuanto a número de palabras que contiene como es el caso del corpus CMR. Es decir, en el primer caso con el corpus CMR el vocabulario positivo y negativo son sumamente distintos en cuanto a número de palabras. Siendo el vocabulario positivo mayor. En este caso, los vocabularios positivo y negativo en COAH son muy similares, por lo que hacer una normalización por vocabularios no provoca mejoras fuertes. En el caso del corpus CMR hay 4554 más palabras positivas que negativas. Por otro lado en el corpus COAH el vocabulario mayor es el negativo pero solamente por 11 palabras. Los datos de vocabularios fueron presentados y analizados en el capítulo 5 en la sección de *colecciones*.

El resultado del estado del arte reportado es un método basado en diccionarios. El diccionario que utilizaron surgió de la misma base de datos mediante un análisis de frecuencias de las palabras en cada una de las clases.

6.2. Resultados: Método basado en aprendizaje

Las siguientes gráficas muestran los resultados de las pruebas descritas en la tabla 5.4. El eje x muestra los diferentes métodos de clasificación y el eje y la exactitud alcanzada por cada método. Las barras son los niveles de exactitud alcanzados mientras los puntos verdes representan a la medida F1. Todas las barras son del mismo color excepto aquella con la más alta exactitud, o cuando se presenta el resultado de un método del estado del arte. La primer barra es el Naive Bayes Multinomial con unigramas, es decir, ese clasificador no tiene cambios por lo que será tomado como baseline para el resto de las variaciones. Esta base está representada por la línea roja para observar más fácilmente si algún método lo supera. La elección del *baseline* se debió a qué el objetivo principal de esta tesis es demostrar que haciendo un tratamiento a la negación o a la intensificación es posible mejorar los niveles de exactitud en la clasificación de polaridad. En algunos casos se mencionan los resultados obtenidos por otros trabajos que han utilizado las mismas colecciones para su experimentación, esos resultados están representados por una barra de color verde.

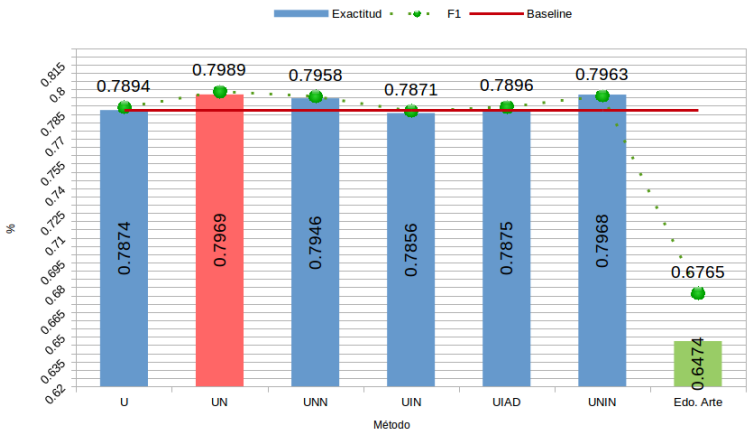


Figura 6.1: Resultados en corpus CMR: Críticas completas.

En las gráficas 6.1 y 6.2 están los resultados de los experimentos hechos en el corpus CMR. Son dos gráficas porque se realizaron experimentos con las críticas completas y con los títulos de esas críticas. Las pruebas separadas se hicieron con la finalidad de observar el comportamiento de los clasificadores con documentos de distintos tamaños. Las críticas completas tienen aproximadamente 500 palabras cada una y los títulos

sólo 30 palabras. Ambas gráficas siguen el mismo comportamiento, excepto por la última variación donde en el caso de las críticas completas tiene un buen resultado pero en el caso de los títulos tiene una caída. La última variación es la de añadir negación e intensificación. Además vemos como los resultados de los títulos tienen valores más bajos, esto se debe al tamaño de las opiniones, al ser más pequeñas hay menos evidencia de pertenecer a cualquiera de las clases. Sin embargo, es importante notar que los valores tienen una diferencia de entre tres y cuatro puntos porcentuales de exactitud.

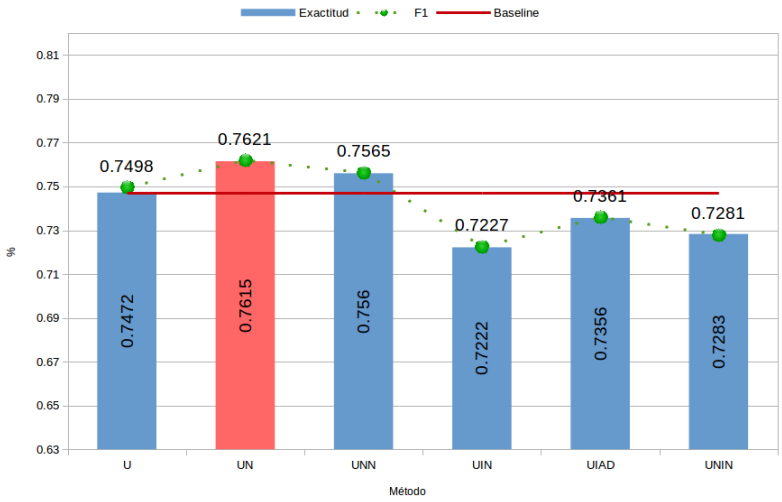


Figura 6.2: Resultados en corpus CMR: Títulos de las críticas.

Otro dato importante es la barra de color verde,ésta representa al resultado del estado del arte. Podemos observar que está hasta 10 puntos por debajo de los resultados obtenidos incluso sin utilizar el tratamiento de la negación. El resultado del arte reportado es el de [Jiménez et al., 2015], el cual fue descrito en los resultados del método híbrido. Es importante mencionar que en la gráfica 6.2 no se encuentra una barra verde debido a que no se encontraron trabajos que hicieran experimentación solamente con los títulos de las críticas de CMR.

El siguiente corpus de experimentación fue el COAH que contiene opiniones sobre hoteles. En la gráfica 6.3 podemos ver que tiene resultados bastante altos. Ésta fue la base que tuvo mejores resultados. Se obtuvo un 93.33 % de exactitud con la variación que incluye negación y normalización especial para cada tipo de vocabulario. Los

buenos resultados en esta colección pueden deberse a los tamaños de los vocabularios. En el caso de COAH el vocabulario general es más pequeño, además el vocabulario compartido entre ambas clases también es menor. Por ser menos palabras puede que la mayoría de las palabras tengan bien definida la clase a la que pertenecen. Además, debido a que las opiniones son más pequeñas, es muy probable que contengan menos palabras de estilo.

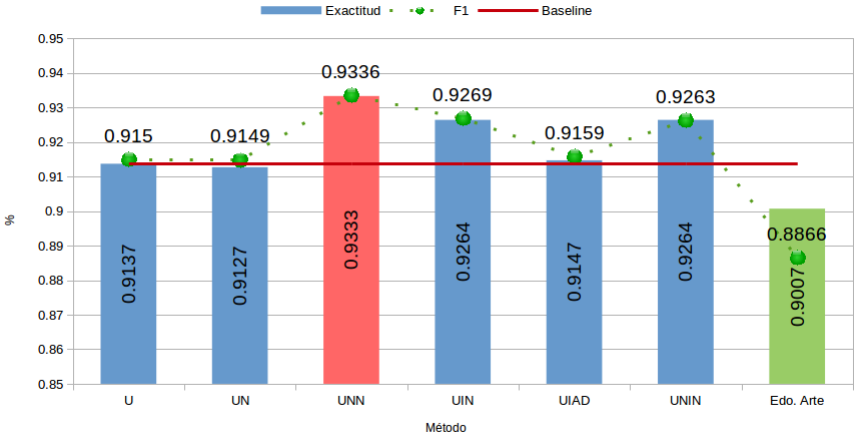


Figura 6.3: Resultados en corpus COAH.

Se promediaron los resultados de los experimentos realizados para Español en las colecciones de CMR y COAH. Los resultados se encuentran en la gráfica 6.4. Las mejores exactitudes se lograron con las variaciones que incluyen sólo información sobre la negación. Por otro lado, los métodos que incluyen la intensificación resultan de poca ayuda para la clasificación. Esto puede deberse a que se necesita un estudio más profundo sobre cuáles son los niveles de afectación que tiene la intensificación y sobre todo identificar claramente cuáles son aquellos que son utilizados en textos de opinión.

6.2.1. Otro idioma, diferentes dominios

Los siguientes experimentos se realizaron en las colecciones escritas en Inglés. La finalidad de realizar estos experimentos fue probar el método de clasificación basado en aprendizaje en el idioma Inglés y en dominios distintos para conocer su compor-

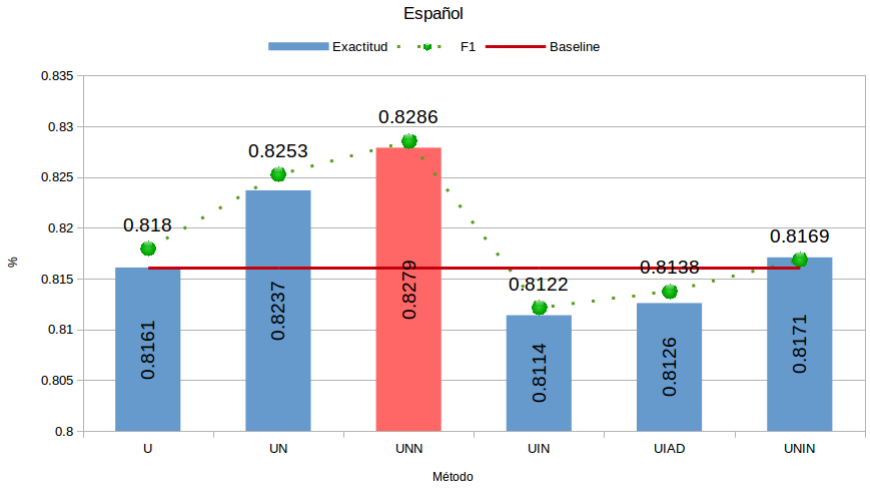


Figura 6.4: Promedio de resultados en colecciones en Español.

tamiento. Se utilizaron las colecciones de Blitzer y de IBM v1.0. Los resultados del corpus Blitzer se muestran en la gráfica 6.5.

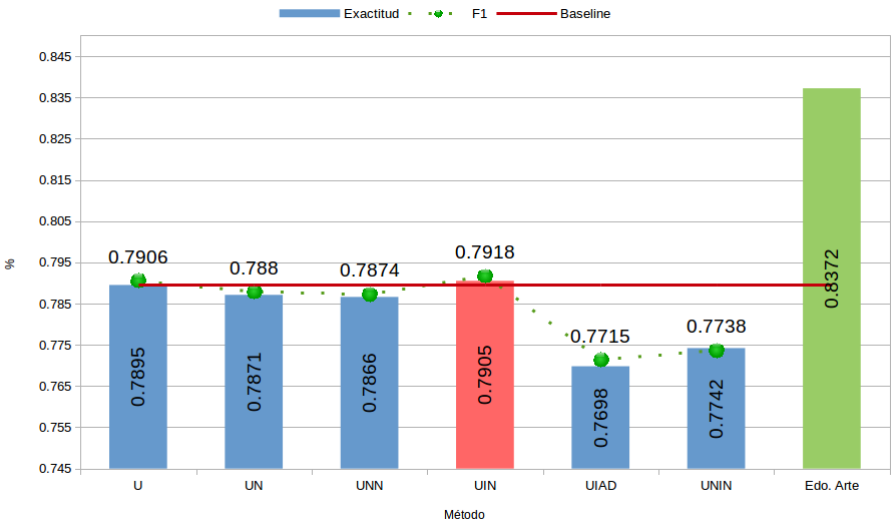


Figura 6.5: Promedio de resultados en corpus Blitzer.

Es importante mencionar que este corpus está formado por opiniones de cuatro dominios distintos. Para realizar esta gráfica de calcularon los promedios de los re-

sultados de cada uno de los dominios. Del mismo modo se calcularon los promedios de los resultados del estado del arte reportados en el trabajo dónde se presentó esta colección. Es importante mencionar que se calcularon las desviaciones estándar. La desviación estándar de los resultados en el estado del arte es de 3.11 mientras la de los mejores resultados (UIN) es menor con aproximadamente un punto siendo 2.67. Notamos que la barra de los resultados más altos obtenidos en el estado del arte es algunos puntos mejor que los mejores resultados obtenidos con los métodos propuestos en este trabajo. Esos resultados se encuentran en el artículo [Blitzer et al., 2007], en ese artículo se realizaron con un clasificador SVM usando como atributos unigramas y bigramas de palabras. Se aplicaron algunas técnicas de caracterización de documentos y de selección de atributos. La única comparación con ese trabajo es que utilizamos la misma colección para la experimentación.

En la gráfica 6.6 se muestran los resultados de cada uno de los dominios con la variación que dio mejores resultados en promedio.

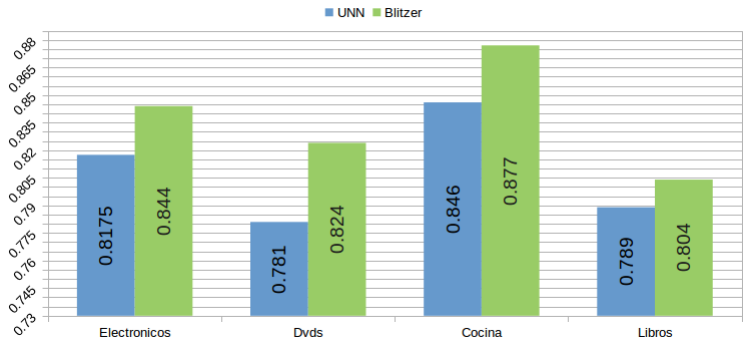


Figura 6.6: Resultados en corpus Blitzer por cada dominio.

La última base de datos de experimentación fue la de IBM v1.0 cuyos resultados se encuentran en la tabla 6.7.

Vemos que los resultados en los diferentes métodos de clasificación son poco variantes. Al ver este comportamiento se pensó en que podríamos haber quitado algunas características importantes al eliminar las palabras vacías. Para saber que tan cierta podría ser esta teoría se hicieron experimentos sin hacer el filtrado de las palabras

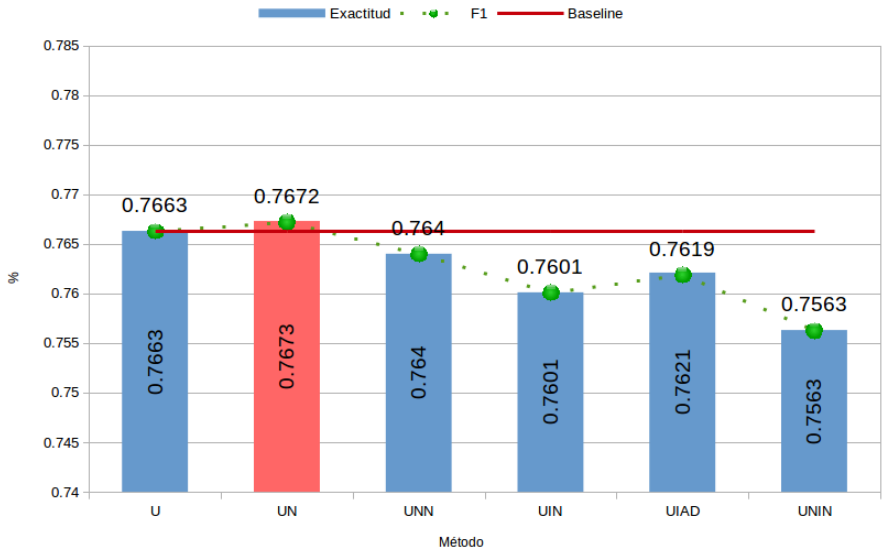
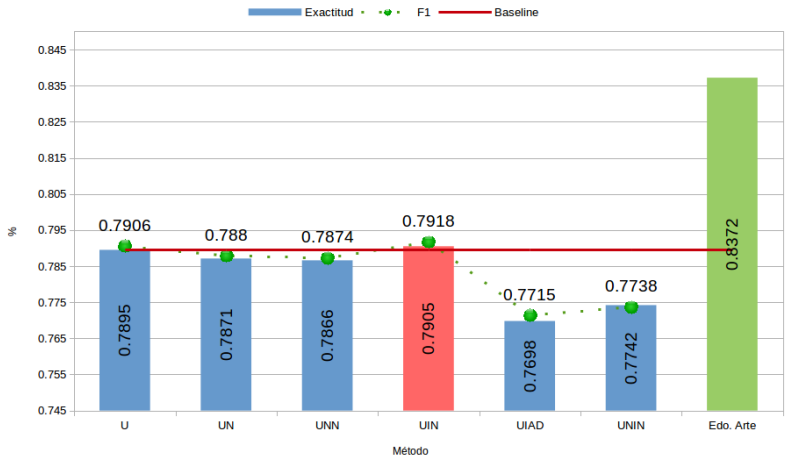


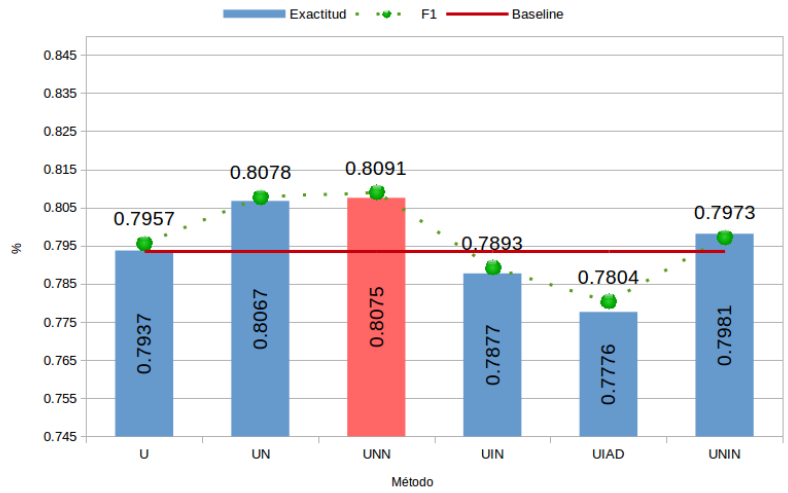
Figura 6.7: Promedio de resultados en corpus IBM v1.0.

vacías en los documentos. Los resultados comparativos se encuentran en la gráficas 6.8 y 6.9. En el idioma Inglés las palabras vacías resultan ser de ayuda para la clasificación dando hasta 2 puntos porcentuales extra en la exactitud. En algunas tareas de clasificación las palabras vacías se consideran atributos de ayuda por ser características estilísticas. En este caso, puede deberse a esa misma razón o bien a que en Inglés las palabras vacías suelen ser también auxiliares que permiten dar el tiempo de una oración (pasado, presente o futuro) o dar algunos otros detalles de la expresión. Se abundará respecto a este punto en las siguientes secciones.

En ambas gráficas de comparación notamos que los resultados de los experimentos sin eliminar las palabras vacías resultan mejores que haciendo la limpieza de stopwords. Recordando, los experimentos en Español se realizaron eliminando la lista de palabras vacías, pero después de ver los resultados de estos experimentos se realizó también la prueba sin hacer el filtrado de stopwords. En el caso del Español los resultados son mejores eliminando las palabras vacías de los documentos pero en promedio las diferencias de exactitud entre dejar o eliminar las palabras vacías son menores a 0.5 puntos porcentuales. Los resultados completos se encuentran en el apéndice A.



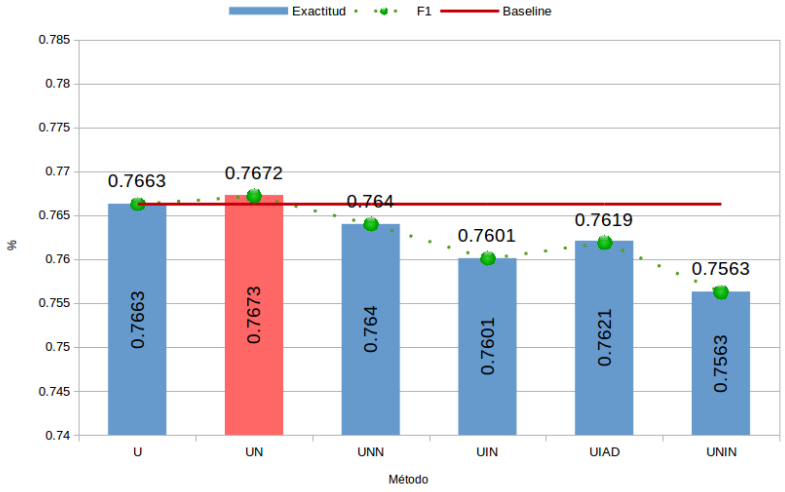
(a) Eliminando palabras vacías.



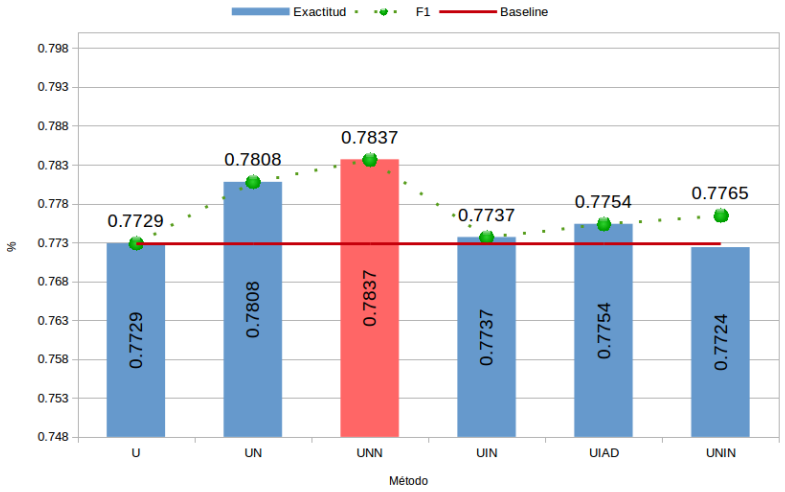
(b) Usando palabras vacías.

Figura 6.8: Resultados comparativos en corpus Blitzer.

En la gráfica 6.10 se encuentran los resultados promediados de todos los experimentos realizados en ambos idiomas y en todos los dominios. En los resultados generales la mejor variación es usando unigramas + bigramas de negación con normalizado por cada tipo de vocabulario, tanto usando como sin usar palabras vacías. Es notorio que dejando las palabras vacías dentro de los documentos las mejoras son más significativas. Los métodos que incluyen intensificación no dan buenos resultados.



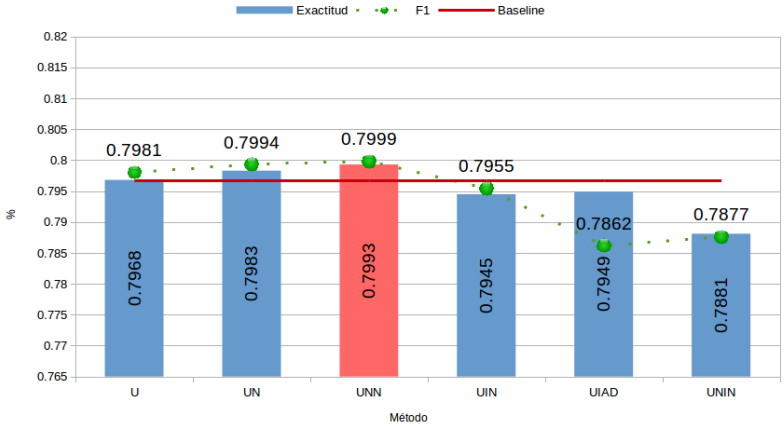
(a) Eliminando palabras vacías.



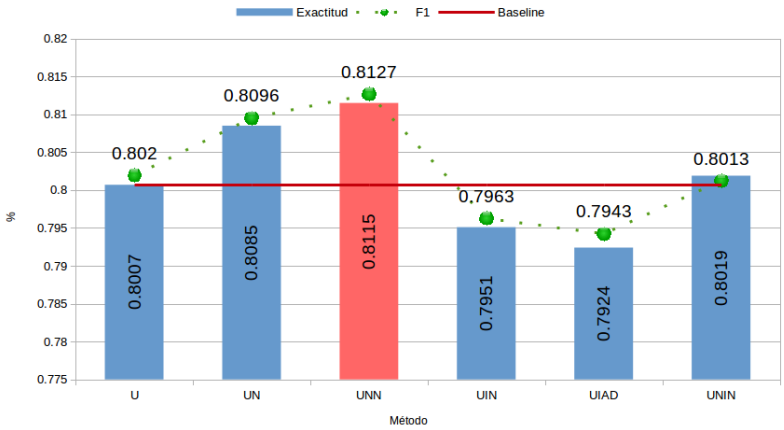
(b) Usando palabras vacías.

Figura 6.9: Resultados en corpus IBM v1.0.

Se aplicó la prueba estadística de Wilconxon sobre los métodos que incluyen información de negación en relación al método baseline los resultados indicaron que los resultados tienen significancia estadística del 95 % en las pruebas que incluyen palabras vacías en los documentos. En el caso de las pruebas sin incluir las palabras vacías en los documentos las variaciones en los resultados son menores por lo que no se alcanzan superar la prueba de significancia estadística.



(a) Eliminando palabras vacías.



(b) Usando palabras vacías.

Figura 6.10: Promedios de los resultados en ambos idiomas.

Las tablas completas de resultados por clase de cada uno de los experimentos se encuentran en el apéndice A.

6.3. Análisis de Resultados

Con la presentación de los resultados surgieron algunas preguntas que se exponen en los siguientes puntos:

¿Qué pasó después de utilizar la negación?

- **Los que permanecieron correctos.** Muchas de las opiniones que fueron clasificadas como correctas sin utilizar la negación continuaron como correctas después de utilizar los bigramas de negación. La principal característica de estos documentos es que no tienen negaciones o tienen muy pocas, por lo que no eran afectados con la inclusión de estas nuevas características. En promedio del total de palabras de estas opiniones sólo el 2 % son bigramas de negación.
- **Los que permanecieron incorrectos.** Este tipo de documento comparten la característica del caso anterior de no tener presencia de negaciones. Se trata de documentos grandes que niegan palabras significativas de la clase contraria a la que pertenecen pero sin hacer uso de ninguna partícula negativa. Un ejemplo es el review #8 del corpus CMR, tiene un tamaño de 984 palabras: *"...y es que cuándo los paisajes te parecen más bonitos e interesantes que la historia de amor que se está contando, es algo que no funciona, falla la pasión, falla la emoción, falla lo conmovedor..."*
- **Los que cambiaron de incorrectos a correctos.** En promedio, el 13 % de las palabras de estos documentos son bigramas de negación. Al contener muchas negaciones es fácilmente explicable que sean corregidos después de tomar en cuenta este fenómeno.
- **Los que cambiaron de correctos a incorrectos.** Este tipo de casos se dió en documentos que tienen muchas partículas negativas pero que esas partículas no actuan como negaciones, por ejemplo, expresiones fijas, dichos o incluso expresiones neutrales:
  - *"No tiene ni ton ni son"*
  - *"Ni más ni menos"*
  - *"No es mala ni buena"*

¿Por qué los métodos que incluyen intensificación no obtuvieron buenos resultados en Español?

La lista de intensificaciones y atenuantes puede ser la mejor respuesta a esta pregunta. Ambas listas fueron desarrolladas por estudiantes hablantes de Inglés y después traducidas al Español. Palabras como "incluso", "real" y "menudo" son consideradas como intensificadores y no queda muy claro que realmente cumplan esa función en los textos. Además, palabras como "bien", "gran" y "medio" están también incluidas, estas palabras pueden representar un problema si no hay un sistema de desambiguación, puesto que esos términos funcionan también como adjetivos calificativos o también como objetos.

En el caso del Inglés la lista tiende a ser más clara, principalmente porque fue desarrollada por personas cuya lengua materna es el Inglés. Sin embargo, se trata también de una lista de palabras fuera de contexto, donde las interpretaciones pueden cambiar dependiendo de su uso.

¿Por qué los métodos que incluyen negación e intensificación no funcionaron?

La idea de ingresar información de estos dos fenómenos lingüísticos en un sistema de clasificación de opinión no tuvo el alto impacto que se esperaba. Esto puede deberse a varias razones, la principal es que la lista de intensificadores y atenuantes puede no ser la ideal para ninguno de los idiomas. Estas listas de términos de intensificación fueron desarrolladas por estudiantes hablantes de Inglés y después traducidas al Español. Además es necesario explorar otros valores de afectación para los intensificadores y atenuantes para poder concluir al respecto.

Por otro lado, el unir la negación y la intensificación es un proceso complicado. En primer lugar, porque al hablarse de textos de opinión con alta carga de subjetividad, el grado de polaridad puede depender completamente del contexto sobre el cuál se esté hablando. En segundo lugar existen casos donde la opinión no es del todo clara siendo incluso difícil de comprender por un lector. Es el caso del siguiente ejemplo en Español, dónde hay palabras afectadas por intensificaciones y por negaciones. Además este ejemplo debería resolverse como neutral, sin embargo, los métodos propuestos parten de un problema binario. Es interesante mencionar que esta frase fue obtenida

de una opinión etiquetada como positiva.

- *"...prefiero situarme en un término medio,* ***ni tan*** *increíble como estoy leyendo por ahí,* ***ni tan*** *mala como se podría esperar..."*

Definir el grado de polaridad de expresiones con negaciones e intensificaciones o atenuaciones no es sencillo ni para los humanos. Si hubiera que ordenar en una escala de menos positivo a más positivo las expresiones: es buena, es muy buena, no es buena, no es muy buena, es poco buena, no es poco buena. ¿Cuál sería el orden?. La intensificación y la negación necesitan ser mayormente estudiadas para poder ser incluidas de manera conjunta en un sistema de clasificación automático.

¿Por qué al utilizar las palabras vacías mejora la exactitud de clasificación en Inglés?

Ésto podría deberse al comportamiento de las palabras vacías en el idioma. En ambos idiomas las palabras vacías suelen ser conectores, preposiciones o incluso algunos verbos muy frecuentes. Sin embargo, en el caso del Español esas palabras suelen ser sólo de estilo y una expresión puede ser totalmente comprendida si se eliminan. En el caso del Inglés, en muchas ocasiones, estas palabras suelen dar el sentido a la oración o incluso el tiempo verbal.

En Español decir oraciones sin el sujeto de la oración es completamente entendible y hasta cierto punto correcto, ya que la conjugación incluye a la persona: "estás","cantarás" o "bailaste". En Inglés por el contrario es obligatorio escribir el sujeto de las oraciones y auxiliares de tiempo. Para escribir las mismas palabras o expresiones serían "you are","you will sing" y "you danced". Con estos ejemplos simples vemos que mientras en un idioma sólo necesitamos una palabras para expresar algo, en Inglés podemos hasta necesitar tres términos y que incluso dos de esos términos son etiquetados como stopwords. Ésto afecta en la clasificación de polaridad, al menos en el caso de las colecciones presentadas en este trabajo. Es probable que ciertas palabras vacías sean más utilizadas en algunas de las clases y con ello ayuden a la clasificación.

6.3.1. Discusión

Los mejores resultados de clasificación entre los métodos presentados en este trabajo de tesis se obtuvieron con la inclusión de bigramas de negación. Aunque los resultados obtenidos en algunos casos no superaron a los trabajos del estado del arte es importante notar que los métodos presentados son métodos sencillos de clasificación, mientras que otros trabajos tienen sistemas que incluyen un fuerte preprocesamiento de datos (*v. gr.* POS taggers, análisis sintáctico, análisis de árboles de dependencias, etc.). La selección de atributos relevantes es también un tema importante abordado en trabajos del estado del arte, en varios de los trabajos presentados se usan técnicas de disminución de características. En este trabajo sólo se eliminan las palabras vacías de los documentos. Aunque una menor selección de atributos permite tratar los documentos en su estado más natural, si es necesario añadir métodos de selección de características para hacer mejoras en el sistema de clasificación presentado en este trabajo.

El tratamiento de la negación ha sido mayormente estudiada en sus campos de identificación y definición del alcance. En este trabajo se busca poder utilizar la información de negación para mejorar los niveles de clasificación. La mayoría de los métodos de clasificación que tratan de utilizar la negación para clasificación de polaridad basan su funcionamiento en diccionarios. En esta tesis se buscó hacerlo desde métodos de aprendizaje, específicamente, tomando como base el NBM.

Cabe mencionar que un trabajo anterior [Narayanan et al., 2013] se exploró una aproximación usando el algoritmo NB y la inclusión del tratamiento de la negación. Sin embargo, dicho trabajo se orientó en la modificación del vocabulario utilizado para mejorar la clasificación. A diferencia de dicho trabajo, en nuestro trabajo se hicieron modificaciones a la forma de calcular las probabilidades de pertenencia a las clases de cada término y se modificó también el algoritmo así como a la representación del método de clasificación.

En cuanto a la utilización de las palabras marcadas con intensificación, los resultados muestran que aún hay trabajo por realizar.

Se hicieron aproximaciones a métodos basados en diccionarios con enfoque híbrido, demostrándose que este tipo de enfoques son mejores que los enfoques tradicionales de diccionarios. Además, se proponen métodos de clasificación basados en NBM que integran información de la negación y la intensificación. Con los resultados obtenidos podemos comparar el método híbrido contra los métodos basados en NBM, concluyendo que los métodos basados en aprendizaje son mejores para la clasificación de documentos según su polaridad, al menos en las colecciones utilizadas para la experimentación. Aunque la negación y la intensificación han sido estudiados e incluidos en métodos de clasificación basados en diccionarios, no han sido igualmente abordados dentro de métodos basados en aprendizaje computacional. Es por ello, que en este trabajo se demostró que es posible mejorar niveles de exactitud añadiendo información de estos dos fenómenos en métodos de clasificación que tienen como base el aprendizaje computacional. No obstante, es necesario buscar nuevos modelos para el tratamiento conjunto de la negación y la intensificación.

CAPÍTULO 7

CONCLUSIONES

Este trabajo realizó aportaciones a varias áreas que rodean el análisis de sentimientos al considerar los fenómenos de la negación y la atenuación. Entre las conclusiones a las que se llegaron están las siguientes:

- Los métodos basados en diccionarios con enfoques híbridos tienen mejor desempeño que los enfoques basados en diccionarios. Esto se debe a que con los enfoques híbridos se ingresa información del dominio sobre el cual se este trabajando. Al aprender las listas de palabras positivas y negativas desde los documentos de entrenamiento puede disminuirse la problemática de que una palabra sea positiva o negativa al considerar aquello que se está criticando.

- Los métodos basados en aprendizaje brindan mejores resultados que los enfoques híbridos, en este trabajo. Realizar clasificadores que sean más complejos que simples conteos de palabras puede dar mejores resultados sobre todo si en los conteos se incluye información de distintos fenómenos del lenguaje.

- Añadir información de fenómenos lingüísticos como la negación y la intensificación ayuda, en cierta medida, a mejorar la exactitud de clasificación en los distintos métodos de clasificación. Es necesario continuar trabajando en formas de añadir la información de esos fenómenos a cada uno de los métodos para incrementar los resultados favorables.

- El método utilizado para calcular el pesos o el valor de pertenencia de cada término a las clases puede mejorar los niveles de exactitud en la clasificación dependiendo del tipo de documentos tratados e incluso del dominio del que hablen esos documentos.

- El tratamiento de la negación si ayuda a mejorar la clasificación de documentos de opinión sobre todo en textos en Español. Incluso el enfoque propuesto puede extenderse al Inglés, donde a pesar de que las diferencias entre los idiomas son importantes, es posible generalizar el tratamiento de las partículas negativas.

- Se comprobó que la intensificación si genera afectaciones en la clasificación de polaridad aunque no es necesariamente de manera positiva. Creemos que esto de debe al tratamiento que se le dio en este trabajo. Se pretendió aplicar ideas parecidas a las aplicadas en métodos basados en lexicones. Resulta obvio que es necesario abundar en el fenómeno de intensificación para proponer su tratamiento en métodos de clasificación basados en aprendizaje.

Las tareas de análisis de sentimientos y especialmente la clasificación de polaridad son difíciles. La dificultad está en la tarea misma puesto que se trata de documentos subjetivos. Incluso es probable que un humano clasifique mal estos documentos debido a que la interpretación de la opinión depende de cada persona. Lo que a alguien puede parecerle positivo, otro lo puede ver como negativo. Si a esto aunamos que la negación y la intensificación pueden incluir grados de apreciación entonces tenemos más niveles de polaridad para confundirnos.

Podemos concluir, de manera general, que este trabajo es una investigación sobre análisis de sentimientos y el sinfín de problemáticas que hay alrededor de los documentos de opinión. Se prestó mayor atención al tratamiento de la negación y a su comportamiento en documentos en Inglés y Español. Del análisis hecho a la negación notamos que dependiendo del autor de cada documento una partícula negativa puede cambiar el sentido completo del texto o sólo cambiarlo un poco, e incluso no modificar en absoluto el sentido (como cuando la negación está incluida en una expresión fija). También notamos que la negación se presenta de manera distinta en cada uno de los dominios sobre los que se trabajaron. Además se agregaron métodos de tratamiento de la intensificación para ver su comportamiento en métodos de aprendizaje. En este

caso se concluye que es necesario un estudio profundo de como añadir éste y otros fenómenos a métodos de clasificación de este tipo.

Para finalizar se exponen algunos puntos a realizarse en futuras investigaciones.

7.1. Trabajo Futuro

El análisis de sentimientos es un área que tiene muchas problemáticas por resolver y es también una investigación que tiene un sinfín de aplicaciones en el mundo real. Seguir estudiando los diferentes fenómenos lingüísticos que interfieren o se encuentran presentes en los documentos de opinión es algo fundamental. Como trabajo futuro a este trabajo se propone lo siguiente:

- Profundizar en la descripción del fenómeno de la negación en documentos de opinión con la finalidad de proponer mejoras a los métodos propuestos.
- Colaborar con lingüistas para hacer un estudio profundo de los intensificadores en los distintos idiomas, su comportamiento y los niveles de afectación que pueden tener en documentos de opinión.
- Identificar el resto de los fenómenos lingüísticos que afectan directamente en la clasificación de documentos según su polaridad como pueden ser ironía, sarcasmo, utilización de expresiones fijas, etc.
- Profundizar en cómo tomar ventaja del uso de las palabras vacías en tareas de clasificación de sentimientos.

Por otro lado, el desarrollo de sistemas funcionales en la vida real es importante, por lo que se plantea como trabajo futuro desarrollar una herramienta de clasificación de opiniones que incluya reconocimiento de entidades y aspecto. Es decir, un clasificador que tenga como salida la etiqueta de polaridad de un grupo de documentos y además brinde información sobre que entidades se nombraron en esos documentos y que aspectos de esas entidades fueron criticados de manera positiva o negativa.

En ese sistema planteado se agregan elementos de clasificación de grupos, identificación de entidades nombradas y reconocimiento de aspectos. Todas esas áreas investigadas desde puntos de vista de tratamiento de lenguaje natural y de aprendizaje computacional.

BIBLIOGRAFÍA

[Abu-Jbara y Radev, 2012] Abu-Jbara, A., & Radev, D. (2012, June). Umichigan: A conditional random field model for resolving the scope of negation (pp. 328-334). In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. Association for Computational Linguistics.

[Agarwal y Mittal, 2013] Agarwal, B., & Mittal, N. (2013). Optimal feature selection for sentiment analysis (pp. 13-24). In Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg.

[Aggarwal y Zhai, 2012] Aggarwal, C. C., & Zhai, C. (2012). Mining text data. Springer Science & Business Media.

[Amir et al., 2014] Amir, S., Almeida, M., Martins, B., Filgueiras, J., & Silva, M. J. (2014). Tugas: Exploiting unlabelled data for Twitter sentiment analysis (pp. 673-677). Proceedings of SemEval.

[Arlot y Celisse, 2010] Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection (pp.40-79). Statistics surveys.

[Barranquero et al., 2015] Barranquero, J., Díez, J., & del Coz, J. J. (2015). Quantification-oriented learning based on reliable classifiers (pp. 591-604). Pattern Recognition, 48(2).

[Blanco y Moldovan, 2011] Blanco, E., & Moldovan, D. I. (2011, March). Some Issues on Detecting Negation from Text (pp. 228-233). In FLAIRS Conference.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation (pp. 993-1022). Journal of Machine Learning Research, 3(Jan).

[Blitzer et al., 2007] Blitzer, J., Dredze, M., & Pereira, F. (2007, June). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification (pp. 440-447). In ACL.

[Brooke et al., 2009] Brooke, J., Tofiloski, M., & Taboada, M. (2009, September). Cross-Linguistic sentiment analysis: From English to Spanish (pp. 50-54). In RANLP.

[Chinchor y Robinson, 1997] Chinchor, N., & Robinson, P. (1997, September). MUC-7 named entity task definition (p. 29). In Proceedings of the 7th Conference on Message Understanding.

[Councill et al., 2010] Councill, I. G., McDonald, R., & Velikovich, L. (2010, July). What's great and what's not: learning to classify the scope of negation for improved sentiment analysis (pp. 51-59). In Proceedings of the Workshop on Negation and Speculation in Natural Language Processing. Association for Computational Linguistics.

[Cover y Thomas, 1991] Cover, T. M., & Thomas, J. A. (1991). Information theory and statistics (pp. 279-335). Elements of Information Theory.

[Cruz et al., 2008] Cruz, F. L., Troyano, J. A., Enriquez, F., & Ortega, J. (2008). Clasificación de documentos basada en la opinión: experimentos con un corpus de crıticas de cine en espanol (pp. 73-80). Procesamiento de Lenguaje Natural, 41.

[Ding y Peng, 2005] Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data (pp. 185-205). Journal of Bioinformatics and Computational Biology, 3(02).

[Dobre et al., 2007] Dobre, O. A., Abdi, A., Bar-Ness, Y., & Su, W. (2007). Survey of automatic modulation classification techniques: classical approaches and new trends (pp. 137-156). Communications, IET, 1(2).

[Española, 2009] Española, R. A. (2009). Asociación de Academias de la Lengua Española. Nueva gramática de la lengua española. Volumen I y II.

[Esuli y Sebastiani, 2006] Esuli, A., & Sebastiani, F. (2006, May). Sentiwordnet: A publicly available lexical resource for opinion mining (pp. 417-422). In Proceedings of LREC.

[Fach, 2012] Flach, P. (2012). Machine Learning: The art and science of algorithms that make sense of data. Cambridge University Press.

[Forman, 2003] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification (pp- 1289-1305).The Journal of Machine Learning Research,3.

[Fuchs y Peres, 1996] Fuchs, C. A., & Peres, A. (1996). Quantum-state disturbance versus information gain: Uncertainty relations for quantum information (p. 2038-2045). Physical Review A, 53(4).

[Gamallo y Garcia, 2014] Gamallo, P., & Garcia, M. (2014). Citius: A naive-bayes strategy for sentiment analysis on english tweets (pp. 171-175). Proceedings of SemEval.

[Gao y Sebastiani, 2015] Gao, W., & Sebastiani, F. (2015, August). Tweet Sentiment: From Classification to Quantification (pp. 97-104). In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. ACM.

[Ghorbel y Jacot, 2011] Ghorbel, H., & Jacot, D. (2011). Sentiment analysis of French movie reviews (pp. 97-108). In Advances in Distributed Agent-Based Retrieval Tools. Springer Berlin Heidelberg.

[González et al., 2015] González, M. D. M., Cámara, E. M., & Valdivia, M. T. M. (2015). CRiSOL: Base de conocimiento de opiniones para el Español (pp. 145-150). Procesamiento del Lenguaje Natural, 55.

[González et al., 2015] González, M. D. M., Cámara, E. M., Valdivia, M. T. M., & Zafra, S. M. J. (2015). eSOLHotel: Generación de un lexicón de opinión en español adaptado al dominio turístico (pp. 21-28). Procesamiento del Lenguaje Natural, 54.

[Guyon y Elisseeff, 2003] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection (pp. 1157-1182). The Journal of Machine Learning Research, 3.

[Han et al., 2011] Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Elsevier.

[Hernández et al., 2004] Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004). Introducción a la Minería de Datos. Editorial Pearson Educación SA, Madrid.

[Hogenboom et al., 2011] Hogenboom, A., Van Iterson, P., Heerschop, B., Frasincar, F., & Kaymak, U. (2011, October). Determining negation scope and strength in sentiment analysis (pp. 2589-2594). In Systems, Man, and Cybernetics (SMC), 2011. IEEE.

[Jiménez et al., 2015] Jiménez Zafra, S. M., Martínez Cámara, E., Martín Valdivia, M. T., & Molina González, M. D. (2015). Tratamiento de la negación en el análisis de opiniones en Español (pp. 37-44). Procesamiento de Lenguaje Natural, 54.

[Kennedy y Inkpen, 2006] Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters (pp. 110-125). Computational Intelligence, 22(2).

[Kent, 1983] Kent, J. T. (1983). Information gain and a general measure of correlation (pp. 163-173). Biometrika, 70(1).

[Kibriya et al., 2004] Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004). Multinomial naive bayes for text categorization revisited. In AI 2004: Advances in Artificial Intelligence (pp. 488-499). Springer Berlin Heidelberg.

[Kloumann et al., 2012] Kloumann, I. M., Danforth, C. M., Harris, K. D., Bliss, C. A., & Dodds, P. S. (2012). Positivity of the English language. PloS one, 7(1), e29484.

[Kramer y Gordon, 2014] Kramer, J., & Gordon, C. (2014). Improvement of a Naive Bayes sentiment classifier using MRS-based (pp.22-29). Lexical and Computational Semantics (* SEM 2014), 22.

[Lapponi et al., 2012] Lapponi, E., Read, J., & Ovrelid, L. (2012, December). Representing and resolving negation for sentiment analysis (pp. 687-692). In Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference. IEEE.

[Le y Mikolov, 2014] Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents (pp. 1188-1196). In ICML (14).

[Lee et al., 2004] Lee, J. W., & Baik, D. K. (2004). A model for extracting keywords of document using term frequency and distribution. In Computational Linguistics and Intelligent Text Processing (pp. 437-440). Springer Berlin Heidelberg.

[Lewis, 1991] Lewis, D. D. (1991, February). Evaluating text categorization I (pp. 312-318). Proceedings of Speech and Natural Language Workshop. Morgan Kaufmann, California, USA, 91.

[Liebrecht et al., 2013] Liebrecht, C. C., Kunneman, F. A., & van den Bosch, A. P. J. (2013). The perfect solution for detecting sarcasm in tweets not (pp. 29-37).Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.

[Lin et al., 2010] Lin, C., He, Y., & Everson, R. (2010, July). A comparative study of Bayesian models for unsupervised sentiment detection (pp. 144-152). In Proceedings of the Fourteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics.

[Liu y Zhang, 2012] Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis (pp. 415-463). In Mining Text Data. Springer US.

[Liu, 2012] Liu, B. (2012). Sentiment analysis and opinion mining (pp. 1-167). Synthesis Lectures on Human Language Technologies, 5(1).

[Llorente et al., 2015] Roberto, J. A., Salamó Llorente, M., & Martí Antonín, M. A. (2015). Polarity analysis of reviews based on the omission of asymmetric sentences (pp. 77-84). Procesamiento del Lenguaje Natural. 2015, 54.

[Maas et al., 2011] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis (pp. 142-150).

In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics.

[Martínez-Cámara et al., 2014] Martínez-Cámara, E., Martín-Valdivia, M. T., Molina-González, M. D., & Perea-Ortega, J. M. (2014). Integrating Spanish lexical resources by meta-classifiers for polarity classification (pp. 1-17). Journal of Information Science.

[McCallum y Nigam, 1998] McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification (pp. 41-48). In AAAI-98 Workshop on Learning for Text Categorization, 752..

[Mendoza et al., 2011] Mendoza, M., Ortiz, I., & Rojas, V. (2011). Categorización de texto en bases documentales a partir de modelos computacionales livianos (pp. 251-274). Revista signos, 44(77).

[Michalski et al., 2013] Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). Machine learning: An artificial intelligence approach. Springer Science & Business Media.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space (pp. 3111-3119). Advances in Neural Information Processing Systems.

[Molina et al., 2013] Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M. T., & Perea-Ortega, J. M. (2013). Semantic orientation for polarity classification in spanish reviews (pp. 7250-7257). Expert Systems with Applications, 40(18).

[Morante y Blanco, 2012] Morante, R., & Blanco, E. (2012, June). * SEM 2012 shared task: Resolving the scope and focus of negation (pp. 265-274). In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. Association for Computational Linguistics.

[Nadali et al., 2010] Nadali, S., Murad, M. A. A., & Kadir, R. A. (2010, June). Sentiment classification of customer reviews based on fuzzy logic (pp. 1037-1044). In Information Technology (ITSim), 2010 International Symposium, 2. IEEE.

[Narayanan et al., 2013] Narayanan, V., Arora, I., & Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced Naive Bayes model (pp. 194-201). In Intelligent Data Engineering and Automated Learning–IDEAL 2013. Springer Berlin Heidelberg.

[Nigam et al., 2000] Nigam, K., McCallum, A. K., Thrun, S.,& Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM (pp. 103-134). Machine Learning, 39(2-3).

[Ohana et al., 2011] Ohana, B., Tierney, B., & Delany, S. J. (2011, March). Domain independent sentiment classification with many lexicons (pp. 632-637). In Advanced Information Networking and Applications (WAINA), 2011 IEEE.

[Packard et al., 2014] Packard, W., Bender, E. M., Read, J., Oepen, S., & Dridan, R. (2014, June). Simple negation scope resolution through deep parsing: A semantic solution to a semantic problem (pp. 69-78). In ACL (1).

[Pang et al., 2002] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques (pp. 79-86). In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics.

[Pang y Lee, 2005] Pang, B., & Lee, L. (2005, June). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales (pp. 115-124). In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics.

[Peng et al., 2005] Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy (pp. 1226-1238). Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(8).

[Pérez et al., 2012] Perez-Rosas, V., Banea, C., & Mihalcea, R. (2012, May). Learning sentiment lexicons in Spanish (p. 73). In LREC, 72.

[Peng y Schuurmans, 2003] Peng, F., & Schuurmans, D. (2003). Combining naive Bayes and n-gram language models for text classification (pp. 335-350).In European Conference on Information Retrieval. Springer Berlin Heidelberg.

[Pluim et al., 2003] Pluim, J. P., Maintz, J. A., & Viergever, M. A. (2003). Mutual-information-based registration of medical images: a survey (pp. 986-1004). Medical Imaging, IEEE Transactions on, 22(8).

[Polanyi y Zaenen, 2006] Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. In Computing attitude and affect in text: Theory and applications (pp. 1-10). Springer Netherlands.

[Qiang, 2010] Qiang, G. (2010, May). An effective algorithm for improving the performance of naïve bayes for text classification (pp. 699-701). In 2010 Second International Conference on Computer Research and Development.

[Ramage et al., 2010] Ramage, D., Dumais, S. T., & Liebling, D. J. (2010). Characterizing microblogs with topic models (pp. 130-137). ICWSM, 10.

[Read et al., 2012] Read, J., Velldal, E., Ovrelid, L., & Oepen, S. (2012, June). Uio 1: Constituent-based discriminative ranking for negation resolution (pp. 310-318). In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation. Association for Computational Linguistics.

[Rennie et al., 2003] Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003, August). Tackling the poor assumptions of naive bayes text classifiers (pp. 616-623). In ICML (3).

[Reza, 1961] Reza, F. M. (1961). An introduction to information theory. Courier Corporation.

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization (pp. 1-47). ACM computing surveys (CSUR), 34(1).

[Shen y Jiang, 2003] Shen, Y., & Jiang, J. (2003). Improving the performance of Naive Bayes for text classification. CS224N Spring.

[Shimodaira, 2014] Shimodaira, H. (2014). Text classification using Naive Bayes. Learning and Data Note, 7.

[Schneider, 2005] Schneider, K. M. (2005, February). Techniques for improving the performance of naive bayes for text classification (pp. 682-693). In International Conference on Intelligent Text Processing and Computational Linguistics. Springer Berlin Heidelberg.

[Shoukry y Rafea, 2012] Shoukry, A., & Rafea, A. (2012, May). Sentence-level Arabic sentiment analysis (pp. 546-550). In Collaboration Technologies and Systems (CTS), 2012 International Conference. IEEE.

[Siedlecki y Sklansky, 1993] Siedlecki, W., & Sklansky, J. (1993). On automatic feature selection (pp. 63-87). In Handbook of Pattern Recognition and Computer Vision. World Scientific Singapore.

[Singh et al., 2010] Singh, S. R., Murthy, H. A., & Gonsalves, T. A. (2010). Feature selection for text classification based on gini coefficient of inequality (pp. 76-85). FSDM, 10.

[Socher et al., 2013] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank (pp. 1642). In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 1631.

[Taboada et al., 2008] Taboada, M., Voll, K., & Brooke, J. (2008). Extracting sentiment as a function of discourse structure and topicality. Simon Fraser Univeristy School of Computing Science Technical Report.

[Taboada et al., 2011] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis (pp.267-307). Computational Linguistics, 37(2).

[Tripathy et al., 2016] Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach (pp. 117-126). Expert Systems with Applications, 57.

[Vapnik, 1999] Vapnik, V. N. (1999). An overview of statistical learning theory (pp. 988-999). Neural Networks, IEEE Transactions on, 10(5).

[Vinodhini y Chandrasekaran, 2012] Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey (pp. 282-292). International Journal of Advanced Research in Computer Science and Software Engineering, 2(6).

[Wiegand et al., 2010] Wiegand, M., Balahur, A., Roth, B., Klakow, D., & Montoyo, A. (2010, July). A survey on the role of negation in sentiment analysis (pp. 60-68). In Proceedings of the workshop on negation and speculation in natural language processing. Association for Computational Linguistics.

[Wilson et al., 2005] Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis (pp. 347-354). In Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics.

[Witten y Frank, 2005] Witten, I. H., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[Zhang et al., 2009] Zhang, C., Zeng, D., Li, J., Wang, F. Y., & Zuo, W. (2009). Sentiment analysis of Chinese documents: From sentence to document level (pp. 2474-2487). Journal of the American Society for Information Science and Technology, 60(12).

[Zhu et al., 2014] Zhu, X., Guo, H., Mohammad, S., & Kiritchenko, S. (2014, June). An Empirical Study on the Effect of Negation Words on Sentiment (pp. 304-313). In ACL (1) .

[Zou et al., 2014] Zou, B., Zhou, G., & Zhu, Q. (2014). Negation Focus Identification with Contextual Discourse Information (pp. 522-530). In ACL.

APÉNDICE A

En este apéndice se añaden algunas tablas de los experimentos realizados con distintos tamaños de ventanas, y comparaciones de experimentos con distintas configuraciones en los algoritmos de definición del alcance de la negación y de la intensificación. Además se muestran las tablas de resultados que se muestran como gráficas en el capítulo 6 en la sección de resultados.

A.1. Algoritmo de negación: configuraciones

Los experimentos con distintos tamaños de ventana en el algoritmos de definición del alcance de la negación se hicieron en el corpus CMR con un enfoque híbrido y con el pesado propuesto que incluye información de vocabulario de cada una de las clases. Los resultados se muestran en la tabla A.1.

	Ventana	**2**	**4**	**6**	**8**
Con nexos	**F1**	0.7332	**0.7342**	0.7306	0.7295
adversativos	**Exactitud**	0.7369	**0.7373**	0.7334	0.7321
Sin nexos	**F1**	0.7040	0.7089	0.7157	0.7009
adversativos	**Exactitud**	0.7174	0.7203	0.7172	0.7126

Tabla A.1: Resultados con distintos tamaños de ventana en el algortimo de negación.

Otra variación estudiada en el algoritmo de definición del alcance de la negación fue estudiar el finalizar el alcance de la partícula negativa por encontrarse un nexo

adversativo. Los resultados se muestran también en la tabla A.1. Se concluyó que el mejor tamaño de ventana es 4 con inclusión de nexos adversativos como bandera de término del alcance de las palabras negativas.

Entre las variaciones en el tratamiento de la negación se encontró la forma de modificar los documentos y marcarlos cuándo apareciera una negación. En el capítulo 4 se explicaron dos variaciones posibles: 1) Modificar el documento añadiendo bigramas de palabras negadas y 2) Cambiar las palabras del documento agregando una bandera de negación si fueron afectadas. Teniendo la frase "La película no me gustó" las dos opciones de modificación del documento serían 1)"La película no me no_me gustó no_gustó" y 2)"La película no no_me no_gustó". En la tabla A.2 Se encuentran los resultados en algunas colecciones con estas variaciones. La primera variación se nombra "dejando unigramas" y la segunda es llamada "quitando unigramas". Después de estos resultados se decidió hacer el todos los experimentos con la variación 2.

	Dejando unigramas		**Quitando unigramas**	
Corpus	**F1**	**Exactitud**	**F1**	**Exactitud**
CMR	0.7899	0.7887	0.7910	0.7899
COAH	0.9056	0.9029	0.9137	0.9117
Blitzer kitchen	0.8352	0.8350	0.8463	0.8460
IMBd	0.7793	0.7793	0.7837	0.7837
Promedio	0.8275	0.8264	**0.8336**	**0.8328**

Tabla A.2: Resultados variaciones en la modificación de los documentos.

A.2. Algoritmo de intensificación: configuraciones

Las variaciones en el algoritmo de definición del alcance de la intensificación se realizaron en los tamaños de ventana. Se hicieron experimentos en colecciones en Español e Inglés. Se decidió usar ventana tamaño dos por tener los mejores resultados en promedio. Los resultados se encuentran en la tabla A.3.

A.3. Resultados: Método híbrido

En esta sección se añaden las tablas de resultados que son presentados en las gráficas del capítulo 6, de la sección del método híbrido.

Ventana Corpus	1	2	3
CMR	0.7881	0.7878	0.7878
COAH	0.9197	0.9237	0.9126
Blitzer DVDs	0.7672	0.7701	0.7645
Blitzer Electronics	0.7912	0.7937	0.7919
Promedio	0.8165	**0.8188**	0.8142

Tabla A.3: Resultados con distintos tamaños de ventana en el algoritmo de intensificación.

Los resultados en la tabla A.4 son los resultados por clase en el corpus CMR. Podemos ver que las mayores mejoras se obtienen en la clasificación de los documentos positivos después de utilizar el tratamiento de la negación.

Pesado	Atributos	Clase	Precisión	Recuerdo	F1	Exactitud
Frecuencia relativa	U	**POS**	0.9884 (±0.088)	0.0247 (±0.009)	0.0481 (±0.017)	0.5116 (±0.004)
		NEG	0.5058 (±0.002)	0.9984 (±0.002)	0.6715 (±0.002)	
	UN	**POS**	0.8780 (±0.043)	0.3095 (±0.050)	04893 (±0.058)	0.6470 (±0.028)
		NEG	0.5916 (±0.020)	0.9535 (±0.287)	0.7300 (±0.0181)	
Pesado propuesto	PP	**POS**	0.9177 (±0.029)	0.3774 (±0.111)	0.5276 (±0.111)	0.6727 (±0.054)
		NEG	0.6114 (±0.044)	0.9680 (±0.008)	0.7486 (±0.032)	
	PPN	**POS**	0.7947 (±0.037)	0.5872 (±0.064)	0.7096 (±0.028)	**0.7373 (±0.014)**
		NEG	0.7029 (±0.027)	0.8288 (±0.054)	0.7589 (±0.015)	

Tabla A.4: Resultados: Enfoque híbrido en CMR.

A.4. Resultados: Método de aprendizaje

Las tablas presentadas están ordenadas cómo se presentaron los resultados en el capítulo de Resultados. Primero se muestran los resultados de los corpus en Español. Después, se agregan las tablas de las colecciones en Inglés. Por último, se encuentran las tablas de los resultados promediados de cada uno de los idiomas y los totales.

A.4.1. Colecciones en Español

En las tablas A.5, A.6 y A.7 se encuentran los resultados de los métodos basados en aprendizaje haciendo un filtrado de palabras vacías en los documentos.

Variación	Prec POS	Rec POS	Prec NEG	Rec NEG	F1	Exactitud
UNI	0.8022	0.748	0.7679	0.8165	0.7822	0.7836
UNI+Neg	0.81433	0.7623	0.7803	0.8261	**0.7942**	**0.7957**
UNI+Neg+N	0.7798	0.8072	0.8046	0.7726	0.7899	0.791
UNI+Int+N	0.7603	0.8166	0.8062	0.7427	0.7796	0.7814
UNI+Int+AD	0.811	0.745	0.7682	0.8269	0.7859	0.7878
UNI+Neg+Int+N	0.7669	0.837	0.8248	0.7448	0.7909	0.7901

Tabla A.5: Resultados en los textos completos de CMR.

Variación	Prec POS	Rec POS	Prec NEG	Rec NEG	F1	Exactitud
UNI	0.7674	0.7448	0.7531	0.7732	0.7590	0.7596
UNI+Neg	0.7423	0.7827	0.7724	0.7270	0.7549	0.7561
UNI+Neg+N	0.7689	0.7505	0.7584	0.7718	**0.7611**	**0.7624**
UNI+Int+N	0.7377	0.7308	0.7334	0.738	0.7344	0.735
UNI+Int+AD	0.7517	0.7387	0.7439	0.7537	0.7462	0.7470
UNI+Neg+Int+N	0.7429	0.7362	0.7392	0.7425	0.7393	0.7389

Tabla A.6: Resultados en los títulos de CMR.

Variación	Prec POS	Rec POS	Prec NEG	Rec NEG	F1	Exactitud
UNI	0.8976	0.9411	0.9392	0.8921	0.9166	0.9175
UNI+Neg	0.8766	0.9588	0.9548	0.8647	0.9117	0.9137
UNI+Neg+N	0.9503	0.9019	0.9085	0.9529	**0.9274**	**0.9284**
UNI+Int+N	0.9444	0.898	0.9053	0.947	0.9225	0.9237
UNI+Int+AD	0.9033	0.945	0.9436	0.898	0.9215	0.9225
UNI+Neg+Int+N	0.9512	0.8784	0.8904	0.9549	0.9166	0.9163

Tabla A.7: Resultados en los documentos del corpus COAH.

A.4.2. Colecciones en Inglés

Los resultados del corpus Blitzer se encuentran en las tablas A.8, A.9, A.10 y A.11.

Variación	Prec POS	Rec POS	Prec NEG	Rec NEG	F1	Exactitud
UNI	0.7977	0.754	0.7677	0.807	0.7805	0.7816
UNI+Neg	0.7974	0.776	0.7823	0.802	0.789	0.7894
UNI+Neg+N	0.8298	0.719	0.7523	0.852	0.7855	0.7882
UNI+Int+N	0.8062	0.709	0.7403	0.829	0.769	0.7711
UNI+Int+AD	0.7951	0.726	0.7482	0.811	0.7685	0.77
UNI+Neg+Int+N	0.8205	0.716	0.7483	0.843	0.7795	0.7785

Tabla A.8: Resultados del corpus Blitzer en el dominio de libros.

Variación	Prec POS	Rec POS	Prec NEG	Rec NEG	F1	Exactitud
UNI	0.8342	0.662	0.7212	0.868	0.765	0.7713
UNI+Neg	0.832	0.702	0.7432	0.859	0.7805	0.784
UNI+Neg+N	0.8319	0.704	0.7442	0.858	0.781	0.7845
UNI+Int+N	0.8162	0.687	0.7315	0.846	0.7665	0.7701
UNI+Int+AD	0.8239	0.636	0.7057	0.865	0.7505	0.7576
UNI+Neg+Int+N	0.8207	0.684	0.73	0.851	0.7675	0.7656

Tabla A.9: Resultados del corpus Blitzer en el dominio de dvds.

Por último los resultados del corpus IMB estan en la tabla A.12.

Por último, la tabla A.13 muestra los promedios de los resultados en los corpora en Español e Ingles respectivamente. Además hacen una comparación de la diferencia de exactitud y medida F1 alcanzada eliminando o no las palabras vacías de los documentos.

Variación	Prec POS	Rec POS	Prec NEG	Rec NEG	F1	Exactitud
UNI	0.8045	0.805	0.8052	0.803	0.804	0.8044
UNI+Neg	0.8096	0.82	0.8175	0.806	0.813	0.8132
UNI+Neg+N	0.824	0.8008	0.812	0.827	0.8175	0.8177
UNI+Int+N	0.8027	0.779	0.7854	0.808	0.7935	0.7937
UNI+Int+AD	0.8114	0.718	0.7477	0.833	0.7755	0.7775
UNI+Neg+Int+N	0.8285	0.776	0.7893	0.838	0.807	0.8067

Tabla A.10: Resultados del corpus Blitzer en el dominio de electrónicos.

Variación	Prec POS	Rec POS	Prec NEG	Rec NEG	F1	Exactitud
UNI	0.8292	0.821	0.823	0.83	0.8255	0.8258
UNI+Neg	0.8395	0.855	0.8522	0.834	0.8445	0.8446
UNI+Neg+N	0.8459	0.847	0.8473	0.845	0.846	0.8463
UNI+Int+N	0.8203	0.826	0.8248	0.818	0.822	0.8223
UNI+Int+AD	0.8366	0.786	0.7984	0.846	0.816	0.8167
UNI+Neg+Int+N	0.8504	0.822	0.8281	0.855	0.8385	0.8384

Tabla A.11: Resultados del corpus Blitzer en el dominio de cocina.

Variación	Prec POS	Rec POS	Prec NEG	Rec NEG	F1	Exactitud
UNI	0.775	0.7692	0.7709	0.7767	0.7729	0.7729
UNI+Neg	0.7792	0.7838	0.7825	0.7778	0.7808	0.7808
UNI+Neg+N	0.7899	0.7731	0.7778	0.7943	0.7837	0.7837
UNI+Int+N	0.7744	0.7726	0.7731	0.7748	0.7737	0.7737
UNI+Int+AD	0.7652	0.7947	0.7864	0.756	0.7754	0.7753
UNI+Neg+Int+N	0.7764	0.7769	0.7767	0.7761	0.7765	0.7765

Tabla A.12: Resultados del corpus IMB v1.0.

	Español				Inglés			
	Dejando vacías		Eliminando vacías		Dejando vacías		Eliminando vacías	
Variación	**F1**	**Exac**	**F1**	**Exac**	**F1**	**Exac**	**F1**	**Exac**
UNI	0.8200	0.8192	0.8180	0.8161	0.7912	0.7895	0.7861	0.7852
UNI+Neg	0.8218	0.8202	0.8253	0.8237	0.8024	0.8015	0.7838	0.7831
UNI+Neg+N	**0.8272**	**0.8261**	**0.8286**	**0.8279**	**0.8040**	**0.8015**	**0.7827**	**0.7821**
UNI+Int+N	0.8133	0.8114	0.8122	0.8114	0.7861	0.7849	0.7854	0.7844
UNI+Int+AD	0.8191	0.8178	0.8138	0.8126	0.7794	0.7771	0.7696	0.7683
UNI+Neg+Int+N	0.8151	0.8156	0.8169	0.8171	0.7931	0.7938	0.7703	0.7706

Tabla A.13: Comparación de resultados eliminando y sin eliminar palabras vacías.