Método Semisupervisado para la Clasificación Automática de Textos de Opinión

por

Nadia Patricia Araujo Arredondo

Tesis sometida como requisito parcial
para obtener el grado de

Maestría en Ciencias en el área de
Ciencias Computacionales

en el

**Instituto Nacional de Astrofísica,
Óptica y Electrónica**
Febrero 2009
Tonantzintla, Puebla

Supervisada por:

Dr. Luis Villaseñor Pineda, INAOE
Dr. Aurelio López López, INAOE

Resumen

Hoy en día se encuentra disponible una gran cantidad de información a través de distintos medios electrónicos, en bibliotecas digitales, en colecciones de documentos o en Internet. La necesidad de acceder a esta información para su extracción y análisis, ha llevado a la creación de diversas formas de manipulación de información, entre las que se encuentra la clasificación de textos. Sin embargo, el crecimiento constante de información hace que la tarea de clasificar documentos de forma manual sea costosa y que requiera de mucho tiempo, por lo que ha surgido el interés por realizar la clasificación de manera automática. Podemos decir entonces que la clasificación automática de textos consiste en colocar un documento dentro de un grupo de clases previamente definidas. La mayor parte del trabajo en esta área se ha enfocado en la clasificación de textos por su tema o tópico. Sin embargo, en los últimos años se ha puesto gran interés en la tarea de clasificación no temática. Algunos ejemplos de esta última son la detección de plagio, la atribución de autoría, la clasificación por género y la clasificación de opiniones. Este trabajo de tesis se enfoca en la tarea de clasificación de opiniones, específicamente se aborda el problema de determinar la polaridad de opiniones, es decir, clasificar aquellas opiniones que expresan algo a favor de aquellas que expresan algo en contra, a nivel de oración, bajo un enfoque de Aprendizaje Computacional utilizando características léxicas. Cabe mencionar que una de las contribuciones de este trabajo es la caracterización de opiniones, necesaria para su clasificación automática. Además, en la actualidad, no existe un *corpus* etiquetado en idioma español, lo que dificulta el proceso de aprendizaje. Es por ello que en este trabajo se dan los primeros pasos para la creación de este *corpus*. Específicamente se propone un enfoque de aprendizaje semi-supervisado de clasificación de textos de opinión, disminuyendo la necesidad de un gran *corpus* ya etiquetado.

Abstract

Today a large amount of information is available through different electronic resources, such as digital libraries, collections of documents or Internet. The need to access this information for its extraction and analysis has led to various forms of information handling, among which is the classification of texts. However, the constant growth of information turns the task of classifying documents by hand expensive and time consuming, requiring to automate the classification process. The automatic classification of texts involves placing a document within a group of predefined classes. Most of the work in this area has focused on the classification of texts by their subject or topic. However, in recent years there has been an increasing interest in the task of non-thematic classification. Examples of non-thematic classification are the detection of plagiarism, authorship attribution, gender classification, and the classification of opinions. This thesis focuses on the task of opinion classification. Specifically, it considers the problem of determining the polarity of opinion in sentences by a Machine Learning approach using lexical features. It is worth mentioning that one of the contributions of this thesis is the characterization of opinions necessary for automatic classification. In addition, currently, there is no tagged *corpus* in Spanish, complicating the learning process. In this work we present the first steps towards the creation of this *corpus*. Specifically it proposes an approach for semi-supervised classification of opinions, reducing the need for a large *corpus* and manual tagging.

Agradecimientos

Mi más sincero agradecimiento al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico otorgado durante mis estudios de maestría a través de la beca No. 201824

De igual manera al Instituto Nacional de Astrofísica, Óptica y Electrónica (IINAOE), especialmente a la Dirección de Formación Académica y a la Coordinación de Ciencias Computacionales por todas las facilidades proporcionadas durante mis estudios de maestría.

También un agradecimiento muy especial a mis asesores de Tesis,
Dr. Luis Villaseñor Pineda y Dr.Aurelio López López por el apoyo
que me brindaron, así como sus consejos, su paciencia y su buen humor.

Quiero agradecer también a los doctores Eduardo Morales Manzanares,
Manuel Montes y Gómez y Saúl Pomares Hernández por su valiosa participación
como miembros del jurado y sus acertadas observaciones
durante el proyecto de investigación.

Finalmente agradezco de todo corazón a mi madre, Juanita, a mi hermana, Ligia por todo su apoyo incondicional. A mi esposo, Juan por todo su apoyo, comprensión y sus palabras de aliento. A todos(as) mis amigos(as) y compañeros(as) de la maestría (octava y novena generación) por su apoyo y su amistad.

Dedicatoria

A Dios,
por ser quien ha estado a mi lado en todo momento dándome la fortaleza necesaria para continuar luchando día tras día

A mi madre, Juanita
por todos tus esfuerzos, sacrificios y apoyo incondicional en todo momento a lo largo de mi vida

A mi hermana, Ligia,
que has estado a mi lado en los momentos más felices, y también en los más difíciles de nuestras vidas

A mi esposo, Juan, el amor de mi vida,
por todo el apoyo incondicional que me has brindado, por tu comprensión,
por tus consejos y palabras de aliento, y sobretodo, por el amor que siempre me has brindado

Índice General

Lista de Figuras

Lista de Tablas

Capítulo 1
Introducción

La comunicación humana ha sido una constante en la historia de la humanidad. Desde tiempos remotos han surgido muchas y variadas formas de expresión, una de ellas y la más importante ha sido la escritura. El hombre ha plasmado sus conocimientos, pensamientos, anhelos, etc., a través de ella, creando documentos, libros, y muchas otras formas de concentración de información. Con el paso del tiempo la tecnología ha hecho posible que todo este acervo se concentre de forma digital creando la fuente de información más grande de nuestra época. Para que estos documentos digitales sean de fácil acceso se han tenido que organizar estas colecciones de tal manera que se permita su recuperación y análisis por medios automáticos. Sin embargo, el problema es complejo y es necesaria la continua investigación en la búsqueda de métodos y representaciones apropiadas. Es por ello que se han creado diversas líneas de investigación para el tratamiento automático de textos, entre las que se encuentran: la recuperación de información, la extracción de información, la búsqueda de respuestas y la clasificación de textos entre otras.

La clasificación de textos es una tarea que facilita la organización de información y la cual consiste en determinar la categoría de un texto, de entre varias categorías predefinidas, de acuerdo a ciertas características presentes en dicho texto. Inicialmente la clasificación de textos se realizaba de manera manual, sin embargo, realizar esta tarea de esta forma es una tarea difícil, costosa

y que requiere de mucho tiempo. Por ello nace la idea de realizar la clasificación de estos documentos de forma automática, y surge el área de Clasificación Automática de Textos. La mayor parte de los trabajos realizados para la Clasificación Automática de Textos se ha enfocado en clasificar textos por su tema, es decir, determinar a que tema o temas pertenece un documento, de entre varios temas dados. Sin embargo, en la última década ha surgido un interés por realizar clasificación no-temática de textos, en donde lo que se busca es caracterizar el texto por cómo se escribe y no por lo qué se escribe. Ejemplos de este tipo de clasificación son: determinar el autor de un texto, determinar el estilo de redacción del autor [1, 2], determinar a qué género se refiere un texto (si es novela o cuento por ejemplo) [3, 4, 5], o a determinar la opinión del autor[1] [6, 7, 8]. En particular, el presente trabajo de investigación se coloca en esta última área.

Dentro de la clasificación de opiniones se encuentran tres subtareas principales. Una de estas subtareas consiste en determinar si un texto dado contiene información objetiva[2] o subjetiva[3]. Otra es determinar la orientación de una opinión, es decir, si el texto dado expresa una opinión a favor o en contra (positiva o negativa, respectivamente). Y la otra subtarea es determinar la fuerza o grado de la opinión. En particular, el presente trabajo se centra únicamente en determinar si un texto dado expresa una opinión a favor o en contra.

En esta tesis se propone un método para la creación de un clasificador automático que sea capaz de distinguir opiniones a favor o en contra. Para ello es necesario encontrar las características que mejor conduzcan a una adecuada identificación de ambas formas de expresión.

[1] Determinar si el autor expresa algo a favor o en contra acerca de algo o alguien.
[2] La información objetiva es aquella en la cuál se describe una situación o evento sin exageración y sin involucrar juicios propios.
[3] La información subjetiva es aquella en la cuál se expresa una opinión acerca de algo o alguien, e involucra juicios o sentimientos del autor.

En la siguiente sección se describe el problema abordado por esta tesis. Después, en la sección 1.2 se presentan los objetivos de la tesis, y por último en la sección 1.3 se expone brevemente la organización de la tesis.

1.1 Descripción del Problema

Dentro del comercio, quien ofrece un producto o servicio, le gustaría saber si su producto o servicio cubre las necesidades de la mayoría de los consumidores, o en su caso, tener la retroalimentación adecuada para mejorar dicho producto suscitando en consecuencia un mayor consumo. De igual forma, un personaje público necesita recabar información de su imagen. Para ello es necesario saber qué opina la gente para poder mejorar o cambiar ciertas conductas. En la actualidad gracias a la Web, se puede acceder a opiniones escritas por innumerables personas acerca de diferentes temas, productos, eventos, personas, etc. Estas opiniones se pueden encontrar en correos, editoriales o en sitios especializados para recabar opiniones, entre otros. Un primer paso en el análisis de estas opiniones es determinar su polaridad, es decir, separar aquellas opiniones que expresan algo a favor de las opiniones que expresan algo en contra. Debido al elevado número de documentos disponibles, la tarea de determinar manualmente qué opiniones expresan algo a favor y qué opiniones expresan algo en contra (polaridad), se convierte en una tarea muy costosa, por lo que es necesario el uso de métodos automáticos.

En este trabajo de tesis se propone un método de clasificación que permite distinguir automáticamente entre oraciones de opiniones que expresan algo a favor y oraciones de opiniones que expresan algo en contra. Este método está basado en técnicas de Aprendizaje Computacional.

Uno de los puntos importantes para la clasificación es saber cuáles son las características o atributos adecuados para poder distinguir entre opiniones a favor y opiniones en contra.

En trabajos previos se ha propuesto el uso de diferentes conjuntos de atributos para realizar la clasificación de textos de opinión como bolsa de palabras[4], bigramas[5] o partes de la oración[6] [6, 7, 8]. En este trabajo, se explora el uso de las Secuencias Frecuentes Maximales como atributos[7], cabe mencionar que este es el primer trabajo que usa esta caracterización en la clasificación de textos de opinión.

Otro punto relevante de nuestro trabajo es que se propone evaluar el método en el idioma español. Prácticamente, todos los trabajos previos que se han realizado dentro del área de clasificación de textos de opinión, se ha evaluado en textos en el idioma inglés. Por supuesto, esto conlleva la necesidad de conformar un *corpus* para el español. Actualmente no existe un *corpus* etiquetado para el idioma español, indispensable para utilizarse en el proceso de Aprendizaje Computacional. Como una consecuencia de lo anterior, en este trabajo se dan los primeros pasos para la creación de este *corpus*, definiendo una serie de criterios para el etiquetado del *corpus*. Dado el costo de la creación de este *corpus* el presente trabajo también exploró la posibilidad de un enfoque de aprendizaje semi-supervisado, disminuyendo la necesidad de un gran *corpus* de entrenamiento etiquetado.

[4] La bolsa de palabras es la representación de un texto (por ejemplo, una frase o un documento) como una desordenada colección de todas las palabras que contiene el texto, haciendo caso omiso de la gramática e incluso del orden de las palabras. [23]
[5] Los bigramas son la representación de un texto como una colección de los conjuntos posibles de dos palabras consecutivas.
[6] Partes de la oración es la representación de un textos como una colección de las palabras acompañadas de una etiqueta que indique a qué parte de la oración pertenece (sustantivo, verbo, adjetivo, adverbio, etc.)
[7] Las secuencias frecuentes maximales se explican en la sección 3.1.2.3

1.2 Objetivos

Objetivo general.

Desarrollar un método de clasificación automática que permita distinguir la polaridad de textos que expresan una opinión.

Objetivos específicos.

- Encontrar los atributos más adecuados para realizar la clasificación de polaridad de opiniones.
- Construir un *corpus* de opiniones etiquetado para el idioma español para la investigación en el área de clasificación de opiniones.
- Desarrollar un método semi-supervisado para la clasificación de polaridad de opiniones.

1.3 Organización de la Tesis

El contenido de esta tesis está estructurado de la siguiente forma. En el capítulo 2 se explican los conceptos básicos necesarios para la comprensión de la tesis, los cuales incluyen conocimientos de Aprendizaje Computacional y Clasificación Automática de Textos.

En el capítulo 3 se presentan conceptos de Clasificación de Textos de Opinión, y la revisión de los trabajos previos referente al área de interés.

En el capítulo 4 se describe el método de caracterización, previo a la implementación del método semi-supervisado propuesto.

En el capítulo 5 se detalla el método semi-supervisado para la clasificación de polaridad, así como los criterios usados para el etiquetado del *corpus* en español. En este capítulo se muestran los experimentos realizados y los resultados obtenidos, así como el *corpus* que fue empleado para dichos experimentos.

Finalmente en el capítulo 6 se presentan las conclusiones y el trabajo futuro.

Capítulo 2
Conceptos Básicos

En este capítulo se explican los fundamentos básicos necesarios para entender los conceptos utilizados a lo largo de la presente tesis. En la primera parte se dan a conocer conceptos de Aprendizaje Computacional. En la segunda parte se proporciona una breve explicación acerca de Clasificación Automática de Textos. Finalmente en la tercera parte de este capítulo se describe en qué consiste la Clasificación de Opiniones y los trabajos previos realizados en esta área.

2.1 Aprendizaje Computacional

Cuando el ser humano adquiere conocimientos, habilidades, actitudes o valores a través del estudio, de la experiencia o la enseñanza, decimos que aprende. Este proceso es fácil para el humano, sin embargo, lograr que una máquina aprenda como lo hace el ser humano es una interrogante que existe desde los inicios de las computadoras. Actualmente no existe una máquina capaz de aprender de la misma manera que lo hace el hombre, sin embargo, se han creado algoritmos eficaces para algunas tareas de aprendizaje.

En términos muy generales, podemos decir que un programa aprende si el desempeño obtenido para realizar alguna tarea, mejora con la experiencia.

De manera formal. Se dice que un programa de computadora **aprende** de la experiencia *E* con respecto a una clase de tareas *T* y una medida de desempeño *P*, si su desempeño en las tareas *T*, medido con *P*, mejora con la experiencia *E* [9].

Podemos decir entonces que el Aprendizaje Computacional estudia los procesos computacionales que hay detrás del aprendizaje en humanos y en las máquinas. Esta disciplina juega un papel importante en muchas áreas de la ciencia.

2.1.1 Tipos de Aprendizaje

Un aspecto importante que influye en el aprendizaje es el grado de supervisión. En algunos casos, un experto en el dominio proporciona al aprendiz retroalimentación acerca de lo que es apropiado para su aprendizaje. En otros casos, a diferencia del aprendizaje supervisado, ésta retroalimentación está ausente, dando lugar al aprendizaje no-supervisado. Y en otros casos, se combinan el aprendizaje supervisado y el no supervisado, dando lugar al aprendizaje semi-supervisado.

En la siguiente sección describiremos brevemente en que consiste cada tipo de aprendizaje.

Aprendizaje supervisado

El aprendizaje supervisado es aquel en donde se intenta aprender de ejemplos como si estos fueran un maestro. Se asume que cada uno de estos ejemplos incluye características o atributos que especifican o definen a qué categoría o clase pertenece, de un conjunto de categorías o clases predefinidas,

de esta manera cada ejemplo se asocia con su clase. Este tipo de aprendizaje es llamado supervisado por la presencia de los ejemplos para guiar el proceso de aprendizaje. Al conjunto de ejemplos del cual se intenta aprender se le llama conjunto de entrenamiento.

Usando estos datos se construye un modelo de predicción, o aprendiz, el cuál nos permitirá predecir la clase para nuevos objetos no vistos por el aprendiz. La construcción del modelo se logra gracias a un algoritmo de aprendizaje.

Los algoritmos comúnmente utilizados son Naive Bayes, Máquinas de Vectores Soporte (MVS), Vecinos más cercanos, J48 entre otros. En particular en el área de Clasificación de Textos los algoritmos típicamente utilizados son Naive Bayes y Máquinas de Vectores Soporte. Para nuestro propósito no explicaremos todos los algoritmos mencionados, sólo los mayormente usados en el área de Clasificación de Textos, estos se describen en la siguiente sección.

Este tipo de aprendizaje tiene la ventaja de que no es necesario que al aprendiz se le muestren todos los ejemplos existentes, es decir que puede clasificar un ejemplo sin haberlo visto nunca. La desventaja es que a pesar de lo anterior, sí es necesaria una gran cantidad de ejemplos para el entrenamiento.

Aprendizaje no supervisado

Este tipo de aprendizaje no presupone ningún conocimiento previo sobre lo que se quiere aprender. Tampoco existe un maestro que conozca los conceptos a aprender, por esta razón a este tipo de aprendizaje se le denomina Aprendizaje no-supervisado.

A diferencia del aprendizaje supervisado, en el aprendizaje no-supervisado, los ejemplos sólo incluyen los atributos, es decir, no se encuentran asociados a una clase. Para este caso la tarea se enfoca en descubrir patro-

nes comunes entre los datos, que permitan separar los ejemplos en clases o jerarquías de clases. De éstas se podrán extraer caracterizaciones, o permitirán predecir características, o deducir relaciones útiles, a lo que se denomina como agrupación (clustering). Algunos de los algoritmos más comunes son: Cobweb, EM, y Kmeans. Siendo este último el más utilizado en el área de Clasificación de Textos.

Este tipo de aprendizaje tiene la ventaja de que no es necesaria la presencia de un maestro para el aprendizaje o de un conjunto de entrenamiento.

Aprendizaje semi-supervisado

El aprendizaje semi-supervisado es la combinación del aprendizaje supervisado y el no-supervisado. En éste se aprende con la ayuda de dos conjuntos. Uno que contiene datos asociados a una clase, y el otro que contiene datos no asociados a una clase. La idea es aprender con los datos asociados a su clase y asociar una clase a los datos que no contienen asociada una clase. Algunos de los algoritmos más comunes son: Co-training, ASSEMBLE y self-training.

2.1.2 Evaluación de Clasificadores

Una vez creado el o los clasificadores para alguna tarea, es importante conocer el desempeño de éstos, por lo que existen medidas de evaluación, entre las más comunes se encuentran la *precisión*, el *recuerdo* y la *exactitud*. La *precisión* es la probabilidad de que un documento etiquetado con la clase *i* corresponda realmente a esa clase. El *recuerdo* es la probabilidad de que un documento que pertenece a la clase *i* es etiquetado dentro de esa clase. La *exactitud* representa el porcentaje de las predicciones que son correctas.

Además cuando hablamos de aprendizaje supervisado, es necesario tener un conjunto de datos independiente del conjunto de aprendizaje, para probar el clasificador, a este conjunto de datos se le llama conjunto de prueba. Sin embargo, muchas veces para evitar posibles sesgos en la selección de los conjuntos de entrenamiento y de prueba se utiliza la validación cruzada, que explicaremos brevemente a continuación.

Validación Cruzada

Este método es el más utilizado para estimar errores de predicción. Consiste en dividir el conjunto disponible en *k* subconjuntos, y se repite el proceso de entrenamiento y prueba *k* veces, cada vez utilizando una partición diferente de datos de entrenamiento y de prueba, y al final los resultados son promediados.

Más específicamente los *m* ejemplos disponibles son divididos en *k* subconjuntos, cada uno de tamaño *m/k*; por ejemplo, cuando *k*=5, la partición sería de la siguiente forma:

	1	2	3	4	5
1ª vez	Prueba	Entrenamiento	Entrenamiento	Entrenamiento	Entrenamiento
2ª vez	Entrenamiento	Prueba	Entrenamiento	Entrenamiento	Entrenamiento
…	…	…	…	…	…
5ª vez	Entrenamiento	Entrenamiento	Entrenamiento	Entrenamiento	Prueba

El procedimiento de validación cruzada es entonces ejecutada *k* veces, cada vez usando un subconjunto diferente como conjunto de validación y combinando los otros *k-1* subconjuntos como conjunto de entrenamiento.

2.2 Clasificación Automática de Textos

La clasificación de textos surge de la necesidad de separar documentos de un tema o clasificación específica de un conjunto de documentos de diferentes temas. Al lograr clasificar los documentos por temas, la búsqueda de información se puede realizar de manera más sencilla.

Debido al elevado número de documentos que pueden pertenecer a una colección de documentos, sobretodo en formato electrónico, realizar la clasificación en forma manual, provoca que la tarea sea complicada, costosa y que requiera mucho tiempo, por lo que surge la idea de hacerlo automáticamente.

Así es como surge el área de Clasificación Automática de Textos, en la cual se han utilizado diferentes métodos estadísticos y más recientemente técnicas de Aprendizaje Computacional.

El primer paso para realizar la tarea de Clasificación Automática de Textos utilizando técnicas de Aprendizaje Computacional, consiste en obtener los atributos que describan el texto a clasificar, así como transformarlos a una representación adecuada para ser utilizados por los algoritmos de Aprendizaje Computacional. A este paso previo se le llama extracción de características. En la siguiente sección se explica con mayor detalle como se realiza la extracción de características en la Clasificación Automática de Textos. Posteriormente se presentan los algoritmos más utilizados en el área de Clasificación Automática de Textos.

2.2.1 Extracción de Características

La extracción de características generalmente consiste en tres etapas:

- Pre-procesamiento
- Indexado
- Reducción de dimensionalidad

Pre-procesamiento

El pre-procesamiento consiste fundamentalmente en eliminar aquellos elementos que generalmente no contienen información para la tarea de la clasificación. Consta de tres posibles fases básicas:

- Eliminación de etiquetas. Si los documentos utilizados contienen algún tipo de etiquetas o cabeceras (ej. etiquetas de html o xml), estás podrán ser removidas, debido a que en algunos casos no proporcionan información útil para la clasificación.
- Eliminación de palabras vacías. Las palabras vacías son palabras que son muy frecuentes y que por lo general no contienen información, por ejemplo: pronombres, preposiciones, conjunciones, artículos, etc.
- Lematización de palabras. Por lematización nos referimos al proceso de remover los sufijos para reducir una palabra a su lema o raíz. Por ejemplo, comprender, comprenderlo y comprendió tienen la raíz comprend.

Indexado

Quizás la representación de documentos más comúnmente usada es la llamada modelo vectorial. En el modelo vectorial, los documentos son representados por vectores de palabras y una colección de documentos son representados por una matriz *A* (palabra por documento), donde cada entrada representa las ocurrencias de una palabra en un documento, i.e.,

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix},$$

Donde a_{ik} es el peso de la palabra *i* en el documento *k*.

Existen muchos caminos para determinar el peso a_{ik} de la palabra *i* en el documento *k*, pero muchas de las aproximaciones están basadas en dos observaciones empíricas:

- Entre más ocurre una palabra en un documento, más relevante es en el tema del documento.
- Entre más veces ocurre la palabra en los documentos de la colección, será menos relevante.

A continuación se describen los 3 esquemas de ponderado más comúnmente usados.

Ponderado Booleano

Este es el esquema más simple y consiste en asignar 1 a a_{ik} si la palabra ocurre en el documento y 0 en otro caso:

$$a_{ik} = \begin{cases} 1 & si \quad f_{ik} > 0 \\ 0 & en \quad otro \quad caso \end{cases}$$

Donde f_{ik} es la frecuencia de la palabra *i* en el documento *k*.

Ponderado por frecuencia de palabra

Otro esquema simple es usar la frecuencia de la palabra en el documento:

$$a_{ik} = f_{ik}$$

Donde f_{ik} es la frecuencia de la palabra *i* en el documento *k*.

Ponderado *TFxIDF*

Los esquemas previos no toman en cuenta la frecuencia de la palabra en todos los documentos en la colección. Una aproximación bien conocida para calcular pesos de palabras es el *TFxIDF* (Term Frequency x Inverse Document Frequency), el cual asigna el peso a la palabra *i* en el documento *k* en proporción al número de ocurrencias de la palabra en el documento, y en proporción inversa al número de documentos en la colección en la que la palabra ocurre al menos una vez:

$$a_{ik} = f_{ik} \times \log\left(\frac{N}{n_i}\right)$$

Donde f_{ik} es la frecuencia de la palabra *i* en el documento *k*, N el número de documentos en la colección y n_i el número de documentos en los que *i* aparece.

Reducción de Dimensionalidad

Un problema en la clasificación de textos es la alta dimensionalidad en el espacio de atributos, lo que hace que el procesamiento sea extremadamente costoso en términos computacionales. De ahí, que existe la necesidad de reducir el conjunto original de atributos, a este proceso se le llama reducción de dimensionalidad [10]. Existen diversos métodos de reducción de dimensionalidad, a continuación se explican dos de ellos:

Umbral de frecuencia de documento

La frecuencia de documento para una palabra es el número de documentos en los cuales las palabras ocurren. Dado un umbral de frecuencia de documento, se calcula la frecuencia de documento para cada palabra en el *corpus* de entrenamiento y se remueven las palabras donde la frecuencia de documento es menor que el umbral determinado. Este método se basa en el supuesto de que las palabras raras generalmente no tienen información para la predicción de categorías.

Ganancia de Información

La Ganancia de Información consiste en medir el número de bits de información para predecir la categoría por medio de la presencia o ausencia de una palabra en el documento.

Sea $c_1,\ldots,c_K$ el conjunto de posibles categorías. La ganancia de información de una palabra w es definida como:

$$IG(w) = -\sum_{j=1}^{K} P(c_j)\log P(c_j) + P(w)\sum_{j=1}^{K} P(c_j \mid w)\log P(c_j \mid w) + P(\overline{w})\sum_{j=1}^{K} P(c_j \mid \overline{w})\log P(c_j \mid \overline{w})$$

Donde $P(c_j)$ es la probabilidad de la clase c_j (fracción de documentos en la colección que pertenece a la clase c_j) y $P(w)$ es la probabilidad de la palabra (fracción de documentos en los cuales la palabra w ocurre). $P(c_j|w)$ es la probabilidad de la clase dada la palabra (fracción de documentos de clase c_j que tiene al menos una ocurrencia de la palabra w) y $P(c_j \mid \overline{w})$ es la probabilidad de la clase c_j dada la no ocurrencia de la palabra w (fracción de documentos de clase c_j que no contienen la palabra w).

La ganancia de información se calcula para cada palabra del conjunto de entrenamiento, y se eliminan aquellas palabras que tengan menor ganancia de información de acuerdo a un umbral.

2.2.2 Aprendizaje Computacional en Clasificación de Textos

La clasificación automática de textos consiste en asignar automáticamente una o más categorías predefinidas a documentos de textos libres. En años recientes se han aplicado diversas técnicas de Aprendizaje Computacional a la clasificación automática de textos.

A continuación describimos algunos de los algoritmos de Aprendizaje Computacional que han sido propuestos para la clasificación de textos y que se utilizaron para nuestro propósito:

Definimos la siguiente notación:

Sea $d = d_1,\ldots,d_M$ el vector de documentos y $c_1,\ldots,c_K$ las posibles categorías o clases. Se asume que tenemos un conjunto de entrenamiento que consiste de N vectores de documentos $d_1,\ldots,d_N$ con las clases $y_1,\ldots,y_N$.

Naive Bayes

El clasificador Naive Bayes [10] se construye usando el conjunto de entrenamiento para estimar la probabilidad de cada clase dados los valores de atributos (palabras) del documento de una nueva instancia. Usamos el Teorema de Bayes para estimar las probabilidades:

$$P(c_j \mid d) = \frac{P(c_j)P(d \mid c_j)}{P(d)}$$

El denominador en la ecuación anterior no distingue entre categorías y puede ser eliminado. Este método asume que los atributos son condicionalmente independientes, dada la clase. Esto simplifica los cálculos.

$$P(c_j \mid d) = P(c_j) \prod_{i=1}^{M} P(d_i \mid c_j)$$

Una estimación $\hat{P}(c_j)$ para $P(c_j)$ puede ser calculada de la fracción de documentos de entrenamiento que es asignada a la clase c_j:

$$\hat{P}(c_j) = \frac{N_j}{N}$$

Donde N_j es el número de documentos de entrenamiento para los cuales la clase es c_j y N es el número total de documentos de entrenamiento.

Una estimación $\hat{P}(d_i \mid c_j)$ para $P(d \mid c_j)$ está dada por:

$$\hat{P}(d_i \mid c_j) = \frac{1 + N_{ij}}{M + \sum_{k=1}^{M} N_{kj}}$$

Donde N_{ij} es el número de veces de la palabra *i* ocurrida dentro de los documentos de la clase c_j en el conjunto de entrenamiento. Para evitar el problema de la probabilidad cero se utiliza Laplace (agregar un 1). *M* es el número de términos en el vocabulario.

A pesar de que la suposición de independencia condicional es generalmente falsa para la aparición de la palabra en documentos, el clasificador Naive Bayes es sorprendentemente efectivo.

K Vecinos más cercanos

Para clasificar un vector de documentos *d*, el algoritmo de *k* vecinos más cercanos [10] ordena los vectores de documentos vecinos del conjunto de entrenamiento, y usa las clases etiquetadas de los *k* vecinos más similares para predecir la clase del documento de entrada. Las clases de estos vecinos son pesadas usando una similaridad de cada vecino de *d*, por ejemplo usando, la distancia Euclidiana o el coseno entre los dos vectores documento.

La distancia Euclidiana entre dos vectores $d=d_1,\ldots d_n$ y $e=e_1,\ldots e_n$ se obtiene con:

$$d(d,e)=\sqrt{(d_1-e_1)^2+(d_2-e_2)^2+\cdots+(d_n-e_n)^2}=\sqrt{\sum_{i=1}^{n}(d_i-e_i)^2}$$

Para encontrar el coseno entre dos vectores utilizamos:

$$d(d,e)=\frac{\sum_{i=1}^{n}(d_n\times e_n)}{\sqrt{\sum_{i=1}^{n}{d_n}^2+\sum_{i=1}^{n}{e_n}^2}}$$

Máquinas de Vectores Soporte

Máquinas de Vectores Soporte han mostrado un buen desempeño en general en una gran variedad de problemas de clasificación, más recientemente en clasificación de textos.

En términos geométricos, el problema que resuelve las MVS es identificar una frontera de decisión linear entre dos clases, a través de una línea que los separe, maximizando el espacio del hiperplano. Sin embargo, las MVS incluyen una función llamada *kernel*, la cual permite realizar separaciones no

lineales de los datos, proyectando la información a un espacio de características de mayor dimensión. Esto se logra cambiando la representación de la función, mapeando el espacio de entradas *D* a un nuevo espacio de características $F = \{\phi(d) \mid d \in D\}$. Esto es:

$$d = \{d_1, d_2, \cdots, d_n\} \rightarrow \phi(d) = \{\phi(d)_1, \phi(d)_2 \cdots, \phi(d)_n\}$$

En la figura 2.1 se muestra un mapeo de un espacio de entradas de dos dimensiones a un nuevo espacio de características de dos dimensiones, donde la información no puede ser separada por una máquina lineal mientras que en el nuevo espacio de características esto resulta sencillo.

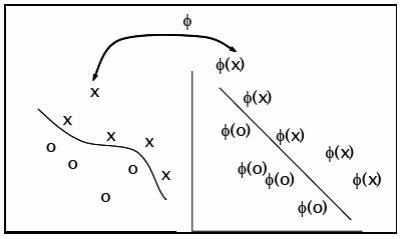

Figura 2.1 Mapeo del espacio de entradas a un espacio de mayor dimensión

La función real ϕ no necesita ser conocida, es suficiente tener una función kernel *k*, la cual hace posible realizar el mapeo de la información de entrada al espacio de características de forma implícita y entrenar a la máquina lineal en dicho espacio.

Capítulo 3
Estado del Arte: Clasificación Automática de Opiniones

La clasificación automática de opiniones es una disciplina reciente del procesamiento de información y lingüística computacional que concierne no al tema de un documento, sino a la opinión expresada.

Dentro de la clasificación de opiniones, se pueden identificar tres subtareas principales [11]:

1. Determinar la subjetividad de un texto
2. Determinar la polaridad u orientación de un texto
3. Determinar la fuerza o grado de una opinión

Además, este tipo de clasificación se puede realizar a nivel de:

1. Documento
2. Oración
3. Término

Dentro de los trabajos previos realizados se han propuesto diferentes tipos de aprendizaje para realizar las subtareas de clasificación de textos de opinión, algunos han utilizado aprendizaje supervisado, otros han propuesto aprendizaje no supervisado y algunos otros han propuesto aprendizaje semi-supervisado.

Para ubicar los trabajos previos realizados dentro de esta área y dependiendo de la subtarea que atacan, si lo hacen a nivel de documento, de oración o de término y que tipo de aprendizaje proponen para resolverlo, se muestra la siguiente tabla:

Subtarea	**A nivel de:**	**Tipo de Aprendizaje**	**Referencia**
Determinar la subjetividad	Documento	Supervisado	[13]
	Oración	Supervisado	[13], [6],[15]
	Término	Supervisado	[11]
Determinar la polaridad	Documento	Supervisado	[8]
		No supervisado	[7]
	Oración	Supervisado	[Nosotros]
		No supervisado	[7]
		Semi-supervisado	[Nosotros]
	Término	Semi-supervisado	[11]
Determinar el grado	Término	Supervisado	[12]

En particular, nuestro trabajo se centra en determinar la polaridad y orientación de un texto a nivel de oración. Y el tipo de aprendizaje que utilizamos es semi-supervisado.

A continuación se describen los trabajos previos que se han propuesto en cada subtarea.

3.1 Subjetividad de un Texto

La tarea de determinar la subjetividad de un texto dado, se refiere a determinar si ese texto contiene información objetiva (i.e., si describe una situación o evento, sin expresar una opinión) o expresa una opinión acerca de un tema.

Hong Yu y Hatzavassiloglou presentan [13] un clasificador bayesiano para diferenciar entre documentos con opiniones y documentos con noticias. Los datos que utilizaron fueron artículos del periódico *Wall Street Journal* que son identificados por Editoriales, cartas al editor, negocios y noticias. Estas etiquetas fueron usadas para proveer las etiquetas correctas de clasificación durante la fase de entrenamiento y evaluación. Como atributos utilizaron bolsa de palabras, sin remover palabras vacías. Para entrenar el clasificador utilizaron los artículos de Editoriales y cartas al editor como opiniones y los artículos de negocios y noticias como documentos de información objetiva. Reportan como medida de evaluación F-measure de 0.97, además proponen tres métodos para la clasificación de documentos con opiniones y documentos con hechos a nivel de oración. El primero se basa en la hipótesis de que dentro de un tema dado, las oraciones de opiniones serán más similares a otras oraciones de opinión, utilizan un sistema llamado SIMFINDER [14] para medir la similitud de oraciones, basada en palabras, frases y conjuntos de sinónimos de WordNet. Primero seleccionan los documentos que son del mismo tema que la oración en cuestión. Calculan la similitud con cada oración en estos documentos y asigna la categoría del documento que dé mayor valor. Los resultados que reportan para este método son 0.61 de recuerdo y 0.34 de precisión para la clase de documentos con información objetiva y 0.30 de recuerdo y 0.49 de precisión para la clase de opiniones. El segundo método es un clasificador entrenado con

Naive Bayes, usando las oraciones de opiniones y oraciones de hechos como ejemplos de entrenamiento. Los atributos incluyen, palabras, bigramas, trigramas, etiquetas *Part of Speach*, secuencias de palabras orientadas semánticamente y adjetivos con orientación (positiva y negativa). Los resultados reportados de este método son 0.15 de recuerdo y 0.43 de precisión para la clase objetiva y 0.91 y 0.69 de recuerdo y precisión respectivamente para la clase subjetiva. El tercer método es el uso de un algoritmo con clasificadores múltiples, cada uno con diferentes conjunto de atributos. Es decir, crean 5 conjuntos de atributos F1, F2, …, F5. Entrenan un clasificador con F1 sobre todo el conjunto de entrenamiento. El clasificador resultante predice las etiquetas para el conjunto de entrenamiento. Las oraciones que reciben una etiqueta diferente a la que le corresponde son removidas. Entrenan entonces otro clasificador con F2 y el conjunto de oraciones que permanecieron. El proceso se repite iterativamente hasta que ya no se puedan remover oraciones. Los resultados que reportan de este método fueron 0.13 y 0.42 de recuerdo y precisión, para la clase objetiva y 0.92 y 0.70 de recuerdo y precisión para la clase subjetiva.

Riloff y Wiebe proponen un método para realizar clasificación de subjetividad a nivel de oración [6, 15]. Proponen una variante de un proceso *self-training*[8] que consta de dos etapas. En la primera etapa, crean un clasificador al cual lo llaman *clasificador basado en reglas*. En la segunda etapa crean un clasificador utilizando Naive Bayes, el cual se basa en el uso de patrones de extracción[9] y partes de la oración (Part of Speach). El *clasificador basado en reglas* utiliza listas de elementos léxicos (palabras, o n-gramas) que fueron recopiladas de trabajos previos, a estos elementos le llaman pistas y se dividen en

[8] *Self-training* es un proceso de aprendizaje semi-supervisado, donde se inicia con un conjunto de datos etiquetados para crear un clasificador con el que posteriormente se etiquetan datos no etiquetados. Con este nuevo conjunto se crear un nuevo clasificador y así hasta que se cumpla alguna condición.

[9] Un *patrón de extracción* es un patrón que sintetiza un conjunto de restricciones sintácticas que debe satisfacer una oración[26].

dos: fuertemente subjetivas y débilmente subjetivas; el *clasificador basado en reglas* clasifica una oración como subjetiva si tiene 2 o más pistas fuertemente subjetivas y clasifica una oración como objetiva si tiene ausencia de pistas fuertemente subjetivas y cuando mucho una débilmente subjetiva. La segunda etapa de este método consiste en obtener patrones de extracción con AutoSlog-TS (Riloff, 1996), y con estos patrones se obtienen más pistas que se agregan a las obtenidas previamente y se repite el proceso una vez más. Los datos que utilizaron para la clasificación fue la colección FBIS (Foreign Broadcast Information Service) que son artículos periodísticos de varios países y publicación que cubren diferentes temas. Para la evaluación utilizaron la colección de datos etiquetados empleados previamente por Wilson y Wiebe, 2003 que consta de 13 documentos y 210 oraciones. Reportan una exactitud máxima de 73.8%.

Esuli y Sebastiani [11] proponen una manera singular para determinar la subjetividad de términos. Primero obtienen dos conjuntos de términos con polaridad positiva y negativa, respectivamente, este proceso se explica al final de la siguiente sección, y después simplemente unen estos conjuntos para obtener un nuevo conjunto de términos subjetivos. Para crear el conjunto de términos objetivos utilizan el conjunto léxico GI (General Inquirer) [16], toman todos los términos que aparecen en el conjunto léxico GI y que no aparecen en el conjunto de términos subjetivos.

3.2 Polaridad de Opiniones

Determinar la polaridad u orientación de un texto se refiere a decidir si un texto dado que contiene opiniones expresa una opinión a favor o en contra (positiva o negativa, respectivamente).

En [7] se presenta un algoritmo de aprendizaje no supervisado para determinar la orientación de documentos. El algoritmo utiliza un etiquetador POS (Part-of-speach) para identificar las frases que contienen adjetivos o adverbios. Una vez que se identificaron estas frases, el algoritmo obtiene la Información Mutua para medir la asociación semántica entre los adjetivos o adverbios de cada frase y las palabras "excellent" y "poor". La Información Mutua entre dos palabras es definida de la siguiente manera [24]:

$$IM\left(palabra_1, palabra_2\right) = \log_2\left(\frac{P\left(palabra_1 \quad y \quad palabra_2\right)}{P\left(palabra_1\right)P\left(palabra_2\right)}\right)$$

Donde *P(palabra)* es la probabilidad de que la palabra ocurra y *P(palabra$_1$ y palabra$_2$)* es la probabilidad de que la palabra$_1$ y la palabra$_2$ ocurran como una secuencia de ambas palabras. Si la información mutua entre el adjetivo o adverbio de una frase y la palabra "excellent" es mayor que la información mutua entre el adjetivo o adverbio de la misma frase y la palabra "poor" se marca como positiva, y viceversa. Por último se calcula el número de frases marcadas como positivas y negativas y se clasifica el documento como positivo si el número de frases marcadas como positivas es mayor al número de frases marcadas como negativas y como negativo en caso contrario. Los documentos que utilizaron fueron 410 documentos de opiniones (170 negativos y 240 positivos) obtenidos de www.epinions.com de automóviles, bancos, películas y destinos turísticos. La exactitud que reportan utilizando las opiniones de películas fue de 65.83%.

Bo Pang, Lilian Lee y Shivakumar Vaithyanathan en el 2002 [8] también realizan clasificación de la orientación de textos a nivel de documentos. Experimentan con diferentes métodos de aprendizaje supervisado y con diferentes tipos de atributos. Los métodos que utilizan son Naive Bayes, Maximum Entropy y Support Vector Machines. Como atributos utilizan unigramas, unigramas+bigramas, bigramas, unigramas+POS y adjetivos. Los datos que utili-

zaron fueron críticas de películas de *IMDb Internet Move Database*, las cuales consisten de 752 documentos negativos y 1301 documentos positivos. De los resultados obtenidos el mayor porcentaje de exactitud lo obtuvieron utilizando unigramas (16,165) como atributos y Naive Bayes como método de clasificación y lograron 82% de exactitud.

Podemos observar que de los trabajos previos, Pang et al. [8] es el único que logra llegar a un 82% de exactitud para determinar la polaridad de opiniones a nivel de documento. Y a nivel de oración [7] logró obtener 65.83% de exactitud utilizando un proceso que implica el uso de atributos sintácticos. Uno de los objetivos de esta tesis es determinar la polaridad de opiniones a nivel de oración utilizando métodos que no hagan uso de atributos sintácticos, sino atributos léxicos.

Andrea Esuli y Fabrizio Sebastiani piensan que para poder identificar la orientación de documentos, primero se debe identificar la orientación de los términos. Ellos proponen un método semi-supervisado para determinar la subjetividad y la orientación de términos [11]. Parten de dos conjuntos de términos iniciales, Lp y Ln. Cada conjunto lo expanden agregando términos del grafo WordNet, más específicamente relaciones de sinonimia y antonimia. Este proceso está basado en la hipótesis de que dos sinónimos tienen la misma orientación y dos antónimos tienen orientación opuesta. El método es iterativo y en cada iteración expanden de la siguiente manera: al conjunto Lp se agregan los sinónimos de cada término de Lp y los antónimos de cada término del conjunto Ln. El conjunto Ln se expande agregando los sinónimos de cada término del mismo conjunto y los antónimos de cada término del conjunto Lp. Continúan expandiendo ambos conjuntos de la misma manera hasta cumplir con 4 iteraciones. De los conjuntos obtenidos se crea un clasificador utilizando representaciones vectoriales basadas en su definición textual (WordNet). Con este método obtienen 66% de exactitud.

3.3 Grado de Opinión

Determinar la fuerza o grado de una opinión se refiere a si la opinión expresada acerca de algo es débil, media o fuerte, i.e. que tan negativa o que tan positiva es.

En esta dirección los trabajos reportados son escasos. Quienes se han dedicado a esta investigación son Esuli y Sebastiani quienes han propuesto una herramienta basada en *WordNet*, que mide el porcentaje de fuerza positiva, negativa u objetividad de cada término, y lo representan gráficamente, a esta herramienta la llaman *SentiWordNet* [12]. *SentiWordNet* es un recurso léxico en el cual a cada *synset*[10] de *WordNet* se le asocian tres puntuaciones numéricas (Objetivo, Positivo y Negativo). Cada puntuación tiene un rango de 0 a 1 y su suma es 1 para cada *synset*. Por ejemplo: el término estimable tiene 3 *synsets* (obtenidos de *WordNet*). Uno de ellos tiene el significado de "poder ser calculado o estimado" y tiene las siguientes puntuaciones: objetivo = 1, positivo = 0 y Negativo = 0. Otro *synset* del mismo término tiene el significado de "que merece aprecio o reconocimiento" y tiene las siguientes puntuaciones: objetivo = 0.25, positivo = 0.75 y negativo = 0. El método depende del análisis de las glosas asociadas a *synsets*. Las tres puntuaciones son derivadas de combinar los resultados producidos por un comité de ocho clasificadores. Cada clasificador es capaz de decidir si un *synset* es positivo, negativo u objetivo. Cada clasificador difiere de otro por el conjunto con que fue entrenado y el método de aprendizaje utilizado para entrenarlo. Las puntuaciones para un *synset* son determinadas por la proporción (normalizada) de los clasificadores. Si todos están de acuerdo, el *synset* tendrá la máxima puntuación. Cada clasificador es generado usando el método semi-supervisado descrito en [11]. Dado que la suma de las puntuaciones de un *synset* es 1 es posible representar estos valo-

[10] Si una palabra tiene diferentes significados, a cada definición de esa palabra se le llama *synset*

res en un triángulo. Ellos mencionan que la forma de probar la exactitud del método es imposible debido a que necesitaría un etiquetado manual de *WordNet* de acuerdo a sus tres etiquetas. Este recurso es accesible públicamente y puede utilizarse en el siguiente sitio: http://sentiwordnet.isti.cnr.it/.

Capítulo 4
Caracterización

Uno de los objetivos de esta tesis es la implementación de un método semi-supervisado de clasificación de polaridad. Previamente a esta implementación es necesario realizar la caracterización de los datos, para obtener el conjunto de atributos más adecuados para la clasificación. En este capítulo se muestran los resultados obtenidos de esta caracterización. El método semi-supervisado de clasificación propuesto se describe en el siguiente capítulo.

Caracterizar un objeto, en este caso un texto o un fragmento de texto, consiste en extraer un conjunto de atributos que describan el objeto. Dicha descripción deberá ser apropiada dependiendo de su uso final. En nuestro caso buscamos una descripción que nos permita distinguir las opiniones positivas de las negativas. Por otro lado, también esperamos que dicha descripción sea fácil de extraer o construir, y que sea lo más concisa posible. Es decir, que el número de atributos no sea excesivamente grande.

4.1 Experimentos

Para probar las diferentes caracterizaciones propuestas se utilizó Naive Bayes como método de Aprendizaje Computacional. Como conjunto de atributos se utilizó la representación por medio de bolsa de palabras, que es la representación tradicional utilizada para la clasificación temática. Sin embargo el usar bolsa de palabras deja a un lado el orden en que se puedan encontrar las palabras, por lo cual, se realizaron otros experimentos tratando de capturar la estructura del lenguaje. Para ello se consideraron *n*-gramas, es decir, secuencias de *n* palabras sucesivas.

Los *n*-gramas permiten obtener no sólo palabras, sino secuencias de palabras y de esta forma considerar el orden en que se encuentran, sin la necesidad de llevar a cabo un costoso análisis sintáctico. Desafortunadamente este tipo de caracterización nos lleva a una explosión combinatoria, por lo que sólo empleamos secuencias de una y dos palabras (unigramas y bigramas). Pero incluso tomando en cuenta sólo los unigramas y bigramas el número de atributos es elevado y existe la limitante de que sólo obtendremos secuencias de máximo dos palabras.

En este trabajo de tesis se experimentó además de las representaciones mencionadas en el párrafo anterior, el uso de un conjunto de atributos formado por Secuencias Frecuentes Maximales, que se describen en la sección de resultados de éste capítulo. El uso de Secuencias Frecuentes Maximales para caracterizar los datos, se basa en la idea de que utilizar secuencias proporcionan mayor información que el usar sólo palabras y a diferencia de los *n*-gramas el número de atributos que se obtienen no es tan grande.

Además como vimos en la sección 2.2.1 existen formas de reducción de dimensionalidad, para ello aplicamos el algoritmo de ganancia de información y se seleccionaron aquellos atributos que obtuvieron una ganancia mayor a cero.

Para validar las diferentes caracterizaciones se aplicó validación cruzada en 10 pliegues, con las herramientas que proporciona WEKA. Sin embargo, la forma en que WEKA realiza la validación cruzada no es exactamente como se describe en la sección 2.1.2. Esto se debe a que es necesario determinar los atributos antes de entregarlo a WEKA. De esta forma, a pesar de excluir una fracción de los datos en cada pliegue, el algoritmo de aprendizaje utiliza el conjunto completo de atributos en la fase de entrenamiento. Para comprobar el alcance de las caracterizaciones en textos nunca antes vistos, también se utilizó un conjunto de prueba.

4.1.1 Conjunto de Datos

Como se mencionó en el capítulo 2 para crear un clasificador se requiere de un conjunto de entrenamiento y un conjunto de prueba.

El *corpus* que se seleccionó fue el utilizado por Pang & Lee en [17], el cual consta de 10662 oraciones de opiniones en idioma inglés, separadas en dos clases: 5331 oraciones con opiniones positivas y 5331 oraciones con opiniones negativas. De este *corpus* se utilizaron 8 530 oraciones (4 264 positivas y 4 265 negativas) para entrenamiento y 2 132 oraciones (1 066 positivas y 1 066 negativas) como conjunto de prueba.

En la tabla 4.1 y 4.2 se muestran ejemplos de oraciones positivas y negativas, respectivamente, del *corpus* en idioma inglés.

Opiniones positivas
• This is a film well worth seeing, talking and singing heads and all.
• A pleasant enough movie, held together by skilled ensemble actors.
• A pleasant enough movie, held together by skilled ensemble actors.
• This is the best American movie about troubled teens since 1998's whatever.
• A feel-good picture in the best sense of the term.
• Run; don't walk, to see this barbed and bracing comedy on the big screen.
• Enormously likable, partly because it is aware of its own grasp of the absurd.
• This is simply the most fun you'll ever have with a documentary!
• A fascinating and fun film.
• At its best early on as it plays the culture clashes between the brothers.
• Daring, mesmerizing and exceedingly hard to forget.
• This is one of polanski's best films.

Tabla 4.1 Ejemplos de opiniones positivas del *corpus* en idioma inglés.

Opiniones negativas
• Unfortunately the story and the actors are served with a hack script.
• Interesting, but not compelling.
• This 100-minute movie only has about 25 minutes of decent material.
• On its own, it's not very interesting. As a remake, it's a pale imitation.
• I didn't laugh. I didn't smile. I survived.
• Please, someone, stop Eric Schaeffer before he makes another film.
• Deadly dull, pointless meditation on losers in a gone-to-seed hotel.
• The movie is a mess from start to finish.
• Some episodes work, some don’t.
• Journalistically dubious, inept and often lethally dull.
• This movie . . . doesn't deserve the energy it takes to describe how bad it is.
• This 10th film in the series looks and feels tired.
• The whole affair is as predictable as can be.
• Star trek was kind of terrific once, but now it is a copy of a copy of a copy.
• The film can depress you about life itself.

Tabla 4.2 Ejemplos de opiniones negativas del *corpus* en idioma inglés.

4.1.2 Resultados

Bolsa de Palabras

La bolsa de palabras se refiere a utilizar como atributos todas las palabras obtenidas del conjunto de oraciones de entrenamiento, también se le llaman unigramas.

En la siguiente tabla, se muestra la exactitud obtenida en la clasificación utilizando bolsa de palabras como atributos. En ambos experimentos se eliminaron las palabras vacías. Sin embargo, en el segundo además se eliminaron todas aquellas palabras que aparecieran una o dos veces en todo el conjunto de entrenamiento.

Frecuencia de las palabras	Número de atributos	Validación cruzada	Conjunto de prueba
1 o más veces	16 259	63.07%	60.92%
3 o más veces	5 827	63.03%	60.78%

Tabla 4.3 Resultados utilizando bolsa de palabras

Se puede observar que a pesar de obtener una ligera baja en la exactitud de los experimentos, el número de atributos se reduce considerablemente, lo que indica que las palabras raramente utilizadas no proporcionan información relevante para la clasificación de la polaridad de las oraciones de opiniones.

Aplicando como método de selección de atributos se le aplicó ganancia de información a los conjuntos de atributos utilizados en el experimento anterior, y se seleccionaron aquellos que obtuvieran una ganancia de información mayor a cero. Los resultados fueron los siguientes:

Frecuencia de las palabras	Número de atributos	Validación cruzada	Conjunto de prueba
1 o más veces	14 477	63.27 %	60.92 %
3 o más veces	1 435	63.05 %	60.13 %

Tabla 4.4 Resultados utilizando bolsa de palabras y GI > 0

En la tabla 4.4 podemos observar que el resultado con mayor exactitud utilizando bolsa de palabras como atributos, es utilizando todas las palabras (frecuencia de 1 o más veces), logrando una exactitud de 63.27% con una validación cruzada y 60.92% en el conjunto de prueba. Sin embargo el número de atributos es muy elevado, 14 477, comparando con la segunda línea de la tabla 4.4, vemos que el porcentaje no varía mucho, y el número de atributos se reduce considerablemente.

Haciendo un análisis de los experimentos de la tabla 4.4, se encontró que existen instancias que no logran ser caracterizadas, es decir, instancias con 0 en todos los atributos. Pensando en el supuesto de que estas instancias estén haciendo ruido a la hora de entrenar, se eliminaron y se creo un nuevo clasificador. El porcentaje de instancias eliminadas fue de 1.5 %. Los resultados obtenidos fueron los siguientes:

Ganancia de Información	Número de atributos	Validación cruzada	Conjunto de prueba
no	5 827	62.94 %	60.65 %
si	**1 435**	**66.69 %**	**63.11 %**

Tabla 4.5 Resultados eliminando instancias sin caracterizar

El clasificador creado reporta entonces una exactitud de 66.69 % con validación cruzada y 63.11% en el conjunto de prueba. Lo que demuestra que eliminando las instancias no caracterizadas del conjunto de entrenamiento permite mejorar la clasificación.

Unigramas + Bigramas

El número total de bigramas obtenidos del conjunto de entrenamiento es de 95 263. Sin embargo debido a la alta dimensionalidad que provocaría esta representación se eliminaron aquellos bigramas que aparecen una o dos veces en todo el conjunto de entrenamiento, quedando un total de 8 460 bigramas. Estos bigramas se agregaron a los atributos (unigramas) obtenidos en el experimento anterior, y los resultados fueron los siguientes:

Número de unigramas	Número total de atributos	Validación Cruzada	Conjunto de prueba
5 827	14 287	66.06 %	63.36 %
1 435	9 895	66.98 %	63.64%

Tabla 4.6 Resultados utilizando unigramas + bigramas

Se puede observar que al agregar los bigramas a los atributos obtenidos previamente, existe un pequeño incremento en el resultado, sin embargo no es significativo. Asimismo el número de atributos es mucho mayor.

Secuencias Frecuentes Maximales

Primeramente y antes de mostrar los resultados obtenidos utilizando Secuencias Frecuentes Maximales como conjunto de atributos, definiremos lo que son las Secuencias Frecuentes Maximales (SFM).

Sea D un conjunto de textos (un texto puede ser un documento completo o solo una frase), cada texto consiste de una secuencia de palabras. Entonces tenemos las siguientes definiciones [18].

Definición 1. Una secuencia $p = a_1 \ldots a_k$ es una *subsecuencia* de una secuencia q si todos los elementos a_i, $1 \leq i \leq k$, ocurren en q y ocurren en el mismo orden como en p. Si una secuencia p es una subsecuencia de una secuencia q, también decimos que p ocurre en q.

Definición 2. Una secuencia p es *frecuente* en D si p es una subsecuencia de al menos σ textos de D, donde σ es un umbral de frecuencia dada.

Definición 3. Una secuencia p es una *secuencia frecuente maximal* en D si no existe ninguna otra secuencia p' en D tal que p es una subsecuencia de p' y p' es frecuente en D.

Podemos observar de acuerdo a la definición anterior que las Secuencias Frecuentes Maximales a diferencia de unigramas, bigramas, trigramas, etc. son secuencias que se distinguen por su repetida aparición y no por el tamaño de la secuencia. Para determinar el umbral de frecuencia apropiado para nuestro problema se realizaron diferentes experimentos:

Sigma	Número de atributos	Validación Cruzada	Conjunto de prueba
3	12 646	70.15 %	71.10 %
4	8 942	71.55 %	**72.09 %**
5	6 783	71.43 %	71.95 %
6	5 445	**72.03 %**	71.43 %
10	2 921	70.93 %	69.51 %
15	1 800	68.55 %	67.40 %

Tabla 4.7 Resultados utilizando SFM

De manera experimental encontramos que utilizando Secuencias Frecuentes Maximales con un umbral de frecuencia de 4 obtenemos un 72.09 % al validar en el conjunto de prueba. Por lo que para los siguientes experimentos utilizaremos sigma igual a 3, 4 y 5. Ya que consideramos que mientras se extraigan secuencias más largas (i. e. sigmas más pequeños), éstas serán más apropiadas para la caracterización de un texto de opinión.

Para los experimentos previos se obtuvieron las SFM (Secuencias Frecuentes Maximales) de todo el conjunto de entrenamiento (instancias de ambas clases, positivo y negativo), en un segundo experimento probamos el calcular las SFM por clase para comprobar si los atributos obtenidos podrían dar mayor información para la clasificación.

Al obtener las SFM por clase se observó que existen atributos que se encuentran en ambas clases (intersección) por lo que estos atributos se eliminaron. A continuación se muestran los resultados obtenidos incluyendo los resultados obtenidos aplicando Ganancia de Información:

Sigma	GI > 0	Número de atributos	Validación Cruzada	Conjunto de prueba
3	no	9 397	76.86 %	70.40 %
4	no	5 788	76.72 %	71.06 %
	si	4 441	75.52 %	69.98 %
5	no	4 157	74.49 %	70.21 %
	si	3 263	79.21 %	69.79 %

Tabla 4.8 Resultados utilizando SFM atributos obtenidos por clase

Comparando los resultados observamos una mejoría al utilizar SFM con frecuencia 5. Llegando a 79.21% con validación cruzada y 69.79% en el conjunto de prueba. Además podemos observar que el número de atributos que se utilizan es menor (3 263) que en cualquiera de los resultados anteriores.

Algunos ejemplos de los atributos que obtuvieron mayor ganancia de información se muestran en la siguiente figura:

and boring	distinctive	the whole
it could	undercut by	well worth
the importance of	not exactly	a small
compelling story	in spite of	remotely
has become	older	nothing in
is only	more interesting	a matter of
brilliant	steeped	point of view
do not	alive	handsome
lived	a fun	gifted
of our	winds up	and powerful

Figura 4.1 Ejemplos de SFM con mayor GI

Para comprobar la caracterización usando SFM se crearon clasificadores utilizando otros métodos de clasificación. De los métodos utilizados, con los que se reportaron mejor resultados fueron los siguientes:

	BAYES		**SVM**		**KNN1**	
Número de atributos	**Validación cruzada**	**Conjunto de prueba**	**Validación cruzada**	**conjunto de prueba**	**Validación cruzada**	**Conjunto de prueba**
3 263	**79. 21 %**	**69.79 %**	75.55 %	69.37 %	61.27 %	60.92 %

Tabla 4.9 Resultados utilizando SFM y tres técnicas de Aprendizaje Computacional

Como puede observarse los resultados son similares para Naive Bayes y SVM. De ahí que mantuvimos este clasificador para los experimentos subsecuentes.

Pesado por Frecuencia

De los atributos obtenidos, podemos observar que existen SFM que se utilizan:

1. Exclusivamente en opiniones con polaridad positiva
2. Exclusivamente en opiniones con polaridad negativa
3. Con mayor frecuencia en opiniones con polaridad positiva que en opiniones con polaridad negativa
4. Con mayor frecuencia en opiniones de polaridad negativa que en opiniones de polaridad positiva y
5. Con la misma frecuencia en ambas clases.

Con este análisis surge la idea de realizar una selección de atributos eligiendo aquellos atributos que aunque se utilicen en ambas clases, sean más frecuentes en una de ellas, tomando como punto de referencia un umbral de razón entre las dos frecuencias. Por ejemplo, si un atributo aparece 2 veces en una clase y 5 veces en la otra clase, la razón de estas dos cantidades sería 2/5 o 0.4, siempre anteponiendo la cantidad menor, de este modo, el resultado siempre será un número entre 0 y 1. Con esta idea pesamos los atributos y a la vez podemos hacer una selección de ellos. Por ejemplo, si utilizamos un umbral de 0.4 significaría que aquellos atributos donde la razón de frecuencia es mayor a 0.4 serán eliminados.

De esta forma, si un atributo aparece relativamente con mayor frecuencia en una clase que en otra, se considera un atributo útil para la discriminación entre las clases. Por el contrario, si el atributo aparece con la misma frecuencia en ambas clases, el atributo se elimina.

Por otro lado, esta razón le dará un peso a cada atributo, de manera que si es un atributo muy frecuente utilizado en una clase tendrá un mayor peso que uno que sea frecuente en ambas. Para obtener el peso de cada atributo se utiliza la siguiente fórmula:

$$p_{a1} = 1 - \frac{v1}{v2} \quad v1 < v2$$

Donde *v1* es el valor de la frecuencia de aparición del atributo en una de las clases, y *v2* es el valor de la frecuencia de aparición del atributo en la otra clase.

Debido a que algunas instancias son muy cortas, no se logran caracterizar, encontrando instancias representadas con ceros que pertenecen a la clase positiva e instancias representadas con ceros que pertenecen a la clase negativa, estas instancias provocan problemas al momento de entrenar al clasificador; por lo que se eliminaron antes de la fase de entrenamiento.

Los resultados de los experimentos realizados con el pesado por frecuencia propuesto son los siguientes:

Frecuencia SFM	Umbral de razón de frecuencia	Número de atributos	Validación cruzada	Conjunto de prueba	Instancias no caracterizadas en el conjunto de prueba
3	0.40	5 071	79.8238%	68.8086%	204
	0.34	4 968	80.0915%	68.7617%	235
	0.25	3 049	**90.4883%**	66.8386%	597
	0.22	3 012	79.6093%	64.1182%	634
4	0.40	2 876	83.3028%	69.9343%	269
	0.34	2 774	83.4757%	69.6060%	318
	0.25	1 845	89.8656%	67.4015%	641
	0.22	1 799	82.4699%	66.6510%	689
5	0.40	2 413	82.4480%	**70.0281%**	**247**
	0.34	2 266	84.4263%	69.0901%	328
	0.25	1 750	89.0060%	69.5591%	567
	0.22	1 686	89.8582%	66.8856%	627

Tabla 4.10 Resultados utilizando SFM y con pesado por frecuencia

De los resultados obtenidos vemos que utilizando SFM con frecuencia 3, con un umbral de razón de frecuencia de 0.25 se logra una exactitud de 90.48% en validación cruzada, sin embargo en el conjunto de prueba sólo se obtiene 66.83%. Además el número de instancias no caracterizadas es de 597.

Utilizando SFM con frecuencia 5, con un umbral de razón de frecuencia de 0.40 obtenemos 70.02% validando en el conjunto de prueba, el número de atributos es de 2 413, y el número de instancias no caracterizadas es de 247 lo cual significaría que los atributos obtenidos son atributos que abarcan una gran cantidad de secuencias frecuentemente utilizadas y que además dan buen resultado en la clasificación.

Después de haber realizado estos experimentos elegimos las SFM con frecuencia 5 como los atributos más adecuados para la clasificación de textos de opinión.

Capítulo 5
Método Semi-supervisado de Clasificación de Polaridad

En este capítulo se explica el método semi-supervisado propuesto para la clasificación de opiniones por su polaridad.

5.1 Descripción del Método

Uno de los objetivos de esta tesis es proponer un método para la clasificación de opiniones en español. Sin embargo, aún no se cuenta con un *corpus* etiquetado para tal efecto. Así que parte de nuestro trabajo consistió en la recolección y etiquetado de un *corpus* relativamente pequeño y la aplicación de un método semi-supervisado que facilite la construcción de un clasificador adecuado para esta tarea en español.

La figura 5.1 muestra un diagrama del método semi-supervisado utilizado, el cuál está basado en el esquema *self-training* [25]. Como puede observarse se parte de un conjunto de datos etiquetados y otro de datos no etiquetados, a los cuales se les aplica un preprocesamiento, el cual consiste en cambiar los signos de puntuación por etiquetas para evitar errores en el procesamiento de los datos. Con el conjunto de datos etiquetados se obtienen un clasificador,

utilizando Naive Bayes como algoritmo de aprendizaje. Con el clasificador creado, se etiquetan los datos no etiquetados.

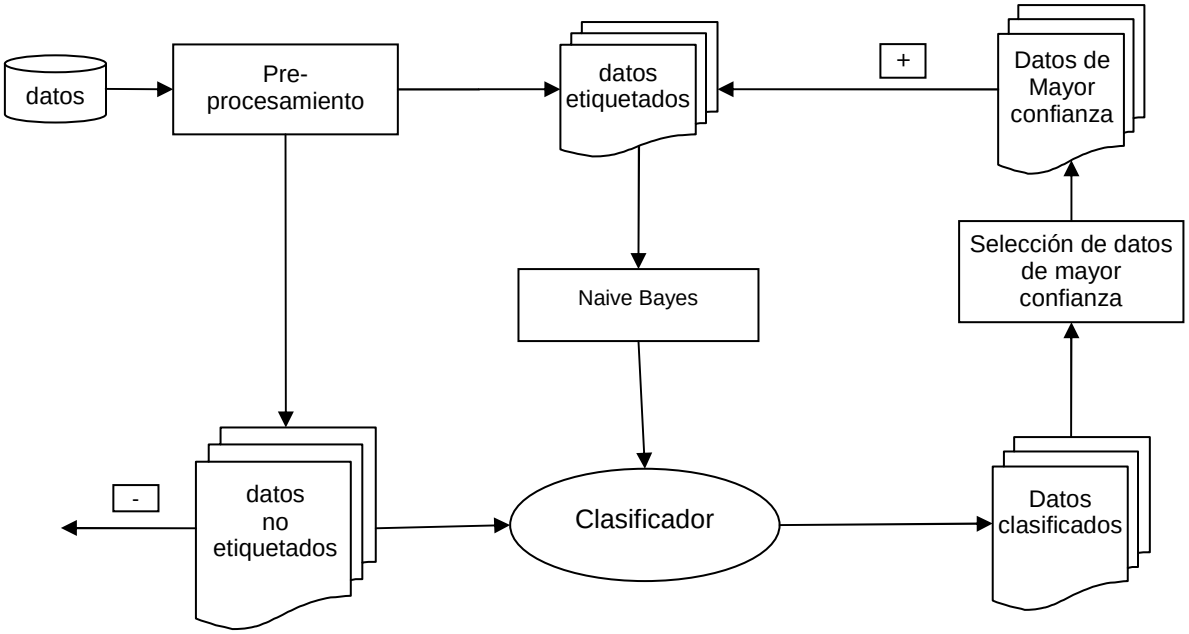

Figura 5.1 Método semi-supervisado

Las instancias con mayor confianza, de acuerdo a un umbral de confianza se agregan a los datos etiquetados y se eliminan de los datos no etiquetados. Con los nuevos datos, se crea un nuevo clasificador y se etiquetan los datos no etiquetados y así sucesivamente hasta que todos los datos no etiquetados se agreguen a los etiquetados o que las nuevas instancias clasificadas no cumplan el umbral de confianza. Para comprobar el método semi-supervisado se utilizó inicialmente el mismo *corpus* del capítulo anterior, posteriormente se aplicó al segundo *corpus*, el *corpus* recolectado para el español y que se detalla en la

siguiente sección. En ambos casos el conjunto de atributos fue el que se determinó en el capítulo anterior.

5.2 Recolección del Corpus en Español

Debido a que no existe un *corpus* etiquetado para el idioma español, se realizó la recolección de un *corpus* que se pudiera utilizar para la clasificación de textos de opinión.

Para ello, se realizaron búsquedas de textos de opinión provenientes de sitios Web. Un primer caso fueron los editoriales de periódicos de circulación nacional como La Jornada y El Universal; otro caso fueron las opiniones sobre productos electrodomésticos y sobre películas. En el primer caso, se observó que la mayoría de las opiniones obtenidas, eran negativas, lo cual provocaría que las clases estuvieran desbalanceadas. Probablemente esto sea debido a que los temas que abarcan dichos editoriales son a cerca de política y gobierno, y se puede observar que en general las críticas en éste ámbito son en su mayoría negativas.

Por otro lado también se observó que las opiniones expresadas de electrodomésticos eran demasiado cortas y en muchas ocasiones eran simplemente conjuntos de adjetivos (e.g. muy grande y pesada).

Finalmente se decidió trabajar con opiniones de películas. Estas opiniones fueron seleccionadas del sitio de Internet www.ciao.es, en el cual permite a los usuarios plasmar su opinión acerca de diferentes productos y servicios, entre los que se encuentran películas.

5.3 Criterios para Etiquetar el Corpus

Para el etiquetado manual de nuestro *corpus* se realizó un análisis inicial en el que se encontraron oraciones de diferente tipo, de las cuales podemos encontrar oraciones con información objetiva y opiniones. A diferencia de los trabajos previos encontramos que dentro de las opiniones no sólo se encuentran opiniones a favor y en contra, sino que existen otro tipo de opiniones. Para el etiquetado de las oraciones con opiniones se propusieron siete clases. Sin embargo, en nuestro trabajo utilizamos solamente al igual que en los trabajos previos las oraciones positivas y negativas, pero mencionamos las demás clases para que en un futuro no se descarte la posibilidad de utilizar las demás clases. A continuación se explica cada una de ellas y mencionamos un ejemplo de cada una:

- **Positivo.** Opinión que expresa algo a favor acerca de algo o alguien.

  Ej. El actor lo hace muy bien, yo no lo conocía (más quisiera, mmmm) pero ha estado a la altura de esta gran producción.

- **Negativo.** Opinión que expresa algo en contra de algo o alguien.

  Ej. Y mira que he visto películas malas eh… Pero ésta se lleva la palma.

- **Neutro**. Opinión que no expresa claramente su sesgo a favor o en contra.

  Ej. A mi eso no me parece malo.

- **Recomendación**. El autor da una recomendación.

  Ej. El combate a muerte en el Ministerio y la múltiple caída de otras profecías tras el ejército de Potter, son efectos que se pueden disfrutar mejor en la pantalla grande y con sonido IMAX.

- **Interrogativo**. El autor hace una crítica, en forma de pregunta.

  Ej. Lo dicho, cada cual aportó su granito de arena, pero, ¿qué podía ofrecernos un recién llegado al mundo del cine como David Yates?

- **Prescriptivo**. El autor da su opinión acerca de cómo le hubiera gustado que fuera algo o alguien.

  Ej. El problema es que era la hora de empezar a ponerse un poco más tenebrosos e incidir en los peligros que conllevan las aventuras de un Potter cada vez más adulto, pero no ha sido así.

- **Combinado**. Oraciones donde el autor combina opiniones de ambas clases: positivas y negativas.

  Ej. Vemos como hay gags excelentes, también los hay buenos, a secas, que te dejan con la sonrisa, pero también los hay que te dejan cara de interrogante.

En nuestro caso, como se mencionó, sólo se utilizaron aquellas oraciones positivas y negativas. Se etiquetaron 154 oraciones positivas y 185 oraciones negativas. De las cuales sólo se utilizaron 154 oraciones positivas y 154 oraciones negativas para mantener las clases balanceadas.

5.4 Experimentos y Resultados

5.4.1 *Corpus* en Inglés

Como se mencionó al principio de este capítulo, para esta fase se utilizaron dos *corpus*, el *corpus* utilizado en los experimentos del capítulo anterior, el cual se encuentra en idioma inglés, y el *corpus* recopilado de www.ciao.es que se encuentra en español.

El *corpus* en inglés consta de 10662 oraciones de opiniones (5331 oraciones con opiniones positivas y 5331 oraciones con opiniones negativas).

De este *corpus* se utilizaron 1706 oraciones (853 positivas y 853 negativas) como conjunto de oraciones etiquetadas, y que se utilizaron para la primera iteración de entrenamiento del método semi-supervisado, 6 824 oraciones (3412 positivas y 3412 negativas) como el conjunto de las oraciones no etiquetadas y 2132 (1066 positivas y 1066 negativas) para conjunto de prueba.

5.4.2 Resultados Utilizando el Corpus en Inglés

De acuerdo a la caracterización realizada y descrita en el capítulo anterior, el conjunto de atributos con los que la clasificación proporciona mejores resultados son Secuencias Frecuentes Maximales de frecuencia 5, es por ello que se realizaron experimentos utilizando este conjunto de atributos. Sin embargo, dado que el tamaño del *corpus* de entrenamiento es mucho menor también se realizaron algunos experimentos con frecuencias menores.

Por razones prácticas y dado que el método Naive Bayes reporta buenos resultados las pruebas se realizaron con este clasificador.

Recordemos que en la caracterización, gracias al análisis realizado al conjunto de datos, encontramos que existen oraciones que no logran ser caracterizadas, es decir, que estas oraciones no contienen ningún atributo del conjunto de atributos, lo cual resulta en una instancia con sólo ceros. Y al igual que en el capítulo anterior, estas oraciones que no logran ser caracterizadas, fueron eliminadas en la fase de entrenamiento. La solución propuesta es eliminar las instancias que no logran ser caracterizadas por los atributos antes de la fase de entrenamiento.

Para realizar los experimentos se implementó el método semi-supervisado con la representación de pesado propuesta, descrita en el capítulo 4.

Se realizaron experimentos utilizando diferentes umbrales de confianza y diferentes umbrales de razón de frecuencia.

I	**Número de atributos**	**Instancias eliminadas**	**Exactitud en:**			**Instancias agregadas**			**Instancias no caracterizadas en el conj. de prueba**
			Validación cruzada	**Conjunto de prueba**	**Conj. Datos no etiq**	**pos**	**neg**	**total**	
1	278	44	68.76%	58.34%	58.54%	25	145	170	648
2	283	47	**80.67%**	**59.84%**	58.67%	0	187	187	725
3	347	50	67.04%	52.15%	52.63%	0	206	206	601
4	416	46	66.94%	50.14%	50.84%	0	226	226	464
5	534	49	69.17%	50.00%	50.68%	0	249	249	365
6	679	49	71.19%	50.00%	50.75%	0	274	274	258

Tabla 5.1 Resultados del método semi-supervisado

En la primera columna de la tabla 5.1 se muestra el número de iteración del proceso semi-supervisado, en la segunda columna se muestra el número de atributos obtenidos en el conjunto de datos etiquetados en cada iteración, en la tercera columna se muestra el número de instancias no caracterizadas y que se han eliminado previamente a la fase de entrenamiento, en la cuarta columna se muestra la exactitud obtenida utilizando validación cruzada (10 pliegues), en la siguiente se muestra la exactitud obtenida en el conjunto de prueba, en la siguiente la exactitud obtenida validando en el conjunto de datos no etiquetados, posteriormente se muestran el número de instancias agregadas, positivas, negativas y el total, y finalmente en la última columna se muestra el número de instancias que no logran ser caracterizadas en el conjunto de prueba. Con respecto a esta última columna se busca que el número de instancias no caracterizadas sea el menor posible, es decir, que el número de atributos sea suficiente para caracterizar la mayor cantidad de instancias.

Observamos que en la primera iteración se obtiene 58.34% de exactitud validando en el conjunto de prueba y en la segunda aumenta a 59.84% lo cual significa que el método funciona al menos para la segunda iteración, a partir de la tercera la exactitud disminuye.

Si se observa la tabla 5.1, se puede ver que con el método semi-supervisado se va agregando un número mayor de instancias de la clase negativa que de la clase positiva, lo que provoca que las clases queden desbalanceadas. Lo anterior se debe a que el clasificador creado predice con mayor confianza instancias de la clase negativa que cuando predice instancias de la clase positiva.

Para mantener las clases balanceadas se propone que el método agregue una instancia de una clase por cada instancia que se agregue de la otra clase.

Para que el método agregue instancias de ambas clases es necesario disminuir el umbral de confianza, de manera que abarque ambas clases. Ade-

más de lo anterior, de manera experimental encontramos que el número de instancias agregadas al disminuir el umbral de confianza se elevaba, de manera que el clasificador empeoraba por lo que se propuso que el número total de instancias agregadas no rebasara el 10% de instancias del conjunto de ejemplos etiquetados. Por otro lado, el número de atributos es bajo, de ahí que se decidiera disminuir el umbral de frecuencia para el cálculo de las SFM para incrementar el número de atributos.

I	Número de atributos	Instancias eliminadas	Exactitud en:			Instancias agregadas			Instancias no caracterizadas en el conj. de prueba
			Validación cruzada	Conjunto de prueba	Datos no etiq.	pos	neg	total	
1	1994	36	83.49%	**60.55%**	62.03%	85	85	170	324
2	2068	37	86.52%	59.75%	61.00%	88	88	176	354
3	2192	42	87.39%	59.56%	60.31%	17	17	34	402
4	2228	42	**87.50%**	59.00%	60.21%	16	16	32	406

Tabla 5.2 Resultados del método semi-supervisado, con clases balanceadas

En la tabla 5.2 se muestran los resultados obtenidos utilizando un umbral de confianza de 58%, un umbral de razón de frecuencia de 0.40 y el conjunto de atributos utilizados fue SFM de frecuencia 2. Utilizando validación cruzada observamos que la exactitud mejora, sin embargo en el conjunto de prueba sucede lo contrario, comenzando en la primera iteración con 60.55% y a partir de la segunda iteración la exactitud en el conjunto de prueba disminuye.

Si disminuimos el umbral de razón de frecuencia, la selección de atributos sería más estricta. En la siguiente tabla se muestran los resultados obtenidos utilizando un umbral de razón de frecuencia de 0.05. El conjunto de atributos es SFM con frecuencia 2 y el umbral de confianza es de 58%, de igual forma que el experimento mostrado anteriormente, las clases permanecen balanceadas en cada iteración.

I	Número de atributos	Instancias eliminadas	Validación cruzada	conjunto de prueba	Datos no etiq	Instancias agregadas			Instancias no caracterizadas en el conj. de prueba
						pos	neg	total	
1	1685	36	99.24%	60.36%	59.84%	6	6	12	676
2	1691	36	98.07%	60.41%	59.82%	6	6	12	676
3	1703	36	98.77%	60.55%	59.95%	6	6	12	682
4	1700	37	98.37%	60.74%	59.75%	6	6	12	693
5	1721	36	**99.39%**	**60.88%**	59.87%	7	7	14	691
6	1727	35	98.46%	60.69%	59.67%	6	6	12	697

Tabla 5.3 Método semi-supervisado, con umbral de razón de frecuencia de 0.05

En la tabla 5.3 se observa que bajo estas condiciones, se logra un 99.39% de exactitud con validación cruzada. Si embargo validando en el conjunto de prueba, vemos que se logra un 60.88% de exactitud en la 5ta iteración. Después de estos experimentos podemos concluir que dado que tenemos pocos datos los umbrales deben relajarse, sin embargo esto nos lleva a una situación en que la mejora al aplicar el método semi-supervisado es mínima.

5.4.3 *Corpus* en Español

Este *corpus* fue recopilado del sitio www.ciao.es de la sección de películas en cartelera (sección 5.3). Consta de 308 oraciones de opiniones en idioma español las cuales fueron etiquetadas manualmente y de 997 oraciones de opiniones no etiquetadas.

De estas 308 oraciones, aproximadamente el 10% de las oraciones conforman el conjunto de datos de prueba (30 oraciones, 15 positivas y 15 negativas).

5.4.4 Resultados Utilizando el Corpus en Español

Aplicando el método semi-supervisado propuesto en el *corpus* en español, obtenemos los siguientes resultados. El conjunto de atributos utilizado fue SFM con frecuencia 2. En la siguiente tabla se muestran los resultados obtenidos utilizando un umbral de razón de frecuencia de 0.70

I	**Num. de atributos**	**Instancias eliminadas**	**Validación cruzada**	**Conj. de prueba**	**Instancias agregadas**			**Instancias no caracterizadas en el conj. de prueba**
					pos	**neg**	**total**	
1	675	3	66.66%	66.66%	13	13	26	2
2	761	3	**66.88%**	**70.00%**	15	15	30	1
3	845	3	63.82%	66.66%	16	16	32	1
4	920	1	64.64%	65.58%	18	18	36	1
5	1024	3	65.31%	63.33%	20	20	40	1
6	1133	4	64.05%	60.00%	16	16	32	2

Tabla 5.4 *Corpus* en español, umbral de razón de frecuencia 0.70

Como podemos observar utilizando un umbral de razón de frecuencia de 0.70, validando el método semi-supervisado propuesto en el conjunto de prueba se logra un 66.66% de exactitud en la primera iteración, en la segunda se logra un **70%** de exactitud y a partir de la 3ª iteración la exactitud disminuye, validando en el conjunto de prueba.

Si cambiamos el umbral de razón de frecuencia y utilizamos ahora un umbral de 0.60 los resultados obtenidos son los siguientes:

I	**Num. de atributos**	**Instancias eliminadas**	**Validación cruzada**	**Conj. de prueba**	**Instancias agregadas**			**Instancias no caracterizadas en el conj. de prueba**
					pos	**neg**	**total**	
1	638	3	67.89%	63.33%	13	13	26	2
2	717	3	70.03%	70.00%	15	15	30	1
3	778	3	68.80%	70.00%	16	16	32	1
4	853	3	67.03%	70.00%	18	18	36	1
5	945	3	68.78%	**73.33%**	20	20	40	3
6	1034	2	**71.19%**	66.66%	22	22	44	3

Tabla 5.5 *Corpus* en español, umbral de razón de frecuencia 0.60

Observamos que, en comparación con los resultados mostrados anteriormente, cambiando el umbral de razón de frecuencia, siendo más estrictos, obtenemos un incremento en los resultados, logrando un 73.33% de exactitud validando en el conjunto de prueba en la quinta iteración.

Con el método semi-supervisado basado en *self-trainning*, sólo se hace uso de un clasificador, pero podríamos pensar que haciendo una combinación de diferentes clasificadores individuales en lugar de uno sólo, podría complementarse la decisión de clasificar y de esta manera mejorar la ejecución de la tarea de clasificación con mayor exactitud. Teniendo en mente esta idea, se propuso el uso de una técnica de aprendizaje utilizada comúnmente en aprendizaje supervisado, y adaptarlo al método semi-supervisado. En la literatura de Aprendizaje Computacional este enfoque es conocido como *stacking, staked generalization* o simplemente combinación de clasificadores. Este proceso semi-supervisado funcionaría de la siguiente manera:

De igual forma que el método propuesto se tiene inicialmente un conjunto de datos etiquetados y otro de datos no etiquetados. Con el conjunto de datos etiquetados se obtienen varios clasificadores c_1, c_2, c_3,… c_n . En nuestro caso sólo utilizaremos 2 clasificadores. El resultado de estos clasificadores se usan como atributos para un nuevo clasificador, utilizaremos el método de Naive Bayes para crear este clasificador. A este tipo de clasificadores que aprenden de las salidas de otros clasificadores, se les llama meta-algoritmo (meta learner). Con el nuevo clasificador creado, se clasifican los datos no etiquetados. Y el proceso semi-supervisado continúa como se explicó al principio de este capítulo, con la diferencia del nuevo módulo agregado, donde se crea la combinación de clasificadores. El siguiente diagrama muestra el nuevo módulo agregado al método semi-supervisado.

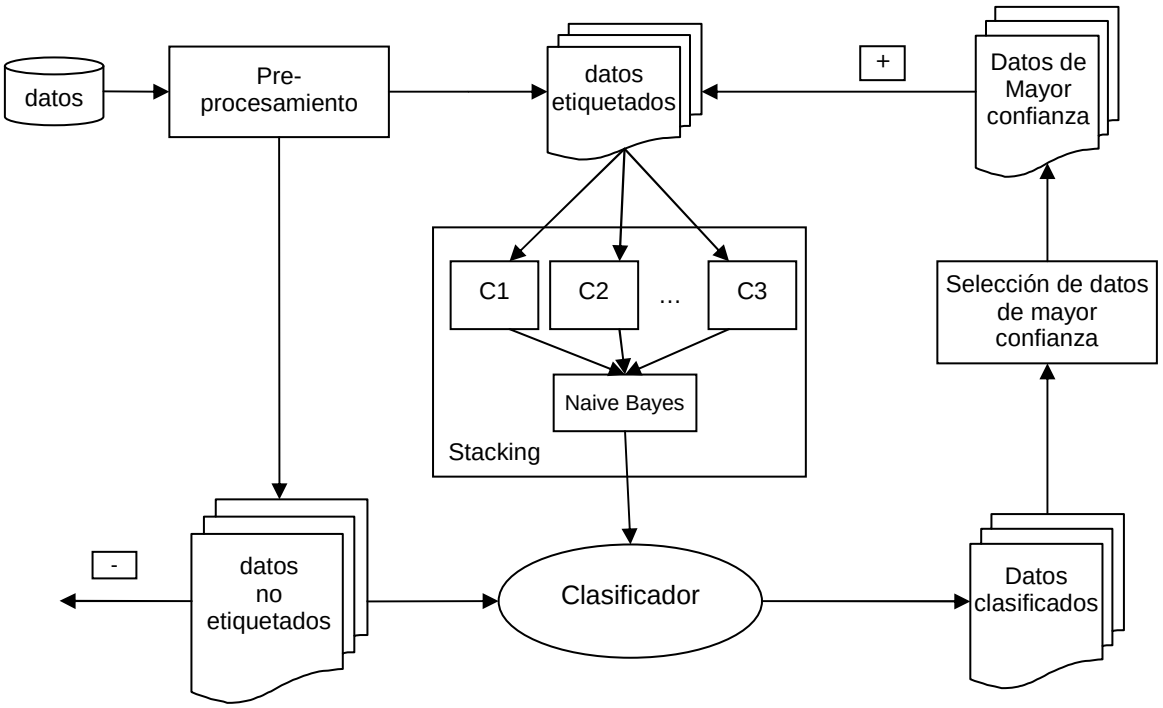

Figura 5.2 Método semi-supervisado con Stacking

Para los siguientes experimentos se utilizaron como métodos base para el stacking: NaiveBayes y SVM y el conjunto de atributos utilizado fue SFM con frecuencia 2.

I	Número de atributos	Instancias eliminadas	Validación cruzada	Conjunto de prueba	Instancias agregadas			Instancias no caracterizadas en el conj. de prueba
					pos	neg	total	
1	675	3	70.69%	63.33%	13	13	26	2
2	731	3	71.47%	66.66%	15	15	30	1
3	801	3	**71.77%**	**70.00%**	16	16	32	2
4	871	3	67.69%	56.66%	18	18	36	1
5	979	2	66.49%	60.00%	19	19	38	1
6	1073	3	68.92%	66.66%	21	21	42	2

Tabla 5.6 Resultados del método semi-supervisado en idioma español

En la tabla 5.6 se muestran los resultados obtenidos utilizando un umbral de razón de frecuencia de 0.70. Como podemos observar utilizando este umbral de razón de frecuencia, aplicando el método semi-supervisado propuesto se logra obtener **70%** de exactitud en la tercera iteración y a partir de la 4ª iteración la exactitud disminuye.

Si cambiamos el umbral de razón de frecuencia y utilizamos un umbral de 0.60 los resultados obtenidos son los siguientes:

I	Número de atributos	Instancias eliminadas	Validación cruzada	Conjunto de prueba	Instancias agregadas			Instancias no caracterizadas en el conj. de prueba
					pos	neg	total	
1	638	3	71.58%	63.33%	13	13	26	2
2	687	3	**74.07%**	70.00%	15	15	30	3
3	746	3	72.92%	70.00%	16	16	32	2
4	825	4	70.50%	**73.33%**	18	18	36	3
5	925	3	69.97%	66.66%	20	20	40	3
6	999	3	70.23%	56.66%	22	22	44	3

Tabla 5.7 Método semi-supervisado con Stacking, idioma español y umbral de razón de frecuencia de 0.60

Observamos que, en comparación con los resultados mostrados anteriormente, cambiando el umbral de razón de frecuencia, siendo más estrictos obtenemos de nuevo un incremento en los resultados, logrando en la segunda y tercera iteraciones una exactitud de 70% y en la cuarta iteración hasta 73.33%. Lo anterior nos indica que existen atributos que se utilizan para ambas clases, sin embargo son más frecuentes en una clase que en la otra, y estos atributos si

contienen información importante para la clasificación. Con este último experimento comprobamos que agregando el módulo de Stacking, no se logra incremento en la exactitud. Sin embargo, llegamos al 73.33% de exactitud con una iteración menos y el número de atributos se reduce de 945 a 825. Podemos decir entonces que el agregar el módulo de Stacking no mejora la clasificación significativamente. Además de que se utilizan mayores recursos computacionales, puesto que el tiempo de procesamiento aumenta al aumentar el número de clasificadores (módulo Stacking). Por otra parte con el método semi-supervisado basado en *self-training*, llegamos a obtener hasta un 73% de exactitud, que comparado con nuestra caracterización en donde logramos un 79% de exactitud, es sólo 7% menos, tomando en cuenta que es un proceso semi-supervisado, donde existen mayor número de datos no etiquetados. Cabe mencionar también que hasta la fecha no existe trabajo previo que proponga un método semi-supervisado para la tarea de clasificación de opiniones a nivel oración, además de orientarse únicamente a recursos léxicos.

Dentro del análisis global de los resultados podemos mencionar también que el umbral de confianza con el que predice el clasificador Naive Bayes, puede variar extremadamente entre las clases, por lo que es mejor utilizar un porcentaje de instancias agregadas en cada iteración, o ambas condiciones como en nuestro caso.

Capítulo 6

Conclusiones y Trabajo Futuro

En este trabajo de tesis se presentó un método de caracterización para la clasificación automática de textos de opinión.

Este método consiste en el uso de Secuencias Frecuentes Maximales como atributos. Es importante recalcar que el uso de SFM como atributos es algo novedoso, pues en la literatura no existe evidencia de trabajos que hagan uso de este tipo de atributos para la tarea de determinar la polaridad de oraciones de opiniones.

Tras realizar los experimentos necesarios se concluyó que efectivamente el uso de SFM como atributos permite al clasificador alcanzar mejores resultados en comparación con los tipos de atributos tradicionales, para la tarea de clasificación de la polaridad de oraciones de opiniones. A pesar de alcanzar mejores resultados utilizando las SFM, no se logran los resultados deseados, ya que dichos resultados están a un 10% de exactitud por debajo de los resultados obtenidos en trabajos previos [19]. Sin embargo, estos trabajos previos utilizan recursos sintácticos lo que los hace dependientes a un cierto idioma. En contraste, nuestro trabajo al utilizar sólo recursos léxicos es fácilmente adaptable a otros idiomas.

El método de caracterización tiene la desventaja de que no existe un criterio claro para determinar el conjunto de atributos más adecuado para la clasificación, incluyendo el umbral de frecuencia adecuado para obtener las SFM, de ahí que se requiera de un gran número de experimentos.

Posteriormente se presentó un método semi-supervisado, el cual se basa en el uso de técnicas de Aprendizaje Computacional. Este método permite obtener oraciones etiquetadas automáticamente con pocos datos etiquetados para la fase de entrenamiento.

El método se evaluó con dos *corpus*, el primero es un *corpus* que consta de oraciones de películas en idioma inglés, el cual ha sido utilizado en [17] y el segundo de igual forma consta de oraciones de películas pero en idioma español, dichas oraciones fueron escritas por diferentes críticos de cine y fueron recopiladas de www.ciao.es.

La razón de incluir el módulo de stacking dentro del método semi-supervisado propuesto responde a que utilizar un conjunto de clasificadores en lugar de uno sólo, puede complementar la decisión de clasificar y de esta manera obtener una clasificación con mayor exactitud. Sin embargo, para la tarea de clasificación de opiniones a nivel de oración no mejora significativamente la clasificación.

Dentro de las desventajas de estos métodos podemos mencionar que el umbral de confianza no es el mismo para cualquier conjunto de datos, debido a que depende del resultado de la fase de entrenamiento. Sin embargo otra forma de obtener los datos de mayor confianza es por medio de un porcentaje de instancias agregadas.

6.1 Trabajo Futuro

Tomando en cuenta las restricciones y desventajas que presenta el método, se han identificado algunas tareas que se desprenden de este trabajo sobre las cuales se pretende explorar a futuro. Estas tareas se presentan a continuación:

1. Evaluar el desempeño del método en los siguientes escenarios:
    a. En otro tipo de tareas de clasificación como determinar la objetividad-subjetividad de oraciones.
    b. En otros idiomas. Dado que el método se basa sólo en atributos léxicos, no está ligado a un idioma específico, por lo que puede ser adaptado de manera sencilla para aplicarse a otros idiomas.
    c. En dominios, como por ejemplo editoriales y opiniones de productos.

2. Aumentar el número de oraciones etiquetadas del conjunto inicial para el método semi-supervisado para el español. El número de oraciones etiquetadas fue bastante limitado, por lo que se propone que el número inicial de oraciones sea mayor al actual. De esta manera el clasificador inicial tendrá más ejemplos de donde aprender.

Referencias

[1] S. Argamon, S. Levitan, Measuring the usefulness of function words for authorship attribution. *In Proceedings of ACH/ALLC 2005, Association for* Computing and the Humanities, Victoria, BC, 2005

[2] J. Diederich, J. Kindermann, E. Leopold, and G. Paass. Authorship attribution

with support vector machines. *Applied Intelligence, v.19, n.1-2, pp.109-123*, 2003

[3] B. Kessler, G. Numberg, H. Schütze. Automatic Detection of Text Genre. *In proceedings of the 35th annual meeting on Association for Computational Linguistics, Madrid, pp. 32-38*, Julio 1997.

[4] A. Finn, N. Kushmerick, B. Smyth. Genre Classification and Domain Transfer for Information Filtering. *In Proceedings of 24th European Colloquium on Information Retrieval Research, pp. 353-362*, 2002

[5] J. Karlgren, D. Cutting. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. *In Proceedings of the 15th International Conference on Computational Linguistics, Kyoto, Japan, pp. 1071-1075*, October 1994

[6] E. Riloff, J. Wiebe. 2003. Learning extraction patterns for subjective expressions. *In Proceedings of the EMNLP-03, pp. 105-112*, 2003

[7] P. D. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 417-424*, Julio 2002

[8] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment Classification Using Machine Learning Techniques. *In Proceedings of the EMNLP-02, pp. 79-86,* 2002

[9] T. Mitchell. Machine Learning, *McGraw Hill*, 1997

[10] K. Aas, L. Eikvil. Text Categorisation: a Survey. *Technical Report, Norwegian Computing Center,* 1999

[11] A. Esuli, F. Sebastiani. Determining term subjectivity and term orientation for opinion mining. *In Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 193–200,* 2006

[12] A. Esuli and F. Sebastiani. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. *In Proceedings of LREC 2006 - 5th Conference on Language Resources and Evaluation, pp. 417-422,* 2006

[13] H. Yu, V. Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing 2003, pp. 129-136,* July 2003

[14] V. Hatzivassiloglou, J. Klavans, M. Holcombe, R. Barzilay, M. Kan, K. McKeown. SIMFINDER: A Flexible Clustering Tool for Summarization. *In Proceedings of the Workshop on Automatic Summarization in NAACL-01, pp. 41-49,* 2001.

[15] J. Wiebe, E. Riloff. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. *In Proceedings of the Sixth International Conference on Intelligent Texto Processing and Computational Linguistics 2005,pp. 486-497,* 2005

[16] P.J. Stone, D.C. Dunphy, M.S. SNT, D.M. Ogilvie. The General Inquirer: A Computer Approach to Content Analysis. *MIT. Press, Cambridge, US,* 1966

[17] B. Pang, L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *In Proceeding of ACL 2005, pp.115-124,* 2005

[18] H. Ahonene. Discovery of Frequent Word Sequence in Text. *In Proceedigns of the ESF Exploratory Workshop on Pattern Detection and Discovery, pp. 180-189,* Septiembre, 2002

[19] B. Erikson. Sentiment Classification of Movie Reviews using Linguistic Parsing. *Technical Report of the University of Wisconsin-Madison. EEUU.*

[20] P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *In Proceedings of ACM Transactions on Information Systems (TOIS), v. 21, n. 4 pp. 315-346,* Octubre, 2003

[21] J. Kamps, M. Marx, R. J. Mokken, M. De Rijke. Using WordNet to measure semantic orientation of adjectives. *In Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, volume IV, pp. 1115-1118, Lisbon, PT,* 2004

[22] B. Liu, M. Hu, J. Cheng. Opinion Observer: Analyzing and Comparing Opinions on the Web. *In Proceedings of the 14th International Conference on World Wide Web, ACM, Press, 2005, pp. 342-351,* May 2005

[23] D.D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. *In Proceedings of the 10th European Conference on Machine Learning*, 1998, pp. 4-15, May 1998

[24] K. Church, P. Hanks. Word association norms, mutual information and lexicography. Proceedings of the 27th Annual Conference of the ACL. pp. 76-83, New Brunswick, 1989

[25] V. Ng, C. Cardie. Weakly Supervised Natural Language Learning Without Redundant Views. *In Proceedings of the 2003 Meeting of the North American chapter of the Association for Computational Linguistics (NAACL-03).* pp. 173–180, 2003

[26] N. Castell, N. Català. Construcción Automática de Diccionarios de Patrones de Extracción de Información. *Procesamiento del Lenguaje Natural. N° 21*, pp. 123-136, Julio, 1997