Capítulo 1

Introducción

1.1. Planteamiento del problema

El crecimiento de la Web ha traído consigo la aparición de diversos sitios como foros de opinión y sitios de ventas donde cualquier usuario puede escribir un comentario acerca del producto o servicio que ofrece dicho sitio. Estos comentarios son de gran importancia ya que tendemos a tomar decisiones basadas en lo que los demás piensan acerca de algo o alguien.

La cantidad de opiniones que está disponible a cualquier persona que cuente con acceso a la Web es muy grande. Por ello, resulta imposible analizarlas todas y determinar la tendencia, orientación o polaridad general sobre algo. Ya sea positiva, negativa o neutra, o algo intermedio. El resultado para el usuario es que es imposible realizar una clasificación de forma rápida y/o automática.

El análisis de sentimientos o minería de opiniones es el tratamiento computacional de opiniones, sentimientos y subjetividad en textos. Es el área de la lingüística computacional y de la recuperación de información que hoy en día atiende este problema [1]. Además, la minería de opiniones no busca determinar el tema de un documento, sino la orientación de la opinión que se expresa en él.

La clasificación de opiniones es una tarea de la minería de opiniones que se encarga de asignar a un documento una etiqueta de acuerdo al tipo de opinión que en él se expresa: positiva, negativa o neutral.

La minería de opiniones encuentra sus aplicaciones en diversas áreas. En la recuperación de información, puede ser útil eliminar las opiniones de un documento para obtener mejores resultados a las consultas realizadas [1]. En el área comercial, es muy útil conocer la percepción de un producto o servicio y estar al tanto de lo que la gente opina, como las características negativas o positivas comentadas con mayor frecuencia. Poder contar con un sistema que pueda encontrar y condensar de alguna forma todas estas opiniones para mejorar o replantear un producto o servicio disminuiría el tiempo invertido por un analizador humano que tendría que leer cientos o miles de opiniones, muchas veces iguales.

En el área de inteligencia gubernamental resultaría útil conocer las fuentes de hostilidades, los temas que causan más polémica, entre qué tipo de población surgen opiniones negativas, e inclusive los temas que generan opiniones subversivas. Asimismo, sería útil conocer los temas que causan malestar entre la población para poder atenderlos y así mejorar la vida de los habitantes.

1.2. Objetivos

El objetivo general de este trabajo de tesis es crear un sistema capaz de clasificar, en positivas o negativas, oraciones[1] provenientes de reseñas en inglés sobre películas, según la opinión que contengan. Este clasificador será supervisado, entrenado con un corpus de reseñas de películas recabado desde la Web. Al entrenar el clasificador, se proponen y se ponen a prueba cuatro métodos que tienen por objetivo mejorar el desempeño del clasificador.

Los objetivos específicos de este trabajo son:

- Extraer de la Web reseñas u opiniones acerca de películas. Preprocesar esa información y dejarla lista para ser usada por las siguientes etapas del sistema.

[1]En este trabajo, se usa oración y enunciado como sinónimos; aunque se entiende la diferencia en el significado de estas dos palabras: mientras que una oración es una estructura que contiene necesariamente un verbo, un enunciado puede ser cualquier cosa que ocurra entre dos puntos. Como en este trabajo no se ocupa de estructuras oracionales, al hablar de oraciones me estaré refiriendo a enunciados.

- Agrupar automáticamente los enunciados, con el fin de separar los enunciados con algún tipo de opinión de los enunciados descriptivos, usando selección de rasgos[2].
- Elegir las oraciones que cuentan con adjetivos o adverbios con el fin de entrenar el sistema, ya que típicamente estos se utilizan para hacer juicios, tanto negativos como positivos [2, 3] .
- Elegir las oraciones que cuentan con disparadores de presuposición, ya que las oraciones que contienen estos disparadores presuponen otro tipo de información que podría ofrecer un juicio.
- Elegir las oraciones que cuentan con adjetivos o adverbios o con disparadores de presuposición. La selección de este tipo de enunciados aumentará la cantidad de oraciones elegidas, manteniendo solo las más relevantes. Este método es la unión de los dos métodos anteriores.
- Crear un clasificador de tipo Bayes ingenuo (*naïve Bayes*) que reciba las oraciones agrupadas o las oraciones subjetivas (encontradas con alguna de las técnicas mencionadas) y que cuenten ya con una clase asignada (separadas en dos conjuntos, uno de oraciones negativas y el segundo de oraciones positivas). Probar el clasificador con ejemplos nuevos. Como último paso, evaluar por medio de medidas de desempeño, el comportamiento del clasificador.

1.3. Estado del arte

Hoy en día, en el área de minería de opiniones, se trabaja principalmente en resolver los dos principales problemas de esta área, la identificación del texto subjetivo y la clasificación de la opinión contenida en ese tipo de textos. Sin embargo, existen otras aplicaciones, de acuerdo a [4, 5, 6], que representan también nuevos problemas y retos. Estas aplicaciones son tres principalmente:

[2]La selección de rasgos es un método que utiliza sólo los rasgos que ofrezcan la mayor información de acuerdo a cierto criterio.

1. **Comparación de productos**: para poder ofrecer una comparación acerca de un producto con otro producto, es necesario conocer qué se opina sobre las características que definen a ese producto. Asimismo, si se conocen las opiniones sobre las características del producto, el usuario podría leer solo las opiniones que conciernen a las características en las cuales él está interesado. En [7] se realizó un sistema capaz de comparar las características de diferentes productos que compiten entre sí. Primero, se usaron técnicas de minería de patrones de lenguaje para identificar las características sobre las cuales los consumidores han expresado alguna opinión. Segundo, por cada característica identificada, se averiguó si la opinión de cada usuario es positiva o negativa. Con esta información realizaron una interfaz que permite visualizar y comparar las opiniones de diferentes productos.

2. **Resumen de opiniones**: el número de opiniones generadas en línea crece rápidamente, y más para productos populares. Para un usuario es complicado leer todas las opiniones, más aún cuando las oraciones juiciosas están contenidas en un texto largo, donde la mayoría de las oraciones no ofrecen opinión alguna. En [8] se realizaron resúmenes de opiniones mediante los siguientes tres pasos:

   a) Se identifican las opiniones de las características del producto,

   b) Se determina la polaridad de esas opiniones, y

   c) Se genera un resumen con la información obtenida.

   Con un resumen, los potenciales usuarios pueden observar fácilmente como otros usuarios se sienten respecto al producto.

3. **Minería de motivaciones de la opinión**: Además de conocer la orientación de la opinión de un producto, es de gran utilidad conocer las razones por las cuales el autor de la opinión se expresó positiva o negativamente. En [9] se detectaron expresiones con opiniones mediante la búsqueda de oraciones que explícitamente indican las ventajas y desventajas de los productos. Posteriormente se entrenó un sistema de reconocimiento de oraciones, para

obtener los enunciados con las características que generan la opinión final de los usuarios.

Además de estos nuevos retos, se ha desarrollado, siguiendo la clasificación positiva o negativa de opiniones, la clasificación de emociones. Las emociones humanas son múltiples y depende de la investigación qué conjunto de ellas usar. En [10], se utiliza un modelo de emociones distinguibles ya verificado empíricamente y se sugiere su uso potencial en el procesamiento del lenguaje natural para la clasificación automática de emociones escritas en textos en inglés en ocho niveles. En [11] se utilizan seis emociones básicas: ira, desagrado, miedo, alegría, tristeza y sorpresa. En este trabajo se construyó un corpus formado con encabezados de noticias. Posteriormente se etiquetó manualmente con las seis emociones mencionadas. Finalmente se probaron distintas técnicas para detectar automáticamente las emociones en los encabezados.

Otra área de investigación desarrollada, muy ligada a la minería de opiniones, es la detección de ironía[3]. En [13] se enfocaron en detectar la ironía en enunciados (escritos en portugués) que contienen predicados positivos, dado que presumen que estos son los más expuestos a la ironía. Esto lo lograron explorando ciertos indicios orales y gestuales en los comentarios de usuarios de un sitio web de noticias. Mediante características lingüísticas, [12] identificó tres grupos diferentes de ironía. También, en ese trabajo, se plantea la posibilidad de identificar automáticamente la ironía de dos de los grupos identificados.

En lo que concierne a la identificación de frases subjetivas y a la clasificación de las opiniones contenidas en esas frases, al ser problemas de clasificación, se pueden resolver por medio de aprendizaje supervisado.

Para la identificación de oraciones subjetivas, en [14], se seleccionaron como rasgos de entrenamiento elementos influenciadores de opinión contextuales, tales como la negación (no, nunca) y la contradicción (pero, sin embargo). En [15, 16] se usaron elementos léxicos, para clasificar oraciones en subjetivas u objetivas. Otro enfoque usado es la aproximación por similitud; este método se basa en la hipótesis de que las frases subjetivas se parecen más a otras frases subjetivas que

[3]La deinición de ironía usada es la encontrada en [12]: la ironía como palabras que expresan lo contrario de lo que se quiere decir.

a frases sin opinión alguna. En [17] se mide la similitud entre oraciones mediante rasgos como palabras compartidas, frases y *synsets*[4] de *Wordnet*.[5]

Para la clasificación de opiniones, en [18] se utilizó el análisis de conjunciones entre adjetivos para detectar la orientación de las frases subjetivas. Al analizar pares de adjetivos (unidos por *y, o, pero,* etc.) extraídos de un conjunto de documentos, la intuición es que el hecho de unir adjetivos está sujeto a limitaciones lingüísticas que definen la orientación de los adjetivos unidos. (por ejemplo, *y* usualmente une dos adjetivos de la misma orientación, mientras que *pero* une regularmente dos adjetivos de orientación opuesta). En [19] se sigue la hipótesis de que dos palabras tienden a ser de la misma orientación semántica si existe entre ellas una fuerte asociación semántica. Haciendo uso de las relaciones léxicas encontradas en *Wordnet*, se pudo calcular una cierta distancia entre adjetivos y definir la orientación de cada uno de ellos.

1.4. Descripción de la tesis

Este trabajo de tesis esta formado por cinco capítulos. En el segundo capítulo se presentan los conceptos relacionados a la clasificación de opiniones, o minería de opiniones, como la minería de datos, de donde se desprende la minería de textos. Asimismo, la minería de textos, como la de datos, aplica técnicas y algoritmos que provienen del aprendizaje de máquinas. Estos temas se abordan en ese capítulo. Se concluye con la minería de opiniones, tema principal de esta tesis, donde se conjuntan todos los conceptos revisados.

El tercer capítulo aborda las herramientas usadas y la metodología seguida para la creación del sistema de clasificación de opiniones. Se presentan las etapas que dieron forma al sistema y también ejemplos de cómo la información recabada fue procesada y almacenada.

[4]Un conjunto de sinónimos o synset es un grupo de datos que son considerados equivalentes semánticamente.

[5]Wordnet es una base de datos léxica para el inglés. Agrupa palabras en synsets, provee definiciones cortas y contiene las diferentes relaciones semánticas entre estos synsets. http://wordnetweb.princeton.edu/perl/webwn

En el cuarto capítulo se presentan y analizan los resultados obtenidos por el clasificador de opiniones. Se muestran tablas de resultados relevantes y diagramas que ayudan a visualizar rápidamente el desempeño de los diferentes métodos utilizados.

En el quinto capítulo se encuentran las conclusiones a las que se llegaron con base en los objetivos planteados y los resultados obtenidos. Se presentan las ventajas y desventajas del sistema y se abordan las posibles mejoras a realizar en el futuro.

En el apéndice A se presentan las cinco tablas con las matrices de confusión para cada uno de los métodos usados para entrenar el clasificador bayesiano. También se presenta la tabla con los resultados promediados de precisión, exhaustividad y medida F.

En el apéndice B se describen los módulos programados que componen al sistema creado en esta tesis. También se presenta un diagrama que muestra la relación que existe entre cada uno de estos módulos.

Finalmente, en los apéndices C y D se describen los módulos de agrupamiento y clasificación, respectivamente. Se indican los parámetros de entrada, los objetos que regresa y los archivos que guarda en disco cada uno de estos módulos.