Capítulo 5

Conclusiones

En este trabajo se desarrolló un sistema capaz de clasificar enunciados dependiendo de la opinión que cada uno expresa acerca de una película. Se cumplió entonces con el objetivo principal y con los objetivos específicos:

- Se extrajo la información desde la Web y se preprocesó para ser usada posteriormente en los procesos de agrupamiento, búsqueda de oraciones subjetivas y clasificación.
- Se agruparon automáticamente los enunciados usando un vector de rasgos constituido de unigramas, trigramas y adjetivos. Esto ocasionó que los enunciados se agruparan por compartir una opinión similar y no por el tema del que hablan (la película comentada).
- Se hizo un esfuerzo por separar las oraciones objetivas de las subjetivas mediante la detección de adjetivos y adverbios. También mediante la detección de disparadores de presuposición y mediante la detección de adjetivos, adverbios o disparadores.
- Se entrenó un clasificador bayesiano ingenuo con las oraciones cuya clase ya era conocida previamente (provenientes del agrupamiento FNM y de la detección de oraciones subjetivas). Después se probó con oraciones no vistas antes por el clasificador. Finalmente se evaluaron los resultados arrojados por las pruebas de cada clasificador.

Los resultados entregados por el clasificador superaron, como se mencionó antes, el baseline de una clasificación aleatoria.

Los resultados de exactitud obtenidos en este trabajo se podrían comparar con los obtenidos en [47] para un clasificador de tipo Bayes ingenuo. Sin embargo, los métodos seguidos en ese trabajo difieren de los seguidos en esta tesis. En ese trabajo la mejor exactitud obtenida para este tipo de clasificadores fue de 81.6 %, trabajando, como se dijo antes, con unigramas concatenados a su etiqueta PoS.

Aunado a la diferente metodología usada en [47], otras diferencias importantes entre este trabajo y aquel son:

- No se utilizó el mismo corpus de entrenamiento. Se extrajo uno desde el sitio web de IMDb y se procesó de forma que no es igual a la forma usada en el trabajo de referencia.

- Se clasificaron oraciones y no reseñas completas. A pesar de que, computacionalmente, una oración se podría considerar igual a una reseña completa (considerando una reseña completa como una sola línea de texto), una reseña completa contiene, obviamente, más palabras. Esta mayor cantidad de palabras podría ofrecer más información acerca de la orientación de la película, mientras que la cantidad reducida de palabras dentro de una oración podría ofrecer menos información acerca de la opinión de esa oración.

- No se realizaron algunos pasos tomados en el trabajo referido, como es el uso de palabras negadas y el uso de signos de puntuación como palabras separadas. También se evito el uso de listas de paro.

A lo largo de la metodología de esta tesis se realizaron pasos que podrían mejorarse en el futuro.

El primero de estos casos se encuentra en el proceso de extracción de información desde IMDb mediante el Web crawler. El Web crawler, al no ser una implementación propia de IMDb, está a merced de la continuidad del estilo y del formato HTML del sitio web. De hecho, en diciembre y enero de 2010 y 2011, el sitio de IMDb cambió considerablemente. De esta manera, si se quisiera repetir este experimento ahora, se tendrían que realizar modificaciones al Web crawler

creado en esta tesis. Sin embargo este problema es común en todos los módulos de este tipo y realmente se escapa del control del programador.

Durante el proceso de agrupamiento con FNM, se tomó un paso poco convencional: el uso exclusivo de trigramas y más aun, de aquellos sin sustantivo en la segunda posición. Como se dijo en la metodología, esta medida parece arbitraria, pero surge de la etapa de exploración del corpus, ya que durante la experimentación se vio que los trigramas, con esa limitación, son los que ofrecen mayor cantidad de frases juiciosas. De hecho, en los resultados experimentales, se observó una ligera mejoría en la exactitud con el uso de estas limitantes. Aun así, convendría hacer un análisis con mayor profundidad acerca de las diferencias existentes en el uso de diferentes n-gramas.

También en el mismo proceso, se decidieron experimentalmente los parámetros del algoritmo de factorización FNM. Estos valores son:

- `r`: Indica el número de grupos deseados.
- `tolerancia`: Indica qué tanto pueden ser diferentes el producto de las dos matrices encontradas y la matriz original.
- `Número máximo de iteraciones`: Indica cuántas iteraciones el algoritmo podrá realizar antes de detenerse.
- `Límite de tiempo`: Indica cuánto tiempo podrá llevarse a cabo la ejecución del algoritmo antes de detenerse.

Sería conveniente realizar pruebas más extensas acerca de la selección de parámetros. Sin embargo, para realizar estas pruebas es necesario dividir los datos de entrenamiento no en dos, entrenamiento y pruebas, sino en tres partes: entrenamiento, estimación de parámetros y pruebas. Este paso resultó imposible de realizar debido al poco tiempo disponible.

Se intentó detectar oraciones subjetivas por medio de disparadores de presuposición, lo cual no funcionó como se esperaba.

A pesar de estas limitantes, el sistema creado en este trabajo posee varias ventajas.

Como se mencionó, la detección de oraciones subjetivas por medio de disparadores de presuposición no funcionó como se esperaba, pero con este experimento se podría descartar su uso en la detección de opiniones. De todas maneras, antes de descartarlo valdría la pena explorar otros tipos de disparadores o si alguno de los usados es mejor que los demás.

Se usaron los adjetivos y adverbios y se confirmó que sí contienen información importante acerca de la orientación de una opinión. Aunque los adjetivos ciertamente dependen del contexto de la oración en la que se encuentran, ya que, como se observo en los resultados de [47], solos no son igual de efectivos.

La información generada y analizada durante la ejecución de los procesos del sistema puede ser reutilizada para otro tipo de proyectos o sistemas ya que a lo largo de las etapas del sistema, se almacena en archivos XML la información procesada, como es el caso de los archivos `peliculas.xml` y `reseñas.xml`. También se utilizaron archivos planos para almacenar las oraciones ya separadas por orientación.

El desarrollo de este sistema requirió de la elaboración de dos procesos que podrían ser utilizados en trabajos futuros, estos procesos son el agrupamiento mediante FNM y el clasificador bayesiano ingenuo. Cabe destacar que el método de FNM es de reciente aplicación con documentos y se ha demostrado ser más útil que otras técnicas de factorización de matrices [39]. Estos dos procesos son módulos independientes, es decir, no requieren de otras partes del sistema para funcionar y se pueden alimentar directamente con otros datos que necesiten ser agrupados o clasificados. En los apéndices C y D se explican los detalles de estos dos módulos de Python.

Este trabajo sirvió también como una introducción a los procesos de pruebas y validación, necesarios en casi cualquier tipo de sistema de minería de textos.

Asimismo, existen varias mejoras que se pueden realizar al sistema: cambiar el algoritmo de clasificación, utilizar uno más robusto como una máquina de vectores de soporte, el de esperanza-maximización, entre otros. También sería ideal ahondar mucho más en la identificación de las oraciones subjetivas dentro de un texto, y también encontrar, dentro de una oración subjetiva, aquellas que realmente hablen del tema del cual se ofrece la opinión. La detección de ironía también sería de gran utilidad para la clasificación de opiniones, ya que ayudaría a

asignar la clase “negativo” a comentarios negativos que podrían parecer positivos (irónicos). Realizar una clasificación de opiniones no solo binaria (negativa o positiva), sino agregando también la clase neutral, podría aumentar la exactitud en la prediccíon de casos positivos y negativos y también permitiría identificar casos neutrales.

Finalmente, la minería de opiniones es un área de investigación relativamente nueva que hoy en día goza de alta popularidad. Es usada ya en múltiples sitios en la Web dado que existen intereses personales, empresariales y hasta gubernamentales por conocer lo que se opina sobre algún tema. Sin embargo, la minería de opiniones está lejos de ser resuelta, aún representa múltiples retos para el procesamiento del lenguaje natural y para la minería de textos.