Creado el 20/02/2014
Última versión: 20/02/2014
José A. Troyano, Universidad de Sevilla
Comunidad moocTLH
Esta obra está bajo una licencia de Creative Commons Reconocimiento-CompartirIgual 3.0 Unported.

Aplicaciones de las TLH/ANÁLISIS DE LA OPINIÓN

INTRODUCCIÓN AL ANÁLISIS DE OPINIÓN

José A. Troyano, de la Universidad de Sevilla nos muestra los objetivos y los problemas a los que se enfrenta la minería y análisis de la opinión, ahora mismo en total auge dada la inmensa cantidad de información disponible en las redes sociales. Cómo evaluar la reputación de tu comercio y hasta cómo detectar tendencias en el futuro más cercano. Pero antes de todo eso hay que conocer a qué nos enfrentamos.

Transcripción del vídeo

Hola, me llamo José Antonio Troyano y soy profesor del departamento de lenguajes y sistemas informáticos de la universidad de Sevilla.

Esta charla estará dedicada al análisis de la opinión, un conjunto de técnicas que permiten extraer información valiosa de los comentarios escritos por usuarios en redes sociales y otros canales de la web 2.0.

La charla estará estructurada en tres bloques.

En un primer lugar motivaremos la utilidad de este tipo de técnicas mediante tres historias reales que ponen de manifiesto la importancia de la opinión, y la viabilidad de extraer indicios de opinión desde contenidos sociales.

En el segundo bloque, identificaremos las tareas más importantes que se pueden abordar a la hora de procesar textos con opiniones.

Por último, explicaremos los retos y dificultades que presenta el análisis de opiniones. Algunos de ellos son propios del procesamiento de textos en lenguaje natural, mientras que otros son específicos de la naturaleza de los textos de opinión y de la calidad de los contenidos que se suelen encontrar en los medios sociales.

Pasemos a la primera parte en la que motivaremos la utilidad del análisis de opiniones.

En el año 2006 el FC Barcelona era uno de los pocos grandes clubes que no lucían publicidad. En aquel momento se estimaba que eso suponía una pérdida de ingresos de unos 15M€ anuales, pero el

hecho de no llevar publicidad se consideraba un orgullo para parte de su masa social.

Cuando se especulaba que se iba a firmar algún contrato con una marca importante, se llegó a un acuerdo con UNICEF por el que no sólo no se recibía nada, sino que el club donaba el 0,7% de sus ingresos a UNICEF.

¿Qué ganaba el club?

Básicamente dos cosas:

- Fundamentalmente una imagen excelente que seguramente contribuyó a ganar simpatías hacia el club en todo el mundo.
- En segundo lugar, romper con un bloqueo de 100 años sin publicidad, lo que le permitió años después firmar un contrato por el que sí recibía ingresos. Actualmente, el club lleva la publicidad de Qatar Fundation lo que le supone unos ingresos de 30M€ al año.

En 2007, un cliente ante la duda de si en los paquetes de sopas de letras de pastas gallo estaban representadas todas las letras (no había encontrado ni la U ni la W), envió una carta a la empresa preguntándole directamente.

Ante la falta de respuesta, decidió escribir el blog “me faltan letras” y promocionarlo en la web a través de menéame.

Al cabo de unos pocos días, y gracias a la repercusión de la iniciativa (primero en la red y luego en medios convencionales) la empresa respondió al cliente e incluyó en sus moldes las letras U, W, Ñ, Ç y la @.

En 2011 fue el estreno de la cuarta entrega de piratas del caribe.

En la industria cinematográfica, antes del lanzamiento de una película se suelen hacer encuestas para intentar predecir el comportamiento en taquilla los primeros días, y tomar decisiones de cara a la distribución en salas y la publicidad previa al estreno.

En aquellas fechas, la empresa Fizziology iniciaba su andadura. En lugar de encuestas, trabajaba con datos extraídos de twitter, fundamentalmente menciones a las películas, para realizar las predicciones.

La taquilla del primer fin de semana fue de 90M$.

Los métodos tradicionales predijeron una taquilla de 110M$ mientras que Fizziology predijo un tope de 95M$, con un coste de evaluación muchísimo menor.

Desde entonces Fiizziology se ha hecho un hueco en el sector de análisis de audiencia usando como fuentes de datos los contenidos generados por los usuarios en redes sociales.

Podemos sacar tres conclusiones de las historias anteriores:

1. La imagen es un aspecto muy importante para muchas entidades. En especial, para las empresas, tener una buena o mala imagen afecta de manera muy sensible en sus resultados

económicos.

2. Hay muchos contenidos en la red en los que se expresan opiniones sobre infinidad de temas y entidades. Y al mismo tiempo, internet tiene una capacidad de propagación y amplificación que en ocasiones puede llegar a dañar la imagen de una entidad.
3. El coste de utilizar la información generada por usuarios es menor que el de muchas técnicas de análisis de mercado basadas en encuestas y, además, se pueden realizar seguimientos casi en tiempo real.

En esta segunda sección, introduciremos cinco tareas típicas del análisis de opiniones. Lo que nos servirá para hacernos una idea del tipo de información que podremos obtener del procesamiento de textos que contienen opiniones.

La primera de las tareas es la de detección de subjetividad. Básicamente consiste en determinar si un texto contiene o no frases subjetivas e identificarlas para su posterior procesamiento.

Se trata de una tarea de clasificación de textos con dos categorías (objetivo, subjetivo). A pesar de que sólo hay que elegir entre dos categorías no es una tarea en absoluto sencilla. Lo que se demuestra por el bajo índice de acuerdo que suele haber cuando diferentes personas anotan manualmente los mismos textos.

El cálculo de la polaridad es, seguramente, la tarea que más se ha aplicado tanto en el ámbito científico como en el de las empresas que ofrecen servicios de análisis de opiniones.

Consiste en determinar si un texto contiene opiniones positivas o negativas con respecto a la entidad analizada.

De nuevo se trata de un problema de clasificación, en principio binario (con las categorías positiva y negativa) y en ocasiones ternario (introduciendo la categoría neutra).

A pesar de que el número de categorías es bajo, el problema suele ser más duro que la clasificación de textos clásica. Especialmente por que no hay una relación tan evidente entre conjuntos de términos y categorías como sí ocurre en la clasificación por temáticas.

En este caso, se pretende hilar más fino y no sólo se persigue decidir si una opinión es negativa o positiva, sino que además se busca obtener algún valor que mida la fuerza de dicha opinión.

Puede plantearse como un problema de regresión, utilizando algoritmos que en lugar de decidir entre categorías ofrezcan como salida un valor numérico.

Para esta tarea, aparte de las palabras que cargan con la opinión (fundamentalmente adjetivos) son especialmente importantes las palabras que actúan de modificadores o intensificadores.

Es la tarea más compleja. En ella se pasa del documento o frase como unidad de análisis a realizar el procesamiento al nivel de palabras.

Se trata de identificar en primer lugar las palabras de opinión (por lo general adjetivos) y las de características (por lo general nombres) y establecer una conexión entre ellas.

Con esto se consigue una mayor profundidad en la interpretación de la opinión. Pasamos de decidir si

una frase es positiva o negativa en general, a especificar el aspecto concreto sobre el que se posiciona.

Para resolver adecuadamente esta tarea serán necesarios recursos y analizadores lingüísticos complejos que permitan trabajar con representaciones sintácticas de los textos.

Esta última tarea plantea la obtención de resúmenes de una colección de textos desde la perspectiva de la opinión expresada en ellos.

Una primera posibilidad es la de generar informes o calcular indicadores a partir de la agregación de los resultados obtenidos por alguna de las tareas anteriores, aplicadas de forma individual a los documentos.

Otra manera de enfocar el resumen de opiniones es obtener como salida un texto que contenga la esencia de las opiniones del conjunto de documentos de entrada. En este caso, el camino para resolver la tarea pasa por la aplicación y adaptación de las técnicas clásicas de resúmenes de documentos.

En esta última sección intentaremos transmitir las dificultades específicas del análisis de opiniones, consecuencia en la mayoría de los casos de las particularidades del tipo de textos con los que habrá que trabajar.

La primera dificultad estriba en la infinidad de fuentes posibles: redes sociales, webs de opinión y foros, son sólo algunos de los tipos de fuentes en las que podemos encontrar información relevante para evaluar opiniones.

Independientemente de los problemas de procesamiento de los textos, la cuestión técnica de acceder a tantas fuentes distintas supone, en sí, un problema considerable.

Una vez superada esta fase, nos queda el problema realmente complejo: ser capaz de decidir si un texto alberga una opinión positiva o negativa. En las siguientes transparencias ilustraremos mediante una serie de ejemplos el tipo de retos y fenómenos a los que nos tendremos que enfrentar.

Las soluciones léxicas (es decir, basadas sólo en palabras y sus frecuencias) no son siempre aplicables al análisis de opinión. En una clasificación temática, por ejemplo, podríamos decidir que un documento trata sobre economía si encontramos términos como “beneficios”, “inflación” o “Banco Central Europeo”.

Pero como vemos en el segundo ejemplo, la presencia de las palabras “favorita” y “bien” no asegura que una frase exprese una opinión positiva.

En esta ocasión, el cambio de sentido se debe a la ironía, uno de los fenómenos más difíciles de tratar de forma automática.

Si antes vimos que con palabras de opinión positivas como “favorita” se podían expresar opiniones negativas, en este ejemplo vemos otro fenómeno curioso: incluso con palabras aparentemente no cargadas de opinión se pueden expresar opiniones.

El hecho de afirmar sobre una cámara que es de plástico, seguramente implicará que pensamos que

no es de buena calidad.

A este tipo de expresiones se les denomina polar facts y son muy difíciles de identificar mediante sistemas que se apoyen en listas de palabras, a priori, cargadas de opinión.

Otro problema es la dependencia del dominio. Como se ve en el ejemplo, una palabra que para un determinado producto supone algo positivo, para otro producto puede ser negativa.

La solución a este problema pasa por la construcción de sistemas entrenados para dominios específicos, que capturen la semántica particular de las palabras en cada dominio.

Lo comentado anteriormente para distintos dominios, es aplicable en ocasiones incluso para objetos de un mismo dominio.

Esto puede ocurrir cuando se analizan textos sobre entidades u objetos de los que nos interesan muchas características. En estos casos, puede que incluso cada característica pueda configurar una carga de opinión distinta para ciertas palabras.

La negación es el mecanismo por excelencia para el cambio del significado de las palabras.

Al igual que la ironía, la negación cambia la polaridad de las palabras de opinión.

El tratamiento de la negación suele ser más abordable. Aunque como vemos en el segundo ejemplo, no siempre es trivial encontrar palabras que nos den pistas de que se está aplicando una negación.

Otra dificultad del análisis de la opinión es cerciorarse de que expresiones de opinión que encontremos en un texto se refieren a la entidad que nos interesa.

En los ejemplos de esta transparencias vemos casos de expresiones claramente negativas que no se corresponden con una opinión negativa sobre la entidad objeto de análisis.

Las formas para expresar opiniones son innumerables. Como vemos en este ejemplo, hablando bien de un libro cuando comentamos una película puede transmitir la sensación de que la película no ha llegado a la altura y que por tanto nos parece mala.

A la hora de desarrollar un sistema de análisis de opiniones, viene bien tener claros los límites de lo abordable y ser conscientes de que siempre habrá opiniones que se nos escaparán.

Hay expresiones tan complejas y que requieren un conocimiento tan profundo del significado de los textos que, por ahora, están fuera de nuestro alcance.

Los textos que encontramos en medios sociales muchas veces son escritos por personas para las que la ortografía y las reglas sintácticas no son su prioridad.

Si a eso se une que es probable que sean escritos desde dispositivos poco cómodos para escribir (como teléfonos o tablets) nos encontramos con una calidad gramatical bastante baja.

Esto supone una dificultad añadida que hace que haya que adaptar, o que directamente no sean aplicables, muchas de las herramientas de procesamiento lingüístico disponibles.

A pesar de las dificultades, el análisis de opiniones es una tarea que tiene un indudable interés comercial y académico.

Mucho del conocimiento y las herramientas del procesamiento del lenguaje natural son aplicables, aunque en ocasiones requerirán de una adaptación a las particularidades propias de los textos de opinión.

Referencias

Estas son las fuentes en las que me he basado para confeccionar esta lección:

- Opinion Mining and Sentiment Analysis, Bo Pang, Lillian Lee; Journal Foundations and Trends in Information Retrieval archive Volume 2 Issue 1-2, January 2008, Pages 1-135, Now Publishers Inc. Hanover, MA, USA doi>10.1561/1500000011
- Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, Peter D. Turney; Proceeding ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics Pages 417-424 Association for Computational Linguistics Stroudsburg, PA, USA ©2002 doi>10.3115/1073083.1073153
- SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining Andrea Esuli Stefano Baccianella, and Fabrizio Sebastiani. Proceedings of the Seventh conference on International Language Resources and Evaluation LREC'10, Valletta, Malta, European Language Resources Association ELRA, (May 2010)