PLN aplicado a internet; en busca de la subjetividad y valoración automática de los textos

NLP techniques & the Internet: Searching for opinions and automatic sentiments analysis

Jorge García Betanzos
Sarenet S.A.
Parque Tecnológico de Zamudio, edificio 103
jorge.garcia@sarenet.es

Resumen: Las redes sociales han cambiado por completo las técnicas de análisis y rastreo de información en Internet. El análisis automático de los textos tratando de conseguir la opinión y la valoración de los mismos es el objetivo a futuro de muchas empresas.
Palabras clave: Sentimiento, automático, internet, subjetividad, polaridad, PLN,...

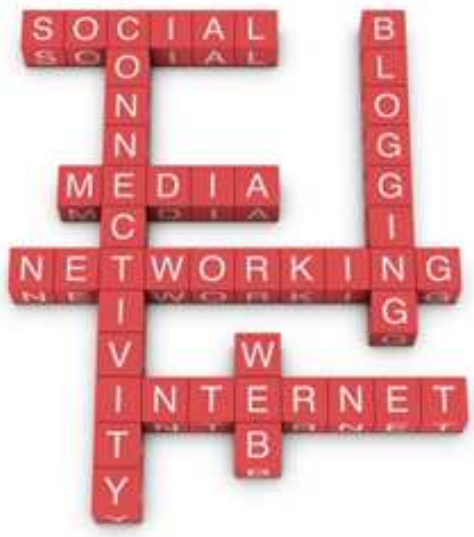
Abstract: The rise of social media networks has entirely transformed the searching and analysis techniques used until today. Many companies aim at analyzing the opinions found on the Internet therefore shaping the future of Sentiment Analysis Tools.
Keywords: Sentiment, automatic, internet, subjectivity, polarity, NLP,…

1 Esquemas pasados, esquemas presentes

Tradicionalmente, en los medios de comunicación escritos, la opinión descubierta y mostrada sin tapujos siempre se ha ubicado en unas pocas páginas, las de opinión.

Dichas opiniones y editoriales siempre se veían condicionadas por la línea política de cada uno de los medios y estaban (y siguen estando) muy focalizadas en aspectos plenamente políticos. Ahora, el papel focalizador de todas esas opiniones está en las redes sociales, en los blogs y en los foros. Se ha ganado en diversidad de formatos y con ello se han incrementado las dificultades técnicas para procesar una ingente cantidad de información.

Otro gran cambio ha venido dado por la importancia que ahora tiene la marca, el producto o el servicio dentro de estos medios. Mientras que las opiniones de los medios tradicionales versan sobre aspectos más puramente políticos, en las redes sociales la marca es importante, en muchas ocasiones, es la clave. Se focaliza sobre una empresa, un servicio concreto e incluso un producto. Estas opiniones son vitales para una empresa. Pueden marcar líneas a seguir en la preventa, producción y atención sobre un producto o servicio. También pueden indicar fallos en las políticas de comunicación o de producción de una compañía. Por tanto, tienen que conocer estas opiniones, escucharlas y valorarlas.



Hasta ahora, en Iconoce, nos hemos centrado en sacar lo que nos demandaban los clientes, la localización de información de actualidad. Si alguien quería obtener la opinión, lo teníamos fácil puesto que ya teníamos segmentada la información en las secciones clásicas de los periódicos. Hoy en día, todo es mucho más heterogéneo, la información está mucho más dispersa y el volumen de datos a tratar es infinitamente superior.

Asimismo, es imperiosa la necesidad de ir un par de pasos más allá, diferenciarse de la competencia y desgranar los resultados obtenidos. Esto pasa por valorar automáticamente los resultados, sacar el sentido y la subjetividad de los textos.

Para conseguir alcanzar este objetivo la aplicación de técnicas de Procesamiento del Lenguaje Natural nos parecen claves. Los objetivos están claros y se podrían desglosar en tres apartados:

- Obtener automáticamente las opiniones y su polaridad positiva, negativa o neutra
- Obtener una relación de conceptos vinculados a una marca, producto o servicio
- Obtener una relación de protagonistas vinculados a una marca, producto o servicio

Además, cuando tratamos de analizar el sentimiento o polaridad de las opiniones sobre una marca, producto o servicio también tenemos que ir un poco más allá de la individualidad. Una opinión aislada vale lo que vale y deber ser analizada en función del interlocutor que la señale. Pero lo realmente importante, lo que buscan y nos solicitan las empresas es poder establecer tendencias, gráficas a corto, medio y largo plazo, opiniones vinculadas a entidades o conceptos.

2 Combinar rapidez y eficacia

Las soluciones de extracción automática de opinión, conceptos y entidades se pueden diseñar tan complejas como queramos: algoritmos, métodos de ponderación, búsquedas cruzadas de diccionarios, etc. Dicha complejidad, ya de por si inherente a estos sistemas, se enfrenta adicionalmente a dos problemas a la hora de aplicarlos en Internet:

1. Combinar un análisis automatizado lo más preciso posible con una necesidad de rapidez de ejecución. La empresa necesita ser rápida, ágil y tenerlo antes que la competencia para poder tomar decisiones. Cuanto más se aproximen los resultados al tiempo real, mejor. Y para ello, cuanto más automático sea el sistema mejor.

2. Afrontar el análisis de los nuevos métodos de escritura. "*dnde kdamos?*", "*ste tfno s 1mierDDDa*". Ninguna máquina lo puede entender, por ahora…

Las opciones de escribir mal son más amplias que las de escribir bien. Y sin embargo cualquier persona entiende esto:

> *Sgeun etsduios raleziaods por una Uivenrsdiad Ignlsea,no ipmotra el odren en el que las ltears etsen ecsritas,la uicna csoa iporntnate es que la pmrirea y la utlima ltera esetn ecsritas en la psiocion cocrreta.*
> *El retso peuden etsar ttaolmntee mal y aun pordas lerelo sin pobrleams, pquore no lemeos cada ltera en si msima snio cdaa paalbra en un contxetso.*

3 Estructuras cambiantes

No solo nos encontramos con los problemas antes mencionados. La situación de la web actual permite múltiples posibilidades de interacción. No vale con sacar “me gusta este producto”, ahora también hay que localizar si alguien tiene el icono del “me gusta” de Facebook. El abanico de posibilidades ya no es tan reglado como lo era hasta hace un par de años.

Siendo sinceros no podemos abarcar previamente todas las opciones disponibles de estructuración de páginas web porque, además

de ser infinitas, son cambiantes. Asimismo, lo que hoy es fundamental conocer (*Tweets*) mañana puede que haya caído en el olvido (¿*MySpace*?). Por tanto se hace importante establecer un método de trabajo diferente, en el que se deje el modelo muy abierto para tratar todos los casos particulares y tratar de aplicarlos a casuísticas generales.

Por ejemplo, los grandes *crawlers* de internet siempre hemos absorbido toda la información para nuestras bases de datos adquiriendo y tratando el código fuente original de las páginas a las que lanzábamos nuestros *spiders*. Pero ahora, cada vez es más común que ciertos aspectos de gran interés, como pueden ser los comentarios de las noticias, se saquen mediante AJAX y no estén directamente publicados en el código fuente de la página web. Hay que hacer procesos especiales para tratar de obtener estos contenidos.

Adicionalmente, hoy en día, el mayor reto reside no solo en sacar la información vinculada a resultados textuales, también debemos ser capaces de aplicar dichos sistemas de PLN al reconocimiento automático de imágenes, audio y vídeo.

4 Carta a los Reyes Magos. Buscando el sistema ideal

Actualmente en Sarenet estamos trabajando en desarrollar un sistema que:

- Recoja todas las apariciones de una determinada marca, servicio o producto filtrando los "ruidos", las apariciones no deseadas.
- Clasifique automáticamente la aparición con los conceptos a los que está vinculado.
- Clasifique automáticamente la aparición con los protagonistas (personas y entidades) a los que está vinculado.
- Marque automáticamente la polaridad positiva, negativa o neutra de dicha aparición y el porqué de esa clasificación.
- Sea adaptable a cualquier marca, producto o servicio. No podemos establecer unos diccionarios vinculados a un sector, debemos ser capaces de adaptarnos a todos y acaparar automáticamente su jerga particular.
- Sea adaptable a los diferentes idiomas que queramos abarcar, catalán, gallego, valenciano y euskera.

Pueda analizar textos, imágenes, audios y vídeos.

- Que sea rápido.
- Que se amolde a las nuevas formas de escritura.
- Que entienda la ironía ;-)