instituto de ingeniería del conocimiento

LYNGUO

Análisis de opinión y contenido en los medios sociales

Autor: Antonio Moreno Sandoval

Investigador senior del IIC en Social Business Analytics
antonio.msandoval@iic.uam.es

www.iic.uam.es

Análisis de opinión y contenido en los medios sociales

Abstract

Dentro del campo del Sentiment Analysis donde se miden aspectos como la opinión, intención, emoción o concienciación de los usuarios de las redes sociales, el análisis automático de los contenidos en estas redes tiene como objeto principal conocer la opinión de los usuarios acerca de productos, servicios, marcas, personas e instituciones. Las nuevas plataformas de expresión permiten a los usuarios bien expresar de manera inmediata valoraciones espontáneas que reflejan intenciones futuras, o bien transmitir simplemente su posición hacia un tema concreto. Trataremos la monitorización de estas expresiones y los principales aspectos cubiertos en el análisis de opinión, así como la evolución sufrida por las distintas generaciones de estas aplicaciones para conseguir proporcionar análisis más sofisticados y con un mayor nivel de cobertura en toda la red.

Palabras clave:

Redes sociales, monitorización, opinión, análisis, Sentiment Analysis.

Autor: *Antonio Moreno Sandoval*

Investigador senior del IIC en Social Business Analytics
antonio.msandoval@iic.uam.es

instituto de ingeniería del conocimiento

LYNGUO

Análisis de opinión y contenido en los medios sociales

instituto de ingeniería del conocimiento

Autor: Antonio Moreno Sandoval

Introducción

El análisis de opinión, sentimiento y contenido (Sentiment Analysis) es un área de investigación del **Procesamiento del Lenguaje Natural** que cuenta con más de 15 años de desarrollo. Los primeros sistemas y modelos están recogidos en la monografía de Pang y Lee de 2008, fecha en la que se producen también las primeras aplicaciones comerciales. El objetivo es conocer lo que opinan los usuarios de las redes sociales sobre productos, servicios, marcas, personas e instituciones. El método tradicional para conocer la opinión de los usuarios son las encuestas y los estudios de mercado.

La diferencia esencial de la información vertida en las redes frente a la obtenida por encuestas y estudios de mercado es su inmediatez. Esta información es espontánea y poco estructurada, y refleja lo que opinan los usuarios, no los expertos. En muchas ocasiones, la red se utiliza para difundir quejas o recomendaciones que llegan a afectar directamente a la imagen corporativa. Las compañías y agencias de comunicación entienden el valor de esta información para sus estrategias comerciales, la atención al cliente y la detección de tendencias. Además, las organizaciones que contratan empresas de comunicación quieren conocer el impacto social de sus noticias y campañas para valorar la estrategia comercial.

Hay dos tipos de aplicaciones comerciales que monitorizan los medios sociales: las que se centran principalmente en la red de usuarios y las que se enfocan hacia el contenido de los mensajes:

- **Red de usuarios**: Proporcionan análisis con métricas de la presencia de una determinada marca y sus competidores, los usuarios más influyentes y los mensajes más difundidos.
- **Contenido de los mensajes**: Ofrecen análisis con polaridad (positivo o negativo) de la opinión, clasificación por categorías temáticas y detección de palabras clave.

La primera generación de herramientas que trataban la red de usuarios y el contenido de los mensajes estaba compuesta por **aplicaciones independientes** que se centraban solo en uno de estos dos aspectos. En un primer momento, los sistemas de análisis de la red de usuarios proporcionaban una rica información sobre los usuarios y sus relaciones, aunque el análisis semántico era muy limitado. Por otra parte, los sistemas de análisis de contenido apenas incorporaban información sobre análisis de la red. Pronto se vio que ambas aproximaciones eran incompletas.

La evolución de estas aplicaciones ha terminado por incorporar progresivamente el aspecto complementario, al tiempo que se ha sofisticado el tratamiento, tanto de la presencia como del contenido. Las últimas versiones de esta segunda generación de aplicaciones incluyen además presentaciones gráficas mejoradas en forma de *dashboards* (paneles de control) y herramientas para generar informes, dos funcionalidades muy apreciadas por los responsables de comunicación. Otras mejoras añadidas son la configuración de alertas para la rápida detección de crisis de marca y las distintas parametrizaciones de las búsquedas, para adaptar la herramienta a las nuevas necesidades.

La parte más compleja de este proceso es la semántica, que está evolucionando de la medición de mensajes positivos, negativos y neutros a la detección de tendencias de compra o de temas relevantes para los medios de comunicación. Para ello, a veces se recurre al análisis del perfil del usuario, que incluye un análisis sociológico y demográfico (sexo, edad, profesión, lugar). Por supuesto, todo análisis se hace en tiempo real, para que la empresa pueda reaccionar inmediatamente a los mensajes.

Los desafíos de las redes sociales

El **Análisis de contenido en redes sociales** forma parte del proceso de recuperación de información. Por lo tanto, el acierto de la aplicación está sujeto a la evaluación tanto de la cobertura como de la precisión de dicho análisis.

- La **cobertura** calcula el número de menciones relevantes a una marca.
- La **precisión** mide el acierto en el análisis del contenido, ya sea mediante una valoración de positivo/negativo, una clasificación temática o una detección de palabras clave.

La **cobertura** depende básicamente del número de canales que se monitoricen y de la capacidad para reunir las menciones a la marca. Algunas aplicaciones solo se centran en unos pocos canales (Twitter o Facebook), mientras que otras recuperan menciones también en blogs, foros y RSS. Naturalmente, depende de las necesidades del cliente, pero una mayor cobertura es siempre mucho más costosa y puede suponer una fuente de «ruido» (información no relevante para la marca seleccionada y, por tanto, inservible). Es el precio a pagar por estar al tanto de (casi) todo lo que se comenta en Internet.

Mejorar la cobertura implica seleccionar los mensajes apropiados para búsquedas ambiguas. Por ejemplo, la palabra «Santillana» puede referirse a una ciudad o calle, a una editorial, a un jugador de fútbol o incluso a un poeta (el Marqués de Santillana). La sinonimia (conceptos, marcas o personas con el mismo nombre) es dependiente en cada caso de la consulta y es el principal factor de ruido. Para reducir el número de respuestas no relevantes a la consulta hay que introducir filtros en la búsqueda. Estos filtros los pueden crear los usuarios de la aplicación pero requieren, en algunos casos, mucha destreza en el manejo de la herramienta (de hecho, muchos usuarios se quejan de la complejidad y la lenta curva de aprendizaje en la definición de filtros).

La evaluación de la **precisión** también tiene matices, pues no es lo mismo clasificar los mensajes en positivos o negativos, que proporcionar una valoración numérica dentro de una escala. A esto hay que añadir que los mensajes pueden ser ambiguos y tener distintas interpretaciones, lo que dificulta su valoración.

Factores lingüísticos como la ironía, usos figurados del lenguaje (como la metáfora) o el mismo contexto interactivo de muchas conversaciones en red hacen que el acierto interpretativo se complique. Por ejemplo, un mensaje como «He tenido un día fantástico: mi 'modelo de coche X' me ha dejado justo donde quería, en medio de la M-30.» no es en absoluto positivo para la marca del coche. Sin embargo, las palabras y expresiones utilizadas en el comentario son positivas: «día fantástico», «justo donde quería». Es el contexto («en medio de la M-30») y el conocimiento del mundo lo que nos permite interpretar la ironía: nadie está contento cuando el coche se le estropea en una vía de circunvalación saturada. Sin embargo, los sistemas actuales de procesamiento del lenguaje natural son incapaces de tratar de manera fiable dicha ironía.

Para paliar estos problemas de precisión se debe mejorar el **motor de análisis semántico**. El significado de los mensajes se distribuye en tres niveles: léxico, oracional y discursivo. Las palabras se encuentran en el primer nivel. Podemos clasificar las palabras en función de su contenido positivo o negativo, y obtener una valoración general del mensaje. Sin embargo, las palabras se combinan entre sí en sintagmas, que a su vez forman oraciones. Esas combinaciones estructurales modifican, restringen e incluso cambian el significado global del mensaje respecto al contenido individual de cada palabra: el todo es más que la suma de las partes. Por ejemplo, la negación puede cambiar el signo de polaridad de un mensaje. «El servicio es malo» (mensaje negativo) frente a «El servicio **no** es **nada** malo» (mensaje positivo). Finalmente, los mensajes están dentro de una conversación y, por lo tanto, hay información que se ha dicho previamente y que no tiene por qué ser mencionada de nuevo en cada mensaje. Esta información elidida suele ser fundamental para interpretar el mensaje.

En resumen, para entender el contenido se debe realizar un análisis en los sucesivos niveles. El **grado de acierto** en la interpretación depende de forma decisiva de contar con módulos especializados que traten todos los aspectos semánticos.

La tecnología subyacente a la primera generación

En Análisis de la opinión se dan dos paradigmas (Zhang et ál., 2011): uno basado en recursos léxicos (diccionarios de polaridad) y otro en técnicas de aprendizaje automático.

Las **técnicas basadas en diccionarios** centran su estrategia en buscar en los textos palabras que tengan asignada alguna polaridad y en realizar una métrica de lo encontrado. Su principal limitación es que no se analizan aquellos mensajes que no contengan las palabras de los diccionarios. Por lo general, estos sistemas tienen baja cobertura y, además, mantener los diccionarios es costoso, pues solo pueden realizar dicho mantenimiento los lingüistas que lo han desarrollado, y no los usuarios de la aplicación interesados en la opinión sobre su producto, marca, etc.

Las **aplicaciones basadas en aprendizaje automático** son más populares: los modelos están entrenados con ejemplos analizados por personas que anotan su valor positivo/negativo. La limitación de este enfoque es la falta de datos etiquetados. En concreto, cuando se cambia de dominio o de canal, el lenguaje empleado puede ser muy diferente, y el modelo previo no recoge información relevante, lo que afecta tanto a la cobertura como a la precisión. Por ejemplo, una herramienta entrenada con textos recogidos en un blog profesional sobre automóviles no sirve para analizar textos de Twitter sobre telefonía móvil. Algunas aplicaciones proponen al usuario que entrene él mismo su propio modelo con una serie de anotaciones sobre ejemplos del nuevo dominio. Pero esto puede convertirse en una tarea iterativa sin fin y hacer trabajar al cliente no está muy bien considerado. Por supuesto, se han intentado aproximaciones híbridas que combinan diccionarios con aprendizaje automático, pero su gran limitación ha sido ya expuesta: estos enfoques solo tratan la semántica léxica, a nivel de las palabras.

Los textos presentan también estructura sintáctica y discursiva, que tienen un papel crucial en la interpretación final del contenido. Por lo tanto, se necesitan módulos de análisis gramatical y discursivo para recuperar todo el significado.

Veámoslo con ejemplos. Emplearemos distintos colores para analizar la información relevante: en negro, el tema del que se habla; en rojo, la palabra negativa; y, en verde, la palabra o locución positiva.

> El *motor* es *escaso de potencia,*
> aunque el *precio* es *muy asequible*.

¿Qué puntuación se podría otorgar a este comentario? Para empezar, no es lo mismo asignar una puntuación global (por ejemplo, neutro) que una dividida por segmentos:

- *motor (potencia):* valoración positiva (+)
- *precio:* valoración muy positiva (++)

Si adoptamos una estrategia puramente léxica, tenemos que eliminar las palabras no marcadas y quedarnos con:

> *motor escaso de potencia, precio muy asequible*.

¿Qué puntuación obtendríamos en este caso?

- *Escaso:* valoración negativa (-)
- *Muy asequible:* valoración muy positiva (++)

Como se puede comprobar, la estructura sintáctica es determinante en la interpretación, pues nos indica qué elementos están afectados por las palabras con polaridad. Los corchetes nos muestran las agrupaciones sintácticas (dos oraciones simples coordinadas por «aunque»):

> [El *motor* es *escaso de potencia*],
> [aunque el *precio* es *muy asequible*].

Este sencillo análisis nos permite dar una valoración acertada como la mostrada arriba. Si solo se empleara una tecnología basada en diccionarios o en aprendizaje automático, sería difícil llegar a este nivel de precisión y detalle. Obtendríamos solo una valoración global (por ejemplo, «neutro»), que es la que encontramos en la mayoría de los sistemas de análisis de opinión de primera generación —aquellos que solo hacían análisis léxico.

Para casos más complejos, como el uso de palabras de negación (*no, nada, nunca)* o construcciones comparativas («Me gusta más el acabado de X que el de Y»), la interpretación es sencillamente incorrecta si no se realiza un análisis gramatical previo que determine las entidades sobre las que se expresa la opinión. Los problemas planteados dejan constancia de la imperiosa necesidad de una segunda generación de aplicaciones que realice un análisis basado en el **procesamiento semántico oracional y discursivo** y que incluya aportaciones tanto del análisis de la red de usuarios como del contenido.

Una nueva generación de aplicaciones

En la actualidad hay una clara conciencia de que los sistemas de análisis de opinión de primera generación han llegado a un punto en el que no hay mejoras posibles. No es cuestión de permitir al usuario entrenar nuevos datos o incluir entradas nuevas al diccionario: sencillamente, hay que abordar el procesamiento semántico oracional y discursivo. En otras palabras, hay que avanzar hacia una **tecnología semántica profunda**, no centrada en las palabras, sino en estructuras superiores.

Además, hay que añadir otro aspecto esencial: el tratamiento avanzado de los **dominios temáticos**. Por una parte, es imprescindible seguir construyendo diccionarios **más completos y actualizados** (continuamente se están incorporando palabras nuevas o expresiones en las redes sociales). Por otra parte, es necesario restringir en las búsquedas los sinónimos que no interesan. Ambas tareas requieren el conocimiento especializado de los lingüistas. Por tanto, estas aplicaciones tienen que seguir invirtiendo en tecnología lingüística contrastada. La actualización continua del componente léxico es imprescindible.

Al mismo tiempo se empieza a perfilar la tercera generación de este tipo de aplicaciones: las nuevas versiones incorporarán métricas más avanzadas de la presencia, podrán generar gráficos e informes y permitirán añadir filtros de alertas para situaciones de emergencia comunicativa.

Ha empezado la nueva era del Análisis de contenidos en la red.

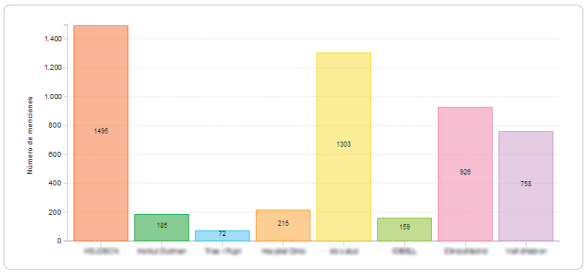
Lynguo: la solución para el análisis de contenidos en español

El Instituto de Ingeniería del Conocimiento ha desarrollado una aplicación específica de tercera generación para el procesamiento del español en las redes sociales: Lynguo, que combina **conocimiento lingüístico** profundo con **algoritmos de monitorización** y **análisis de *Big Data*.** El equipo de lingüistas del IIC adapta los diccionarios al dominio del cliente, interaccionando con él para conocer los elementos clave de su negocio. Los ingenieros informáticos proporcionan el acceso a la información en tiempo real, consiguiendo una cobertura casi total de las menciones para una determinada marca.

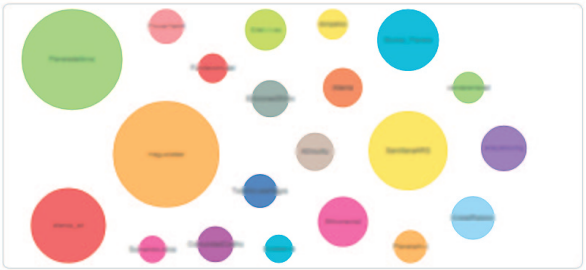
La capacidad analítica de Lynguo comienza con una **óptima monitorización**. Los expertos lingüistas configuran las consultas y los filtros para conseguir la máxima cobertura minimizando el ruido. Esta tarea, especializada y tediosa, no se deja bajo la responsabilidad del cliente; aunque, por supuesto, este tiene la posibilidad de configurar algunos filtros para palabras y usuarios prohibidos, y de definir los usuarios oficiales y las nuevas entidades relacionadas con la marca.

Lynguo aporta diferentes funcionalidades distribuidas en varios módulos:

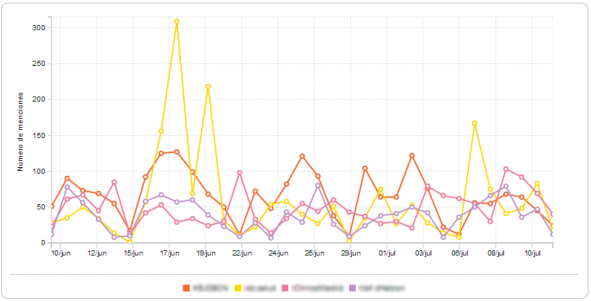
Lynguo Observer realiza un análisis completo de la presencia de la marca, basado en el recuento de menciones, usuarios, temas, usuarios más activos, comentarios más difundidos, así como de la evolución de estas variables en el tiempo. También recurre a la geolocalización para mostrar los comentarios distribuidos por país, y, dentro de España, por comunidades autónomas.



Resumen de la presencia

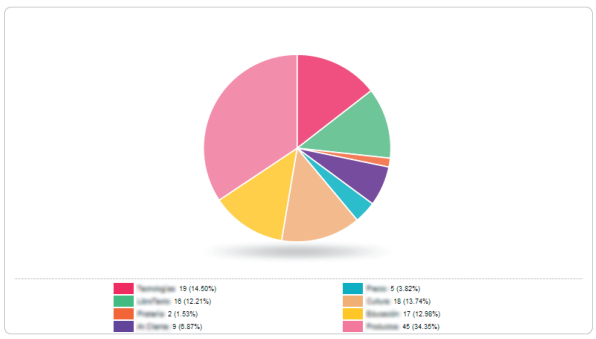


Usuarios más activos

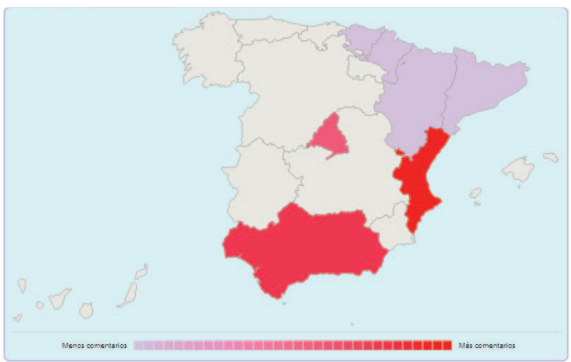


Evolución de la presencia

Comentarios más difundidos



Desglose temático

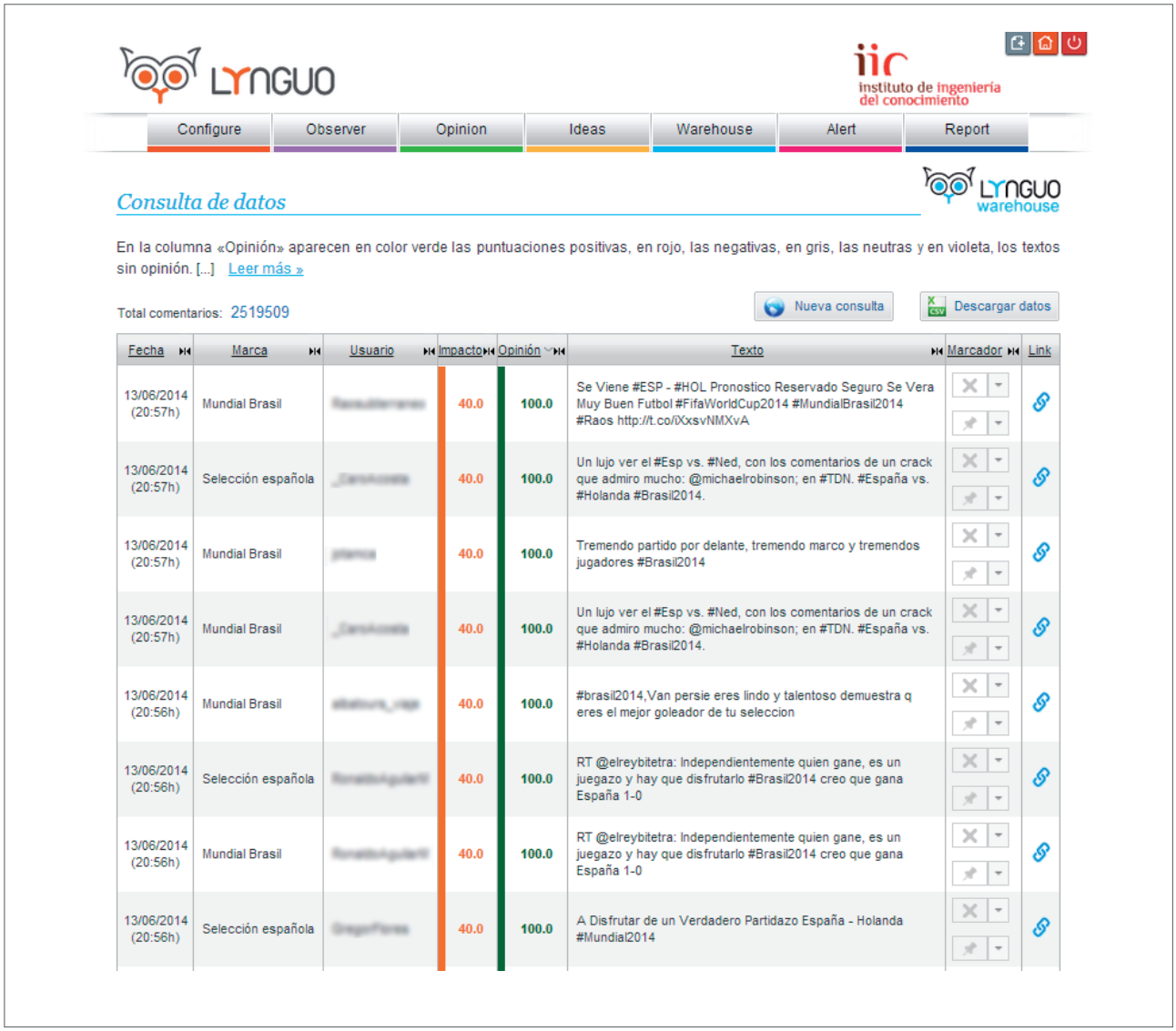


Presencia por país

Se han añadido tres nuevas métricas para mejorar el análisis de la presencia: *Share of voice* (que muestra el porcentaje de conversación que corresponde a cada marca, evento o entidad); *Share of users* (que muestra el porcentaje de usuarios que comentan sobre una marca, entidad o evento); y *Contenidos más virales* (que muestra los contenidos más difundidos). Estas métricas permiten ver la cuota de presencia en la red y qué contenidos externos están asociados a la marca.

Lynguo Opinion se centra en el análisis de la opinión de los comentarios. Se divide en cuatro secciones: *Resumen de la opinión* (un gráfico general que muestra la proporción de mensajes positivos, negativos y neutros para la marca y sus competidores); *División por marca* (distribución de la opinión en cada marca); *Evolución de la opinión* (representación temporal de la opinión por marcas) y el *Desglose de la opinión* (por marca y por área temática).

Desde cada uno de estos módulos se puede acceder al detalle de cada mención en *Lynguo Warehouse*, que permite consultar los datos con todos sus atributos, como el usuario, su impacto y el acceso a la fuente original en Twitter, Facebook y blogs.



Fecha	Marca	Usuario	Impacto	Opinión	Texto	Marcador	Link
13/06/2014 (20:57h)	Mundial Brasil		40.0	100.0	Se Viene #ESP - #HOL Pronostico Reservado Seguro Se Vera Muy Buen Futbol #FifaWorldCup2014 #MundialBrasil2014 #Raos http://t.co/iXxsvNMXvA		
13/06/2014 (20:57h)	Selección española		40.0	100.0	Un lujo ver el #Esp vs. #Ned, con los comentarios de un crack que admiro mucho: @michaelrobinson; en #TDN. #España vs. #Holanda #Brasil2014.		
13/06/2014 (20:57h)	Mundial Brasil		40.0	100.0	Tremendo partido por delante, tremendo marco y tremendos jugadores #Brasil2014		
13/06/2014 (20:57h)	Mundial Brasil		40.0	100.0	Un lujo ver el #Esp vs. #Ned, con los comentarios de un crack que admiro mucho: @michaelrobinson; en #TDN. #España vs. #Holanda #Brasil2014.		
13/06/2014 (20:56h)	Mundial Brasil		40.0	100.0	#brasil2014,Van persie eres lindo y talentoso demuestra q eres el mejor goleador de tu seleccion		
13/06/2014 (20:56h)	Selección española		40.0	100.0	RT @elreybitetra: Independientemente quien gane, es un juegazo y hay que disfrutarlo #Brasil2014 creo que gana España 1-0		
13/06/2014 (20:56h)	Mundial Brasil		40.0	100.0	RT @elreybitetra: Independientemente quien gane, es un juegazo y hay que disfrutarlo #Brasil2014 creo que gana España 1-0		
13/06/2014 (20:56h)	Selección española		40.0	100.0	A Disfrutar de un Verdadero Partidazo España - Holanda #Mundial2014		

El análisis del contenido semántico se completa con *Lynguo Ideas*. Este módulo muestra las palabras que se asocian con las diferentes marcas, con la indicación de su valor de distintividad. Esta funcionalidad permite conocer aquellas palabras o expresiones idiomáticas representativas de una marca en comparación con su competencia, bien porque sean exclusivas de los mensajes sobre la marca o porque tengan una aparición mucho mayor que en los otros competidores.

Lynguo Alert permite configurar alertas sobre comentarios que tengan un impacto o interés potencial para el cliente. Además, se pueden establecer alertas por volumen cuando se producen situaciones excepcionales por acumulación de comentarios con ciertas características comunes.

Para completar las funcionalidades, Lynguo dispone de una amplia variedad de gráficos y un generador de informes. Destacamos el *dashboard* (panel de control) al inicio de la aplicación: muestra en una sola pantalla *Comentarios, Share of Voice, Share of Users, Opinión, Nube de Ideas, Contenidos más virales y Desglose temático,* accesibles en tres opciones temporales: 24 horas, 7 días y 1 mes.

Finalmente, el módulo *Lynguo Report* ofrece la posibilidad de generar informes personalizados a partir de la información proporcionada por los módulos de Lynguo. El cliente tiene a su disposición una serie de gráficas en función de la naturaleza de los datos para poder generar informes y gráficos y descargar datos sin la necesidad de usar una aplicación distinta, ya que esta funcionalidad está plenamente integrada en Lynguo.

Conclusiones

La tercera generación de aplicaciones se centra en un análisis semántico más profundo del contenido de los mensajes vertidos en las redes sociales. No solo valora una opinión vertida en el mensaje como positiva o negativa, sino que también apuesta por la detección de tendencias y temas. Las nubes de palabras se hacen más ricas, agrupando términos formados por más de una palabra y relacionándolos con entidades y personas. Se trata de ayudar a descubrir el significado oculto mediante las interrelaciones conceptuales. Es decir, abarca más redes y analiza mejor su contenido. En definitiva, mejora la cobertura.

Al mismo tiempo, la última tecnología semántica afina las valoraciones para eliminar mensajes irrelevantes o errores de interpretación, es decir, contribuye a una mejora de la precisión.

Todo ello debe ir integrado en una herramienta como Lynguo: **una aplicación sencilla de usar** que proporcione acceso directo a los mensajes y permita generar informes gráficos atractivos. En definitiva, una aplicación que permita a sus usuarios gestionar con fiabilidad las redes sociales para localizar líderes de opinión, reorientar campañas, ver las ideas más difundidas o detectar crisis en tiempo real.

Agradecimientos

Gonzalo Martínez y Marta Guerrero han leído versiones previas de este trabajo y agradezco sus comentarios y sugerencias al texto.

Referencias

- Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*. 1-2 (2), 1–135. ACM Digital Library. Hanover, MA, USA. (2014) http://dl.acm.org/citation.cfm?id=1454712

- Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis. *Hewlett-Packard Laboratories. Technical Report HPL-2011. (89). (2014)* http://www.hpl.hp.com/techreports/2011/HPL-2011-89.pdf

- Liu, B. (2012): *Sentiment Analysis and Opinion Mining*. 1 (5), 1-167. Morgan and Claypool. Chicago, USA. (2014) http://www.morganclaypool.com/doi/abs/10.2200/s00416ed1v01y201204hlt016

- Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *Knowledge-based Approaches to Concept-level Sentiment Analysis*. 2 (29) IEEE. (2014) http://www.computer.org/intelligent

©ADIC

Siguenos en:

C/ Francisco Tomás y Valiente, nº 11
EPS, edificio B, 5ª planta
UAM Cantoblanco
28049 Madrid, España.

Tel.: (+34) 91 497 2323
Fax: (+34) 91 497 2334
iic@iic.uam.es
www.iic.uam.es