Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 177-180 recibido 21-03-2018 revisado 13-04-2018 aceptado 03-05-2018

# Monitorización de Social Media

## *Social Media Monitoring*

**Rosa Montañés, Rocío Aznar, Saúl Nogueras, Paula Segura, Rubén Langarita, Enrique Meléndez, Paula Peña, Rafael del Hoyo**
Grupo de Big Data y Sistemas Cognitivos
ITAINNOVA (Instituto Tecnológico de Aragón)
C/ María de Luna, nº 7. 50018 Zaragoza
{rmontanes,raznar,snogueras,psegura,rlangarita,emelendez,ppena,rdelhoyo}@itainnova.es

**Resumen:** El sistema desarrollado tiene como objetivo la integración y monitorización de la información en castellano de las redes sociales de un usuario (Facebook, Twitter y noticias web de interés) a través de una única aplicación web. El sistema se sustenta en tres componentes principales: un módulo que implementa una gran variedad de tareas de Procesamiento del Lenguaje Natural (PLN), un módulo software de recuperación de datos de redes sociales mediante crawlers y almacenamiento de resultados, y una aplicación web que presenta una interfaz de usuario para la visualización de la información de forma sugestiva e interactiva. De esta forma, la solución propuesta permite a los usuarios estar actualizados y tener un control de sus redes sociales, pudiendo estar al día de la información, tanto de sus publicaciones como de sus intereses, en una única interfaz sencilla e intuitiva.
**Palabras clave:** Redes sociales, crawlers, PLN, interfaz de usuario

**Abstract:** The developed system aims to integrate and monitor information in Spanish of a user's social networks (Facebook, Twitter and web news of interest) through a single web application. It is based on three main components: a module that implements a wide variety of Natural Language Processing tasks (NLP), an information retrieval module which capture social networks data by means of crawling and stores processing results, and an application web that presents a user interface through which visualizing the information obtained in a suggestive and interactive way. Therefore, the proposed solution allows users to be updated and control their social media networks, being to able to be up-to-date about the information of their publications and their interests, in a single, simple and intuitive graphical interface.
**Keywords:** Social networks, crawlers, NLP, user interface

### *1 Introducción*

Hoy en día, el uso de redes sociales (Twitter, Facebook, blogs, etc.) está ampliamente extendido en todos los ámbitos de la sociedad. Los usuarios, sea cual sea su perfil, comparten gran cantidad de información multimedia en la red, especialmente información escrita, lo que implica la generación de datos textuales de forma masiva. Esto ha llevado en los últimos años al estudio y desarrollo de aplicaciones capaces de explotar estos datos disponibles para extraer analíticas y conocimiento implícito de gran valor (He et al., 2015; Batrinca y Treleaven, 2015; Chang, 2017; Stieglitz et al., 2018).

No obstante, esta generación masiva de información conlleva a su vez a que los usuarios, en su motivación por permanecer informados, encuentren problemas a la hora de entender, clasificar y reconocer la información más relevante de su entorno.

Este prototipo pretende solucionar el problema anterior por medio de la monitorización de la actividad en castellano de las redes sociales del usuario, en concreto Facebook, Twitter y noticias de su interés (publicadas a través de RSS, *Really Simple Sindication* en inglés), mediante el desarrollo de una aplicación web interactiva e intuitiva que permita al usuario la visualización y filtrado de la información más relevante de su entorno, apoyándose en el uso de un amplio abanico de técnicas de procesamiento del lenguaje natural (PLN).

En las siguientes secciones se describe en

ISSN 1135-5948. DOI 10.26342/2018-61-26

detalle la metodología seguida, así como las conclusiones del trabajo realizado y posibles líneas de trabajo futuro.

## *2 Sistema de monitorización de Social Media*

El sistema propuesto se estructura sobre tres módulos funcionales: ingesta y almacenamiento de datos (crawler), procesamiento del lenguaje natural y aplicación web. En la

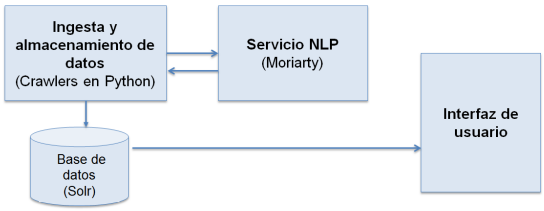


Figura 1: Arquitectura del sistema

Figura 1 se muestra la arquitectura del sistema. En primer lugar, se captura la información mediante crawlers en Python, después se invoca un servicio web a través de su interfaz REST, encargado de realizar el procesado y explotación de la información con la aplicación de técnicas de PLN y, por último, se almacena la información relevante en la base de datos, que es sobre la que se alimenta la interfaz de usuario. El módulo de PLN ha sido implementado en *Moriarty*[1], que es una herramienta de diseño e implementación de soluciones avanzadas de software de Big Data e Inteligencia Artificial desarrollada por ITAINNOVA.

### 2.1 Ingesta y almacenamiento de información

El proceso de ingesta y almacenamiento de información se realiza según una periodicidad configurable y de forma altamente eficiente mediante paralelización. La implementación se compone de tres módulos implementados en Python:

1. **Recuperación de la información**
   Se han creado sendas cuentas de usuario en Facebook y Twitter sobre las que se ha generado actividad variada y se ha recopilado un listado de RSS de interés. La recuperación de la información de las publicaciones realizadas en dichas fuentes se ha realizado mediante crawling. En el caso de Twitter y Facebook haciendo uso de sus APIs públicas (Twitter, 2018; Facebook, 2018), y en el de los RSS mediante consulta directa a la lista configurada. Si las publicaciones recuperadas contienen links de páginas a terceros, el proceso de crawling también extrae el contenido de dichas páginas.

   Como texto origen a analizar, se distingue entre lo que se denomina *perfil* e *interés*. El *interés* hace referencia a las publicaciones de los usuarios a los que se sigue (páginas en caso de Facebook) y el *perfil* a las propias publicaciones del usuario. En el caso de las RSS, todos los documentos se consideran de *interés*.
2. **Procesamiento de la información**
   Una vez se tiene extraído el texto, se preprocesa para eliminar metadatos, como cabeceras, títulos o pies de página para la posterior aplicación de técnicas de PLN. Asímismo, debido a la diversidad de fuentes de información, es necesario realizar un proceso de normalización y estructuración de la información que se recupera.
3. **Almacenamiento de la información**
   Por último, la información resultante de la aplicación de técnicas de PLN se incorpora a la información origen extraída de las fuentes de información y se almacena en una base de datos NoSQL. Se ha elegido el uso de *Apache Solr*, por ser una base de datos orientada a documentos que proporciona funcionalidades avanzadas de indexación y búsqueda, y se ha diseñado un esquema que permite unificar todos los datos de forma sencilla y eficiente.

### 2.2 Procesamiento del Lenguaje Natural

Recuperados y procesados los datos textuales de las publicaciones se hace una invocación a los servicios de PLN, integrados en el framework *Moriarty*.

Puesto que la información de las redes sociales puede aparecer en cualquier idioma y el prototipo se ha enfocado en el castellano, se aplica en primera instancia un algoritmo de detección del languaje que permite filtrar la información.

El servicio de PLN integra una gran variedad de técnicas de PLN dándole un gran po-

[1]"Moriarty". Información disponible en: `http://www.ita.es/moriarty/`


Figura 2: Pantalla principal de la aplicación web

tencial al sistema. En las siguientes secciones se explican las diferentes técnicas utilizadas.

### 2.2.1 Síntesis de la información

Una de las funcionalidades del sistema es la representacion visual de los conceptos más relevantes. Puesto que el dominio de trabajo son las redes sociales, se considera tanto la generación de nube de palabras como la de hashtags. Para la generación de la **nube de palabras** se realiza un preprocesado del texto de forma que se normalice la información y se eliminen conceptos no relevantes. Eliminación de stopwords o lematización son algunas de las técnicas de PLN que son aplicadas para ese propósito.

Además, para ofrecer mayor grado de detalle, el sistema es capaz de generar **resúmenes** de las publicaciones mediante la aplicación de un algoritmo de ranking basado en grafos (Erkan y Radev, 2004) que permite obtener sus frases más significativas.

### 2.2.2 Reconocimiento y clasificación de entidades nombradas

Otra de las funcionalidades es el reconocimiento y clasificación de entidades nombradas (NERC, siglas en inglés) mediante la aplicación de algoritmos basados en redes neuronales (Chiu y Nichols, 2015). Esta tarea permite al usuario tener un conocimiento acerca de las personas, organizaciones y localizaciones a las que hace referencia la publicación.

### 2.2.3 Análisis de sentimiento

Otra de las tareas de PLN que se aplican es el análisis de sentimientos que permite una clasificación de los documentos en diferentes categorías según la opinión que se expone en ellos. Tal como se presenta en la competición TASS (Martınez-Cámara et al., 2017), se diferencian hasta cinco categorías: muy malo, malo, neutro, bueno, muy bueno.

Previo a la aplicación del modelo entrenado, se aplica un preprocesamiento del texto que facilita la clasificación. Algunas de las técnicas que se usan son el reemplazamiento por sinónimos, la eliminación de stopwords o la lematización.

### 2.2.4 Categorización semántica

Además de la clasificación de las publicaciones según la opinión que manifiestan, el servicio PLN integra también una clasificación semántica de las publicaciones según su contenido. Esta categorización se realiza mediante un tesauro. Partiendo de las categorías genéricas que ofrece el estándar de "*Iptc newscodes*"[2], se ha poblado y creado un diccionario propio.

## 2.3 Aplicación web

Extraída la información, procesada y almacenada en la base de datos de Solr se visualizan los resultados en una interfaz gráfica mediante conexión a la base de datos. El desarrollo de la interfaz gráfica se ha realizado pensando en la usabilidad de cualquier usuario, ofreciendo la información de forma sencilla y atractiva. Además, los módulos implementados en la interfaz son interactivos, lo que permite al usuario navegar a través de la infor-

[2] https://iptc.org/standards/newscodes/

mación mostrada. En la Figura 2 se muestra la pantalla principal de la aplicación web.

En la parte de la izquierda de la interfaz se ofrece al usuario la posibilidad de filtrar por diferentes campos, como por ejemplo el tipo de fuente o el emisor (usuario o página a las que se sigue).

En el resto de la interfaz se visualiza la información más relevante de las redes sociales del usuario en diferentes formatos. En el primer bloque se muestra un resumen de cada publicación, con la opción de visualizar la publicación completa en la propia interfaz o incluso pudiendo navegar hasta la publicación original de la red social. La interfaz permite navegar por diferentes pestañas ofreciendo la siguiente información: una nube de palabras y de hastaghs, unos diagramas de sectores de las personas, organizaciones y localizaciones nombradas, un diagrama de sectores que muestra la frecuencia de las categorías de opinión, una evolución de dichas opiniones a través de un gráfico temporal y dos tipos de visualización de árbol de la distribución de las categorías semánticas en las que se han clasificado las publicaciones.

## 3 Conclusiones y trabajo futuro

La aplicación de diferentes técnicas de PLN ha permitido construir un sistema complejo que permite al usuario estar al día de la información más relevante de sus redes sociales.

Aunque se trabaja con diversidad de fuentes, dominios y registros, el sistema desarrollado presenta, en general, una buena precisión de sus resultados. Además, el uso de diferentes filtros de información y gráficas de visualización a través de la interfaz potencia un mayor grado de usabilidad y utilidad para el usuario final.

Además de su usabilidad y precisión, el sistema desarrollado es altamente escalable. En este sentido, el sistema podría extenderse y adaptarse al uso de otros lenguajes mediante el entrenamiento e integración de nuevos modelos de lenguaje, así como dar apoyo a varios perfiles de usuario independientes añadiendo funcionalidades de log-in.

## Agradecimientos

Este trabajo ha sido patrocinado en parte por el Grupo de Big Data y Sistemas Cognitivos del Instituto Tecnológico de Aragón. La difusión de este trabajo ha sido parcialmente financiada por el Programa Operativo FSE para Aragón (2014-2020).

## Bibliografía

Batrinca, B. y P. C. Treleaven. 2015. Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, 30(1):89–116.

Chang, V. 2017. A proposed social network analysis platform for big data analytics. *Technological Forecasting and Social Change*.

Chiu, J. P. y E. Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. *arXiv preprint arXiv:1511.08308*.

Erkan, G. y D. R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Facebook. 2018. Api overview. facebook developers. Disponible en: https://developers.facebook.com/docs/graph-api/overview/. Recuperado en 2018.

He, W., H. Wu, G. Yan, V. Akula, y J. Shen. 2015. A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, 52(7):801–812.

Martınez-Cámara, E., M. Dıaz-Galiano, M. Garcıa-Cumbreras, M. Garcıa-Vega, y J. Villena-Román. 2017. Overview of tass 2017. En *Proceedings of TASS 2017: Workshop on Semantic Analysis at SEPLN (TASS 2017)*, volumen 1896, páginas 13–21.

Stieglitz, S., M. Mirbabaie, B. Ross, y C. Neuberger. 2018. Social media analytics–challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 39:156–168.

Twitter. 2018. Api overview. twitter developers. Disponible en: https://developer.twitter.com/en/docs. Recuperado en 2018.