TweetAlert: Sistema de Análisis Semántico de la Voz de los Ciudadanos en Redes Sociales en la Ciudad del Futuro

Julio Villena-Román[1,2], Adrián Luna-Cobos[1,3],
José Carlos González-Cristóbal[3,1]

[1] DAEDALUS - Data, Decisions and Language, S.A.
[2] Universidad Carlos III de Madrid
[3] Universidad Politécnica de Madrid
{jvillena, aluna}@daedalus.es, josecarlos.gonzalez@upm.es

Resumen. En este artículo se presenta un sistema automático de almacenamiento, análisis y visualización de información semántica extraída de mensajes de Twitter, diseñado para proporcionar a las administraciones públicas una herramienta para detectar y analizar de una manera sencilla y rápida los patrones de comportamiento de los ciudadanos, su opinión acerca de los servicios públicos, la percepción de la ciudad, los eventos de interés, etc. Además, puede ser usado como un sistema de alerta temprana, mejorando la eficiencia y rapidez de actuación de los sistemas de emergencia.

Palabras clave: Análisis semántico, redes sociales, ciudadano, opinión, temática, clasificación, ontología, eventos, alertas, big data, consola de la ciudad.

1 Introducción

El objetivo final de las decisiones de las administraciones públicas es el bienestar del ciudadano. Sin embargo, no siempre es fácil para los gestores identificar rápidamente los asuntos más importantes que afrontan sus ciudadanos, escalarlos adecuadamente de una manera relativa a la importancia real que los propios ciudadanos les asignan, o simplemente detectar y reconocer lo suficientemente rápido aspectos que aparecen espontáneamente.

El ciudadano se trata desde un punto de vista dual: como el principal usuario de los servicios que presta la ciudad, pero también, como un sensor proactivo, capaz de generar grandes cantidades de datos, por ejemplo en las redes sociales, con información útil acerca de su grado de satisfacción sobre su entorno. Por todo esto, el análisis de la opinión ciudadana es un factor clave dentro de la ciudad del futuro para identificar y resolver los problemas de los ciudadanos.

Sin embargo, toda esta información no es realmente útil a no ser que sea automáticamente procesada y anotada semánticamente para distinguir entre información relevante y no relevante y lograr un mayor nivel de abstracción. En este

proceso de análisis y minería de datos, las tecnologías de análisis y procesamiento de lenguaje natural juegan un papel clave.

En este artículo se presenta un sistema para el análisis en tiempo real de información en Twitter. El sistema permite recopilar y almacenar los mensajes, analizarlos semánticamente y visualizar información agregada de alto nivel. El objetivo del sistema desarrollado es proporcionar a los administradores públicos una herramienta potente para entender de una manera rápida y eficiente las tendencias de comportamiento, la opinión acerca de los servicios que ofrecen, eventos que tengan lugar en su ciudad, etc. y, además, proveer de un sistema de alerta temprana que consiga mejorar la eficiencia de los sistemas de emergencia.

2 Arquitectura del sistema

El sistema está formado por cuatro bloques principales, mostrados en la Figura 1. El componente central es el ***datawarehouse***. Se trata del repositorio de información principal, que es capaz de almacenar el gran volumen de datos a los que hace frente el sistema además de proporcionar funcionalidad avanzada de búsqueda para ser capaces de sacar partido a la información. Este componente se basa en Elasticsearch [1], un motor de búsqueda y análisis en tiempo real, flexible y potente, de código abierto y distribuido. Su buena escalabilidad en sistemas con gran cantidad de datos fue un factor decisivo en la selección de esta arquitectura.

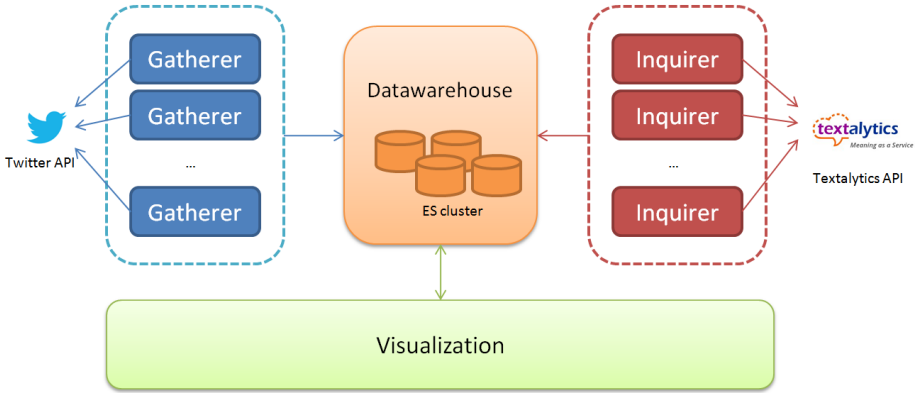

Figura 1. Arquitectura del sistema

El segundo componente lo forman un conjunto de procesos ***recolectores*** que implementan el acceso a los documentos vía consultas a las API de Twitter [2]. Estos recolectores pueden ser configurados para filtrar tweets según una lista de identificadores de usuario, listas de palabras clave a seguir cómo términos o hashtags y localizaciones geográficas a las que restringir la búsqueda.

Un tercer componente, formado por un conjunto de procesos ***consumidores***, tiene como tarea anotar los mensajes de Twitter utilizando las APIs de Textalytics [3].

Se han diseñado dos modelos de clasificación temática (usando la API de clasificación de textos) específicos para este proyecto: *SocialMedia*, que define los temas generales de clasificación, y *CitizenSensor*, que se orienta a características propias del ciudadano como un sensor de eventos de la ciudad. El algoritmo de clasificación de texto utilizado combina [4] una clasificación estadística con un filtrado basado en reglas, que permite obtener un nivel de precisión muy alto para entornos muy diferentes.

También se utiliza la API de extracción de *topics* para anotar entidades nombradas (personas, organizaciones, lugares, etc.), conceptos, expresiones temporales, expresiones monetarias, URIs, etc. Este proceso se lleva a cabo combinando varias técnicas avanzadas de procesamiento de lenguaje natural. Dichas técnicas permiten obtener análisis morfosintáctico y semántico del texto y a través de estas características, identificar distintos tipos de elementos significativos, permitiendo inflexión, variantes y sinónimos.

Con la API de análisis de sentimiento se extrae la polaridad del tweet, para determinar si el texto expresa un sentimiento positivo, neutral o negativo. Este análisis se compone de varios procesos [5]. Primero se evalúa el sentimiento local de cada frase y posteriormente se identifica la relación entre las distintas frases dando lugar a un sentimiento global. Además, empleando el análisis morfosintáctico, se detecta también la polaridad asociada a entidades y conceptos que aparecen en el texto.

Este módulo es capaz de detectar la objetividad del texto analizado, así como detectar marcas de ironía, tanto a nivel global como a nivel de frase, dando al usuario información adicional útil para un análisis exhaustivo de la polaridad.

Por último, se utiliza *user demographics* para extraer características demográficas relativas al usuario que ha generado el texto analizado. Utilizando técnicas de extracción de información y algoritmos de clasificación, se estiman parámetros tales como el tipo de usuario (persona u organización), el sexo del usuario (hombre, mujer o desconocido) y su rango de edad (<15, 15-25, 25-35, 35-65 y >65 años). Para realizar esta estimación, se utiliza la información del usuario en Twitter, el nombre asociado a su cuenta y la descripción de su perfil. El modelo se basa en n-gramas y ha sido desarrollado utilizando Weka.

El cuarto componente es el ***sistema de visualización*** para explotar los datos generados, que se muestra en el apartado siguiente.

La Figura 2 muestra un mensaje anotado por el sistema, en formato JSON:

- Las distintas categorías asignadas del modelo *CitizenSensor* (etiqueta "sensor").
- La ubicación del usuario (una vía pública), y dos posibles alertas: aviso meteorológico por viento e incidencia por congestión de tráfico.
- Según el modelo *SocialMedia* (etiqueta "topic") se clasifica el mensaje dentro de la categoría de "medio ambiente".
- Las entidades (“Gran Vía”) y conceptos ("viento") detectados en el texto.
- La salida del modelo de análisis de sentimiento, que clasifica el mensaje como objetivo, no irónico y con polaridad negativa.
- El análisis del usuario: mujer de edad entre los 25 y los 35 años.

```
{
 "text":"el viento ha roto una rama y hay un atascazo increible
en toda la gran vía...",
 "tag_list":[
   {"type":"sensor",
    "value":"011002 Ubicación - Exteriores - Vías públicas"},
   {"type":"sensor",
    "value":"070700 Alertas meteorológicas - Viento"},
   {"type":"sensor",
    "value":"080100 Incidencia - Congestión de tráfico"},
   {"type":"topic",
    "value":"06 medio ambiente, meteorología y energía"},
   {"type":"entity",
    "value":"Gran Vía",
    "extra":"sumo:Transitway"},
   {"type":"concept",
    "value":"viento"},
   {"type":"sentiment",
    "value":"N"},
   {"type":"subjectivity",
    "value":"OBJECTIVE"},
   {"type":"irony",
    "value":"NONIRONIC"},
   {"type":"user_type",
    "value":"PERSON"},
   {"type":"user_gender",
    "value":"FEMALE"},
   {"type":"user_age",
    "value":"25-35"}
  ]
}
```


Figura 2. Ejemplo de anotación del sistema

En este primer despliegue se analizan tweets en español, pero las APIs utilizadas en el sistema están disponibles para inglés, francés, italiano, portugués y catalán.

3 Consola de la ciudad

El módulo de visualización ofrece una interfaz web, a modo de consola de la ciudad, que permite el filtrado versátil de los datos expresando consultas complejas de una manera estructurada y presenta información de alto nivel, agregada y condensada.

La consola se define mediante elementos denominados “widgets”, configurados en una plantilla específica para los diferentes casos de uso del sistema y adaptada a cada necesidad. Para el desarrollo de los diferentes elementos se han utilizado librerías JavaScript para la creación de gráficos intuitivos e interactivos [6], para la representación de información de la posición por medio de mapas [7], y componentes propios.

La Figura 3 presenta un ejemplo de consola de análisis implementada como primer prototipo para una ciudad española, que hace uso de diferentes widgets. Se muestra el número de tweets y alertas totales en el periodo de filtrado y en el último minuto, y un *timeline* con la evolución temporal de tweets, alertas y polaridad positiva y negativa.

Figura 3. Consola de visualización con filtros, estadísticas y *timelines*

La Figura 4 presenta varias gráficas de tarta con estadísticas de usuario (número de usuarios por tipo, rango de edad y sexo), y la polaridad global de sentimientos, y una lista de las alertas, ubicaciones y eventos más frecuentes detectados en dicho periodo.

Figura 4. Consola de información de usuarios, sentimientos y alertas

La consola también incluye un mapa con las localizaciones de las alertas detectadas en el periodo, junto con los tweets anotados semánticamente que concuerdan con el criterio de filtrado (Figura 5).

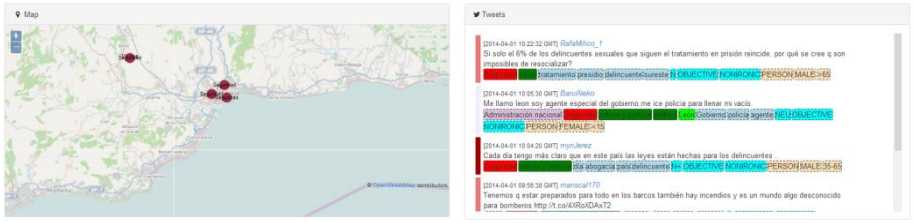

Figura 5. Consola de visualización con tweets y mapa de geolocalización

4 Conclusiones y Trabajos futuros

Actualmente el sistema está en fase beta, acabando la puesta a punto de los diferentes módulos, y estará listo para ser desplegado en distintos escenarios a corto plazo.

Analizando el aspecto tecnológico, las capacidades de almacenamiento del sistema permiten, no sólo analizar los datos en tiempo real, capturando el estado actual de la ciudad, sino también permiten aplicar algoritmos de minería de datos sobre los datos almacenados para, de esta manera, entender mejor las particularidades de la población, mediante técnicas de perfilado y *clustering* que permitan identificar distintos grupos de ciudadanos que se encuentran en la ciudad, comparar singularidades entre los distintos grupos detectados, etc.

Además se está investigando para explorar los siguientes aspectos: análisis de movilidad de la ciudad (cómo, cuándo y por qué los ciudadanos se mueven de un lugar a otro), temas más relevantes a nivel de barrio o zona, reputación de la ciudad, reputación y personalidad de marca, etc.

Agradecimientos

Ese trabajo ha sido financiado parcialmente por el proyecto *Ciudad2020: Hacia un nuevo modelo de ciudad inteligente sostenible* (INNPRONTA IPT-20111006), cuyo objetivo es el diseño de la Ciudad del Futuro, buscando mejoras en áreas como la eficiencia energética, sostenibilidad medioambiental, movilidad y transporte, comportamiento humano e Internet de las cosas.

Referencias

1. Elasticsearch.org. Open Source Distributed Real Time Search & Analytics. 2014. *http://www.elasticsearch.org*.
2. Twitter REST API v1.1. 2014. *https://dev.twitter.com/docs/api/1.1*.
3. Textalytics API. 2014. *http://textalytics.com*.
4. Villena-Román, J., S. Collada-Pérez, S. Lana-Serrano, and J.C. González-Cristóbal. 2011. Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization. *In Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-11), May 18-20, 2011, Palm Beach, Florida, USA. AAAI Press.*
5. Villena-Román, J., S. Lana-Serrano, C. Moreno-García, J. García-Morera, and J.C. González-Cristóbal. 2012. DAEDALUS at *RepLab 2012: Polarity Classification and Filtering on Twitter Data. CLEF 2012 Labs and Workshop Notebook Papers.*
6. Highcharts - Interactive JavaScript charts for your webpage. 2014. JavaScript library website. *http://www.highcharts.com.*
7. Openlayers. 2014 *http://openlayers.org.*