Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 163-166 recibido 27-03-2019 revisado 06-05-2019 aceptado 15-05-2019

Prevención de enfermedades infecciosas basada en el análisis inteligente en RRSS y participación ciudadana

Prevention of infectious diseases based on intelligent analysis in social networks and citizen participation

Óscar Apolinario[1,2], José Medina-Moreira[1], Harry Luna-Aveiga[1], José Antonio García-Díaz[3], Rafael Valencia-García[3] , José Ignacio Estrade-Cabrera[4]

[1] Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Cdla. Universitaria Salvador Allende, Guayaquil, Ecuador

[2] VIAMATICA S.A., Edif. San Francisco 300, Córdova y Av. 9 de Octubre, 090313,Guayaquil, Ecuador

[3] Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Murcia, España

[4] DANTIA Tecnología S.L., Parque Empresarial de Jerez 10, Calle de la Agricultura, 11407, Jerez de la Frontera, Cádiz, España

{oscar.apolinarioa, jose.medinamo.es@ug.edu.ec, harry.lunaa}@ug.edu.ec

{joseantonio.garcia8, valencia}@um.es

jiestrade@dantia.es, oapolinario@viamatica.com

Resumen: Este proyecto consiste en el desarrollo una plataforma inteligente de monitorización de enfermedades infecciosas a partir de la monitorización de redes sociales, fuentes de datos oficiales y la participación ciudadana. Esta información estará disponible para las autoridades sanitarias y gubernamentales a través de un panel de mandos personalizable para que puedan detectar zonas calientes en las que exista algún tipo de brote o focos de infección. También estará disponible un sistema de alertas para avisar a los ciudadanos cuándo se ha detectado cierto nivel de alarma en radio cercano a donde se encuentran. Este proyecto está siendo desarrollado por la empresa de Ecuador VIAMÁTICA, algunos docentes de la Universidad de Guayaquil, la empresa española DANTIA y las Universidades de Murcia y Carlos III de Madrid. La parte española del proyecto está financiada mediante una convocatoria de proyectos unilaterales del CDTI.

Palabras clave: Análisis de sentimientos, Infodemiología, Enfermedades Infecciosas, BlockChain

Abstract: This project consists in the development of an intelligent platform for the monitoring of infectious diseases based on: text written in natural language on social networks, official data sources and citizen participation. This information will be available to the health and governmental authorities through a customizable control panel so they can detect hot areas in which there is some type of outbreak. An alert system will also be available to notify citizens when a certain level of alarm is detected in a nearby radius of where they are. This project is being developed by VIAMÁTICA (Ecuador), the University of Guayaquil (Ecuador), DANTIA (Spain) and the Universities of Murcia and Carlos III of Madrid (Spain). The Spanish part of the project is financed through the unilateral CDTI projects call.

Keywords: Sentiment analysis, Infoveillance, Infectious Diseases, BlockChain

1 Introducción

Los brotes epidémicos son uno de los problemas más graves a los que se enfrenta la especie humana. En los últimos años se han producido cerca de cinco alertas sanitarias internacionales graves: el ZIKA, la Gripe Aviar, el

ISSN 1135-5948. DOI 10.26342/2019-63-21

© 2019 Sociedad Española para el Procesamiento del Lenguaje Natural

virus del Ébola, la Gripe Tipo A y el SARS. Aunque ninguno de ellos tuvo una elevada mortalidad ni comprometió la existencia humana en su conjunto, si que supusieron un grave perjuicio para la sociedad en general y para las instituciones en particular; repercutiendo, además, negativamente en la economía. Además, el virus del Zika, muy presente en América Latina, tiene millones de casos infectados y miles de bebés están siendo afectados con trastornos neurológicos.

Para mitigar los efectos de estos brotes infecciosos, la Organización Mundial de la Salud recomienda diseñar estrategias de detección temprana. El diseño de estas estrategias requiere disponer de suficientes evidencias para predecir, con suficiente probabilidad, un brote epidémico. Aunque la mayoría de estos estudios se realiza a partir de datos clínicos recolectados en hospitales, la comunidad científica está explorando métodos fiables alternativos para la adquisición de información a través de Internet. Sin embargo, estos sistemas tienen todavía un amplio margen de mejora ya que todavía hacen interpretaciones inadecuadas de los datos (Choi et al., 2016).

El presente proyecto consiste en el desarrollo de un sistema de monitorización de enfermedades infecciosas que detecte zonas calientes en las que existe algún tipo de brote o foco de infección y comunique dicha información, de forma comprensible y aprovechable, tanto a los ciudadanos como a las autoridades sanitarias y gubernamentales.

A nivel técnico, esta plataforma extrae datos de tres categorías de fuentes de datos: 1) datos estructurados, a partir de la extracción del conocimiento en fuentes de datos oficiales publicadas en la web; 2) evidencias, a partir de la participación directa de la ciudadanía identificando y notificando directamente casos encontrados por ellos y, finalmente, a partir de 3) datos no estructurados, procedentes de textos escritos en lenguaje natural procedente de redes sociales públicas, tales como Twitter o Facebook.

Con la consecución de los objetivos del proyecto se pretende, por un lado, reducir los costes sociales y económicos derivados del tratamiento de estos brotes infecciosos y, por otro lado, mejorar la percepción de la ciudadanía en temas de salud pública transparentes y participativos, permitiendo el empoderamiento de la ciudadanía[1].

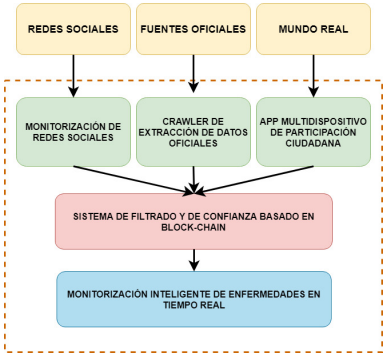

Figura 1: Arquitectura del sistema

2 *Arquitectura del sistema*

Esta plataforma está formado por cinco módulos: 1) Módulo de monitorización de redes sociales, 2) Módulo Crawler de extracción de datos oficiales, 3) Aplicación multidispositivo de participación ciudadana, 4) Módulo de sistema de filtrado y confianza basado en BlockChain y, 5) Módulo para la monitorización inteligente de enfermedades en tiempo real (ver Figura 1).

Una característica clave de esta plataforma es el desarrollo de algoritmos de evaluación de la confianza basados en la credibilidad de la fuente, con objeto de descartar falsos positivos y de priorizar hallazgos relevantes.

Una vez la información ha sido filtrada, es presentada a los usuarios finales mediante un panel de mandos configurable, donde pueden indicar que indicadores de rendimiento (KPIs) necesitan, con el fin de ayudarles a establecer medidas estratégicas acerca de cómo actuar ante posibles nuevos brotes o epidemias.

A continuación, se describen brevemente cada uno de los módulos de la plataforma haciendo mayor hincapié en los relacionados con tecnologías del lenguaje humano.

2.1 Módulo de monitorización de redes sociales

Este módulo se encarga de extraer información escrita en lenguaje natural en redes sociales. En pocas palabras, el sistema funciona de la siguiente manera: En primer lugar, se obtienen todos los mensajes que contienen palabras clave, además de otros indicadores relacionados, de la enfermedad objetivo.

[1]https://www.who.int/csr/labepidemiology/projects/earlywarnsystem/en/

En segundo lugar, se emplean técnicas de reconocimiento de entidades (Ruiz-Martínez et al., 2012) y técnicas de análisis de sentimientos basado en aspectos (Peñalver-Martínez et al., 2014) para obtener la información subjetiva relacionada con los tópicos que tratan, así como son los síntomas, las causas o los medicamentos. Para ello, se ha definido una ontología que describe el vocabulario principal de las enfermedades infecciosas, sus síntomas, posibles focos de infección, medicamentos, etc. Esta ontología será ampliada durante el desarrollo del proyecto incluyendo más información de enfermedades infecciosas ya que actualmente se ha centrado en las enfermedades infecciosas transmitidas por mosquito como dengué, malaria, zika y chuchugua.

Con respecto al análisis de sentimientos, se ha compilado hasta el momento un corpus balanceado inicial de 8.966 tuits y se han realizado pruebas con clasificadores binarios a partir de un modelo lingüístico basado en el análisis y extracción de variables psicolingüísticamente relevantes específica para textos escritos en castellano y en el español de América latina obteniendo unos buenos resultados iniciales (García-Díaz et al., 2018).

2.2 Módulo Crawler de extracción de datos oficiales

Este módulo se encarga de la obtención de datos estructurados a partir de un sistema de extracción de conocimiento (crawler). Para ello, se emplean técnicas de reconocimiento de entidades, extracción de términos y construcción automática de ontologías (Valencia-García et al., 2008) con el fin de extraer información relacionada con el dominio de las enfermedades infecciosas. Debido a que la información se encuentra en la red en distintos formatos, tales como documentos web, imágenes o documentos de procesadores de texto, es necesario el desarrollo de diversos procedimientos específicos capaces de extraer información para cada formato.

2.3 Aplicación multidispositivo de participación ciudadana

Este módulo consiste en una aplicación multidispositivo de participación ciudadana para la identificación activa de factores de riesgo relacionados con enfermedades infecciosas. Los ciudadanos también etiquetarán sus contribuciones y participación mediante etiquetas obtenidas de las ontologías del dominio.

Esta aplicación se ha diseñado con un enfoque basado en Aplicaciones Web Progresivas por dos motivos. En primer lugar, para permitir a los usuarios identificar evidencias en zonas sin conexión a Internet, realizando el envío de datos una vez que recupere la conexión. En segundo lugar, para aumentar el periodo de autonomía del dispositivo a través de minimizar el consumo de su batería (Malavolta et al., 2017). Además, esta aplicación incluye diversas técnicas de gamificación para fomentar la participación ciudadana a través de un sistema de reputación y recompensas.

2.4 Módulo de sistema de filtrado y confianza basado en BlockChain

Este módulo se encarga del valorar la calidad y la confianza de las distintas evidencias encontradas por los módulos anteriormente descritos. Al tratarse de información sensible, se ha optado por usar la tecnología de cadena de bloques (Blockchain) para garantizar la privacidad de la información médica recolectada. Esta aplicación de las cadenas de bloques ya ha sido estudiada en (Zyskind, Nathan, y others, 2015), donde los autores describen un sistema descentralizado encargado de manejar información personal sensible.

Para asignar un nivel de confiabilidad a los datos extraídos en redes sociales, se analizan características tales como si la cuenta del usuario ha sido validada o número de publicaciones relevantes realizadas. Por otro lado, para valorar las evidencias recogidas por los usuarios a través de la aplicación de participación ciudadana, se tienen en cuenta parámetros como el historial del usuario o la frecuencia de las mediciones.

2.5 Módulo para la monitorización inteligente de enfermedades en tiempo real

Este módulo se encarga de la gestión y monitorización de los parámetros relacionados con las enfermedades infecciosas a través de un proceso guiado por ontologías. En concreto, este sistema se encarga de cuatro apartados principales: 1) Configuración, que permite a los operarios del sistema crear áreas de interés, a partir de especificar un área geográfica y las enfermedades objetivo a monitorizar; 2) dashboard, que consiste en el desarrollo de un panel de mandos genérico y configurable donde los operarios podrán indicar y confi-

gurar los KPIs que estimen convenientes; 3) KPIs, que son indicadores de desempeño independientes capaces de leer datos de manera periódica y de mostrar la información en distintos formatos, tales como tablas o gráficas; 4) Sistema de alertas, capaz de configurar y notificar automáticamente a los usuarios suscritos en el momento en que los datos superen cierto umbral establecido para que se tomen medidas preventivas.

3 Trabajo futuro

El presente proyecto se encuentra todavía en una fase temprana y su desarrollo termina en 2020. Para cada uno de los módulos se están realizando estudios básicos para comprobar la viabilidad de ciertas operaciones, especialmente los relativos al módulo de módulo de sistema de filtrado y confianza basado en BlockChain, descrito en la 2.4sección 2.4.

Con respecto a las técnicas de análisis de sentimientos, se han llevado a cabo estudios a la hora de analizar la polaridad de textos en castellano y en español de México. Estos sistemas lingüísticos se han probado para clasificar el sentimiento a nivel general y se está trabajando en la extracción de conceptos a partir de una ontología para poder realizar un análisis más minucioso basado en aspectos.

Durante el tiempo que queda de desarrollo del proyecto está previsto la mejora de las tecnologías de extracción de conocimiento y minería de opiniones basados en aspectos. Por otro lado, se desarrollará la aplicación de participación ciudadana que permitirá etiquetar las contribuciones con vocabulario del dominio, además que se analizarán también sus contribuciones en texto.

Por último, se realizará una integración de los datos estructurados y no estructurados dentro del primer prototipo de la plataforma global. En este sentido, se planificarán distintas pruebas de campo para comprobar la calidad de las mediciones por parte de los usuarios en entornos sin conexión a Internet.

En la segunda anualidad también se desarrollarán las tecnologías de análisis inteligente de datos guiado por ontologías en el que se permitirá seleccionar distintos conceptos de las ontologías para así solamente realizar el análisis inteligente sobre los datos relacionados con esos conceptos.

Agradecimientos

Este trabajo está siendo financiado por el CDTI dentro del proyecto con referencia IDI-20180989 dentro de la convocatoria de proyectos unilaterales.

Bibliografía

Choi, J., Y. Cho, E. Shim, y H. Woo. 2016. Web-based infectious disease surveillance systems and public health perspectives: a systematic review. *BMC Public Health*, 16(1):1238.

García-Díaz, J. A., O. Apolinario-Arzube, J. Medina-Moreira, J. O. Salavarria-Melo, K. Lagos-Ortiz, H. Luna-Aveiga, y R. Valencia-García. 2018. Opinion mining for measuring the social perception of infectious diseases. an infodemiology approach. En *International Conference on Technologies and Innovation*, páginas 229–239. Springer.

Malavolta, I., G. Procaccianti, P. Noorland, y P. Vukmirović. 2017. Assessing the impact of service workers on the energy efficiency of progressive web apps. En *Proceedings of the 4th International Conference on Mobile Software Engineering and Systems*, páginas 35–45. IEEE Press.

Peñalver-Martínez, I., F. García-Sanchez, R. Valencia-García, M. A. Rodriguez-García, V. Moreno, A. Fraga, y J. L. Sánchez-Cervantes. 2014. Feature-based opinion mining through ontologies. *Expert Systems with Applications*, 41(13):5995–6008.

Ruiz-Martínez, J. M., R. Valencia-García, R. Martínez-Béjar, y A. Hoffmann. 2012. Bioontoverb: A top level ontology based framework to populate biomedical ontologies from texts. *Knowledge-Based Systems*, 36:68–80.

Valencia-García, R., J. T. Fernández-Breis, J. M. Ruiz-Martínez, F. García-Sánchez, y R. Martínez-Béjar. 2008. A knowledge acquisition methodology to ontology construction for information retrieval from medical documents. *Expert Systems*, 25(3):314–334.

Zyskind, G., O. Nathan, y others. 2015. Decentralizing privacy: Using blockchain to protect personal data. En *2015 IEEE Security and Privacy Workshops*, páginas 180–184. IEEE.