CollaborativeHealth - Plataforma de monitorización de enfermedades infecciosas en Twitter

CollaborativeHealth - Infectious disease monitoring platform on Twitter

Óscar Apolinario-Arzube[1], Diego Roldán[2], José Antonio García-Díaz[3], Lisardo Prieto González[4], Rafael Valencia-García[3]
[1]VIAMATICA S.A., Edif. San Francisco 300, Córdova y Av. 9 de Octubre, 090313, Guayaquil, Ecuador
[2]DANTIA Tecnología S.L., Parque Empresarial de Jerez 10,
Calle de la Agricultura, 11407, Jerez de la Frontera, Cádiz, España
[4]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, España
[3]Departamento de Informática, Universidad Carlos III de Madrid, Madrid, España
oscar.apolinarioa@ug.edu.ec, droldan@dantia.es, lpgonzal@inf.uc3m.es
{joseantonio.garcia8, valencia}@um.es

Resumen: Brotes de enfermedades infecciosas como el SARS, MERS o el COVID-2019 causan la pérdida de vidas humanas, el incremento de presión en hospitales y pérdidas económicas. La detección temprana permite mitigar estos efectos, ya que permite aplicar medidas cuando estas no son tan lesivas. Aunque esta tarea tradicionalmente se ha llevado a cabo mediante pruebas en lugares concurridos como aeropuertos, gracias a avances en procesamiento del lenguaje natural se está complementado esta tarea con el análisis de redes sociales. En esta demostración mostramos CollaborativeHealth, una herramienta de vigilancia epidemiológica a través de la monitorización activa de redes sociales y la participación ciudadana. Esta herramienta facilita a rastreadores detectar zonas calientes donde los ciudadanos hablan sobre síntomas, enfermedades o riesgos. También permite evaluar la aceptación por parte de los ciudadanos de las medidas tomadas por los organismos competentes.
Palabras clave: Infodemiología, Minería de Opiniones, Recuperación de la Información, Ontologías, Blockchain.

Abstract: Outbreaks of infectious diseases such as SARS, MERS or COVID-2019 cause the loss of human lives, increased pressure on hospitals and economic losses. Early detection allows these effects to be mitigated, as it allows measures to be applied when they are not so harmful. Although this task has traditionally been carried out through tests in crowded places such as airports, advances in Natural Language Processing have allow to include the analysis of what people are talking about on social networks. Here we present CollaborativeHealth, an epidemiological surveillance tool for active monitoring on social networks and citizen participation. This tool makes it easy for trackers to spot geographical areas where citizens talk about symptoms, illnesses, or risks. It also makes it possible to evaluate the acceptance by citizens of the measures taken by the competent authorities.
Keywords: Infodemiology, Opinion Mining, Information Retrieval, Ontologies, Blockchain.

1 Introducción

Las epidemias de enfermedades infecciosas causan graves perjuicios a nivel individual, social y económico. Una de las estrategias para prevenir la propagación de enfermedades infecciosas es la monitorización activa, tarea que realiza el personal sanitario a través de controles en lugares de tránsito tales como aeropuertos o estaciones de tren. Debido a avances en los últimos años y de la gran aceptación por parte de la ciudadanía de las tecnologías de comunicación, la infodemiología surgió como una ciencia para mejorar y analizar aspectos relacionados con la salud pública a través de Información en Internet (Mavragani, 2020). Además, muchas de las medidas que tienen que tomar las autoridades competentes para mitigar estos brotes están relacionadas con medidas de distanciamiento social o confinamientos selectivos. Estas me-

didas son a veces polémicas y en ocasiones no son aceptadas por parte de los ciudadanos. Por lo tanto, poder monitorizar la percepción pública de la ciudadanía hacia estas medidas es importante para entender cuáles están funcionando y cuáles no, y así poder ajustar la estrategia.

Aunque ya existen herramientas enfocadas a la monitorización de enfermedades infecciosas a partir de datos de Internet, tales como HealthMap (Freifeld et al., 2008), en los últimos años se están desarrollando nuevas herramientas para la monitorización y el rastreo, así como otras herramientas de consulta. Sin embargo, uno de las mayores retos tecnológicos de estos sistemas es la explotación de los recursos no estructurados presentes en Internet. Para ello, la empresa DANTIA en colaboración con VIAMÁTICA, y con las universidades Carlos III de Madrid y la Universidad de Murcia, han desarrollado CollaborativeHealth (Apolinario-Arzube et al., 2020), una plataforma de monitorización activa de redes sociales a partir del análisis de redes sociales, noticias en Internet y de evidencias reportadas por la ciudadanía. A nivel tecnológico, esta plataforma hace uso de anotación semántica, ontologías, cadenas de bloques y minería de opiniones orientada a aspectos. Este artículo presenta una demostración de esta plataforma.

2 Arquitectura del sistema

La Figura 1 muestra la arquitectura funcional de CollaborativeHealth. En pocas palabras se puede resumir de la siguiente manera. En primer lugar, el sistema de monitorización se encarga de compilar las evidencias a partir de tres fuentes: redes sociales, documentos web y evidencias reportadas por los ciudadanos a través de una aplicación web progresiva. En segundo lugar, las evidencias compiladas pasan por un proceso que valora el grado de confianza de cada evidencia. Aquellas que superan cierto umbral son selladas y almacenadas en una base de datos distribuida. En tercer lugar, las evidencias son mapeadas a una ontología del dominio a partir de un proceso de anotación semántica y se extrae la polaridad subjetiva de cada una de ellas. A continuación, distintos indicadores de rendimiento (KPIs) semánticos se muestran una interfaz web donde los usuarios finales pueden configurar sus vistas para monitorizar rangos de fechas, tipos de fuentes, o zonas geográficas. A continuación, se detalla cada uno de estos componentes.

2.1 Monitorización de redes sociales

Este componente se encarga de la monitorización en Twitter. Los usuarios finales pueden indicar qué palabras clave y qué zonas geográficas quieren monitorizar. Cabe destacar que un subconjunto de los tweets se empleó para entrenar un modelos de minería de opiniones. Para ello, organizamos equipos de etiquetado con voluntarios, estudiantes y personal del proyecto empleando la herramienta UMUCorpusClassifier para tal fin (García-Díaz et al., 2020).

2.2 Web Crawler

Este componente se encarga de la monitorización de noticias web. Para ello, se diseñó una crawler parametrizable donde los usuarios finales pueden indicar qué sitios web quieren monitorizar. Con el fin de realizar un cribado más selectivo de qué contenidos son relevantes, los usuarios también pueden indicar una o varias expresiones regulares para filtrar URLs específicas. También pueden indicar qué parte del contenido contiene la información relevante. Para ello, se pueden configurar filtros de hojas de estilo en cascada y así eliminar menús, publicidad u otro tipo de contenido.

2.3 Aplicación Web Progresiva de colaboración ciudadana

Se diseñó una aplicación web progresiva para reportar evidencias relacionadas con las enfermedades infecciosas como podrían ser zonas insalubres como criaderos de mosquitos o basura en la calle. Esta aplicación permite a los ciudadanos enviar fotos acompañados de un texto, unas etiquetas y datos de la geolocalización. Para asistir a los usuarios con este proceso de etiquetado se emplean técnicas de visión por computador. También hay que destacar que la aplicación permite funcionar sin conexión a Internet y puede enviar las evidencias en cuanto recupere la conexión. Lo cuál permite emplearse en lugares aislados.

2.4 Filtrado de confianza

Este componente se subdivide en dos módulos. En primer lugar, el módulo de filtrado de confianza tiene como objetivo asignar una puntuación a cada evidencia para priorizar

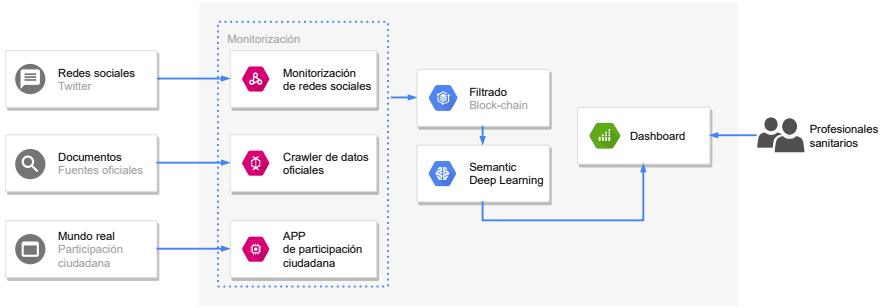

Figura 1: Arquitectura funcional de la plataforma CollaborativeHealth.

las más relevantes. Para ello, se emplean diversos criterios en función de la fuente, tales como la precisión de la geolocalización, la longitud del texto, o la confianza del usuario en función de evidencias previas. En segundo lugar, aquellas evidencias que superar cierto umbral son selladas y almacenadas en una base de datos distribuida basada en BigChainDB. Mediante esta base de datos, se pueden compartir de manera segura datos entre distintas instancias de CollaborativeHealth de manera segura asegurando la integridad de los datos. Además, se puede usar como respaldo de seguridad.

2.5 Semantic Deep Learning

Este componente se basa en el desarrollo de distintas KPIs semánticos. Para ello, se optó por un diseño dirigido por ontologías. Se creó una ontología a partir de combinar y extender ontologías existentes, tales como la Disease Ontology (DO)[1] o la Infectious Disease Ontology (IDO) (Cowell y Smith, 2010) y del empleo de códigos SNOMED-CT[2] para la interoperabilidad. La ontología resultante incluye conceptos tales como enfermedades, síntomas, regiones, riesgos o medidas de prevención.

Cada evidencia recolectada por CollaborativeHealth es mapeada a la ontología mediante un proceso de anotación semántica basado en una adaptación de la técnica Term-Frequency Inverse Document Frequency (TF-IDF) (Rodríguez-García et al., 2014). Para ello, asociamos a cada concepto de la ontología un conjunto de expresiones regulares con términos relacionados. Por cada evidencia se calculaba el TF-IDF de cada concepto de la ontología. El TF-IDF de cada concepto afectaba a otros conceptos de la ontología en base a la distancia entre ambos conceptos. Por ejemplo, si un documento hace mención explícita a uno o varios síntomas de una enfermedad, el valor del TF-IDF se aumentaba para dichos conceptos, aunque el término no apareciera explícitamente en el documento. Además de este proceso, se extrae la polaridad subjetiva de los textos a partir de un modelo de deep-learning basado en BETO (Cañete et al., 2020).

Una vez obtenida la relación semántica y la polaridad, se diseñaron distintos KPIs tales como nubes de palabras, mapas de calor, gráficos de barras y de tendencia. Cada KPI se configuró para que fuera independiente y permite el uso de distintos filtros de geolocalización, confianza y conceptos de la ontología. La Figura 2 muestra un ejemplo de estos KPIs donde se ve la correlación entre los conceptos *COVID-2019* y *vacunas* en los tuits recopilados durante el mes de abril en Madrid.

2.6 Cuadro de mandos

Finalmente, se diseñó una panel de mandos mediante tecnologías web responsivas. Este panel de mandos permite a los usuarios finales poder visualizar los reportes y los KPIs. También se integraron varios de los KPIs desarrollados con la aplicación ciudadana para dar feedback a los usuarios.

3 Trabajo futuro

Como trabajo futuro se pretende mejorar la interpretabilidad de los modelos de minería de opiniones a partir del uso de características lingüísticas para complementar los word embeddings a partir de características lingüísticas empleadas en (García-Díaz et al., 2021; García-Díaz, Cánovas-García, y Valencia-García, 2020). Otra vía de investigación relacionada sería el uso de modelos de word embeddings multilingües para aumen-

[1] `https://disease-ontology.org/`
[2] `https://www.snomed.org/`

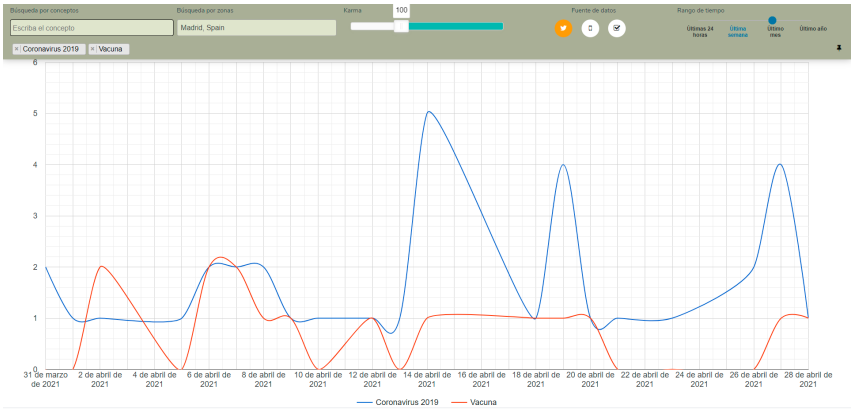

Figura 2: Ejemplo de KPI semántico.

tar la recuperación de la información en distintos idiomas. Con respecto a la ontología, se estudiará el análisis de técnicas que permitan el mantenimiento semiautomático de la ontología y el uso de knowledge graphs. Por último, se están realizando modelos de deep-learning para mejorar los sistemas de filtrado para identificar fake news o mensajes publicitarios.

Agradecimientos

Este trabajo está siendo financiado por el CDTI y el Fondo Europeo de Desarrollo Regional (FEDER / ERDF) a través del proyecto CollaborativeHealth IDI-20180989.

Bibliografía

Apolinario-Arzube, Ó., J. A. García-Díaz, S. Pinto, H. Luna-Aveiga, J. J. Medina-Moreira, J. M. Gómez-Berbis, R. Valencia-Garcia, y J. I. Estrade-Cabrera. 2020. Collaborativehealth: Smart technologies to surveil outbreaks of infectious diseases through direct and indirect citizen participation. En *Computer Science On-line Conference*, páginas 177–190. Springer.

Cañete, J., G. Chaperon, R. Fuentes, y J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *PML4DC at ICLR*, 2020.

Cowell, L. G. y B. Smith. 2010. Infectious disease ontology. En *Infectious disease informatics*. Springer, páginas 373–395.

Freifeld, C. C., K. D. Mandl, B. Y. Reis, y J. S. Brownstein. 2008. Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 15(2):150–157.

García-Díaz, J. A., A. Almela, G. Alcaraz-Mármol, y R. Valencia-García. 2020. Umucorpusclassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks. *Procesamiento del Lenguaje Natural*, 65:139–142.

García-Díaz, J. A., M. Cánovas-García, R. Colomo-Palacios, y R. Valencia-García. 2021. Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506–518.

García-Díaz, J. A., M. Cánovas-García, y R. Valencia-García. 2020. Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america. *Future Generation Computer Systems*, 112:641–657.

Mavragani, A. 2020. Infodemiology and infoveillance: scoping review. *Journal of medical Internet research*, 22(4):e16206.

Rodríguez-García, M. Á., R. Valencia-Garcí, F. García-Sánchez, y J. J. Samper-Zapater. 2014. Creating a semantically-enhanced cloud services environment through ontology evolution. *Future Generation Computer Systems*, 32:295–306. Special Section: The Management of Cloud Systems, Special Section: Cyber-Physical Society and Special Section: Special Issue on Exploiting Semantic Technologies with Particularization on Linked Data over Grid and Cloud Architectures.