Gac Sanit. 2025;39(S2):539

Gaceta Sanitaria

https://www.gacetasanitaria.org

404 - ADAPTACIÓN DE MODELOS DE PROCESAMIENTO DE LENGUAJE NATURAL PARA LA VIGILANCIA EPIDEMIOLÓGICA DE IRAG EN DOS HOSPITALES (VIGILEN-IRAG)

M. Lozano, L. Vigil, A. Rojas-Benedicto, M. Rodríguez, J.A. Marín, J. Garrido, B. Borras-Bermejo, M. Krallinger, S. Monge, et al.

CNE–ISCIII; EIDUNED; CIBERESP; Centro de Supercomputación de Barcelona (BSC); Hospital Universitario Virgen de las Nieves; Hospital Universitario Vall d'Hebron; CIBERINFEC.

Resumen

Antecedentes/Objetivos: La automatización de la vigilancia de las infecciones respiratorias agudas graves (IRAG) puede mejorar su estabilidad, eficiencia, resiliencia, aceptabilidad y temporalidad. Los sistemas de procesamiento de lenguaje natural (NLP), permiten la extracción automática de datos estructurados a partir de los informes médicos. El objetivo de este trabajo es adaptar sistemas de NLP para la extracción de datos relevantes para la vigilancia de IRAG.

Métodos: Se generó un conjunto de datos a partir de casos clínicos de neumología y enfermedades infecciosas, informes clínicos ficticios de IRAG e informes de alta sintéticos generados mediante modelos del lenguaje generativos, los cuales fueron categorizados por sexo, edad, IRAG o no IRAG y por la presencia de infección de los virus de interés en la vigilancia de IRAG. Se emplearon modelos de reconocimiento de entidades nombradas, para la preanotación de cinco clases de conceptos: síntomas, enfermedades, fármacos, procedimientos y especies. Utilizando guías de anotación, un equipo de sanitarios y expertos en vigilancia de IRAG revisó y corrigió las anotaciones para el reentrenamiento de los modelos. Se calcularon los porcentajes de acuerdo entre anotadores y se revisaron las discrepancias.

Resultados: Se generaron 1.248 textos clínicos (1.024 casos clínicos, 100 ficticios y 124 sintéticos): 686 hombres y 497 mujeres (65 sexo desconocido), 880 adultos y 361 niños (7 edad desconocida), 553 de IRAG y 695 no IRAG, 212 de COVID, 58 de gripe y 118 de VRS. Se inició la anotación por los dominios de enfermedades y especies. Tres anotadores revisaron cada dominio, alcanzando un acuerdo inicial en un primer lote de 20 textos de entre 61,3% y 80% para enfermedades y 86,4% y 95% para especies. Tras la revisión de discrepancias, en un segundo lote de 20 textos se mejoraron los porcentajes a entre 72,3% y 84,7% para enfermedades y 88,2% y 100% para especies.

Conclusiones/Recomendaciones: Esta experiencia inicial muestra que los modelos utilizados anotan con alta validez textos de enfermedades respiratorias y que el porcentaje de acuerdo inicial entre revisores fue elevado, posiblemente gracias a seguir guías de anotación estandarizadas. La corrección manual y reentrenamiento de los modelos permiten mejorar la calidad de los sistemas automáticos de forma controlada. Antes de la implementación de estas herramientas en la vigilancia, será preciso la normalización de las menciones anotadas, su mapeo a las variables de interés y su validación mediante la comparación con el estándar de oro de la vigilancia manual.

0213-9111 © 2025 SESPAS. Publicado por Elsevier España, S.L.U. Todos los derechos reservados