Junta de Andalucía
Consejería de Cultura,
Patrimonio Histórico y Deporte
Instituto Andaluz del Patrimonio Histórico

REVISTA PH
ISSN 2340-7565

Sección Debate (***revista PH*** 120, febrero 2027)

Debate 36: IA y patrimonio cultural. ¿Qué deberíamos tener en cuenta para el uso social, ético y sostenible de la IA en el ámbito del patrimonio?

Textos provisionales

Inteligencia artificial y modelos de lenguaje para estudiar el riesgo del patrimonio cultural: cómo extraer datos de vulnerabilidad para su análisis en SIG

Alfonso Hidalgo Mañero, David Díaz Jiménez, Macarena Santos Cárdenas, Laura Toro-Murillo, Mónica Moreno Falcón | Universidad Pablo de Olavide

Introducción y objetivo

Actualmente, gran parte de la información disponible sobre patrimonio cultural no se encuentra estructurada debido a la falta de un esquema predefinido de organización (König et ál. 2025). Además, una parte de esta información se encuentra distribuida en documentos de distinta naturaleza, extensión y estructura que, a su vez, se almacenan en archivos físicos y digitales, así como en diferentes sistemas de información (Abdullah et ál. 2023). Esto dificulta la identificación y extracción sistemática de información relevante.

La información disponible sobre patrimonio cultural presenta estas mismas limitaciones, lo que dificulta su integración en los Sistemas de Información Geográfica (SIG) para el monitoreo y reducción del riesgo. En Andalucía se dispone de información geoespacial sobre amenazas naturales (Junta de Andalucía s.f.) y de la *Guía Digital del Patrimonio Cultural de Andalucía* que facilita el acceso a información sobre bienes culturales (Fernández Cacho et ál. 2017). Sin embargo, la evaluación del riesgo del patrimonio cultural sigue siendo compleja debido a las dificultades para obtener datos de vulnerabilidad (materiales constructivos, patologías y estado de conservación) normalizados y estructurados, que puedan integrarse en los SIG.

En este contexto, la presente contribución al debate de *revista PH* 120 tiene como objetivo explorar el potencial de desarrollar modelos de lenguaje especializados en el análisis de la vulnerabilidad de los bienes patrimoniales para su integración en bases de datos geoespaciales.

Fundamentos de los modelos de lenguaje

El procesamiento del lenguaje natural (PLN) comprende un conjunto de técnicas computacionales orientadas al análisis y la representación del lenguaje humano, así como la identificación de patrones y relaciones lingüísticas (Liddy 2009). Entre sus aplicaciones se encuentran tanto la identificación y clasificación de entidades como la extracción de relaciones y eventos especificados por el investigador, lo que permite transformar texto no estructurado en información estructurada que pueda ser interpretada y utilizada en otras aplicaciones (Grishman 2019).

Metodológicamente estos modelos utilizan representaciones numéricas del texto para facilitar su procesamiento computacional. Para ello, el texto se divide en unidades denominadas *tokens*, que pueden corresponder a palabras, fragmentos de palabras o signos de puntuación (Kudo y Richardson 2018). Estas unidades pueden representarse mediante vectores numéricos continuos, conocidos como *embeddings*, que permiten representar propiedades semánticas y relaciones de similitud entre palabras (Grishman 2019; Kudo y Richardson 2018). Además, estos modelos pueden usar arquitectura *Transformer*, que permite considerar no sólo los elementos individuales de un texto, sino también las relaciones entre ellos. (Vaswani et ál. 2017). Entre los modelos basados en la arquitectura *Transformer* se encuentran BERT (Devlin et ál. 2019) y BETO, un modelo basado en BERT preentrenado exclusivamente con textos en español (Cañete et ál. 2023).

Para comprender mejor el potencial de estos modelos de lenguaje y sus posibles usos en el análisis de la vulnerabilidad del patrimonio cultural, la tabla siguiente recoge tres experiencias y plantea una serie de cuestiones clave que conviene tener en cuenta al iniciarse en esta línea de investigación.

REVISTA PH
ISSN 2340-7565

Tabla 1. *Estudios basados en modelos de lenguaje BERT y potencial para analizar la vulnerabilidad del patrimonio cultural*

Referencia bibliográfica	¿Qué hace el estudio?	Aplicación potencial
Jeon et ál. (2022)	Usa tres modelos de lenguaje preentrenados: Multilingual BERT, KoBERT y KoELECTRA y tesauros para analizar 69.752 reclamaciones sobre patologías en viviendas redactadas en coreano usando jerga y errores ortográficos.	Posibilidad de combinar el uso de tesauros y modelos de lenguaje para identificar vulnerabilidad según patologías, materiales y factores de deterioro en informes técnicos, inventarios y registros de intervenciones.
Brandsen et ál. (2022)	Desarrolla ArcheoBERTje, un modelo BERT en neerlandés adaptado para el análisis de textos arqueológicos. El modelo reconoce y clasifica automáticamente seis tipos de entidades: objetos arqueológicos, periodos cronológicos, lugares, contextos arqueológicos, materiales y especies.	Uso de informes de conservación y planes de salvaguarda para el entrenamiento de modelos de lenguaje especializados en patrimonio.
Li et ál. (2026)	Modelo BERT-BiLSTM-CRF, que combina el análisis contextual de las palabras (BERT), con el procesamiento bidireccional del texto (BiLSTM) y la asignación de etiquetas (CRF) para estudiar las relaciones entre factores de riesgo geológico.	Estudio sistémico del riesgo. Análisis no solo de la vulnerabilidad, sino también de las causas que la motivan.

¿Es posible enseñar a un modelo de lenguaje a reconocer información sobre el estado de conservación de un edificio a partir de tesauros?

El primer ejemplo, desarrollado por Jeon et ál. (2022), aborda un problema habitual en la documentación técnica: la dificultad de extraer información de textos que contienen errores ortográficos, abreviaturas, jerga profesional y expresiones poco estandarizadas. Los autores utilizan modelos de lenguaje preentrenados y técnicas de reconocimiento automático de entidades (NER) para identificar defectos constructivos, elementos afectados y localizaciones en reclamaciones sobre viviendas. Además, incorporan un tesauro especializado, que permite organizar y relacionar términos del sector. Los resultados muestran que los tesauros mejoran la identificación de información en textos con ruido lingüístico, aunque persisten dificultades para reconocer determinadas categorías como los defectos estructurales y los elementos eléctricos.

¿Y si el modelo aprende directamente el lenguaje de una disciplina?

Esta es la cuestión que exploran Brandsen et ál. (2022), quienes desarrollan un modelo BERT especializado a partir de aproximadamente 60.000 documentos arqueológicos en neerlandés. Su objetivo es reconocer automáticamente seis tipos de entidades, como lugares, periodos históricos, objetos, materiales y contextos arqueológicos. Los investigadores evalúan el efecto del pre-entrenamiento con textos arqueológicos y analizan si la incorporación de información de un tesauro especializado mejora adicionalmente el reconocimiento de entidades. Los resultados muestran que el modelo mejora significativamente cuando aprende el lenguaje propio de la arqueología, mientras que el tesauro no aporta mejoras adicionales relevantes. Sin embargo, siguen existiendo dificultades para distinguir conceptos ambiguos y clasificar correctamente las entidades más importantes.

¿Es posible analizar patrones, términos más importantes y relaciones?

El tercer ejemplo, desarrollado por Li et ál. (2026), va un paso más allá: no se limita a reconocer entidades, sino que también busca comprender cómo se relacionan entre sí. Para ello, combina BERT con otros métodos de aprendizaje profundo (BiLSTM, CRF y mecanismos de autoatención) para analizar textos sobre peligros geológicos. El modelo permite identificar información relevante y reconocer relaciones semánticas, incluso cuando aparecen separados por

numerosas palabras dentro de un documento. Esta capacidad resulta especialmente interesante para estudiar fenómenos complejos, en los que es necesario conocer no sólo qué factores intervienen, sino también cómo se relacionan.

¿Qué pueden aportar estas experiencias al estudio de la vulnerabilidad del patrimonio cultural?
Los textos analizados sugieren que el pre-entrenamiento de modelos BERT a partir de lenguaje especializado en patrimonio mejoraría la identificación automática de materiales, patologías, factores de deterioro, estados de conservación y localizaciones en informes técnicos, inventarios y documentos históricos. Además, el reconocimiento de relaciones entre estos elementos permitiría avanzar desde la simple recopilación de datos hacia una interpretación más completa de las condiciones de vulnerabilidad y sus posibles causas. La organización de esta información en bases de datos y su vinculación con los bienes patrimoniales facilitarían su incorporación a Sistemas de Información Geográfica (SIG), abriendo nuevas posibilidades para el análisis espacial de la vulnerabilidad y la identificación de patrones territoriales de riesgo.

BIBLIOGRAFÍA

- Abdullah, M.H.A., Aziz, N., Abdulkadir, S.J., Alhussian, H.S.A. y Talpur, N. (2023) Systematic literature review of information extraction from textual data: Recent methods, applications, trends, y challenges. *IEEE Access,* 11, 10535-10562. Disponible en: https://doi.org/10.1109/ACCESS.2023.3240898 [Consulta: 3/11/2026]
- Brandsen, A., Verberne, S., Lambers, K. y Wansleeben, M. (2022) Can BERT dig it? Named entity recognition for information retrieval in the archaeology domain. *Journal on Computing y Cultural Heritage,* 15(3), Article 51. Disponible en: https://doi.org/10.1145/3497842 [Consulta: 3/11/2026]
- Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H. y Pérez, J. (2023) Spanish pre-trained BERT model y evaluation data [Preprint]. *ArXiv. Disponible en:* https://doi.org/10.48550/arXiv.2308.02976 [Consulta: 03/11/2026]
- Devlin, J., Chang, M.-W., Lee, K. Y Toutanova, K. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. En: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long y Short Papers), pp. 4171-4186. Disponible en: https://doi.org/10.18653/v1/N19-1423 [Consulta: 03/11/2026]
- Grishman, R. (2019) Twenty-five years of information extraction. *Natural Language Engineering*, 25(6), pp. 677-692
- Jeon, K., Lee, G., Yang, S. y Jeong, H. D. (2022) Named entity recognition of building construction defect information from text with linguistic noise. *Automation in Construction,* 143, 104543
- Khalil, A., Stravoravdis, S. y Backes, D. (2021) Categorisation of building data in the digital documentation of heritage buildings. *Applied Geomatics*, 13, pp. 29-54. Disponible en: https://doi.org/10.1007/s12518-020-00322-7 [Consulta: 03/11/2026]
- König, F., Egger, A., Kratsch, W., Röglinger, M. y Wördehoff, N. (2025) Unstructured data in process mining: A systematic literature review. *ACM Transactions on Management Information Systems*, 16 (3), Article 25. Disponible en: https://doi.org/10.1145/3727148 [Consulta: 03/11/2026]
- Kudo, T. y Richardson, J. (2018) SentencePiece: A simple y language independent subword tokenizer y detokenizer for neural text processing. En *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations,* pp. 66-71. Disponible en: https://doi.org/10.18653/v1/D18-2012 [Consulta: 03/11/2026]
- Li, Y., Shao, Y., Luo, L. y Han, Z. (2026) A pipeline approach for entity y relationship extraction in geological hazard texts using BERT-BiLSTM-CRF with self-attention. *Intelligent Geoengineering*, 3(2), 16-178. Disponible en: https://doi.org/10.1016/j.ige.2026.06.008 [Consulta: 03/11/2026]
- Liddy, E.D. (2009) Natural language processing. En: *Encyclopedia of Library y Information Science.* 3.ª ed. CRC Press
- MathWorks (s.f.) *MATLAB Help Center*. Disponible en: https://se.mathworks.com/help/ [Consulta: 03/11/2026]
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. y Polosukhin, I. (2017) Attention is all you need. En: *Advances in Neural Information Processing Systems, v*ol. 30