III Congreso Interdisciplinar sobre Despoblación

Zaragoza, 5-7 de noviembre de 2025

¿De qué hablan los medios cuando hablan de despoblación? Una clasificación temática de noticias de prensa a partir de un modelo de inteligencia artificial

Pla-Bañuls, J. (1); Esparcia Pérez, J. (1); Martínez Rodrigo, A. (2).

(1) Universidad de Valencia. jaupla@uv.es; Javier.esparcia@valencia.edu

(2) Universidad de Castilla-La Mancha. arturo.martinez@uclm.es

Resumen:

La despoblación rural es un fenómeno demográfico y territorial complejo que está presente en la mayoría de los países europeos. La despoblación genera procesos de declive económico y desarticulación social, que a su vez resultan en un deterioro territorial. Este fenómeno se desarrolla dentro de un ciclo vicioso creciente de marginalización y despoblación rural. La despoblación, si bien ha sido ampliamente estudiada desde el ámbito geográfico, no lo ha sido tanto desde la Comunicación. Ante el papel clave de los medios en el fenómeno, se entrevé una potencial colaboración, estableciendo un "puente" entre ambas ciencias. Es por ello que nos acercamos a esta investigación desde el paradigma de la Geografía de la Comunicación, una línea de investigación innovadora que perfectamente puede adaptarse a cualquier investigación cuyo nexo en común sean ambas ciencias. En este sentido, desde hace aproximadamente una década, el fenómeno de la despoblación en España ha estado especialmente vinculado a los medios de comunicación. La cobertura mediática del declive demográfico aumentó considerablemente a partir de la Conferencia de Presidentes (17 de enero de 2017), a la que le siguió el nombramiento de la primera Comisionada del Gobierno para el Reto Demográfico. Todo ello coincidió también con divulgación del libro *España Vaciada* de Sergio del Molino (2016) que, en todo caso, tenía precedentes muy significativos en el tratamiento de diferentes aproximaciones a la desvitalización rural, como había sido el libro *La lluvia amarilla*, de Julio Llamazares (1988), o la serie *Un País en la mochila*, de José Antonio Labordeta. Aportaciones muy significativas fueron el documental *Tierra de nadie*, de Salvados (2017), o los reportajes de Jalis de la Serna sobre la España vaciada y las Highlands de Escocia (La Sexta, 2019). En este contexto, los medios de comunicación se hicieron eco ampliamente de la situación demográfica que experimentaban gran parte de los municipios de la España rural interior. Tanto la cobertura mediática como el creciente nivel de concienciación se realimentaron mutuamente durante aquellos años, de forma que el fenómeno de la -mal llamada- "España despoblada", o "España Vacía" adquirió una presencia especialmente relevante tanto en los medios de comunicación como entre amplias capas de la sociedad española (como también puso de relieve el Barómetro del CIS de febrero de 2019). Todo ello culminó en 2019 con la multitudinaria manifestación de Madrid, bajo el lema *La Revuelta de la España Vaciada*, organizada por distintas plataformas cívicas íntimamente relacionadas con la despoblación rural. Entre

ellas, cabe destacar Teruel Existe, que en las elecciones del mismo año consiguió un decisivo escaño en el congreso de los diputados (que, a la postre, utilizó su diputado como caja de resonancia de las reivindicaciones que venían planteando). Por tanto, queda patente la importancia que los medios han tenido a lo largo de esta última década en situar a la despoblación en la agenda política. Resulta de gran interés científico conocer cuáles son las temáticas clave que atiende la prensa cuando aborda la cuestión general de la despoblación, como han puesto de relieve diversos autores, con relación a las cuestiones rurales. Para este tipo de análisis hay diferentes estrategias, la primera y tradicional es la lectura, análisis y clasificación manual. Sin embargo, eso nos enfrenta a cientos, e incluso miles de piezas periodísticas, lo cual implica una inversión de tiempo muy considerable (lo que nos obligaría a trabajar con muestras más reducidas, con el riesgo de dejar fuera información relevante). No obstante, y dado el potencial de la Inteligencia Artificial (IA) para analizar textos, resulta viable adoptar una segunda estrategia basada en el uso de modelos de aprendizaje automático y representación semántica del lenguaje. En este enfoque, los modelos se entrenan a partir de un conjunto de noticias previamente categorizadas, permitiéndoles identificar patrones lingüísticos y temáticos en los datos. Esto posibilita una clasificación automática de las noticias por tema y, en fases más avanzadas, el desarrollo de herramientas que faciliten el análisis exploratorio del contenido, con el objetivo de ampliar su extrapolación a bases de datos más extensas Por lo tanto, la hipótesis principal de esta aportación es que, a partir de un entrenamiento supervisado basado en técnicas avanzadas de aprendizaje automático, un modelo de IA podría ser capaz de clasificar noticias de prensa según la temática predominante en su contenido. Para desarrollar esta hipótesis se plantean dos objetivos principales: a) Clasificar manualmente una muestra de noticias de prensa para, en primer lugar, poder entrenar al modelo de inteligencia artificial y, en segundo lugar, poder supervisar los resultados de su clasificación; b) Diseñar, entrenar y testar un modelo de inteligencia artificial que sea capaz de clasificar noticias de prensa a partir de la muestra de noticias previamente clasificada. Para ello, se ha realizado una descarga masiva noticias de prensa de periódicos regionales de tres comunidades autónomas: Comunidad Valenciana (*Las Provincias* y *El Periódico Mediterráneo*), Aragón (*El Heraldo de Aragón*) y Castilla y León (*El Norte de Castilla*). La muestra se ha obtenido a partir de la hemeroteca virtual MyNews delimitando las siguientes palabras clave (entre 1-01-1996 a 29-02-2024): *"despoblación" "despoblamiento" "despoblado" "despoblada" "España vacía" "España vaciada" "reto demográfico"*, que han sido utilizadas en trabajos previos (Saiz *et al.*, 2022). El resultado ha sido una muestra total de 4882 noticias que, tras la depuración de noticias duplicadas, ha quedado en 3493. El estudio de literatura científica, así como documentación diversa sobre cuestiones de despoblación nos han llevado, en primer lugar, a establecer una tipología de 11 categorías, que fue contrastada con la lectura y análisis de una muestra aleatoria inicial de noticias (en torno al 25 %). Con la tipología definida de manera sólida, se ha procedido, en segundo lugar, a la clasificación de la muestra completa (igualmente de forma manual). Para abordar el segundo de los objetivos (entrenamiento del modelo de inteligencia artificial), se ha realizado un preprocesado consistente en eliminar del texto de noticias (corpus) los signos de puntuación, números y caracteres especiales, además de convertir todo a minúsculas para evitar que palabras iguales, pero con diferente capitalización, fueran tratadas como distintas. Además, también se han eliminado las *stopwords* (palabras vacías como artículos, pronombres, preposiciones, etc.) en español. Posteriormente, el texto fue tokenizado y convertido en representaciones numéricas mediante un modelo Word2Vec, implementado en Python. Este modelo fue configurado para generar vectores de tamaño 200 para cada palabra, capturando su significado en función del contexto dentro de una

ventana de 10 palabras a la izquierda y a la derecha. Finalmente, cada noticia fue representada como el promedio de los *embeddings* de sus palabras, generando una matriz donde cada fila correspondía a una noticia y cada columna a una dimensión del espacio de *embeddings*. El Word2Vec es un algoritmo de aprendizaje profundo desarrollo por Google (2013), que permite obtener el contexto de cada palabra en el texto, a diferencia de otras técnicas como los *bags of words* (frecuencias de palabras) o Term Frequency - Inverse Document Frequency (importancia de palabras dentro e un texto). Existen otros métodos más actuales y potentes, pero Word2Vec sigue siguiendo muy útil para modelos ligeros, como con el que aquí se trabaja. Una vez las noticias se han convertido en matrices, se ha pasado a la fase de clasificación. Se han testado dos modelos: Random Forest (RF) y Support Vector Machines (SVM), siendo este último el que se adecúa mejor a nuestras necesidades. Se ha optado por una estrategia estable de validación múltiple *Hold-Out*, dividiendo los datos en 80% para entrenamiento y 20% para prueba, repitiendo este proceso en 10 iteraciones (donde en cada una de ellas se realizaba una nueva mezcla de los datos. De esta forma, se refuerza la estabilidad y robustez de cada modelo, reduciendo la posibilidad de que los resultados estén sesgados por una única partición de los datos. Para garantizar una determinada eficiencia del modelo se han usado las tres categorías con mayor representatividad, es decir, aquellas que tuvieran al menos 1000 noticias, en este caso, actores institucionales, acciones contra la despoblación y servicios. Para evaluar el rendimiento de cada modelo se utilizaron la precisión, Recall y F1-score, tanto a nivel global como por categoría, para determinar su capacidad de clasificación. En primer lugar, la precisión indica el porcentaje de veces que el modelo asigna correctamente una categoría cuando decide etiquetar una noticia en ella. Es decir, mide qué tan exacto es el modelo al predecir una categoría y qué tan bien evita asignaciones incorrectas. En segundo lugar, el Recall mide cuántas de las noticias que realmente pertenecen a una categoría fueron correctamente detectadas por el modelo. Un Recall bajo significa que el modelo está dejando muchas noticias sin etiquetar en su categoría correcta, lo que se traduce en falsos negativos. Finalmente, el F1-score es la media armónica entre precisión y Recall, proporcionando un balance entre falsos positivos y falsos negativos. Los resultados obtenidos con el modelo SVM muestran variaciones en el rendimiento según la categoría, lo que confirma que la distribución de noticias entre clases sigue siendo un factor clave en la clasificación. En términos generales, el modelo logra un mejor desempeño en la categoría con mayor cantidad de noticias, mientras que en aquellas con menos datos, el rendimiento es inferior, presentando un mayor número de errores. Esto sugiere que el tiempo de entrenamiento no ha sido suficiente para que el modelo aprenda patrones robustos en las categorías menos representadas. Globalmente, los resultados preliminares son altamente positivos. En promedio, el modelo logra clasificar correctamente alrededor del 75 % de las noticias en una de las tres categorías. Aunque los valores medios de Recall y F1-score son ligeramente inferiores (72 % en ambos casos), destaca el rendimiento en la categoría de actores institucionales, donde el Recall alcanza un 90 % y el F1-score un 81 %, lo que indica que prácticamente no hay falsos negativos en esta clase. Además, los resultados reflejan una notable estabilidad del modelo, ya que la variabilidad de las métricas se mantiene dentro de un margen del 5 %, lo que sugiere que el modelo generaliza bien y su rendimiento no fluctúa significativamente entre iteraciones. Sin embargo, dado que los resultados que se están obteniendo son muy prometedores, la estrategia más adecuada es la de aumentar la base de datos con noticas clasificadas manualmente, para mejorar el entrenamiento del modelo. De esta forma se podría, por un lado, aumentar la precisión y fiabilidad global de los resultados obtenidos y, por otro, seguramente tendríamos más categorías temáticas dentro del umbral que hemos establecido para considerar muy fiable la clasificación

(umbral que, por lo demás, es muy exigente). Se prevé explorar estrategias complementarias más avanzadas, que, aunque son más complejas, pueden ser especialmente útiles para el análisis e interpretación de grandes volúmenes de información, como las noticias en los medios de comunicación. Entre ellas, la clasificación por clústeres permitirá profundizar en el análisis a partir de los resultados obtenidos hasta ahora. Asimismo, se investigarán otras aproximaciones que requieren técnicas y metodologías distintas, como la clasificación sin una tipología temática predefinida, lo que permitiría descubrir patrones emergentes sin depender de categorías establecidas previamente. Para concluir, aunque esta comunicación se centra en la aplicación de modelos de inteligencia artificial al estudio de la despoblación, es evidente que estamos en un proceso de aprendizaje que permitirá desarrollar procedimientos trasladables a otros ámbitos científicos, basados en bases de datos masivas con información textual. En nuestro ámbito de conocimiento, de hecho, podremos trabajar con explotación de información centrada en aproximaciones más globales, pero también más complejas, como los propios procesos de desarrollo y cambio en las áreas rurales.

Palabras clave:

Inteligencia Artificial, Modelos de Aprendizaje Profundo, despoblación, medio rural, medios de comunicación.

Área temática:

Medios de comunicación, Nuevas tecnologías.

¿Cómo realizará la presentación de su comunicación en caso de ser aceptada?

Comunicación oral