eib0915853805

Caracterización de riesgos urbanos en prensa aplicando minería de texto para el enriquecimiento de datos abiertos

Luis M. Vilches-Blázquez*
Diana Comesaña Ocampo**

Artículo recibido:
25 de octubre de 2021
Artículo aceptado:
14 de marzo de 2022

Artículo de investigación

Resumen

Las noticias se difunden libremente y con amplia disponibilidad para los usuarios de internet con mucha más facilidad que con los medios tradicionales. En estas noticias se pueden encontrar infinidad de "datos menores" ocultos que pueden suministrar valiosa información no recogida en otras fuentes de información. En este contexto, en este artículo nos ha interesado analizar y caracterizar los riesgos urbanos recogidos en prensa abierta en el contexto nacional uruguayo utilizando técnicas de minería de texto. Esta propuesta permite conformar un *corpus*

* Centro de Investigación en Computación, Instituto Politécnico Nacional, México
lmvilches.blazquez@gmail.com lmvilches@cic.ipn.mx
* Departamento de Tratamiento y Transferencia de Información, Instituto de Información, Facultad de Información y Comunicación-Universidad de la República de Uruguay
diana.comesana@fic.edu.uy

de noticias que parte de eventos de riesgo contenidos en datos abiertos. El *corpus* abarca el periodo 2003-2019 y proviene de periódicos digitales abiertos *(El Eco Digital, Montevideo Portal* y *La Red 21)*. Sobre este *corpus* se aplican diversas técnicas de minería de texto empleando el software QDA-MinerLite y el lenguaje Python (mediante la librería Scattertext) para identificar, caracterizar y descubrir conocimiento sobre estos eventos. Los resultados obtenidos del procesamiento del *corpus* ayudan a enriquecer los datos abiertos existentes sobre riesgos en Uruguay, incorporando información sobre sus efectos, actores e intervenciones asociadas.

Palabras clave: Riesgo Urbano; Minería de Texto; Prensa Digital Abierta; Datos Abiertos

Characterization of urban risks in the press applying text mining for the enrichment of open data

Luis M. Vilches-Blázquez and *Diana Comesaña Ocampo*

Abstract

News is freely spread and widely available to Internet users much more easily than traditional media. In the news, we can find an infinite number of hidden "minor data," that can provide valuable information not collected in other sources of information. In this context, we have been interested in analyzing and characterizing the urban risks contained in the Uruguayan open newspapers using text mining techniques. This proposal makes it possible to create a news corpus based on risk events included in open data. The corpus covers 2003-2019 and is built from the digital open newspapers *El Eco Digital, Montevideo Portal,* and *La Red 21.* Various text mining techniques are applied to this corpus using the QDA-MinerLite software and the Python language (concretely, through the Scattertext library) to identify, characterize, and discover insights on these events. The corpus processing results help enrich the existing open data on risks in Uruguay, incorporating information on their effects, actors, and associated interventions.

Keywords: Urban Risk; Text Mining; Open Digital Newspapers; Open Data

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58538

Introducción

Actualmente, las noticias se difunden con libertad y están ampliamente disponibles para los usuarios de internet con mucha más facilidad que antes, con los medios tradicionales. Así, "los medios de comunicación son cada vez más un foro de información y debate con un flujo de información no lineal y periodismo de código abierto" (Mhamdi, 2016: 272). Existen infinidad de "datos menores" ocultos en las noticias de los periódicos al ser considerados una fuente de información pública (McCallum, Hammond y Covello, 1991), por lo que pueden suministrar valiosa información no recogida en otras fuentes de información (Vilches-Blázquez, Comesaña y Arrieta-Moreno, 2020).

Para obtener esa riqueza de información oculta en las noticias, la minería de textos se ha convertido en una de las áreas de investigación de moda, incorporando diferentes campos y técnicas en el ámbito de recuperación textual y lingüística computacional, tales como la recuperación de información, procesamiento de lenguaje natural, minería de datos, aprendizaje automático o gestión del conocimiento (Feldman y Sanger, 2007; Salloum-Said *et al*., 2017).

En este contexto, nos interesa analizar y caracterizar los riesgos urbanos utilizando técnicas de minería de texto sobre prensa y datos abiertos de Uruguay. Este interés se enmarca en el hecho de que la mitad de la humanidad, unos 3 500 millones de personas, vive actualmente en ciudades, y se estima que en 2050 70 % de la población mundial residirá en zonas urbanas (ONU, 2018). Este crecimiento urbano está provocando que muchas ciudades sean más vulnerables al cambio climático y a los desastres naturales. De esta situación se hace eco la *Agenda 2030 para el Desarrollo Sostenible* de la ONU, incluyendo entre sus objetivos "lograr que las ciudades y los asentamientos humanos sean inclusivos, seguros, resilientes y sostenibles" y "combatir el cambio climático y sus efectos".

En este nuevo escenario, las ciudades necesitan identificar aquellos riesgos que las hacen vulnerables. No todas las ciudades parecen estar preparadas para esta nueva realidad. Particularmente, en el caso de las ciudades de Uruguay, a pesar de la creación del Sistema Nacional de Emergencia de Uruguay (SINAE) con el objetivo de "planificar, coordinar, ejecutar, conducir, evaluar y entender en la prevención y en las acciones necesarias en todas las situaciones de emergencia, crisis y desastres excepcionales o situaciones similares, que ocurran o sean inminentes" (SINAE, 2020: 12), se reconoce que no se cuenta con un historial exhaustivo de eventos adversos ni con tradición en materia de gestión de riesgos, que existen fallas en la comunicación y falta de conocimiento del público e, incluso, por parte de algunos entes públicos (SINAE,

2020: 19). En este sentido, existe la necesidad de conocer y caracterizar aquellos riesgos que afectan y hacen vulnerables a las ciudades de Uruguay.

La contribución de este artículo se centra en la caracterización y enriquecimiento de riesgos urbanos con "datos menores" extraídos de noticias publicadas en prensa digital abierta. Para ello, se combinan prensa y datos abiertos relacionados con eventos de riesgo en el contexto nacional uruguayo. En esta dualidad, los datos abiertos son el punto de partida para conformar un *corpus* de noticias de periódicos abiertos donde se mencionan estos eventos. Sobre este *corpus*, compuesto por noticias comprendidas entre 2003-2019, se realiza un tratamiento aplicando técnicas de minería de texto cuyos resultados permiten identificar, caracterizar y descubrir conocimiento sobre diversos riesgos presentes en Uruguay, así como enriquecer los datos abiertos inicialmente considerados, con información sobre sus efectos, autores e intervenciones asociadas. Este proceso de enriquecimiento de los datos abiertos, que son la base del inicio de este trabajo, se hace aún más necesario al comprobar la escasez y limitado detalle de los datos abiertos existentes sobre riesgos urbanos en Uruguay. Esta situación y las características de los datos abiertos disponibles fue lo que dio origen al trabajo descrito en este artículo.

Algunos trabajos han abordado el análisis de la dinámica de diferentes riesgos a partir de datos abiertos (Orlecka-Sikora *et al.*, 2020; Paprotny *et al.*, 2020; Lasaponara *et al.*, 2017; Cooper, 2014; Vescoukis y Bratsas, 2014), artículos de prensa (Yagoub *et al.*, 2020; Llasat *et al.*, 2009; Wakefield y Elliott, 2003), o incluso, aunque en menor medida, combinando ambas fuentes (Adelekan, 2020). Aunque la minería de texto ha sido utilizada para diversos propósitos (Da Silva y Godoy Viera, 2021) y algunas de las propuestas mencionadas han analizado los eventos de riesgos urbanos utilizando técnicas de minería de texto, en la literatura relacionada no hemos identificado ningún trabajo donde la aplicación de técnicas de minería de texto sea utilizada para descubrir conocimiento sobre eventos de riesgo y, con ello, enriquecer datos abiertos a través de la información recuperada de un *corpus* de noticias de periódicos digitales abiertos.

Fundamentación teórica

El concepto de minería de texto no es nuevo. Surge a comienzos de la década de 1980 con los primeros intentos de realizar minería de textos y el esfuerzo humano que significaba. Witten, Eibe y Mark (2011) lo definen como el área que busca hallar patrones en el texto. Para alcanzar dicho objetivo, analiza textos para extraer informaciones que sean útiles para un propósito particular.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58538

Según Hearst (1999), este tipo de minería "tiene como objetivo descubrir información y conocimiento que previamente se desconocía y que no aparecía en ninguno de los documentos analizados".

En la literatura existen múltiples definiciones de este campo de investigación, entre las que destacamos la propuesta por Sullivan (2001), afirmando que "la minería textual es el proceso de compilar, organizar y analizar grandes colecciones de documentos para apoyar en la distribución de información a los analistas y a las personas encargadas de tomar decisiones, y para descubrir relaciones entre hechos relacionados que se reparten entre distintos dominios de investigación", y la planteada por Brun y Senso (2004), que define a la minería de texto como "una aplicación de la lingüística computacional y del procesamiento de textos que pretende facilitar la identificación y extracción de nuevo conocimiento a partir de colecciones de documentos o *corpus* textuales. Recoge distintas técnicas formuladas en el ámbito de la recuperación textual o *text retrieval* y la lingüística computacional".

La minería de texto proviene en gran parte de las investigaciones en minería de datos y, por lo tanto, tienen similitudes en su arquitectura de alto nivel; por ejemplo, ambos sistemas se basan en rutinas de preprocesamiento, algoritmos para descubrir patrones y la capa de elementos de presentación que contienen herramientas de visualización para mejorar la navegación en los conjuntos de respuestas (Viera, 2017). Sin embargo, es importante diferenciar la minería de texto de la minería de datos, aunque en algunas ocasiones la primera se considere complemento de la segunda. Viera (2017) describe en detalle las diferencias entre ambos enfoques de minería.

Considerando que la minería de texto se enfoca en el análisis de textos en lenguaje natural, esto hace que incluya otras disciplinas de las ciencias de la computación que trabajan con el manejo de lenguaje natural. Además, este tipo de minería utiliza, principalmente, técnicas y metodologías de las áreas de recuperación de la información, extracción de información y lingüística computacional (Feldman y Sanger, 2007).

De forma generalizada, cuando se aborda el trabajo de minería de texto se llevan a cabo las siguientes tareas (Brun y Senso, 2004; Contreras-Barrera, 2014):

1. Recolección de datos de diferentes recursos documentales. Este proceso puede ser completamente automatizado o semiautomático.
2. Preprocesamiento. Comprende la limpieza de textos para eliminar o depurar la información innecesaria o no deseada.
3. Extracción de características y análisis. La extracción permite la recuperación de características mediante la utilización, por ejemplo, de análisis léxico, tratamiento y separación de palabras vacías (artículos,

preposiciones, conjunciones), tratamiento de palabras compuestas, normalización de palabras, obtención de las raíces de las palabras, etiquetado de palabras, etc. Además, el análisis de clases, relaciones, asociaciones o secuencias permite la identificación de patrones a través de la aplicación de diferentes técnicas.

4. Presentación de resultados. En esta etapa se muestran los resultados a través de diferentes formas (por ejemplo, resúmenes, relaciones, taxonomías, etc.) y se construyen visualizaciones para facilitar su interpretación.

Metodología

La necesidad de identificar y caracterizar los diversos riesgos que hacen vulnerable a Uruguay nos conduce al estudio de diversos recursos de información y a la utilización de técnicas de minería de texto. Para ello, junto a las lecturas técnicas, se utilizan el software QDA-MinerLite y el lenguaje Python, así como la librería *Scattertext*. A continuación, se presentan los detalles de las fuentes de datos consideradas y la metodología aplicada para el desarrollo de este trabajo.

Fuentes de datos

Las fuentes de datos consideradas en este trabajo se caracterizan por ser abiertas y se asocian a datos y prensa digital. Seguidamente, se proporcionan las principales características de las fuentes consideradas.

Datos abiertos. Son datos obtenidos del SINAE (https://www.gub.uy/sistema-nacional-emergencias/gestion-integral-riesgos), del Instituto Uruguayo de Meteorología (INUMET, https://www.inumet.gub.uy/tiempo/historico-alertas-meteorologicas) y del Ministerio de Ambiente (https://www.gub.uy/ministerio-ambiente/politicas-y-gestion/atlas-nacional-inundaciones-drenaje-pluvial-urbano-version-07-2020). Del SINAE se trabajó con un catálogo histórico de eventos meteorológicos correspondientes al periodo 1968-2011 (SINAE, 2016), mientras que del INUMET se recopiló un histórico de alertas meteorológicas (2017- 2021) y una zonificación de riesgo de incendios forestales. Del Ministerio de Ambiente se obtuvo una clasificación de poblaciones por riesgo de inundación.

En estos datos abiertos identificamos que sobre los riesgos no figuran datos específicos, excepto un mapa que clasifica las zonas de riesgo de incendios forestales en determinados periodos de tiempo. Asimismo, en estos

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58538

datos no aparecen detalles de la percepción social sobre la fragilidad ante los riesgos, actores implicados o recursos y población afectada. En definitiva, comprobamos que los datos abiertos existentes son escasos y con limitado nivel de detalle, reflejando la disponibilidad actual de datos sobre riesgos urbanos en Uruguay.

Prensa digital abierta. El *corpus* de noticias correspondió a casi 20 años de acontecimientos sobre riesgos (2003- 2019), conformado por 1 700 artículos de tres periódicos digitales abiertos con cobertura nacional en Uruguay: *Montevideo Portal* (https://www.montevideo.com.uy/), *El Eco Digital* (https://elecodigital.com.uy/) y *La Red 21* (https://www.lr21.com.uy/). Estos medios fueron seleccionados por su carácter abierto y entre ellos se encuentra el primer medio de información enteramente digital que surgió en Uruguay en 1995: *Montevideo Portal.*

Desde una perspectiva general, ambas fuentes de datos presentan cobertura nacional y sobre ellas no se realiza ningún filtrado espacial para la obtención o clasificación de riesgos. Por otro lado, las diferentes fuentes abarcan distintos periodos de tiempo, aunque con una amplia superposición temporal. Esta diferencia se debe a que los datos abiertos presentan información con un componente histórico sobre los riesgos aparecidos en la zona de estudio, lo que resultó de gran interés para ver la presencia y tipologías de riesgos. Por otro lado, la consideración de prensa digital abierta limita la temporalidad de periódicos con los que se podían tratar, dado que el periódico más antiguo de esta tipología (*Montevideo Portal*) en el contexto uruguayo es 1995. Sin embargo, no es sino hasta 2003 cuando empiezan a presentarse noticias relacionadas con riesgos en la prensa digital abierta considerada.

Etapas de la investigación

Esta investigación se compone de tres actividades, orientadas a i) la conformación del *corpus* de noticias, ii) su tratamiento, y iii) la caracterización y enriquecimiento de los eventos de riesgo. A continuación se describen los detalles de cada una de las actividades desarrolladas.

i) Conformación del *corpus*. El *corpus* de noticias se conforma aplicando un vocabulario primario que permite una mejor identificación de las noticias pertinentes para este trabajo. Este vocabulario primario está conformado por la terminología obtenida de los datos abiertos considerados, poniendo especial atención a los datos del SINAE ya que en ellos aparecen los riesgos más probables en Uruguay, y por la terminología meteorológica desarrollada por Vilches-Blázquez, Comesaña y Arrieta-Moreno (2020).

Para obtener este vocabulario se realiza un análisis de dominio (AdD) basado en la propuesta de Hjørland y Albrechtsen (1995). Este paradigma

plantea estudiar los dominios del conocimiento como comunidades discursivas, tomando en consideración el contexto psicosocial y sociolingüístico de la sociología del conocimiento y de la ciencia. En este análisis se combinan los enfoques de estudios terminológicos y estudios de usuarios empíricos propuestos por Hjørland (2002) aplicados al contexto de los riesgos urbanos de Uruguay.

En el vocabulario conformado producto del AdD aparecen términos como dengue, zika, chikungunya, eventos meteorológicos adversos, incendios, leishmaniasis, ola de calor, ola de frío, temporal, tornado, inundación, etc. Esta terminología se utiliza para abordar un proceso de búsqueda manual en los repositorios de la prensa digital de *Montevideo Portal*, *El Eco Digital* y *La Red 21*. Sin embargo, debe tenerse en cuenta que la prensa digital abierta en Uruguay no está indizada sistemáticamente por profesionales de la información y, por tanto, carece de un criterio o normativa para el archivo de sus noticias. Esto significa que un tipo de acontecimiento puede estar etiquetado en un mismo archivo por diferentes sinónimos o por los efectos que provoca. Por ejemplo, un "temporal", puede hallarse etiquetado como "tempestad" o por sus efectos: "corte de rutas", "corte de caminos", "caída de árboles" o "voladuras de techo".

Considerando esta situación, decidimos construir relaciones entre términos para enriquecer y optimizar la búsqueda de resultados en la prensa digital abierta. Para ello, sobre los resultados obtenidos en las búsquedas manuales, se seleccionó una muestra de noticias de dos años (correspondiente a 15 %, aproximadamente, de las noticias del periodo 2003-2019) y se procedió a realizar lecturas técnicas de las noticias. El resultado de estas lecturas permitió enriquecer el vocabulario inicial y establecer los efectos detectables de cada riesgo dentro del *corpus* de noticias. Además, estas lecturas permitieron organizar los riesgos en tres categorías (zoonosis, eventos meteorológicos extremos e incendios forestales) y establecer relaciones entre los términos vinculados a los riesgos para optimizar la búsqueda de artículos en los archivos de la prensa digital abierta considerada.

ii) Tratamiento del *corpus*. Una vez recuperadas las noticias asociadas con la terminología conformada, se aborda el tratamiento del *corpus* de noticias mediante la aplicación de técnicas de minería de texto (Gupta y Lehal, 2009; Feldman y Sanger, 2007) sobre el conjunto de noticias de la prensa abierta considerada.

El *corpus* de noticias conformado fue descargado en formato PDF y se aplicó minería textual para su tratamiento utilizando el software QDA-MinerLite y el lenguaje Python. Es importante destacar que se realizó una recuperación textual de los temas de interés para una posterior codificación,

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58538

prestando especial atención a que no se produjeran duplicados de un hecho. Así, se asignaron códigos a los archivos (noticias) y no a los párrafos de las noticias para facilitar la recuperación de cualquier elemento del *corpus* vía códigos. Los códigos asociados corresponden al vocabulario establecido y la recuperación de los elementos del *corpus* permite extraer la categoría (grupo de términos), código (término descriptivo del riesgo), caso o archivo al que pertenece la noticia de la prensa digital y frase a que se aplica la codificación.

Adicionalmente, realizamos un análisis temporal de la presencia (frecuencias) con la que aparecieron noticias sobre los riesgos considerados en el *corpus* y utilizamos técnicas de minería asociadas con el lenguaje natural (tokenización y lematización) para encontrar palabras y frases que discriminen las categorías de texto presentes en nuestro *corpus*, permitiendo construir diversas visualizaciones aplicando diferentes modelos.

Con este marco, se realiza el tratamiento semiautomático del *corpus* para caracterizar los riesgos en Uruguay a través de su cubrimiento en la prensa y poder contribuir al enriquecimiento de los datos abiertos sobre estos eventos.

iii) Caracterización y enriquecimiento. El tratamiento del *corpus* de noticias nos permite acceder a datos que no están contenidos en los conjuntos de datos abiertos disponibles. En este sentido, con este trabajo, enriquecemos los eventos de riesgo presentes en los datos abiertos con la información obtenida de nuestro *corpus*, incorporando a los datos abiertos información sobre actores que intervinieron, afectados, acciones oficiales, etc.

Resultados y discusión

En esta sección se describen los resultados de esta investigación sobre la caracterización de los riesgos urbanos. Para ello, se abordan los detalles relacionados con la conformación y tratamiento del *corpus*, así como la caracterización y el enriquecimiento de eventos de riesgos urbanos presentes en Uruguay.

Conformación del corpus

Como se mencionó en la sección "Etapas de la investigación", el proceso de conformación del *corpus* permitió organizar los riesgos en tres categorías (zoonosis, eventos meteorológicos extremos e incendios forestales) y establecer relaciones entre los términos vinculados a los riesgos. La *Figura 1* muestra las relaciones establecidas entre las categorías de riesgo estudiadas con los sinónimos hallados y el primer nivel de relación causa-efecto. Las relaciones entre

las diferentes categorías de riesgos y sus términos se obtuvieron para el caso de zoonosis (dengue, chikungunya, zika y leshmaniasis) directamente de sus definiciones, mientras que para los eventos meteorológicos se utilizó la terminología propuesta en Vilches-Blázquez, Comesaña y Arrieta-Moreno (2020).

Este trabajo permitió conformar un *corpus* de 1 700 noticias provenientes de los periódicos abiertos digitales mencionados, las cuales se dividen en 1 231 noticias de *Montevideo Portal*, 387 pertenecientes a *La Red 21* y 82 fueron recuperadas de *El Eco Digital*.

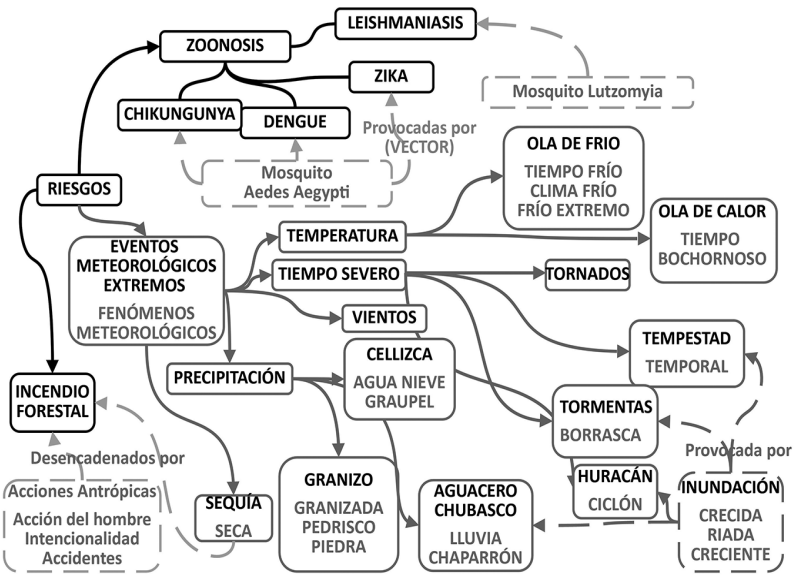

Figura 1. Relaciones entre términos del AdD

Tratamiento del corpus

La aplicación del proceso de codificación a los archivos (noticias) permitió obtener la frecuencia con que la prensa trató cada evento de riesgo. Así, identificamos que las noticias del *corpus* se distribuyen conforme a tres categorías de riesgos identificadas: eventos meteorológicos (85 %), incendios forestales (11 %) y zoonosis (4 %). La *Figura 2* muestra la distribución de artículos por riesgo y periódico considerado. En dicha figura podemos identificar que los riesgos más frecuentes son las tormentas y lluvias *(Montevideo Portal*), sequía,

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58538

inundaciones y tormentas (*El Eco Digital*) y lluvias (*La Red 21*). En el otro extremo, es decir, entre los riesgos menos presentes en el *corpus* de noticias aparecen granizo y ciclón extratropical (*Montevideo Portal*), dengue, zika u ola de calor (*El Eco Digital*) y leptospirosis (*La Red 21*). Asimismo, descubrimos que un elemento relevante en las noticias de los tres periódicos digitales abiertos considerados son los aspectos relacionados con el apoyo ante desastres. Por otro lado, identificamos que la sequía presenta una gran disparidad en la noticias de nuestro *corpus*, ya que en el periódico *El Eco Digital* es uno de los riesgos más mencionados (14 noticias asociadas), mientras en *La Red 21* y *Montevideo Portal* tienen una presencia testimonial (una noticia).

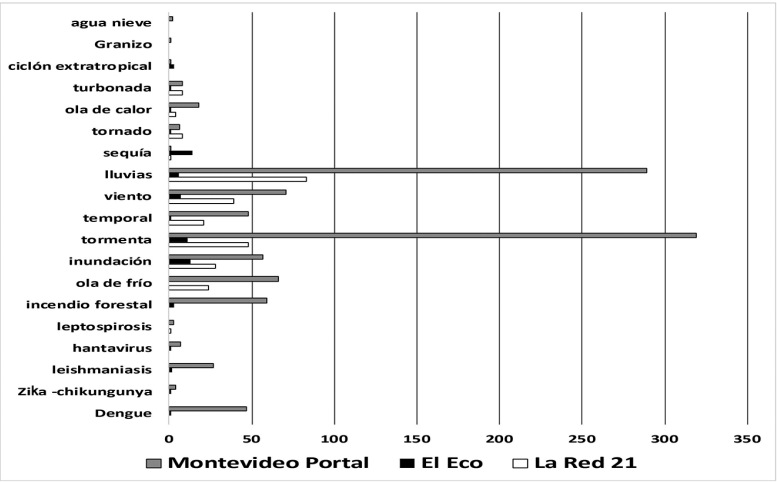

Figura 2. Distribución de artículos por riesgo y prensa

Adicionalmente, realizamos un análisis temporal de la presencia (frecuencias) con la que aparecieron noticias sobre los riesgos considerados en el *corpus* (*Tabla 1*) y al analizar su distribución identificamos que las lluvias, como uno de los riesgos más persistentes, tienen dos hitos importantes, con un máximo en 2004 (55 % de noticias ese año) y un mínimo en 2015 (6.6 % de noticias en el año mencionado). Por otro lado, vientos y tormentas presentan un comportamiento por lo general estable en cuanto al volumen de noticias durante el periodo estudiado. Sin embargo, las inundaciones se caracterizan por un comportamiento de dientes de sierra, mostrando una gran oscilación de noticias año tras año.

Este análisis temporal también nos ayuda a identificar cómo evolucionan los riesgos (*Figura 3*). Así, por ejemplo, vemos que riesgos "tradicionales" como

lluvia tienen una presencia constante y relevante a lo largo del periodo analizado, aunque presentan un importante descenso entre 2013-2016. Asimismo, resulta curioso que, en ese mismo periodo de descenso de las noticias relacionadas con lluvias, se produce un incremento de los riesgos relacionados con inundaciones y tormentas. Por otro lado, en este análisis descubrimos la aparición de nuevos riesgos como el dengue, zika o hantavirus en los últimos años.

Por otro lado, utilizamos técnicas de minería de texto asociadas con el lenguaje natural para encontrar palabras y frases que discriminen las categorías de texto presentes en nuestro *corpus*. En esta área se han utilizado múltiples visualizaciones para resaltar palabras discriminatorias, tales como listas, nubes o burbujas de palabras, así como diagramas de dispersión basados en palabras. Estas técnicas tienen varias limitaciones como, por ejemplo, la dificultad de comparar las frecuencias relativas de dos términos en una nube de palabras o de mostrar etiquetas de términos de forma legible en diagramas de dispersión (Kessler, 2017).

Considerando estas limitaciones, decidimos utilizar *Scattertext*, una librería de visualización de texto de Python que traza un conjunto de unigramas y bigramas (a los que se hace referencia en este artículo como términos) sobre el *corpus* de noticias conformado, asignando los términos a una de dos categorías en un diagrama de dispersión bidimensional. La *Figura 4* muestra el diagrama *Scattertext* resultante de comparar eventos meteorológicos y efectos del *corpus* de noticias. En esta figura, cuanto más arriba está un punto (término) en el eje Y, más se asocia a los efectos (por ejemplo, términos como inundación o vida) y cuanto más a la derecha aparece un punto en el eje X, más se asocia el término con los eventos meteorológicos (por ejemplo, precipitaciones, granizo, etc.). Los términos altamente asociados se sitúan más cerca de las esquinas superior izquierda e inferior derecha del gráfico, mientras que las palabras vacías (menos relacionadas) se encuentran en la esquina superior derecha y las palabras compartidas entre eventos meteorológicos y efectos se sitúan en la parte central del gráfico. Las palabras que ocurren con poca frecuencia en ambas clases caen más cerca de la esquina inferior izquierda (por ejemplo, velocidad, magnitud, temor, etc.).

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Dengue	-	-	1.3	-	-	-	-	-	-	-	-	-	-	16	7.9	1.6	0.8
Zika - Chikungunya	-	-	-	-	-	-	-	-	-	-	-	-	-	1.1	-	-	-
Leptospirosis	-	-	-	-	-	-	-	-	-	1	-	-	-	-	2.6	-	-
Hantavirus	-	-	-	-	-	-	-	-	-	-	-	0.7	-	-	2.6	0.8	0.8
Leishmaniasis	-	-	-	-	-	-	-	-	1	-	-	0.7	-	1.7	-	-	-
Incendio forestal	-	-	1.3	0.9	-	4	0.8	2.5	2.2	4.3	-	1.3	6.6	2.3	4	0.8	1.6
Lluvia	12	55	24	27.1	32	41	34.4	37.8	40.2	32.4	33.3	12.2	6.6	26.4	27.3	34.3	28
Viento	34	10	24.4	28	16	13	14.8	11.8	12	17.2	11.2	14.2	10	12	16	22.2	12
Tormenta	15	15	15	18	12	22	13.2	18	15.2	20.4	26.3	43.2	23.4	24.1	7	36.7	21.6
Temporal	5	12	3.5	0.9	-	-	6.5	0.7	2.2	5.4	-	0.7	-	0.6	5	-	2.4
Tornado	-	-	-	0.9	-	1	-	-	-	1	-	-	-	3.4	-	0.4	-
Turbonada	1.5	-	-	3.7	1.3	-	0.8	1.4	2.2	2.2	2.7	0.7	-	2.3	-	-	0.8
Ola de frío	1.5	5	6.7	11.3	9.5	10	5	13.2	16	-	-	-	-	-	2.6	0.4	7.2
Ola de calor	4	-	1.3	1.8	4	4	1.6	-	1	-	8.6	0.7	10	2	2.6	-	5.6
Granizo	4	-	-	2.8	2	2	0.8	0.7	1	-	2.7	-	-	1.1	4	-	3.2
Inundación	23	-	19	4.6	22	3	21.3	13.2	6	12.9	13.8	25.6	43.4	7	18.4	2.4	15.2
Sequía	-	3	-	-	-	-	0.8	0.7	1	2.2	1.4	-	-	-	-	-	-
Ciclón extratropical	-	-	3.5	-	0.6	-	-	-	-	-	-	-	-	-	-	0.4	-
Agua nieve	-	-	-	-	0.6	-	-	-	-	1	-	-	-	-	-	-	0.8

Tabla 1. Distribución porcentual del registro anual de riesgos

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58538

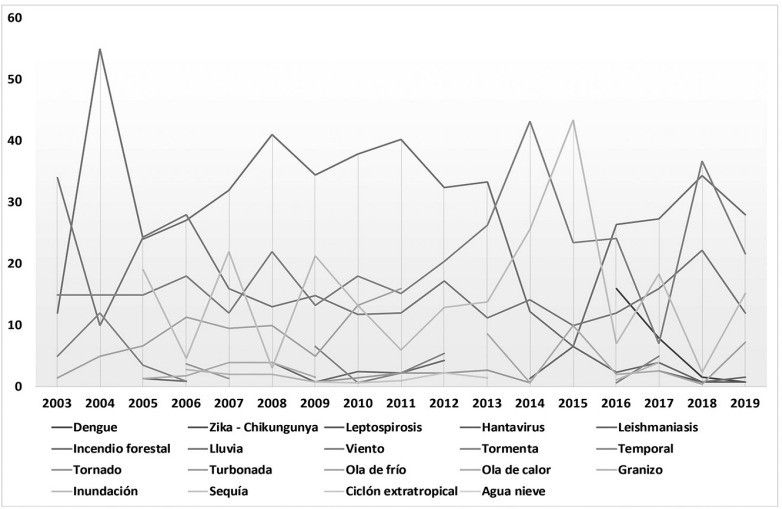

Figura 3. Distribución temporal de riesgos

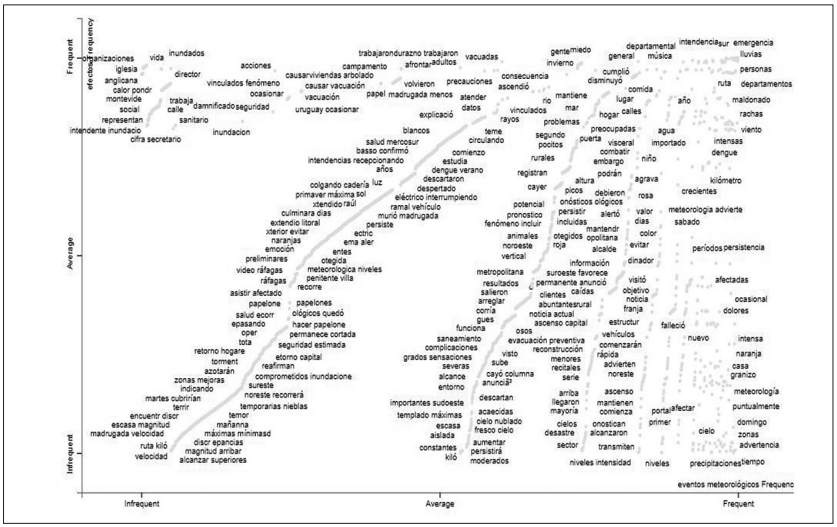

Figura 4. Visualización *Scattertext* del conjunto del *corpus*

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58538

Adicionalmente, aplicamos la visualización *Scattertext* con el modelo Word-2Vec, que permite generar representaciones de palabras distribuidas, conocidas como *word embeddings* (Li *et al*., 2015), a través de la aplicación de dos modelos de redes neuronales (Mikolov *et al*., 2013). La *Figura 5* muestra el resultado de la visualización a partir de la consideración de los eventos ola de calor y ola de frío. Los términos más frecuentes presentes en nuestro *corpus* son territorio, evacuados o precipitaciones para el riesgo ola de calor, mientras que para ola de frío algunos de los términos más frecuentes son tormenta, fuerte o vientos.

Conforme a la distribución en los ejes de la figura, encontramos que los términos más vinculados al fenómeno de olas de calor (margen superior izquierdo) son advertencias, personas o rachas, mientras que semana, oscilar o precipitaciones se vinculan a ola de frío. En el centro de la figura aparecen los términos comunes que aparecen cuando se tratan noticias referentes a estos riesgos, entre ellas aparecen palabras como lluvia, puntualmente, país, situación, Uruguay o Montevideo.

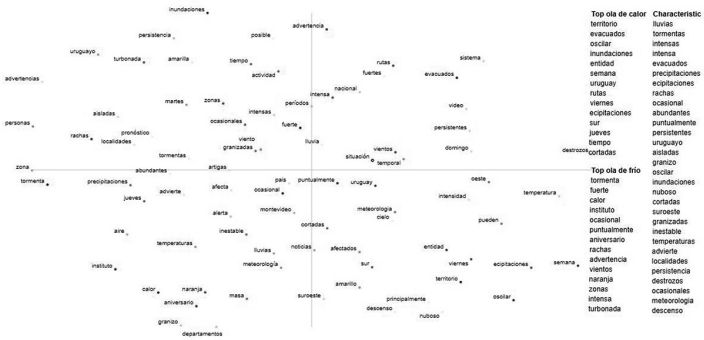

Figura 5. Visualización *Scattertext* utilizando Word2vec (riesgos olas de calor y de frío)

Caracterización y enriquecimiento de eventos

Para reflejar la caracterización y enriquecimiento producidos sobre los eventos de riesgos a partir de las fuentes de información consideradas, seleccionamos las inundaciones. El SINAE las define como el evento más frecuente y de mayor impacto en Uruguay, ya que durante la pasada década se contaron por cientos de miles las personas desplazadas como consecuencia de este evento (SINAE, 2021). Dada la importancia de este riesgo, la prensa se nutre de información proveniente del SINAE y de Comités Departamentales de Emergencia entre

sus principales fuentes. Así, numerosos eventos relacionados con las inundaciones publicados por fuentes oficiales también aparecen reportados en prensa. Por ejemplo, encontramos noticias como:

> En Canelones ayer había (particularmente en Santa Lucía) un total 60 evacuados y se estima que unas 240 personas se auto-evacuaron en todo el departamento, sumando un total de 300 personas en Canelones. Según el CNE al cierre de esta edición en la Ciudad de la Costa se venía trabajando intensamente con maquinaria pesada para mejorar drenaje. (*La Red 21*, 2010)

En estos casos, el motivo de la inundación coincide con lo detectado en los datos abiertos considerados. Sin embargo, el trabajo con nuestro *corpus* permite identificar algunos ejemplos de noticias como el caso de "Después de la lluvia el Comité Nacional evalúa la situación en Rivera", donde se menciona: "Otra de las localidades afectadas por las copiosas lluvias que se vienen registrando es Cerro Chato. En ese lugar trabajan tres intendencias para reparar los daños, pues es jurisdicción de Florida, Treinta y Tres y Durazno" (*Montevideo Portal*, 2009). Este es un ejemplo de evento no recogido en los datos abiertos considerados, ya que no aparecen datos sobre riesgo de inundación en la localidad de Cerro Chato.

Asimismo, encontramos algunas irregularidades en los datos abiertos del SINAE. Si bien éste no considera los tornados como uno de los riesgos presentes en Uruguay, el país se encuentra ubicado en su totalidad en el llamado "callejón de los tornados" de América del Sur (Meteored, 2006). A pesar de esta ausencia de riesgo, en el mencionado catálogo aparecen algunas referencias a ocurrencias de tornados en el país. Por ejemplo, el 29 de diciembre de 2009 se produjeron tornados de categorías F2 y F3 en las localidades de Bolívar y Chamizo, respectivamente, con vientos de 100 km/h y efectos detectables como "voladura de techos, daño a unidades productivas y pérdida de cosechas".

Considerando esta información, se aprecian incoherencias en los registros, ya que la fuerza de un tornado se mide según la escala de Fujita-Pearson, basada en los daños ocasionados y en la velocidad del viento. Por tanto, teniendo en cuenta que ambos eventos producidos en Bolívar y Chamizo tuvieron la misma velocidad del viento y daños ocasionados, no se justifica que fueran calificados de diferente manera. Suponemos que en este y otros puntos del catálogo existe una interpretación errónea de la escala mencionada.

Junto con la caracterización de los eventos de riesgo considerados, procedemos a enriquecer los datos abiertos con los detalles presentes en el *corpus* de noticias, tales como el acontecimiento en sí mismo, sus efectos o quienes

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58538

participaron en la búsqueda de la solución, entre otros aspectos. Un ejemplo del enriquecimiento realizado se muestra a continuación a través del caso específico de los tornados producidos en la localidad de Dolores.

El primer evento de este tipo se produjo en 2012 y se caracterizó por ser de baja intensidad, afectando únicamente a la zona rural. Ese mismo año hubo también inundaciones que ocasionaron que 60 personas fueran evacuadas. Adicionalmente, el 15 de abril de 2016 se produjo un nuevo tornado de mayor intensidad, alcanzando la categoría F3-F4, según la escala de Fujita mejorada, con vientos de entre 251-330 km/h. Este evento dejó cinco muertos, 1 000 viviendas destrozadas, así como los Liceos Nº1 y 2, el hospital y la Escuela Rural Nº74 con graves daños estructurales (Vázquez-Melo, 2019).

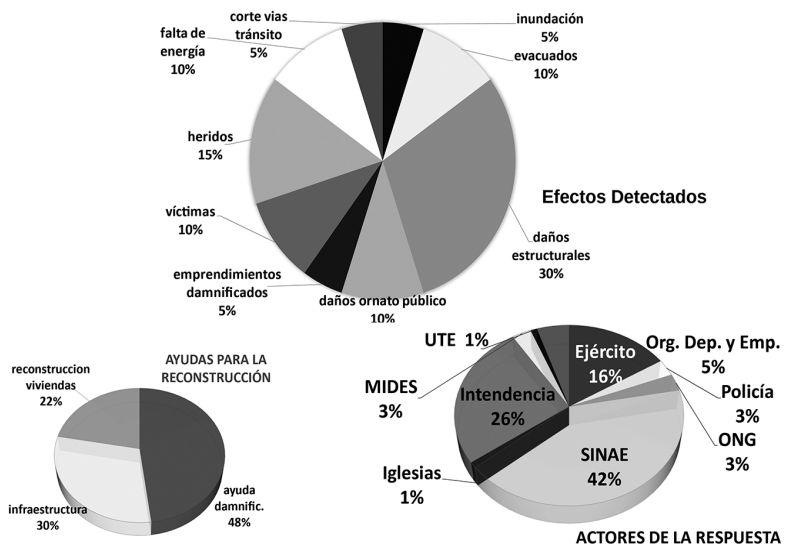

Figura 6. Datos asociados al enriquecimiento obtenido de la prensa digital abierta sobre el tornado de Dolores del 15/abril/2016

El análisis de nuestro *corpus* permitió caracterizar los problemas a los que se enfrentó la localidad de Dolores, identificar en qué se aplicaron las ayudas recibidas para la reconstrucción y cuáles fueron los actores que intervinieron en la misma. En la *Figura 6* se muestra una representación gráfica de los efectos, ayudas y actores intervinientes en la reconstrucción según la información recuperada del *corpus* de noticias para el tornado de 2016. Adicionalmente, en el *corpus* podemos ver que el proceso de reconstrucción de la localidad de Dolores se llevó a cabo en tres años. La *Tabla 2*, cuyos datos

fueron obtenidos de forma semiautomática, muestra los detalles recuperados del *corpus* sobre la reconstrucción de la mencionada localidad.

Código	% noticias codificadas	Afectación	Observaciones	Año
Ayuda damnificados	48.6 %	Población Damnificada	---	2016
Infraestructura dañada	21.7 %	Liceo Nº 1	Daño estructural	2016
		Liceo Nº 2	Daño estructural total	
		Hospital de Dolores	Daño estructural	
		Tendido eléctrico	Daño estructural y del servicio	
		Servicio de comunicaciones	Daño estructural y del servicio	
		Ornato público	Daño estructural total	
		Viviendas y comercios	Daño estructural	
Reconstrucción	29.7 %	Tendido eléctrico	Restablecimiento	2016
		Servicio de comunicaciones	Restablecimiento	2016
		Vivencias	Reconstrucción	2016
		Liceo Nº 1	Remodelación	2019
		Liceo Nº 2	Reconstrucción	2019
		Escuela No. 74	Remodelación	2019
		Hospital de Dolores	Remodelación	2019

Tabla 2. Noticias sobre reconstrucción de la Ciudad de Dolores

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58538

Conclusiones

Este trabajo presenta un proceso para la caracterización y enriquecimiento de datos sobre riesgos en Uruguay utilizando datos abiertos (SINAE, INUMET y Ministerio de Ambiente) y periódicos digitales abiertos (*El Eco Digital*, *Montevideo Portal* y *La Red 21*).

Tomando como referencia los tipos de eventos de riesgo que se presentan en los datos abiertos considerados, se conforma un *corpus* de noticias comprendidas entre 2003-2019. Sobre este *corpus* se aplican técnicas de minería de texto utilizando el software QDA-MinerLite y el lenguaje Python, lo que permite una identificación y caracterización de los eventos. El resultado del mencionado proceso permite enriquecer los datos abiertos con información obtenida del *corpus*, logrando complementar estos datos con información sobre efectos, actores e intervenciones asociadas a los tres grandes grupos de eventos de riesgo considerados (zoonosis, eventos meteorológicos extremos e incendios forestales).

En definitiva, el trabajo realizado permite comprobar que datos y prensa abierta pueden ser fuentes de información complementarias para la caracterización de los eventos de riesgo. Además, los resultados obtenidos demuestran que la aplicación de la minería de texto ayuda al descubrimiento de datos en los recursos de la prensa abierta que contribuyen al enriquecimiento de datos abiertos con esa infinidad de "datos menores" ocultos en las noticias sobre dichos eventos. Esta visión integrada de los riesgos permite generar una nueva fuente de información con los datos obtenidos tras la aplicación de las técnicas de minería de texto. No obstante, cabe destacar que este trabajo se centra en demostrar la viabilidad del enriquecimiento de los datos abiertos a partir de la información de la prensa digital, quedando fuera de su alcance la comprobación de la fiabilidad de los datos incorporados. Sin embargo, asumimos que trabajar con información contenida en otro tipo de fuentes pertenecientes a los agentes activos en los eventos de riesgos urbanos, como, por ejemplo, la recogida por los servicios de urgencias, sanitarios, etc. que realizan la intervención sobre el terreno podría generar un enriquecimiento de mayor fiabilidad y precisión en los datos abiertos.

El resultado de este análisis de riesgos urbanos facilita que gobiernos locales y ciudadanos puedan incrementar su capacidad para conocer y afrontar los desafíos asociados con la vulnerabilidad de los espacios urbanos. Este trabajo también demuestra la pertinencia de la aplicación de la minería de texto en un ámbito de conocimiento diferente, como es el caso de los riesgos urbanos.

Adicionalmente, el trabajo permite vislumbrar la importancia de la colaboración entre diversas organizaciones que tratan con información sobre

riesgos urbanos, tanto de forma directa como indirecta. La combinación de diversas fuentes de información puede ayudar a generar una visión *completa* de los riesgos que asolan a una ciudad o, incluso, a un país. Por tanto, el desarrollo de trabajos en esta línea, del que este trabajo puede ser considerado un inicio, podrán permitir el desarrollo de aplicaciones prácticas para el conocimiento, manejo y gestión de los riesgos urbanos desde diversos ámbitos como centros de investigación de riesgo, medios de comunicación, formación de expertos en temas de riesgos, entre otros.

Referencias

Adelekan, I. O. 2020. "Urban dynamics, everyday hazards and disaster risks in Ibadan, Nigeria". *Environment and Urbanization* 32 (1): 213-232. https://doi.org/10.1177/0956247819844738

Brun, R. E. y J. A. Senso. 2004. "Minería textual". *El profesional de la información* 13 (1): 11-27.

Contreras-Barrera, M. 2014. "Minería de texto: una visión actual". *Biblioteca Universitaria* 17 (2): 129-138. https://www.redalyc.org/articulo.oa?id=2854027 9005

Cooper, R. 2014. "Open data flood mapping of Chao Phraya River basin and Bangkok metropolitan region". *British Journal of Environment and Climate Change* 4 (2): 186. https://doi.org/10.9734/BJECC/2014/11872

Da Silva, M. P. y A. F. Godoy Viera. 2021. "Descubrimiento de conocimientos mediante técnicas de minería de textos aplicadas a documentos textuales de la investigación policial brasileña". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 35 (88): 161-183. https://doi.org/10.22201/iibi.24488321xe.2021.88.58389

Feldman, R. y J. Sanger. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511546914

Gupta, V. y G. Lehal. 2009. "A survey of text mining techniques and applications". *Journal of emerging technologies in web intelligence* 1 (1): 60-76. https://doi.org/10.4304/jetwi.1.1.60-76

Hearst, M. 1999. "Untangling text data mining". En *Proceedings of ACL'99: the 37th annual meeting of the Association for Computational Linguistics*. http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html

Hjørland, B. 2002. "Domain-analysis in Information Science - Eleven Approaches - Traditional as Well as Innovative". *Journal of Documentation* 58 (4): 422-462. https://doi.org/10.1108/00220410210431136

Hjørland, B. y H. Albrechtsen. 1995. "Toward a new horizon in information science: Domain-analysis". *Journal of the Association for Information Science and Technology* 46 (6): 400-425. https://doi.org/10.1002/(SICI)1097-4571(199507)46:6<400::AID-ASI2>3.0.CO;2-Y

Kessler, J. 2017. "Scattertext: a browser-based tool for visualizing how corpora differ". Preprint, presentado: 2 de marzo de 2017, última revisión: 20 de abril de 2017.
https://arxiv.org/abs/1703.00565

Lasaponara, R., B. Murgante, A. Elfadaly, M. Molaei-Qelichi, S. Z. Shahraki, O. Wafia y W. Attia. 2017. "Spatial Open Data for Monitoring Risks and Preserving Archaeological Areas and Landscape: Case Studies at Kom el Shoqafa, Egypt and Shush, Iran". *Sustainability* 9 (4): 572.
https://doi.org/10.3390/su9040572

La Red 21. 2010. "Mejora la situación de los evacuados por las intensas lluvias de Santa Rosa". 6 de setiembre de 2010.
https://www.lr21.com.uy/comunidad/422629-mejora-la-situacion-de-los-evacuados-por-las-intensas-lluvias-de-santa-rosa

Li, Y., L. Xu, F. Tian, L. Jiang, X. Zhong, y E. Chen. 2015. "Word embedding revisited: A new representation learning and explicit matrix factorization perspective", in *Twenty-Fourth International Joint Conference on Artificial Intelligence*.
https://www.aaai.org/ocs/index.php/IJCAI/IJCAI15/paper/viewPa per/10863

Llasat, M. C., M. Llasat-Botija, M. Barnolas, L. López y V. Altava-Ortiz. 2009. "An analysis of the evolution of hydrometeorological extremes in newspapers: the case of Catalonia, 1982-2006". *Natural Hazards and Earth System Sciences* 9: 1201-1212.
https://doi.org/10.5194/nhess-9-1201-2009

McCallum, D., S. L. Hammond y V. Covello. 1991. "Communicating about environmental risks: How the public uses and perceives information sources". *Health Education Quarterly* 18 (3): 349-361.
https://doi.org/10.1177/109019819101800307

Meteored. 2006. "Tornados en América del Sur".
https://www.tiempo.com/ram/2543/tornados-en-amrica-del-sur/

Mhamdi, C. 2016. "Transgressing media boundaries: News creation and dissemination in a globalized world". *Mediterranean Journal of Social Sciences* 7 (5): 272-277.
https://doi.org/10.5901/mjss.2016.v7n5p272

Mikolov, T., K. Chen, G. Corrado y J. Dean. 2013. "Efficient estimation of word representations in vector space". Preprint, presentado: 16 de enero de 2013, última revisión: 7 de septiembre de 2013.
https://arxiv.org/abs/1301.3781

Montevideo Portal. 2009. "Después de la lluvia el Comité Nacional evalúa la situación en Rivera". 7 de noviembre de 2009.
https://www.montevideo.com.uy/Noticias/Comite-Nacional-evalua-situacion-en-Rivera-uc96126

Orlecka-Sikora, B., S. Lasocki, J. Kocot, T. Szepieniec, J. R. Grasso, A. García-Aristizabal, M. Schaming *et al*. 2020. "An open data infrastructure for the study of anthropogenic hazards linked to georesource exploitation". *Scientific data* 7 (1): 1-16.
https://doi.org/10.1038/s41597-020-0429-3

ONU. *Agenda para el Desarrollo Sostenible*. 2018.
https://www.un.org/sustainabledevelopment/es/development-agenda/

Paprotny, D., H. Kreibich, O. Morales-Nápoles, P. Terefenko y K. Schröter. 2020. "Estimating exposure of residential assets to natural hazards in Europe using open data". *Natural Hazards and Earth System Sciences* 20 (1): 323-343.
https://doi.org/10.5194/nhess-20-323-2020

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58538

Salloum-Said, A., M. Al-Emran, A. A. Azza y K. Shaalan. 2017. "A survey of text mining in social media: Facebook and Twitter perspectives". *Advances in Science, Technology and Engineering Systems Journal* 2 (1): 127-133. https://doi.org/10.25046/aj020115

SINAE (Sistema Nacional de Emergencias de Uruguay). 2016. *Catálogo histórico de eventos meteorológicos adversos*. http://enperspectiva.uy/wp-content/uploads/2016/04/Cat%C3%A1logoHist%C3%B3ricoSINAE-Fen%C3%B3menosMeteorol%C3%B3gicos-1.pdf

SINAE. 2020. *Política Nacional de Gestión Integral del Riesgo de emergencias y desastres en Uruguay* (2019-2030). https://www.gub.uy/sistema-nacional-emergencias/sites/sistema-nacional-emergencias/files/2020-03/Poli%CC%81tica%20Nacional%20de%20Gestio%CC%81n%20Integral%20del%20Riesgo%20de%20Emergencias%20y%20Desastres%20en%20Uruguay.pdf

SINAE. 2021. "Inundaciones". https://www.gub.uy/sistema-nacional-emergencias/inundaciones

Sullivan, D. 2001. *Document warehousing and text mining*. New York: Wiley Computer Publishing.

Vázquez-Melo, D. A. 2019. "A casi tres años del Tornado de Dolores, un informe que repasa como fue el fenómeno". *Montevideo Portal*. 8 de abril de 2019. https://www.montevideo.com.uy/Noticias/A-casi-tres-anos-del-tornado-de-Dolores-un-informe-que-repasa-como-fue-el-fenomeno-uc714962

Vescoukis, V. y C. Bratsas. 2014. *Open data in natural hazards management*. European Public Sector Information Platform (EPSIplatform) Topic Report No. 2014/01. https://data.europa.eu/sites/default/files/report/2014_natural_hazards_and_open_data.pdf

Viera, Á. F. G. 2017. "Técnicas de aprendizaje de máquina utilizadas para la minería de texto". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 31 (71): 103-126. https://doi.org/10.22201/iibi.0187358xp.2017.71.57812

Vilches-Blázquez, L. M., D. Comesaña y L. Arrieta-Moreno. 2020. "Construcción de una red de ontologías sobre eventos meteorológicos a partir de periódicos históricos". *Transinformação* 32. https://doi.org/10.1590/1678-9865202032e180077

Wakefield, S. y S. Elliott. 2003. "Constructing the news: The role of local newspapers in environmental risk communication". *The Professional Geographer* 55 (2): 216-226. https://doi.org/10.1111/0033-0124.5502009

Witten, I. H., F. Eibe y A. Mark. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann.

Yagoub, M. M., A. A. Alsereidi, E. Mohamed, P. Periyasamy, R. Alameri, S. Aldarmaki e Y. Alhashmi. 2020. "Newspapers as a validation proxy for GIS modeling in Fujairah, United Arab Emirates: identifying flood-prone areas". *Nat Hazards* 104: 111-141. https://doi.org/10.1007/s11069-020-04161-y

Para citar este texto:

Vilches-Blázquez, Luis M. y Diana Comesaña Ocampo. 2022. "Caracterización de riesgos urbanos en prensa aplicando minería de texto para el enriquecimiento de datos abiertos". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 36 (91): 85-107. http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58538

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2022.91.58538