Evaluación de la Gestión de la Calidad del Aire en Guanajuato con Procesamiento de Lenguaje Natural

Air Quality Management Evaluation with Natural Language Processing

David Salas-Rodríguez
Instituto Tepeyac de León, México
davino66@hotmail.com
https://orcid.org/0000-0001-9144-7067

Recepción: 11 Enero 2023
Aprobación: 29 Diciembre 2023

Acceso abierto diamante

Resumen

El objetivo fue evaluar la Gestión de la Calidad del aire 5 de las 10 ciudades con mayor contaminación del aire en México y que pertenecen al estado de Guanajuato. El método de investigación consistió en medir la información con Inteligencia Artificial orientada con el modelo LART de Gestión ambiental usando el Procesamiento de Lenguaje Natural en las funciones y estrategias para la gestión de la calidad del aire. Se analizó un corpus de 32 enunciados. Como resultado se obtienen una bolsa de 80 palabras y un vocabulario de 82 N-gramas de longitud 1 a 7 para medir la información del proceso de gestión. Los hallazgos revelan que las mejores gestiones están en Celaya, León y Silao de la Victoria. La originalidad del método radica en que la información encontrada por el algoritmo permite validar parcialmente el modelo LART. Se limita a evaluar la gestión y los estudios siguientes se orientarán al desarrollo de un vocabulario más amplio y un corpus mayor para utilizar el modelo W2V que incruste los N-gramas en un modelo n-dimensional.

Palabras clave: Bolsa de Palabras, Calidad del Aire, Gestión Ambiental, PLN, N-grama, Guanajuato. México, Modelo de Gestión LART.

Abstract

The objective was to evaluate Air Quality Management in 5 of the 10 cities with the highest air pollution in Mexico and that belong to the state of Guanajuato. The research method consisted of measuring the information with Artificial Intelligence oriented with the LART model of Environmental Management using Natural Language Processing in the functions and strategies for air quality management. A corpus of 32 statements was analyzed. The result was a bag of 80 words and a vocabulary of 82 N-grams with length 1 to 7 that measure the information of the management process. The findings reveal that the best managements are in Celaya, León and Silao de la Victoria. The originality of the method lies in the fact that the information found by the algorithm allows partial validation of the LART model . It is limited to evaluating the management and subsequent studies will focus on the development of a broader vocabulary and a larger corpus to use the W2V model that embeds the N-grams in a n-dimensional model.

Keywords: Bag Of Words, Air Quality, Environmental Management, Natural Language Processing, Guanajuato. México.

Notas de autor

davino66@hotmail.com

Introducción

El estudio de la contaminación del aire tiene apenas 70 años, su detonante fue la tragedia de Londres a raíz de la catástrofe que fue mal llamada como una "niebla" que cubrió la ciudad en 1952, en un país que apenas se recuperaba de la devastación por la segunda guerra mundial y aun gobernada por Churchill contemplando atónitamente la muerte de más de 3,000 súbditos ingleses. Este episodio dejó claro a los políticos de la época los efectos nocivos para la salud las de las neblinas mortales, dando inicio a una incipiente gestión ambiental que tomo fuerza y forma con la promulgación del Acta para Aire Limpio (AAL) de Londres en 1953 (Pattle & Cullumbine, 1956).

El evento descrito no había sido el primer evento trágico, debido a que en 1930 en la zona industrial del valle Mosa ubicada en Bélgica ocurrió el primer evento de niebla, y el cual mató a unas 60 personas por una combinación mortal de contaminación del aire y gas flúor resultado de los desechos tóxicos de la zona, hecho documentado en su momento por el científico Roholom (1937).

Para le década siguiente ocurrió otro más en Estados Unidos en la ciudad de Donora al suroeste de Pennsylvania entre el 27 y 30 de octubre en 1948, este fue provocado por la empresa US STEEL que se saldó con al menos 17 muertes reconocidas aunque dichos números no son confiables ya que se indemnizó a 80 personas. La empresa que aunque grato de deslindarse del Smog, señalando que había sido tan solo un acto de Dios, tuvo que pagar cuantiosas indemnizaciones luego que las demandas que siguieron le obligaron a cambiar de opinión sobe las causas (Bell & Davis, 2001; Davis et al., 2002; García, 2016).

De lo anterior se destaca que el AAL de Londres de 1953 y la tragedia de Donora que propició la creación de la primera ley de calidad del aire (CA) de 1963 supervisada por la prestigiosa agencia Americana de Protección Ambiental (EPA), y que desde el año 1970 identificó seis contaminantes del aire (NAAQS Table, 2016): Monóxido de carbono (CO), Plomo (Pb), Dióxido de nitrógeno (NO2), Ozono (O3), Partículas contaminantes (PM) y Dióxido de azufre (SO2), determinan los primeros instrumentos de gestión: jurídicos por su carácter regulador, político organizacional por la existencia de una agencia encargada, tecnológico por determinar los contaminantes criterio y económico por las sanciones correspondientes.

Desde entonces y con un enfoque internacional, se han tenido avances como el de la Organización Mundial de la Salud (OMS) que publicó en 1987 su guía para el aire limpio, misma que fue revisada en 1997 (OMS, 2006) y que se considera un instrumento internacional por su carácter de referente para los países. En 2001 la Unión Europea crea el programa Clean Air For Europe (CAFE) con la intención impulsar un programa a largo plazo para la Gestión de la Calidad del Aire (EUR-Lex, 2006). China inició su gestión de la Calidad del Aire (GCA) con su acta identificada como Environmental Protection Law (EPL) de 1972 y que es revisada en 1989 y 2014 (Wang et al., 2014; You, 2015).

En México para 1971 casi 40 años después que la inglesa y 20 después que la americana, surge la Ley Federal para Prevenir y Controlar la Contaminación ambiental (LFPCCA) (Secretaria de Salud, 2018) que aunque tardía, reglamentó la prevención y control de la contaminación atmosférica originada por humos y polvos, fue publicada en el DOF el 17 de septiembre de ese año (Secretaria de Gobernación , 1971). Este es instrumento jurídico orientado al cuidado de la salud humana que se afecta por la exposición a los contaminantes por periodos largos, y que ocasiona enfisemas que frecuentemente se vuelven carcinogénicos, recientemente son de particular interes para preservar la salud humana como lo destaca Arroyo-Hernándes y otros con su estudio sobre los beneficios por la reducción de los contaminantes del aire (2016).

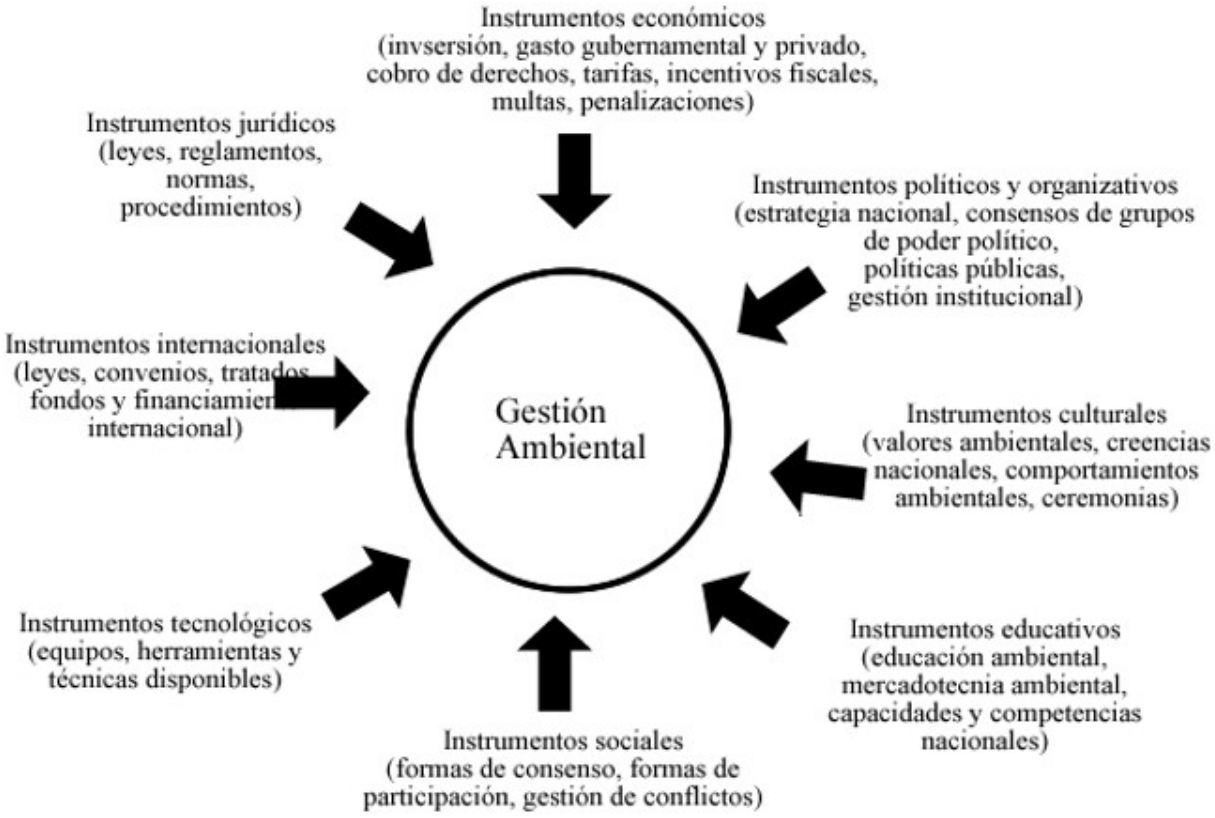
Para evaluar la GCA como un proceso de gestión, este se conceptualiza como un sistema con una multidimensionalidad de enfoques, disciplinas e indicadores de evaluación. De tal manera que su medición requiere el uso de modelos complejos capaces de validar su carácter caótico.

En 20 ciudades de México (INE, 2011), la GCA se encuentra a cargo de una Unidad Responsable de la Calidad del Aire (URCA) que en el peor de los casos, no existe dentro de la estructura de gobierno, algunas

son nacientes y otras con años de existencia originadas por sucesos o episodios de contaminación como los ocurridos en la historia.

La GCA se define como un proceso de gestión que es instrumentado por una serie de enfoques disciplinares con la finalidad de lograr un aire limpio. El modelo científico utilizado para teorizar la heterogeneidad de estos instrumentos es el modelo LART de gestión (Rivas-Tovar, 2009) mostrado en la figura 1. Con este es posible clasificar los enfoques cuantitativos y cualitativos de la GCA clasificándolos en ocho categorías de instrumentos que influyen en el proceso de gestión y que a continuación se sintetizan.

Figura 1. Modelo LART de Gestión Ambiental.



Fuente: Elaboración propia con base en (Rivas-Tovar, 2009; Rivas-Tovar, 2023)

El modelo LART de gestión en un modo de valaoración para un sistema que se encuentra en el borde del caós, y su análisis inatrumental de forma holística tienden al equilibrio (Salas-Rodríguez, 2023).

La teoría de la información es una rama matemática de la probabilidad y estadística que tiene una infinidad de aplicaciones (Kullback, 1968, pp. 1-3), una de sus primeras aplicaciones fue en el campo de las comunicaciones por Shannon y Weaver en donde definen los lenguajes naturales escritos como una fuente discreta que producen texto en una serie de palabras (Words) que son representadas por un proceso estocástico (Shannon & Weaver, 1964, pp. 40-43).

La Inteligencia Artificial (IA) es una rama de las ciencias computacionales cuyo objeto es la creación de sistemas que repliquen la inteligencia humana y la solución de problemas. Su origen se remonta a las aportaciones a este campo de estudio por Alan Turing en 1950 y seis años más tarde por el primer programa de IA presentado por John McCarthy y Marvin Minsky (Rockwell, 2017).

El Procesamiento del Lenguaje Natural (PLN) y con siglas en inglés (NLP) es una rama de la IA que se encarga del estudio de la forma en que las máquinas comuncan y procesan el lenguaje humano (natural). La unidad básica de análisis son los N-gramas que son comprenidos como un conjunto de términos que forman un texto.

Uno de los modelos usados para el PLN es la Bolsa de Palabras (BoW) por sus siglas en inglés que utilizan como técnica de extracción de información la vectorización, consiste en convertir en tokens de longiud n y su representación matricial; el conteo de la frecuencia de ocurrencia de ellos y su

normalización, lo anterior se conoce en la disciplina como incrustración. Esta técnica permite el análisis de texto y medir la información que proporcionan con base en un vocabulario. Es imporante mencionar que es no supervisada y aplica para vocabularios y corpus reducidos, ya que para grándes volúmenes de estos se puede utilizar redes neuronales como el algoritmo W2V (Word to Vector) catalogado dentro del los métodos de IA llamados deep learning.

Los avances tecnologicos en el procesamiento, almacenamiento y reducción de costos en los sistemas de cómputo, han propiciado una vertiginosa evolución del análisis de datos (data analytics) cuyo origen se remonta a 1947 con el estudio de John Tukey sobre métodos estadísticos aplicados en la computación (Boston College, 2023), en la actualidad se conoce como un campo iterdiscplinario en el que convergen las ciencias computacionales, la estadístca y tanto la economía como la gestión llamada Ciencia de Datos (CD).

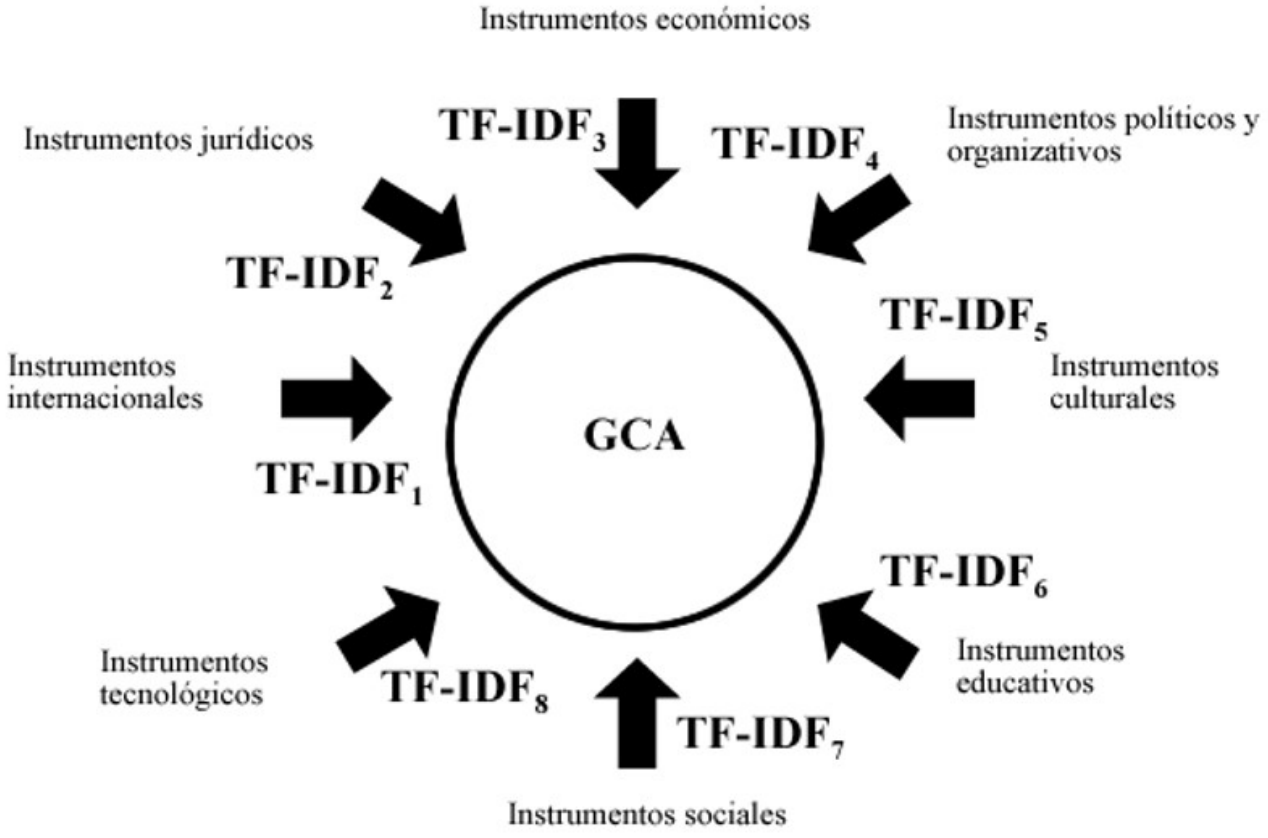
En la revisión de literatura se enconraron novedosos enfoques basados en CD que emplean PLN en diversas disciplinas como las documentadas en el trabajo de Yue Kang (Yue et al., 2022) donde analizaron mediante esta técnica 24 revistas líderes en negocios encontrando asociaciones de conceptos (N-gramas) con teorías de múltiples disciplinas enfocadas a la investigación administrativa.

Chang y otros realizaron un estudio exploratorio para clasificar en el campo del turismo mediante el análisis con PLN, una serie de conceptos incrustrados y medidos mediante la matríz TF-IDF (Term Frequency – Inverse Document Frequency) que fueron encontrados al analizar los resumenes de revistas entre 2010 y 2019 de WoS (Web of Science) relacionados con el turismo. Su hallazgo fue la agrupación de conceptos (N-gramas) en seis categorias: viajes, cultura, sostenibilidad, modelos, comportamiento y hotel (I-Cheng et al., 2022).

Otra novedosa aplicación en el ámbito de la gestión es el estudio realizado por Santos y otros donde utlizan PLN para analizar las ordenes de trabajo con la intención de detectar y extraer información importante de ellas, relacionada con las acciones y los componentes escritos en ellas (Santos et al., 2023).

Con base en la revisión del estado de la cuestión se afirma el hecho del uso novedoso de la IA mediante el PLN como técnica analítica no supervisada para obtener información mediante la incrustración de N-gramas que se asocian a una diversidad de teorías y disciplinas como lo es la GCA. Con la matríz TF-IDF es posible medir los enunciados del corpus con los N-Gramas cuyos valores proveen un mayor información para explicar con su peso la variable estudiada, en la figura 2 se muestra el modelo ex ante.

Figura 2. Modelo PLN Ex Ante para Evaluar la GCA.



Fuente: Elaboración propia

En las ciudades de León, Silao de la Victoria, Irapuato, Salamanca y Celaya pertenecientes al estado de Guanajuato México, cada URCA desempeña una serie de funciones y estrategias para la GCA, con lo anterior se plantea la pregunta: ¿Cómo evaluar la GCA con base en la información obtenida analizando sus funciones y estrategias reportadas por cadad una de ellas?

Este estudio propone el PLN por N-gramas para medir la información referente a la GCA en las funciones y estrategias de las URCA en cinco ciudades del Estado de Guanajuato incluidas en el almanaque con los datos de 18 ciudades en México que se destacan por el problema de la calidad del aire (INE, 2011).

Método

Es una invstigación mixta utilizando el PLN por N-gramas mediante el modelo BoW empleando como medida el TF-IDF (Term Frequency – Inverse Document Frequency). Los sujetos de investigación fueron las funciones y estrategias en GCA de cada URCA en las ciudades de León, Silao de la Victoria, Irapuato, Salamanca y Celaya pertenecientes al estado de Guanajuato México incluidas en las 20 ciudades del país con problemas de calidad del aire.

La frecuencia inversa de documento (IDF) se calcula con una variante de la fórmula tradicional utilizada por la librería scikit-learn 1.3.2 para Python que se muestra en la ecuación 1 (Pedregosa et al., 2011).

$IDF(t) = \log\frac{1+n}{1+df(t)}$ (Ec. 1)

En esta df(t) es el número de documentos en el corpus que contiene el término t. Este corresponde a un N-grama que será modelado y n es el número de enunciados (documentos) que forman el conjunto que será analizado (corpus). Un valor IDF alto indica que t aparece pocas veces y los valores bajos indican que se usa frecuentemente.

La matríz TF (frecuencia de términos) es la vectorización del corpus de tal forma que cada renglón se identifica como una oración i numerada de 0 a n-1 y cada columna corresponde a una palabra j del vocabulario, el valor de cada elemento se calcula con la ecuación 2.

$TF(i,j)=\frac{n(i,j)}{\sum n(i,j)}$ (Ec. 2)

Donde n(i,j) es la frecuencia de ocurrencia del término j en la oración (documento) i del corpus y $\Sigma n(i,j)$ es la suma de la frecuencia de todos los términos del vocabulario presentes en la oración n

La matríz TF-IDF se calcula con la ecuación 3.

$TF-IDF(j,i) = TF(i,j) \times IDF(t)$ (Ec. 3)

Un valor alto de TF-IDF se logra con una alta frecuencia de t en la oración i y con una baja frecuencia en el corpus, de tal manera que t es relevante entre más aparece en una oración pero se compensa por su frecuencia en el corpus.

Para determinar la cantidad de información proporcionada por cada oración i del corpus, se hace una ponderación con base en la suma de los valores TF-IDF mediante los siguientes pasos:

1. Se obtiene la sumatoria de los valores TF-IDF para cada oración i (documento) del corpus.
2. Se obtiene el promedio de la sumatoria de los valores TF-IDF por cada N-grama.
3. Se asigna un valor adicional a cada documento multiplicando el valor del paso 2 por el df(t) de cada oración i (frecuencia del vocabulario en el documento).
4. Se asigna un valor por posición en el corpus a cada oración entre 0 y 1, la oración 0 tiene un valor de 0.01 y la 31 de 0.32.
5. Se asigna un peso a cada oración i sumando el valor del punto 1 con el punto 3 y multiplicado por su valor del punto 4. El resultado indica la cantidad de información que contiene el enunciado i del corpus.

Las fuentes de información fueron la entrevista y cuestionario vía correo con la URCA de la ciudad de Celaya, y para las restantes se recurrió a solicitudes de información pública a través de la Plataforma Nacional de Transparencia (PNT, 2020). El análisis hace en dos pasos, el primero para determinar el vocabulario para el modelo BoW y el segundo para analizar la información mediante la matriz TF-IDF.

Resultados

El corpus fue integrado por 32 sentencias que consisten en los enunciados i reportados por cada URCA como funciones y estrategias para la GCA. Este se analiza en primer lugar identificando un total de 80 palabras (BoW), son ordenadas de acuerdo a la suma de su frecuencia de término y se mustran en la tabla 1. Para el análisis es necesario el preprocesamiento que consiste en la eliminación de las stop-words[1] y uso de minúsculas.

Con lo anterior y con base en el modelo teórico de gestión ambiental LART, utilizando la BoW se crea un vocabulario de 82 N-gramas que se muestra en la tabla 2, este se utilizó para analizar las estrategias y funciones en GCA organizadas como corpus.

Tabla 1. Bolsa de palabras para la elaboración del vocabulario.

N-grama	Σ df(t)	N-grama	Σ df(t)	N-grama	Σ df(t)	N-grama	Σ df(t)
cumplimiento	4	aplicar	1	establecer	1	necesarias	1
verificación	4	así	1	establecidas	1	obligaciones	1
coordinación	3	atención	1	estatal	1	ostensible	1
emisiones	3	atmosféricas	1	expedición	1	ostensiblemente	1
municipal	3	autoridades	1	falta	1	participar	1
vehicular	3	ca	1	formular	1	personas	1
acciones	2	coadyuvar	1	generen	1	poseedoras	1
automotores	2	competente	1	imponer	1	presenten	1
autoridad	2	competentes	1	impulsar	1	prevención	1
competencia	2	condicionantes	1	instrumentación	1	programa	1
contaminantes	2	conducir	1	inventarios	1	propietarias	1
fijas	2	contingencias	1	licencias	1	provenientes	1
fuentes	2	control	1	mantenimiento	1	público	1
gobierno	2	declaradas	1	mecanismos	1	realizar	1
operación	2	detección	1	medidas	1	reducción	1
programas	2	determinar	1	mejoramiento	1	resulten	1
revisión	2	dispositivos	1	monto	1	revisar	1
vehículos	2	ejecución	1	movilidad	1	servicio	1
ambientales	1	emisión	1	multas	1	sistemas	1
aplicadas	1	encargada	1	municipio	1	transporte	1

Fuente: elaboración propia con base en los resultados (Pedregosa et al., 2011).

Tabla 2. Vocabulario de para modelo BoW utilizado para el análisis de la GCA

N-gramas en el vocabulario			
acciones para impulsar	operación de sistemas	coadyuvar	impulsar
acciones de coordinación	programas de prevención	competente	instrumentación
autoridad Municipal	programas para el mejoramiento	competentes	inventarios
competencia Municipal	participación ciudadana	condicionantes	licencias
coordinación con Gobierno del Estado	reducción de emisiones de contaminantes	conducir	mecanismos
coordinación con la autoridad	revisión de cumplimiento de	contingencias	medidas
coordinación con la autoridad competente	vehículos automotores	control	mejoramiento
contingencias ambientales	verificación vehicular	declaradas	monto
cumplimiento de la verificación vehicular	verificación cumplimiento	detección	movilidad
cumplimiento de las obligaciones	cultura ambiental	determinar	multas
cumplimiento del Programa Estatal de Verificación Vehicular	comunicación participación	dispositivos	municipio
emisión ostensible de contaminantes	calidad del aire	ejecución	obligaciones
emisión de contaminantes	ambientales	emisión	participar
emisiones de contaminantes	aplicadas	encargada	prevención
emisiones ostensiblemente	aplicar	establecer	programa
emisiones ostensiblemente de contaminantes	así	establecidas	reducción
falta de verificación	atención	estatal	sistemas
fuentes fijas	atmosféricas	expedición	transporte
inventarios de emisiones	autoridades	falta	vehículos
licencias de operación	ca	formular	
operación de fuentes fijas	ciudadana	imponer	

Fuente: elaboración propia con base en los resultados (Pedregosa et al., 2011).

Análisis

Para el análisis debido al reducido tamaño del corpus y siguiendo la teoría (Pedregosa et al., 2011), se emplea la matríz TF-IDF para determinar la cantidad de información en cada una de las oraciones i del corpus. Cada una es identificada por su renglón de 0 a 31 y 82 columnas que corresponden a cada palabra (N-grama) del vocabulario, cada valor en la matriz corresponde al TF-IDF que se obtiene mediante el producto del vector IDF con la matríz TF.

Para cada N-grama del modelo, se realizó la suma de su valor TF-IDF para determinar la información proporcionada $\Sigma_1^n TF-IDF_j$ en la matríz para identificar aquellos que proporcionan más inforrmación cuya suma excede la unidad, y fueron ordenados de mayor a menor para sintetizarlos en la tabla 3.

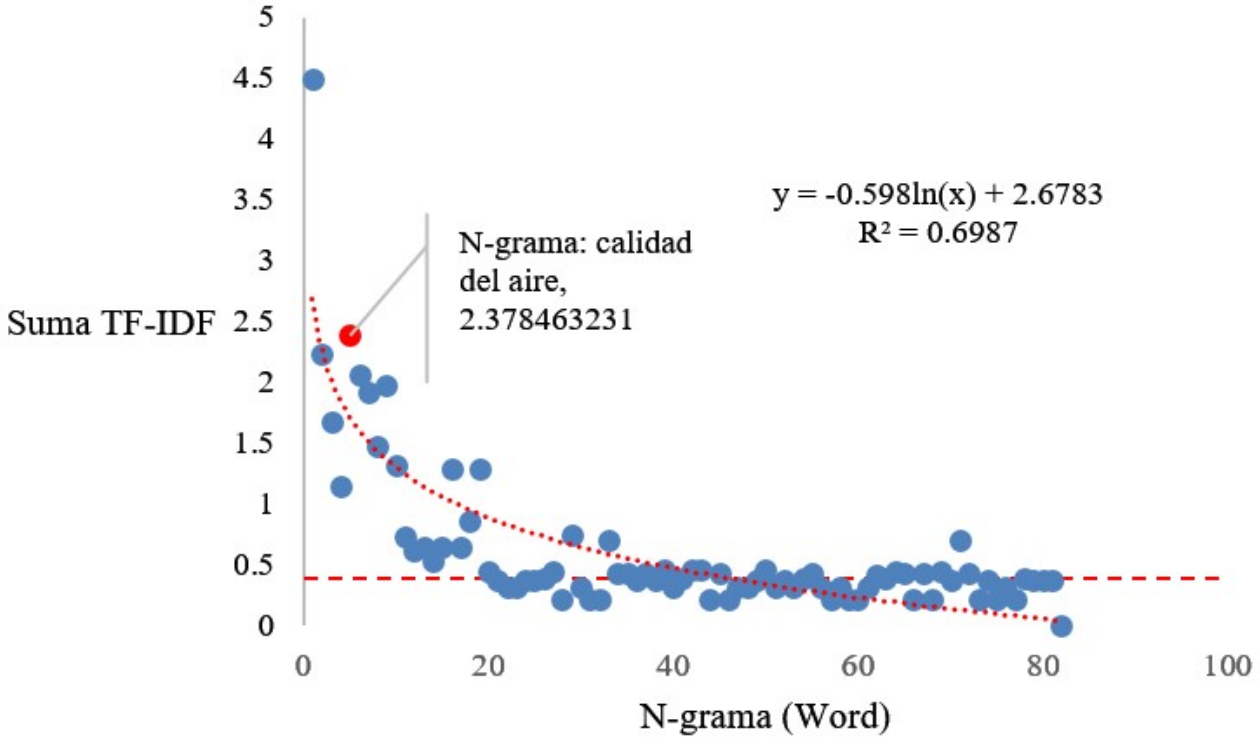
Tabla 3. Top 10 de N-gramas en Categorías Modelo LART

N-gram	*d(ft)*	Suma TF-IDF	Categoría LART
Programa	7	4.489547719	Jurídicos
Calidad del aire	3	2.378463231	Internacionales
Verificación vehicular	5	2.229200628	Tecnológicos
Control	3	2.061068298	Jurídicos
Reducción	3	1.971001851	Internacionales
Municipio	3	1.920341968	Político-organizativos
Fuentes fijas	3	1.671316368	Jurídicos
Prevención	3	1.47352511	Sociales
Vehículos	3	1.316214486	Tecnológicos
Vehículos automotores	3	1.143608168	Tecnológicos

Fuente: elaboración propia con base en los resultados (Pedregosa et al., 2011).

Analizando la cantidad de información para cada N-grama mediante la sumatoria de sus valores TF-IDF, esta presentan una distribución logarítmica con un valor mínimo de 0, una mediana de 0.39324 y un valor máximo de 4.48955 correspondiente al N-grama “programa” mostrado previmente en la tabla 3 con un df(t) de 7. La figura 3 nos muestra la distribución de la información proporcionada por cada N-grama, en esta el eje horizontal cruza en la mediana. Cabe destacar que el N-grama “calidad del aire” se encuentra en el Q3 con un df(t) = 3 y una suma TF-IDF de 2.37846. Esto confirma la teoría de que las oraciones más semejantes al vocabulario y con N-gramas menos frecuentes proporcionan una mayor información.

Figura 3. Distribución Logarítmica de la Información en los N-gramas



Fuente: elaboración propia con base en los resultados.

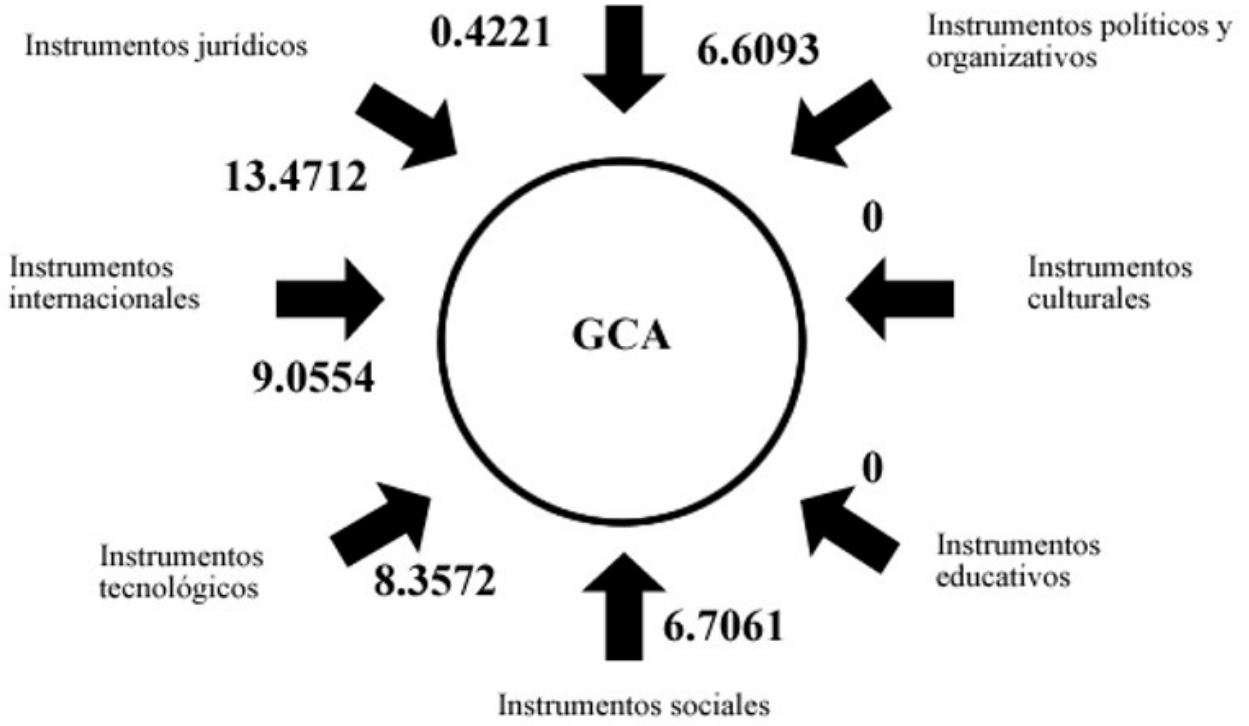
El 88 % de la información corresponde a los N-gramas categorizados por instrumento de GCA y el 12 % a la de contexto, hay que recordar que estos son los que mejoran el contexto del análisis. La síntesis se muestra en la tabla 4 y el modelo ex post facto en la figura 4.

Tabla 4. Información total por Categoría de Instrumento de Gestión

Categoría LART	f	Suma TF-IDF	% Información	Inf. Acumulada
Jurídicos	17	13.471225	27%	27%
Internacionales	14	9.055374	18%	44%
Tecnológicos	12	8.3572243	16%	61%
Sociales	12	6.7061218	13%	74%
Político-organizativos	13	6.6093462	13%	87%
Económicos	2	0.4221473	1%	88%
Culturales	0	0	0%	88%
Educativos	0	0	0%	88%
Contexto	12	6.0358839	12%	100%

Funte: Elaboración propia con los resultados de la matríz TF-IDF.

Figura 4. Modelo Ex Post Facto para evaluar la GCA.



Fuente: Elaboración propia con el análisis de la información con el PLN.

Es relevante destacar que los instrumentos juridcos, los internacionales, los tecnologicos, los sociales y los politicos y organzativos aparecen como los determinates de calidad del aire. Esto supone una validación parcial aunque mayoritaria del modelo LART.

Para determinar la cantidad de información proporcionada por cada oración del corpus, se hace una ponderación con base en la suma de los valores TF-IDF, el df(t), el promedio de la sumatoria del valor TF-IDF por N-grama y el valor por posición de cada una en el corpus (de 0.1 a 0.32). Con esto se determinó la ausencia de información en el corpus con un valor de 0, una mediana de información de 0.17360 y la máxima información con 1.18208. En el Q3 se ubican tres elementos i del corpus que a continuación se describen en orden de mayor a menor peso de la información, para cada uno se identifican los N-gramas del vocabulario y su categoría de instrumento de GCA.

En primer lugar se ubica la oración correspondiente al documento 9 que es una función desempeñada por la URCA de la ciudad de Celaya:

establecer los mecanismos para la ejecución de los dispositivos de revisión de cumplimiento de verificación vehicular y de detección de vehículos que generen emisiones ostensiblemente contaminantes, así como imponer y establecer el monto de las multas aplicadas a las personas propietarias o poseedoras de vehículos automotores por la falta de verificación o emisión ostensible de contaminantes.

Esta función integra los siguientes instrumentos de GCA: Internacionales, Jurídicos, Económicos, Político-Organizacionales y Tecnológicos, considerandose una función holistica con tendencia al equilibrio (Salas-Rodríguez, 2023), debido a su instrumentación holística proporciona una información de 1.18208 y un df(t) = 18 (frecuencia de N-gramas).

En segundo se posiciona una estrategia para la GCA de la URCA de Silao de la Victoria identificada como el documento D31 (último de la lista con un valor por posición de 0.32), y que se considera una gestión naciente debido a que se incorpora al programa PROAIRE en el año 2014 y enuncia lo siguiente: “cultura ambiental, comunicación y participación ciudadana para la calidad del aire”.

Esta estrategia es instrumentada considerando sólo por lo Educativo y lo Social de los instrumentos de gestión, que por su limitada instrumentación se encuentra en los límites del caos con una tedencia al orden

si entegrara más instrumentos (Ídem), sin embargo proporciona una información de 1.00252 y un df(t) = 5, aunque es menor al anterior, se considera excelente por su semejanza con el vocabulario.

Finalmente se ubica la estrategia de la URCA de Celaya con el identificador D23 (valor por posición de 0.24) que dice: "programa de prevención de altos niveles de partículas menores a 10 micras en el municipio de celaya guanajuato (contingencias ambientales) y reforestación en la zona urbana y rural del municipio".

Esta estrategia incluye los instrumentos para la GCA: Jurídicos, Político-Organizacionales, Sociales y Tecnológicos con una información de 0.90062 y un df(t) = 6.

Se encontraron cuatro oraciones que no proporcionan información debido a que no incluyen ningún N-grama del vocabulario y que a continuación se muestran:

En León la función "promoción de la verificación y seguimiento al proaire" y la estrategia "actualizar el inventario de gases de efecto invernadero"; en Silao de la Victoria la función "prevenir y controlar la contaminación atmosférica generada en las zonas o fuentes de la jurisdicción municipal" y la estrategia "regulación de quemas a cielo abierto".

Discusión

La teoria del PLN por N-gramas utilizando el modelado con BoW indicada para vocabularioss y corpus de tamaño reeducido y la inclusión de las stop-words como de, del, la y las, mejoran el contexto del vocabulario para el análisis y obtener la mayor información de las funciones y estrategias de GCA en cada URCA y es medible con la matríz TF-IDF. En este estudio la información de contexto corresponde a un 12 %.

Con un 88 % de información, 10 de los elementos más relevantes del vocabulario observan valores TF-IDF superiores a la unidad predominando los instrumentos jurídicos, internacionales, tecnológicos, sociales y político-organizativos. Para los económicos sólo se observa información referente a las multas por falta de verificación vehicular y finalmente para los culturales y educativos la información encontrada fue nula. Este hallazgo se considera detrminte y mueve a la reflexión sobre los valores que deben ser promovidos entre niños y jovenes, asi mismo nos orienta a urgir una revisióon de los programa de edcuación en calidad del aire en una de las ciudades mas contaminadas de Mexico. El gobierno no es el unico responsable, las asociaciones empresariales, las universidades, la sociedad civil y las comunidades tine un importante reto que enfrentar.

Conclusiones

El PLN utilizando N-gramas prororciona en la matriz TF-IDF información suficiente para medir las funciones y estrategias para la GCA obtenida de un vocabulario y corpus reducido, confirmando lo que indica esta rama teórica de la IA. Los hallazgos confirman una gestión que incluye solo los referentes internacionales relativos a la calidad del aire, enfocada solamente en los instrumentos Jurídicos orientados al control de la contaminación originada por fuentes móviles (vehículos de motor).

Los instrumentos Político-Organizacionales en conjunto con los Económicos y Tecnológicos se orientan solamente en el establecimiento de multas para los vehículos de motor que emiten de forma ostensible contaminantes del aire, la parte tecnológica corresponde a la utilizada por los vehículos de motor para el control de emisiones como el uso de convertidores catalícos, vehículos híbridos o eléctricos. El desarrollo de instrumentos Culturales no proporciona información observándose nulos y los Educativos y Sociales orientados sólo al cumplimiento de los programas de verificación vehicular en pro de un aire limpio y las contingencias ambientales.

Con lo anterior se concluye con la relevancia del uso del PLN para la evaluación de la GCA, ya que una gestión consistene y comprometida, es aquella que presenta una mayor información con referencia a la relevancia de los N-gramas del vocabulario. Se limitó a la evaluación del proceso de gestión, los estudios siguientes se orientarán al desarrollo de un vocabulario más amplio y un corpus mayor para utilizar el

modelo W2V que incrustre los N-gramas en un modelo n-dimensional. Esta invesigación ha sido desarrollada sin ningún financiamiento.

Referencias

Arroyo-Hernández, M., Monraz-Pérez, S., & Pérez-Padilla, R. (2016). Beneficios a la salud debidos a la reducción de la contaminación ambiental. *Neumol Cir Torax, 75*(2), 132-135. Retrieved 10 de Agosto de 2017, from http://new.medigraphic.com/cgi-bin/contenido.cgi?IDPUBLICACION=6576

Bell, M., & Davis, D. (2001). Reassessment of the Lethal London Fog of 1952: Novel Indicators of Acute and Chronic Consequences of Acute Exposure to Air Pollution. *Environmental Health Perspectives, 109*(3), 389-394. Retrieved 3 de octubre de 2016, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1240556/pdf/ehp109s-000389.pdf

Boston College. (2023). *How the History of Data Science Has Led to the Demand for Data Analysts*. Woods College of Avancing Studies: https://appliedeconomics.bc.edu/the-history-of-data-science/

Davis, D., Bell, M., & Fletcher, T. (2002). A Look Back at the London Smog of 1952 and the Half Century Since. *Environmental Health Perspectives, 110*(12), 734-735. Retrieved 1 de octubre de 2016, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1241116/pdf/ehp0110-a00734.pdf

EUR-Lex. (2006). *Clean Air for Europe (CAFE) Programme*. Retrieved 20 de octubre de 2016, from http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV%3Al28026

García, O. (2016). *Gases reactivos y calidad del aire a nivel global.* Seminario: Conceptos Básicos en Composición Atmosférica, Centro de Investigación Atmosférica de Izaña, Santa Cruz de Tenerife. Retrieved 10 de octubre de 2016, from http://repositorio.aemet.es/bitstream/20.500.11765/3515/1/5_Seminario%20Iza%C3%B1a%2018%20junio_Gases%20Reactivos_Calidad%20Aire_Omaira%20Garcia.pdf

I-Cheng, C., Jeou-Shyan, H., Chih-Hsing, L., C., S.-F., & Tai-Yi, Y. (2022). Exploration of Topic Classification in the Tourism Field with Text Mining Technology—A Case Study of the Academic Journal Papers. *Sustainability, 14*(7). https://doi.org/https://doi.org/10.3390/su14074053

Kullback, S. (1968). *Information theory and statistics.* Dover.

OMS. (2006). *Guías de calidad del aire de la OMS relativas al material particulado, el ozono, el dióxido de nitrógeno y el dióxido de azufre. Actualización mundial 2005. Resumen de evaluación de los riesgos.* Suiza: OMS. Retrieved 7 de septiembre de 2016, from http://apps.who.int/iris/bitstream/10665/69478/1/WHO_SDE_PHE_OEH_06.02_spa.pdf

Pattle, R. E., & Cullumbine, H. (20 de octubre de 1956). Toxicity of some atmospheric pollutants. *Toxicity of atmospheric pollutants, 2*(4998), 913-915. Retrieved 11 de octubre de 2016, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2035479/pdf/brmedj03177-0031.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *JMLR, 12*(85), 2825-2830.

PNT. (2020). *Plataforma Nacional de Transparencia*. https://www.plataformadetransparencia.org.mx/

Rivas-Tovar, L. A. (2009). *Efectos de la Teoría de La Complejidad en la Gestión Ambiental en México* (1 ed.). DF, México: Instituto Politécnico Nacional.

Rivas-Tovar, L. A. (2023). *Normas Apa 7ª Edición: Estructura, Citas y Referencias.* Ciudad de México: Instituto Politécnico Nacional.

Rockwell, A. (2017). *The History of Artificial Intelligence*. Harvard Univrsity. Science in the news: https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

Roholom , K. (1937). *Fluorine Intoxication*.https://archive.org/details/FluorineIntoxication

Salas-Rodríguez, D. (2023). Índice Complejo de Gestión de la Calidad del Aire y Sostenibilidad. *Investigación Administrativa, 52*(131). https://doi.org/10.35426/iav52n131.04

Santos, P., Datia, N., Pato, M., Sobral, J., Gomes, N., Leitão, N., & Ferreira, M. (2023). NLP for Enterprise Asset Management: An Emerging Paradigm. *2023 27th International Conference Information Visualisation (IV)*. https://doi.org/10.1109/IV60283.2023.00049

Secretaria de Gobernación . (1971). *DIARIO OFICIAL.* Distrito Federal: Secretaria de Gobernación .

Secretaria de Salud. (14 de octubre de 2018). *1971. Se expide la primera ley sobre control de la contaminación ambiental*. https://www.gob.mx/salud/75aniversario/articulos/1971-se-expide-la-primera-ley-sobre-control-de-la-contaminacion-ambiental?state=published

Shannon, C., & Weaver, W. (1964). *The mathematical theory of communication.* Illinois: University of Illinois Press.

Wang, Y., Ying, Q., Hu, J., & Zhang, H. (2014). Spatial and temporal variations of six criteria air pollutants in 31 provincial capital cities in China during 2013–2014. *Environment international, 73*, 413-422. Retrieved 22 de octubre de 2016, from https://www.researchgate.net/profile/Yungang_Carl_Wang/publication/265050563_Spatial_and_temporal_variability_of_PM2.5_and_PM10_over_the_North_China_Plain_and_the_Yangtze_River_Delta_China/links/54ee9d080cf25238f93ac86f.pdf

You, M. (2015). Changes and challenges of the 2014 revised environmental protection law in the context of China's five fundamental transitions. *Hong Kong Law J, 2*, 621-650. Retrieved 19 de octubre de 2016, from https://www.researchgate.net/profile/Mingqing_You/publication/284586087_Changes_and_Challenges_of_the_2014_Revised_Environmental_Protection_Law_in_the_Context_of_China's_Five_Fundamental_Transitions/links/5655454108ae1ef9297716df.pdf

Yue, K., Zhao, C., Chee-Wee, T., Qian, H., & Hefu, L. (2022). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics, 7*(2), 139-172. https://doi.org/https://doi.org/10.1080/23270012.2020.1756939

Información adicional

Clasificación JEL: Q01