

Minería de texto y aprendizaje automático para identificar prioridades de desarrollo sostenible

Nidia Isabel Molina Gómez
División de Ingenierías
Universidad Santo Tomás
Bogotá, Colombia
<https://orcid.org/0000-0003-4485-262X>

Camilo Andrés Rodríguez
División de Ingenierías
Universidad Santo Tomás
Bogotá, Colombia
<https://orcid.org/0000-0002-8079-2591>

P. Amparo López
Dpto de Ing. del Agua y Medio Ambiente
Universitat Politècnica de València
Valencia, España
<http://orcid.org/0000-0002-7043-3683>

José Luis Díaz Arévalo
Facultad de Ingeniería
Universidad de la Sabana
Chia (Cundinamarca), Colombia
<https://orcid.org/0000-0003-4330-1957>

Resumen—Alrededor del término desarrollo sostenible se han generado planes, programas y estrategias enfocados en la satisfacción de las necesidades de las generaciones del presente sin limitar la satisfacción de las necesidades de las generaciones futuras. Bajo este fin, los gobiernos a nivel mundial han realizado importantes esfuerzos, por un lado, para alcanzar los objetivos del milenio y ahora en el marco de los Objetivos de Desarrollo Sostenible (ODS). El gobierno colombiano asume el desafío de garantizar el logro de una vida sana para sus habitantes, la cual requiere de la interconexión e integralidad de los ODS. A nivel urbano se incluyen, entre otros desafíos, la mejora en los indicadores de salud, asociados a la contaminación del aire y que son alcanzables desde un marco de política y para su logro requiere de la participación de los grupos de interés encabezados por la comunidad.

Este estudio analizó la información publicada en diversos medios de comunicación virtuales, del orden nacional, en el periodo 2009 a 2018; se aplicaron técnicas de minería de texto y herramientas de aprendizaje automático, mediante el uso del software R. Se identificaron temas centrales, sobre los cuales el país y su ciudad capital han prestado relevancia desde los medios de comunicación. Fue posible conocer el alcance en la divulgación de los retos, avances y oportunidades con la implementación de los Objetivos del Milenio, de los ODS y el papel de la calidad del aire.

Se encontró que el desarrollo sostenible no presenta mayor divulgación en los medios y la sostenibilidad se relaciona principalmente a la biodiversidad y las actividades financieras. Los esfuerzos se concentran en información específica y es escasa la divulgación de los retos, metas y avances en ODS. Aunque la calidad del aire es un tema de interés, no presentó un papel relevante como lo soportan los hechos de seguridad, pobreza, educación y calidad de vida. Un análisis de este tipo permite establecer las prioridades temáticas, de información y de desarrollo político tendencial, divulgados a la comunidad, como actor sobre quien recae la implementación de los lineamientos de política

Palabras clave—ODS, ODM, minería de texto, comunicación, desarrollo sostenible, calidad del aire

I. INTRODUCCION

La sostenibilidad en el desarrollo es un objetivo que se ha trazado desde 1987 con el establecimiento de su definición, pasando por la declaración realizada por las naciones unidas en el año 2000, acerca de la reducción de la pobreza y ocho objetivos a lograr para el año 2015 [1], los Objetivos de Desarrollo del Milenio (ODM). Posteriormente, desde enero de 2016 se pusieron en marcha los 17 Objetivos de

Desarrollo Sostenible (ODS), establecidos en virtud de los avances generados con los ODM y con la inclusión de nuevos objetivos, entre estos: cambio climático, la desigualdad económica, la innovación, el consumo sostenible y la paz y la justicia [2]. Los ODS son liderados a nivel gubernamental, estableciendo las acciones y presupuestos para lograr a 2030 las metas de desarrollo sostenible. Su implementación requiere de la participación del sector privado y del sector público; no obstante, los avances y resultados generados recaerán sobre la comunidad, quien del mismo modo podrá soportar su implementación.

Con el ánimo de identificar el grado de divulgación de los ODS, los anteriormente establecidos ODM, así como el alcance en la divulgación de los aspectos asociados a la calidad del aire, se analizó la documentación de artículos de prensa publicada en la web por medios de comunicación tradicionales de alto impacto; para ello se aplicaron técnicas de minería de texto a través del análisis semántico de textos no estructurados.

La minería de textos se ha aplicado en diversos estudios cuyo objetivo se concentra entre otros, en la identificación de énfasis o conjuntos temáticos predominantes en corpus textuales, así como en la identificación de patrones de una colección de textos. Entre la literatura revisada se encontró el desarrollo de estudios orientados a la identificación de vacíos de información y coocurrencias documentales, aplicados específicamente al análisis de planes de cambio climático [3], la identificación de tendencias y prácticas de sostenibilidad en procesos productivos [4], así como las tendencias y temas principales en publicaciones científicas [5]. El estudio más cercano al que se presenta en este trabajo, aplicó técnicas de aprendizaje supervisado y procesamiento de lenguaje natural sobre documentos de noticias comparado con los indicadores reportados en informes de sostenibilidad [6]. Se encontró que, la minería de textos y el aprendizaje supervisado, permitieron identificar en las noticias los indicadores establecidos como importantes en los reportes de sostenibilidad, constituyéndose en una herramienta útil en la identificación de los problemas de sostenibilidad y de soporte a la identificación de indicadores específicos.

Los medios de comunicación son un canal directo con la comunidad, que pueden ejercer cierto grado de influencia en la opinión pública y orientar la formación de conceptos basados en la información publicada. Son un grupo de interés que influencia a la comunidad sobre quien recaen los resultados en la implementación de los ODS y quien puede

soportar las diferentes acciones establecidas en los planes de gobierno. En la revisión de literatura no se encontró la aplicación de herramientas de minería de texto a artículos de prensa para determinar el alcance y especificidad de la información publicada por los medios de comunicación en temáticas relacionadas con los ODM y ODS, o en temas específicos como lo es la calidad del aire, que se vincula al ODS 11: ciudades y comunidades sostenibles. Se evidenciaron dos estudios de análisis tendencial en la gestión de los recursos hídricos [7], [8], en que se analizaron textos en artículos de prensa y su comparación con el comportamiento de eventos relacionados a dicho recurso.

A partir del análisis de la divulgación de los ODM y ODS en artículos de prensa publicados en la web, se buscó identificar, a través de este trabajo, los principales temas tratados en los últimos diez años (2009-2018) por los medios de comunicación, con el fin de identificar el grado de información que se entrega a la comunidad a propósito de los ODS y el papel de la calidad del aire como tópico fundamental en el objetivo ciudades y comunidades sostenibles.

II. METODOLOGÍA

A. Recopilación de Información

Se desarrollaron cuatro pasos para cada año en el periodo de análisis y cada objetivo de desarrollo sostenible (ODS), consistentes en la búsqueda y recuperación de los artículos de prensa publicados en la web, su agrupación, limpieza preliminar de los textos recuperados y el almacenamiento de la información en documentos de formato *txt*.

Se generaron dos conjuntos de información. El primero, correspondiente a los artículos de prensa generados en Colombia y que incluyeran el término Colombia en su texto y, en el segundo conjunto, las noticias generadas en el país, que incluyeran a la ciudad capital: Bogotá. Para ambos conjuntos de datos se realizó búsqueda: a) por año de emisión del artículo según ODS y b) por año e inclusión de las palabras clave: ODS, sostenibilidad y desarrollo sostenible. Para el primer conjunto, se recuperaron artículos informativos en medios de comunicación de alto impacto en el país y a nivel nacional, mediante el uso de palabras clave, relacionadas con cada uno de los 17 ODS orientando la representatividad para cada enfoque de los ODS analizados. En la tabla I se relacionan las palabras clave utilizadas para cada objetivo.

TABLA I. PALABRAS CLAVE SEGÚN ODS PARA LA BÚSQUEDA DE INFORMACIÓN

ODS	Palabras Clave
1	pobreza, desastre, reducción de desastres
2	hambre, inseguridad alimentaria, malnutrición
3	mortalidad, vih, tuberculosis, malaria, hepatitis, cardiovascular, enfermedades respiratorias, diabetes, cáncer, suicidio, sustancias adictivas, accidentes de tráfico, planificación familiar, fecundidad, gastos sanitarios, contaminación ambiental, contaminación intradomiciliar, agua insalubre, saneamiento deficiente, higiene, intoxicación, consumo de tabaco
4	educación, tic, alfabetización
5	igualdad de género, violencia física, psicológica, sexual
6	agua potable, aguas residuales, tratamiento de aguas, uso eficiente del agua, estrés hídrico, saneamiento
7	energía, combustibles, electricidad, eficiencia energética
8	pib, empleo informal, desempleo, trabajo infantil, accidentes de trabajo, enfermedades laborales, derechos laborales, turismo

ODS	Palabras Clave
9	infraestructura resiliente, industrialización inclusiva, innovación, industrialización sostenible
10	desigualdad, discriminación
11	asentamientos informales, transporte público, crecimiento población, gestión urbana, preservación, protección y conservación del patrimonio cultural y natural, desastres, material particulado, partículas finas, contaminación del aire, zonas de esparcimiento, acoso, planes de desarrollo urbano y regional, reducción del riesgo de desastres, edificios sostenibles
12	consumo y producción sostenible, huella ecológica, pérdida de alimentos, desechos peligrosos, productos químicos, reciclaje, informes de sostenibilidad, turismo sostenible, combustibles fósiles
13	cambio climático, adaptación al cambio climático, gases de efecto invernadero, mitigación al cambio climático
14	eutrofización costera, acidez marina, conservación de mares, zonas protegidas, pesca sostenible, investigación, uso sostenible de los océanos
15	bosques, desertificación, pérdida de biodiversidad, degradación de tierras, gestión forestal sostenible
16	conflicto, muertes, violencia, armas, soborno, corrupción, secuestro
17	impuestos, asistencia oficial, inversión extranjera, ciencia y tecnología, internet

Para el ámbito nacional se recuperaron en total 7551 artículos noticiosos para un periodo de diez años comprendidos entre el 2009 a 2018, y 1976 para el ámbito distrital. Los artículos recopilados bajo las claves de búsqueda “ODS”, “sostenibilidad” y “desarrollo sostenible” corresponden con un 6.3% para el caso nacional y 4% para la ciudad capital.

B. Procesamiento y Análisis de la Información

Mediante el uso del software de código abierto R versión 3.5.2, se efectuó la preparación de texto de los documentos recopilados; se realizó limpieza a los textos eliminando términos que generaran ruido, palabras vacías, espacios vacíos, signos de puntuación y números; también se acotó el número de términos haciendo uso de sus raíces, mediante la lematización. Una vez preparado el texto se procede a la generación del corpus del documento y la creación de la nube de términos más frecuentes. Este procedimiento se aplicó para la información recopilada a) por ODS para todos los años, b) al conjunto de ODS por cada año y c) al total de documentos con los términos desarrollo sostenible, sostenible y ODS, en los conjuntos de información nacional y distrital.

Identificados los términos predominantes en cada conjunto y subconjunto de datos, se procede a establecer una matriz de frecuencias y se continúa con un análisis de asociación de términos frecuentes respecto a los demás términos del documento.

Adicionalmente, se hace un análisis de agrupación jerárquica de los términos con base en la distancia existente entre ellos. Se decidió utilizar la distancia “euclídea” y la distancia de “manhattan”; para ambos casos con el método de agrupación “promedio”, como método por defecto utilizado en el análisis de textos en el software R. La elección de la combinación de la distancia y método de agrupación se definió atendiendo al resultado del coeficiente cofenético que se presenta en un rango de 0 a 1, en donde la cercanía a 1 presenta un mejor comportamiento en la agrupación de los datos.

III. RESULTADOS

A. El enfoque de sostenibilidad

Dendrogram illustrating hierarchical clustering of variables. The y-axis represents Height (0.40 to 0.52). The variables are: paz, agua, empresari, reto, ciudad, ambient, empresa, and social. The dendrogram shows that 'paz' and 'agua' are the most distinct variables, joining the main cluster at a height of approximately 0.52. The main cluster consists of 'empresari' and 'reto' joining at ~0.43, 'ciudad' joining at ~0.45, and a sub-cluster of 'ambient', 'empresa', and 'social' joining at ~0.42.

Para el caso de Bogotá, a pesar de la relevancia en las claves de búsqueda, se encuentra dispersión en los términos del corpus documental, siendo las palabras más frecuentes: urbana y premio. Aunque los términos agua y aire son evidentes, se muestran en un mismo nivel que otros términos poco frecuentes. No fue posible aplicar un algoritmo de agrupación dada la alta dispersión de los términos en el conjunto de datos para el ámbito distrital.

Al fusionar el conjunto de artículos de prensa clasificados según las palabras clave definidas para todos los ODS con los artículos de prensa clasificados por los términos desarrollo sostenible, sostenibilidad y ODS, se encontró como términos frecuentes para el ámbito nacional: pobreza, agua y ambiente, seguido de los términos desarrollo sostenible, desempleo, mujer, desigualdad, hambre y violencia (ver Fig. 2). No es frecuente el uso de los términos ODS o milenio; tampoco se encontró alguna correlación del término “desarrollo sostenible” con los términos presentes en el corpus del documento analizado.

reciclaje, arma, colegio y cerros (ver Fig. 3). Aunque el componente atmosférico no se observa como uno de los términos más frecuentes en el ámbito distrital, si presenta una asociación a términos como buses, contaminado, descontaminar, chimenea, integrado (relacionado al sistema de transporte de la ciudad) con un 57% de correlación para cada término.



IV. ANÁLISIS DE RESULTADOS Y DISCUSIÓN

Desde la definición de los ODM y ODS los estados han emprendido diversas acciones desde el ámbito gubernamental; se incluyen agendas de trabajo, productos técnicos con metas e indicadores que han sido divulgados desde dichas entidades a nivel documental y en sus portafolios web. Sin embargo, hace falta la integración de los medios de comunicación con información específica y relacional de los hechos que se enuncian como soporte a la comunidad, con la orientación a la generación de conceptos de manera más informada. Esto se evidencia en el enfoque central de los resultados encontrados para los últimos 10

años en los artículos de prensa en la web, en los ámbitos nacional y distrital.

Las clases generadas en el proceso de clúster, son coherentes dentro de los ámbitos nacional y distrital analizados (ver Tabla II). El ámbito nacional presenta con mayor claridad las temáticas de algunos ODS, en comparación con el ámbito distrital.

TABLA II. RESULTADOS DE AGRUPACIÓN DE TÉRMINOS SEGÚN CORPUS ANALIZADO: ODS Y TÉRMINOS DE SOSTENIBILIDAD

Ámbito	Clase de términos	Coefficiente cofenético	Distancia	Dispersión de términos
Bogotá	1. arma 2. agua 3. plan cerros 4. predial 5. reciclaje 6. transporte	0,72	manhattan	97%
Colombia	1. ambiente 2. desarrollo sostenible 3. Agua- pobreza	0,99	euclídea	95%

De un análisis de la información publicada por entidades gubernamentales, acerca de indicadores relacionados a los ODM y ODS, se encontró que en un periodo de 6 años se redujo el indicador de pobreza en 12,6 puntos porcentuales [9]; estos años hacen parte del grupo de documentos analizados y los avances y actividades asociadas pudieron haber influenciado la información divulgada por los medios de comunicación.

Con relación al ODS 6, agua limpia y saneamiento, se reporta un incremento en 14,7 puntos porcentuales en tratamiento de aguas residuales en el país y una mejora en la calidad y continuidad en el servicio de acueducto y alcantarillado [9]. Aunque la gestión de las aguas residuales ha mejorado en el país, pasando del 27,8 % de aguas residuales urbanas tratadas en 2010 al 42,6 % en el año 2017 [10] el crecimiento poblacional continua incrementándose y con ello el uso y aprovechamiento del recurso hídrico. Comunidades informadas acerca de estos indicadores, así como de su papel en la mejora de las condiciones son el soporte para el alcance de los ODS.

Adicionalmente la información que se suministra es específica de la situación territorial, no encontrando para Bogotá un enfoque claro en la información de los artículos de prensa acerca de la divulgación en materia de desarrollo sostenible, sostenibilidad, ODS y ODM; en el caso del componente atmosférico, con una reducida frecuencia de aparición se presentó para Bogotá correlación con las fuentes móviles. Sin embargo, dada la dispersión de los términos en el conjunto de datos no fue posible identificar conjuntos temáticos a partir de la agrupación no supervisada. Para el caso nacional se generaron tres grupos de términos; no obstante, el término sostenibilidad, a la luz de la información que se divulga en los artículos de prensa en el país, está limitada a la responsabilidad empresarial y el recurso hídrico.

Para ningún conjunto de datos se generó un grupo temático relacionado con la calidad del aire. Sólo al analizar la información compilada en virtud de las palabras clave del ODS 11 ciudades y comunidades sostenibles, correspondiente al 5.7% del total de artículos recuperados y analizados, se encontró que los términos más frecuentes son:

transporte, movilidad, ambiente y aire; además con una dispersión del 95% y un coeficiente cofenético del 0.86 estos mismos términos se presentan como etiquetas principales. No obstante, sólo se encuentra una asociación entre los términos aire y ambiente (46%). Además, pese a que la meta establecida para este ODS en materia de calidad del aire es *reducir el impacto ambiental negativo per cápita de las ciudades, incluso prestando especial atención a la calidad del aire ...* [11], no se encontró en el análisis de textos asociación a los términos gobernanza, políticas, planes, acciones ó medidas; por su parte el término salud se observa de manera general al presentar correlación con términos como ambiente y ciudad.

Las prioridades de desarrollo sostenible identificadas, se ven limitadas a las etiquetas generadas en el proceso de clasificación en cada conjunto y subconjunto de documentos analizado. Se trata de información que suministran los medios de comunicación acerca de hechos noticiosos y sobre los cuales la comunidad enfoca su atención. De un estudio realizado en España [12], se encontró que los jóvenes conceden a la información un alto valor cívico (8.2 en una escala de 0 a 10); siendo importante pues posibilita y garantiza tanto el acceso al debate público como el desarrollo de una conciencia cívica [12].

Finalmente, al realizar una comparación con los resultados presentados por [6] y este estudio, se encontró que la información de prensa digital publicada por los medios de comunicación nacionales y distritales no es suficiente a la luz de identificar, rastrear y reportar indicadores de sostenibilidad. No obstante, coincide en que los medios de comunicación son un instrumento para identificar potenciales intereses de la comunidad y la ausencia de información en comparación con los retos y avances en materia de objetivos de desarrollo sostenible.

V. CONCLUSIONES

El componente atmosférico es un recurso vital, sobre el cual es reducida la información publicada en medios en los últimos 10 años. Llama la atención los términos más cercanos al recurso aire, en donde en su mayoría reportan a débiles condiciones de calidad del mismo, fuentes asociadas a la emisión de contaminantes atmosféricos, y orientaciones para su gestión que incluyen la reforestación. Aun cuando se registra relevancia en la búsqueda de términos relacionados con el ODS 11, es baja la frecuencia de información en el conjunto de los artículos de prensa digital del ámbito nacional.

En materia de ODS, Colombia revela información específica relacionada con agua-pobreza, ambiente, desarrollo sostenible. Para el periodo de análisis, esta agrupación de términos sugerida en la identificación de clases, mediante clúster, coincide con sólo algunos de los ODM y ODS. Por lo tanto, el papel de los medios de comunicación en la información completa requiere de revisión. Del mismo modo este trabajo se consolida en una oportunidad para aquellos medios de comunicación enfocados en la entrega de información ambiental, de responsabilidad empresarial y con intención de informar acerca del desarrollo sostenible. La información se fundamenta en la responsabilidad social y se desarrollan temas entorno a las necesidades suscitadas en el país.

Las herramientas de minería de texto utilizadas, permitieron extraer objetivamente la información relevante

en virtud de la aplicación de métodos de inferencia que incluyeron el análisis de frecuencias, asociación de términos y la clasificación no supervisada con la generación de etiquetas de términos. Para dicho propósito se hizo necesario preparar los documentos eliminando términos de alta frecuencia como las palabras vacías.

Una etapa siguiente del presente estudio corresponde con la comparación de los resultados del presente estudio con las temáticas de ODS en redes sociales. Si bien es cierto las redes sociales y medios conversacionales son un medio preferido de información, son un vehículo a los medios tradicionales con los artículos de prensa. Adicionalmente se prevé una clasificación de documentos basado en las etiquetas generadas en este estudio

REFERENCIAS

- [1] UNDP, "Objetivos de Desarrollo del Milenio | UNDP." [Online]. Available: https://www.undp.org/content/undp/es/home/sdgoverview/mdg_goals.html. [Accessed: 01-Jul-2019].
- [2] PNUD, "Objetivos de Desarrollo Sostenible | PNUD." [Online]. Available: <https://www.undp.org/content/undp/es/home/sustainable-development-goals.html>. [Accessed: 01-Jul-2019].
- [3] M. W. Bickel, "A new approach to semantic sustainability assessment: text mining via network analysis revealing transition patterns in German municipal climate action plans," 2017.
- [4] W. Te Liew, A. Adhitya, and R. Srinivasan, "Sustainability trends in the process industries: A text mining-based analysis," *Comput. Ind.*, vol. 65, no. 3, pp. 393–400, Apr. 2014.
- [5] A. Schober, C. Kittel, R. J. Baumgartner, and M. Füllsack, "Identifying dominant topics appearing in the Journal of Cleaner Production," *J. Clean. Prod.*, 2018.
- [6] S. J. Rivera, B. S. Minsker, D. B. Work, and D. Roth, "A text mining framework for advancing sustainability indicators," *Environ. Model. Softw.*, vol. 62, pp. 128–138, 2014.
- [7] J. Wei, Y. Wei, A. Western, D. Skinner, and C. Lyle, "Evolution of newspaper coverage of water issues in Australia during 1843–2011," *Ambio*, vol. 44, no. 4, pp. 319–331, May 2015.
- [8] Y. Xiong, Y. Wei, Z. Zhang, and J. Wei, "Evolution of China's water issues as framed in Chinese mainstream newspaper," *Ambio*, vol. 45, no. 2, pp. 241–253, Mar. 2016.
- [9] DNP, "Fin de la pobreza - La Agenda 2030 en Colombia - Objetivos de Desarrollo Sostenible," 2019. [Online]. Available: <https://www.ods.gov.co/es/objetivos/fin-de-la-pobreza>. [Accessed: 07-Jul-2019].
- [10] DNP, "CONPES 3948. Concepto favorable a la nación para contratar empréstitos externos con la banca multilateral o bilateral hasta por 40 millones de euros, o su equivalente en otras monedas, destinados al financiamiento parcial del programa para el saneamiento." DNP, Bogotá D.C, pp. 1–38, 2018.
- [11] PNUD, "Objetivo 11: Ciudades y comunidades sostenibles | El PNUD en América Latina y el Caribe." [Online]. Available: <http://www.latinamerica.undp.org/content/rblac/es/home/post-2015/sdg-overview/goal-11.html>. [Accessed: 12-Mar-2018].
- [12] A. Bergström and M. Jervelycke Belfrage, "News in Social Media," *Digit. Journal.*, vol. 6, no. 5, pp. 583–598, May 2018.