ISSN 1669-9106

ARTÍCULO ORIGINAL

MEDICINA (Buenos Aires) 2021; 81: 214-223

MINERÍA DE TEXTO EN PUBLICACIONES CIENTÍFICAS CON AUTORES ARGENTINOS

RICARDO A. DORR, JUAN JOSÉ CASAL, ROXANA TORIANO

Laboratorio de Biomembranas, Instituto de Fisiología y Biofísica Bernardo Houssay (IFIBIO Houssay), Facultad de Medicina, Universidad de Buenos Aires-CONICET, Buenos Aires, Argentina

Resumen En el presente trabajo utilizamos la minería de texto como herramienta de tratamiento de una gran base de datos científica, con el objetivo de obtener nueva información de todas las publicaciones firmadas por autores argentinos e indexadas hasta 2019 en el área de las ciencias de la vida. Se analizaron más de 75 000 artículos, publicados en alrededor de 5000 medios, firmados por cerca de 186 000 autores con lugar de trabajo en la Argentina o en colaboraciones con laboratorios argentinos. Mediante herramientas automatizadas, que fueron desarrolladas *ad hoc,* se analizó el texto de alrededor de 70 800 resúmenes y se buscaron, mediante detección digital no supervisada, los principales temas abordados, su relación con problemáticas de salud en la Argentina y su tratamiento. Se presentan, además, resultados del número de publicaciones por año, las revistas que las publicaron, y sobre sus autores y colaboraciones. Estos resultados, junto con las predicciones que se obtuvieron, podrían constituirse en una herramienta útil para optimizar el manejo de recursos dedicados a la investigación básica y clínica.

Palabras clave: minería de texto, Argentina, publicaciones científicas

Abstract ***Text mining in scientific publications with Argentine authors.*** In the present work we use text mining as a treatment tool for a large scientific database, with the aim of obtaining new information about all the publications signed by Argentine authors and indexed until 2019, in the area of life sciences. More than 75 000 articles were analysed, published in around 5000 media, signed by about 186 000 authors with a workplace in Argentina or in collaborations with Argentine laboratories. Using automated tools that were developed *ad hoc*, the text of around 70 800 abstracts was analysed, seeking, through non-supervised digital detection, the main topics addressed by the authors, and the relationship with health problems in Argentina and their treatment. Results are also presented regarding the number of publications per year, the journals that have published them, and their authors and collaborations. These results, together with the predictions that were obtained, could become a useful tool to optimize the management of resources dedicated to basic and clinical research.

Key words: text mining, Argentina, scientific publications

Recibido: 22-IX-2020 Aceptado: 19-X-2020

Dirección postal: Ricardo A. Dorr, Instituto de Fisiología y Biofísica Bernardo Houssay (IFIBIO Houssay), Paraguay 2155 piso 7, 1121 Buenos Aires, Argentina

e-mail: rdorr@fmed.uba.ar

PUNTOS CLAVE
Conocimiento actual

- La minería de texto (*text mining*), aplicada a los textos de publicaciones científicas, se convirtió en una poderosa herramienta para realizar estadísticas, analizar comportamientos, rastrear tendencias y hacer predicciones. Hasta 2019 (inclusive) aparecen 75 294 artículos con el término "Argentina" en el campo Filiación en la base de publicaciones *Europe PMC*.

Contribución del artículo al conocimiento actual

- Se analizó el texto de estas publicaciones, obteniéndose resultados estadísticos de medios, autores y países participantes. Se correlacionaron los datos con parámetros económicos argentinos. Se analizaron enfermedades mencionadas y sustancias usadas para su tratamiento, agrupándolas por sitio o por mecanismo de acción. Se detectaron mediante algoritmos los principales temas abordados.

La minería de datos (o *data mining* en inglés) agrupa procesos que, de una manera automática, trabajan con grandes conjuntos de datos para encontrar patrones, tendencias o reglas que expliquen el comportamiento de esos datos en un contexto específico. Dentro de la minería de datos, la minería de texto (*text mining* en inglés) es una especialización que, aplicada a bases de publicaciones científicas, se convirtió en una poderosa herramienta para analizar comportamientos, rastrear tendencias y hacer predicciones[1, 2].

Una de las bases de datos especializadas en ciencias de la vida es *Europe PubMed Central*® (ePMC), plataforma abierta que permite acceder a su colección de publicaciones indexadas provenientes de todo el mundo. Está compuesta por alrededor de 36.7 millones de resúmenes, libros y documentos[3].

El objetivo del presente trabajo fue aplicar minería de texto para analizar el contenido de las publicaciones en ePMC que presentan en el campo Filiación (Aff) el término "Argentina" y, a partir de esto, obtener nueva información subyacente. Se examinaron estadísticamente 75 294 artículos publicados hasta 2019 inclusive, en 5063 medios, firmados por 186 410 autores con lugar de trabajo en la Argentina o en colaboración con el país. Se analizó semánticamente el texto de los 70 798 resúmenes obtenidos (no todos los artículos tienen resúmenes), buscando en ellos estructuras y patrones de información no explícitos, a menudo ocultos. También se buscaron, mediante detección digital no supervisada, los principales temas abordados en las publicaciones. La nueva información obtenida pudo relacionarse con indicadores económicos y otros relacionados con problemáticas de salud en la Argentina.

Materiales y métodos

El *corpus* de texto utilizado fue resultado de una búsqueda en el portal *Europe PMC* (*europepmc.org*), usando la palabra "Argentina" en el campo Filiación. Los datos (todos en inglés) se guardaron en formato XML (*Extensible Markup Language*).

Mediante un flujo de trabajo digital se seleccionaron los campos: *Id* (identidad asignada a una publicación a nivel del repositorio), *source* (fuente), *Pmid* (identificador de PubMed de los trabajos indexados), *Pmcid* (identificador de artículo con el texto completo en PubMed Central), *doi* (identificador de objeto digital), *title* (título), *authorString* (autores que firman la publicación), *pubYear* (año de publicación), *abstractText* (texto del abstract), *fullTextUrlList* (link al texto digital completo del artículo), *affiliation* (lugar/es donde se realizó el trabajo; contiene en muchos casos el nombre de la institución, del laboratorio, de la ciudad y del país), *medlineAbbreviation* (abreviatura usada por Medline para referirse al medio donde fue publicado el artículo).

Se utilizaron las siguientes herramientas informáticas: i) Knime 4.1 (*https://www.knime.com/*) como principal plataforma analítica; ii) AntConc 3.5.8 (desarrollado por Laurence Anthony, Facultad de Ciencias e Ingeniería, Universidad de Waseda, Japón) para investigar el uso de la lengua en el *corpus*; iii) Microsoft Excel 365 para estadísticas y gráficos específicos.

Se desarrolló un flujo de trabajo *ad hoc* en Knime, que hace la búsqueda en ePMC, selecciona campos a estudiar y entrega: i) número de artículos por año de publicación; ii) número de documentos con títulos y/o resúmenes; iii) lista de todos los autores; iv) número de artículos publicados por cada autor; v) lista de primeros autores; vi) número de artículos publicados por cada primer autor; vii) número de artículos por publicación; viii) texto de los resúmenes. Posteriormente, los resúmenes fueron tratados con otro flujo de Knime para extraer, en forma no supervisada, los principales temas de investigación abordados.

AntConc se configuró para considerar como unidad gramatical (componente de lenguaje con significado coherente) a las combinaciones de i) solo letras y ii) letras con números y/o letras griegas y/o algunas marcas de puntuación, ya que muchos términos científicos están compuestos por combinaciones de ellos (por ejemplo, Nα-Benzoyl-DL-arginine o GLP-2). Detectadas las unidades gramaticales diferentes (formas o *types*, en inglés), se determinó la frecuencia de su uso. Como la misma unidad gramatical puede aparecer varias veces en un texto, se contabilizó también la suma de todas las unidades gramaticales usadas en el texto (*tokens*). Para el análisis se aplicó la lista de lemas BNC_lemmafile5.txt[4]. La lista de lemas incluye una palabra y sus variaciones, permitiendo representar al conjunto mediante un único término representativo.

En AntConc y Knime se utilizó, cuando fue necesario, una lista de palabras (*stop-list* en inglés) que se decidió excluir de los resultados. La lista contiene palabras funcionales sin significado para el análisis (por ejemplo, determinantes y preposiciones como los términos ingleses *the*, *a*, *in*, *to*, *from*).

En Knime se aplicó la asignación de Dirichlet en paralelo (LDA, por sus siglas en inglés) para el procesamiento del lenguaje natural, con la que se realizó una detección automática de 30 temas en forma no supervisada. A cada tema se adjudicaron, también en forma automática, 10 palabras que lo caracterizaran. El nodo utilizado sigue a Newman y col.[5], con un esquema de muestreo LDA disperso y una estructura de datos de Yao y col.[6], utilizando la biblioteca de modelado de temas de *Machine Learning for Language Toolkit* (MALLET). Las opciones de configuración fueron *seed*: -1593552080; parámetro alfa para temas: 0.1; parámetro beta para palabras:

0.01; iteraciones: 1000. El preprocesamiento incluyó: i) el borrado de signos de puntuación; ii) el filtrado por *stop-list*; iii) el empleo del nodo Abner Tagger con *OpenNLP Whitespace Tokenizer*, que asigna etiquetas a los términos, y permite reconocer entidades con nombre biomédico como genes, proteínas o células; iv) la "lematización" con POS (*part of speech*) Tagger usando el set *Penn Treebank* adicionándose la biblioteca de *PNL Stanford Core*; v) la conversión de los términos a mayúsculas.

Los datos de publicaciones por año se utilizaron para realizar predicción (*forecasting*). La predicción de 2017 a 2025 se generó empleando la versión AAA del algoritmo de Suavizado Exponencial (*Exponential Smoothing*, ETS), incluyendo un intervalo de confianza. El resultado de la predicción se validó comparando las predicciones para 2018 y 2019 con los datos reales existentes de esos años.

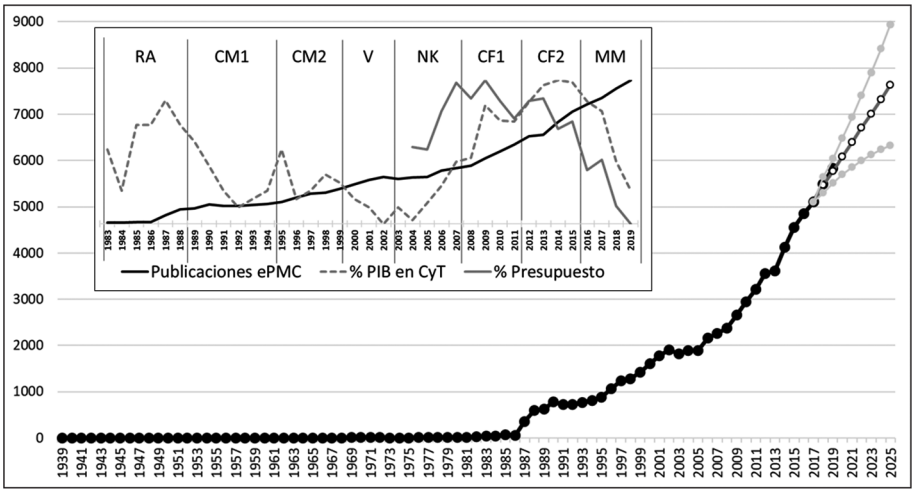
Los datos de publicaciones se cotejaron con índices económicos para la Argentina obtenidos del Banco Mundial, el Instituto de Estadística de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO), el Ministerio de Hacienda (datos 1972-2003, *https://www.minhacienda.gob.ar/onp/documentos/series/Serie6506.pdf*), el Presupuesto abierto (datos 2004-2018, *https://www.presupuestoabierto.gob.ar/sici/datos-abiertos#* (consultado julio 2020), el Presupuesto 2020 (estimado 2019) y *https://www.argentina.gob.ar/economia*.

Resultados

El portal ePMC permite realizar búsquedas avanzadas en publicaciones sobre ciencias de la vida. Desarrollado por el *European Bioinformatics Institute* (EMBL-EBI), está asociado a EPMC Central, *Agricola* y otras fuentes bibliográficas. Para la selección de publicaciones, dentro de la opción "búsqueda" del portal, ubicamos el término "Argentina" en el campo bibliográfico *Affiliation* (Aff), dejando el resto de los campos en blanco. De esta forma se seleccionaron los artículos indexados donde se especifica a la Argentina como el lugar de pertenencia de uno o más de sus autores. El resultado de la búsqueda fue de alrededor de 77 500 trabajos al momento de escribir este artículo. Vale señalar que el número de resultados varía continuamente, ya sea porque siguen agregándose nuevos artículos a la base de datos o porque se realiza alguna corrección en la misma. Al delimitar temporalmente la búsqueda hasta el año 2019 inclusive, el número resultante fue de 75 294 artículos.

La Figura 1 muestra el número de publicaciones indexadas y su variación a lo largo de los años, su proyección, y una asociación a parámetros económicos. Una primera conclusión es que recién pasada la mitad de la década de los 80 comienza una verdadera carga de datos de filiaciones en el repositorio. Se observa un cambio de tendencia en el número de publicaciones a partir de 2006 que, en un principio, podría ser atribuible al mayor porcentaje del Producto Interno Bruto (PIB) dirigido a Ciencia y Técnica en el país (ver inserto). Sin embargo, al considerar períodos largos desde el regreso de la democracia a la Argentina, no aparece una correlación directa de los valores normalizados de este porcentaje y del de fondos del Presupuesto nacional dirigido a Ciencia y Técnica. La Figura 1

Fig. 1.– Variación del número de publicaciones a lo largo del tiempo. Parámetros económicos



Se grafica el número de publicaciones obtenido hasta el año 2019 (círculos negros), y la predicción de 2018 hasta 2025 (círculos blancos) con intervalo de confianza superior e inferior (círculos grises). Los datos reales de 2018 y 2019 se usaron como control para la predicción. Inserto: % PIB en CyT: porcentaje del Producto Interno Bruto (PIB) dirigido a Ciencia y Técnica (CyT); % Presupuesto: porcentaje del Presupuesto Nacional asignado a Ciencia y Técnica. Las iniciales corresponden a períodos presidenciales. RA: Raúl Alfonsín; CM1 y CM2: primera y segunda presidencia de Carlos Menem; V: período institucional con varios presidentes; NK: Néstor Kirchner; CF1 y CF2: primera y segunda presidencia de Cristina Fernández; MM: Mauricio Macri. Los datos fueron normalizados para facilitar la representación.

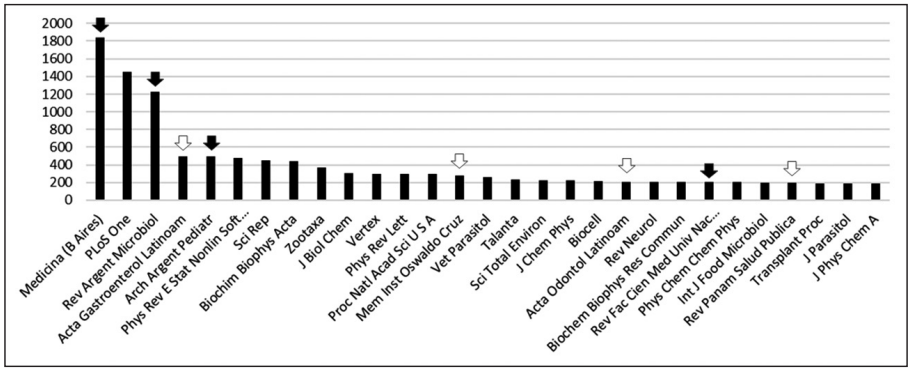
también muestra la predicción sobre futuras publicaciones hasta el año 2025. Se hizo a partir del año 2017, y el número real de publicaciones de los años 2018 y 2019 se tomó como control para la predicción. Como puede observarse, para esos años las curvas real y predictiva se superponen.

Cuando se analiza dónde fueron publicados los artículos, se detectan 5063 sitios de publicación. De las 30 editoras con más publicaciones (Fig. 2), 4 son argentinas (flechas negras) y 4 del resto de Latinoamérica (flechas blancas). La revista argentina *Medicina (Buenos Aires)* encabeza la lista de medios con mayor número de publicaciones (1842 artículos). Junto a *Revista Argentina de Microbiología* (1229 artículos) suman casi un 4% del total de publicaciones.

Respecto de las autorías de las publicaciones, se detectaron 1 067 535 firmas de 186 410 autores únicos de diversas nacionalidades. El promedio es de 14 autores por publicación, un dato *a priori* llamativo. La explicación radica en que varios títulos son firmados por consorcios de autores (como ejemplo extremo, un artículo de 2015 tiene 5159 autores[7]). Si se focaliza el análisis sobre los artículos con un máximo de 11 autores (el 93% del total de artículos de nuestra base), los trabajos con 4 autores son los más frecuentes (Fig. 3). Del análisis de primeros autores con filiación argentina, aparecen 43 186 firmas, que corresponden a 29 171 autores únicos.

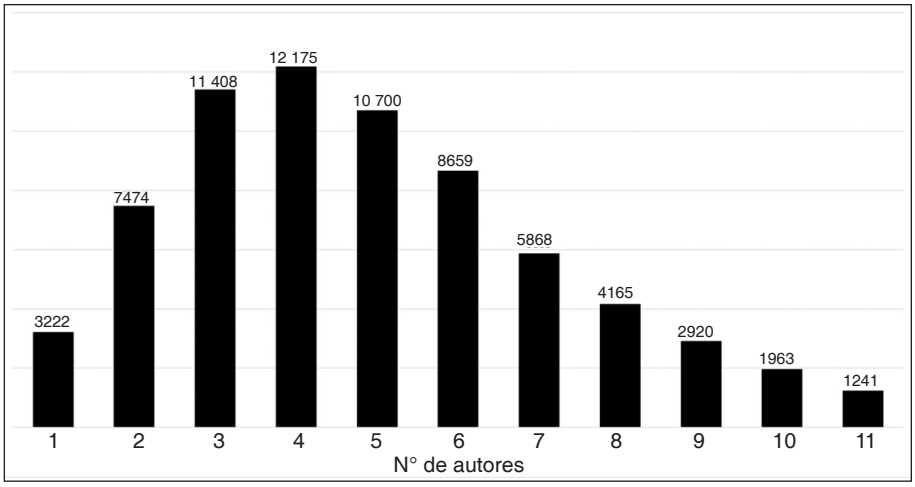
Además de la filiación argentina, se detectaron filiaciones de 161 países, lo que da cuenta de una gran red de colaboraciones de los autores argentinos con

Fig. 2.– Principales sitios de publicación de artículos



El número de artículos se especifica en el eje de ordenadas. Las flechas negras señalan revistas argentinas; las flechas blancas indican revistas latinoamericanas no argentinas

Fig. 3.– Número de autores por artículo



Se hizo el estudio sobre los artículos con un máximo de 11 autores, que representan el 93% del total de publicaciones. Los números sobre las barras indican el número de publicaciones en cada caso

otros trabajando en diferentes instituciones de todo el mundo. Actualmente (2020) hay 193 países soberanos reconocidos por la ONU en el mundo[8]. Cabe aclarar que varios países cambiaron de nombre, o se han dividido o fusionado en el período de tiempo estudiado. Los 15 países que comparten la mayor cantidad de autorías con la Argentina se muestran en la Figura 4. Del total de filiaciones detectadas (812 811), estos 16 países representan el 87%. Quedó sin determinar en forma automática un 0.6% de las filiaciones. EE.UU. encabeza el número de colaboraciones. El único país latinoamericano que aparece en esta lista de colaboradores es Brasil.

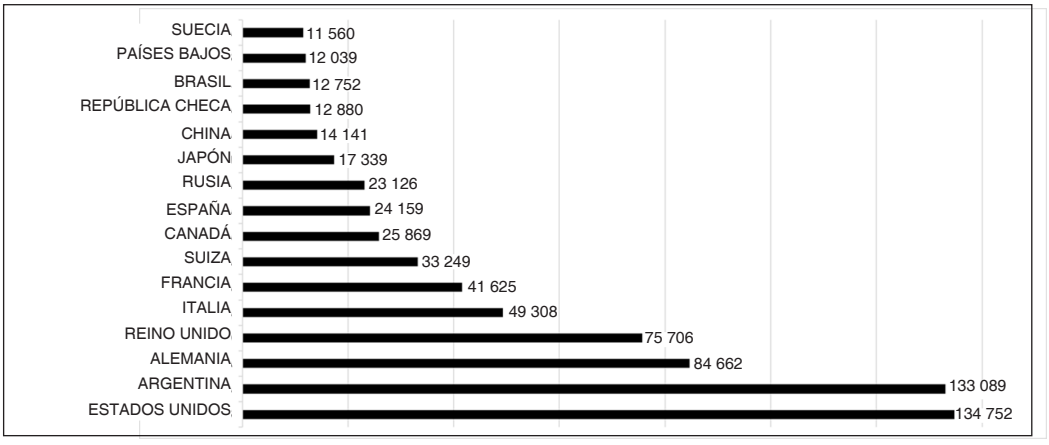
Los 70 798 resúmenes a los que hemos tenido acceso en nuestra búsqueda representan el 94% del total de publicaciones. El primer resumen detectado es el de una publicación de Bernardo Houssay y col. del año 1941[9]. Los resúmenes fueron agrupados por décadas a partir de 1980 para estudiar la evolución temporal de los términos utilizados por los autores (Tablas 1 y 2). Mediante herramientas de análisis semántico obtuvimos el listado de palabras usadas en los resúmenes y su frecuencia.

El número de resúmenes aumenta significativamente a medida que pasan los años y aumenta el número de publicaciones (comparar con Fig. 1). También se incrementa, por lo tanto, el número de unidades gramaticales totales (o *tokens*) contado en cada década. El aumento de unidades gramaticales diferentes (formas o *types*) denota la aparición de términos nuevos en el lenguaje utilizado por los autores.

En una segunda etapa se seleccionaron las primeras 100 palabras más frecuentes del listado para cada década. Nuestra hipótesis es que las palabras más frecuentes dan indicios de tendencias temáticas y técnicas, así como una indicación de la importancia dada a los términos seleccionados en cada década.

Algunas palabras muy usadas en los resúmenes (como por ejemplo *study*, *use*, *result*, entre otras) tienen poco peso para el análisis que proponemos, por lo que fueron excluidas de esta parte del estudio. En la Tabla 2 se muestran ejemplos de palabras específicas del ámbito científico. Los números representan su posición en la lista de palabras más usadas a lo largo del tiempo. Para el

Fig. 4.– Quince países que comparten la mayor cantidad de autorías con la Argentina



Junto a cada barra, el número de autorías

TABLA 1.– *Estadística de resúmenes agrupados por períodos*

Años	Resúmenes	%	*Tokens*	Formas
1980-1989	1693	2.39	152 687	18 565
1990-1999	8945	12.63	932 394	59 126
2000-2009	19 367	27.36	2 185 405	117 470
2010-2019	40 727	57.53	5 001 887	222 219
1941-2019	70 798	100.00	8 277 529	305 014

%: porcentaje del total de resúmenes que corresponde a cada período estudiado; tokens*: total de unidades gramaticales detectadas; formas (*types*): número de unidades gramaticales diferentes*

TABLA 2.– *Uso de palabras en los resúmenes a lo largo del tiempo*

Palabra	Posición en la lista				
	1980-1989	1990-1999	2000-2009	2010-2019	1980-2019
Patient	7	1	2	2	2
Cell	3	2	3	4	4
Treatment	26	17	17	11	13
Activity	6	10	11	15	11
Protein	21	15	13	17	15
Expression	–	–	23	21	26
Model	–	–	47	25	33
Gene	–	–	41	29	38
Infection	–	72	51	44	55
Population	–	–	78	45	59
Strain	20	62	36	59	52
Human	–	68	77	63	66
Acid	12	24	28	78	45
Mouse	46	41	72	96	75
Virus	51	–	–	–	–
Antibody	52	42	–	–	–
Serum	56	46	–	–	–
Rat	1	6	18	–	25
Receptor	35	39	58	–	80
Animal	28	36	68	–	82
Membrane	50	66	91	–	–
Tissue	69	76	94	–	–
Enzyme	39	52	98	–	–

Los números representan la posición de una palabra en la lista de las más usadas a lo largo del tiempo en cada período. El signo – indica que la palabra no se encuentra entre las 100 primeras en ese período de tiempo. Se respeta el idioma inglés del corpus

análisis, las palabras se presentan según su ordenación en la última década (2010-2019). Esto permite observar que muchos términos utilizados conspicuamente en los últimos años no tuvieron la misma preponderancia en otras décadas.

El *corpus* de resúmenes se usó para la identificación no supervisada de 30 temas, mediante un modelo generativo de aprendizaje automático (*machine learning*). Las palabras clave relevantes no fueron especificadas previamente, sino determinadas por el mismo sistema de análisis. El algoritmo probabilístico se basa en supuestos estadísticos como: i) el orden de las palabras y el de los documentos no es importante; ii) el número de temas se estipula por anticipado; iii) una misma palabra puede pertenecer a varios temas. Los temas con sus palabras características, determinados del modo descripto, se muestran en la Tabla 3.

También se detectó en forma automatizada la mención de enfermedades en el texto de los resúmenes. Se partió de una lista original de aproximadamente 16 500 enfermedades[10], curada y reducida por agrupamiento a una lista de poco más de 1700, que fue la utilizada para buscar concordancias en el texto. El resultado fue la detección de 514 afecciones mencionadas en 31 000 resúmenes. La Figura 5 muestra el porcentaje de menciones de enfermedades que agrupan el 60% de las detectadas, lo que señala tendencias en la selección de temas de estudio por parte de los autores.

Con el fin de buscar relaciones entre las enfermedades estudiadas y su tratamiento, el *corpus* de resúmenes se utilizó también para estudiar la mención, por parte de los autores, de alguna de las 2067 sustancias aprobadas por *Food and Drug Administration* (FDA), agencia del gobierno de los EE.UU. responsable de la regulación de medicamentos[11]. De estas sustancias, 744 se mencionan al menos una vez en 15 182 resúmenes. Las mencionadas más de 150 veces se muestran en la Figura 6.

En la lista completa de la FDA aparecen términos que describen componentes comunes a los organismos biológicos y metabolitos (calcio, óxido nítrico, aminoácidos, etc.)[12]. Excluyendo estos términos, dicha lista se redujo a 436 drogas, que se clasificaron empleando dos categorías:

TABLA 3.– *Agrupación temática automatizada*

Tema	Palabras representativas
1	SPECIES, ARGENTINA, HOST, SP, EGG, STUDY, PARASITE, PROVINCE, ADULT, DESCRIBE
2	ACTIVITY, STRAIN, COMPOUND, EXTRACT, EFFECT, ACID, PRODUCTION, GROWTH, CULTURE, STUDY
3	PATIENT, PRESSURE, HEART, CARDIAC, ARTERY, CORONARY, BLOOD, VENTRICULAR, STROKE, STUDY
4	CELL, TUMOR, CANCER, BREAST, LINE, EXPRESSION, TREATMENT, EFFECT, STUDY, GROWTH
5	SAMPLE, METHOD, OBTAIN, USE, TEST, DETECTION, CONCENTRATION, VALUE, SENSITIVITY, ANALYSIS
6	SPECIES, STUDY, POPULATION, CHANGE, RESULT, PATTERN, VARIATION, COMMUNITY, SUGGEST, DIVERSITY
7	PROTEIN, ACTIVITY, BINDING, ENZYME, PEPTIDE, ACID, SITE, RESIDUE, MEMBRANE, AFFINITY
8	STUDY, AGE, RISK, PATIENT, CI, RATE, CHILD, VS, WOMAN, PREVALENCE
9	PATIENT, DISORDER, STUDY, DISEASE, COGNITIVE, MS, TEST, CONTROL, BRAIN, RESULT
10	GENE, SEQUENCE, GENETIC, MUTATION, ANALYSIS, POPULATION, DNA, REGION, STUDY, GENOTYPE
11	PLANT, ROOT, LEAF, SOIL, GROWTH, SEED, SPECIES, ISOLATE, CROP, INCREASE
12	ACID, LIPID, FATTY, DIET, INCREASE, CHOLESTEROL, FOOD, FEED, PROTEIN, CONTENT
13	PATIENT, TREATMENT, STUDY, THERAPY, TRIAL, RECEIVE, DOSE, MONTH, DRUG, CLINICAL
14	EFFECT, RAT, BRAIN, RECEPTOR, NEURON, INCREASE, MEMORY, RESPONSE, RESULT, ACTIVITY
15	ACTIVITY, CONCENTRATION, INCREASE, EFFECT, OXIDATIVE, EXPOSURE, LEVEL, STRESS, STUDY, OXYGEN
16	PATIENT, BONE, SURGERY, SURGICAL, RESULT, PERFORM, PROCEDURE, TECHNIQUE, COMPLICATION, TREATMENT
17	SPERM, FEMALE, MALE, OOCYTE, DAY, EMBRYO, STUDY, INCREASE, REPRODUCTIVE, OVARIAN
18	GENE, PROTEIN, EXPRESSION, CELL, ROLE, MECHANISM, PATHWAY, MUTANT, SIGNAL, INVOLVE
19	CELL, EFFECT, ACTIVITY, INCREASE, ACTIVATION, RECEPTOR, EXPRESSION, PROTEIN, PATHWAY, INHIBITOR
20	RAT, DAY, LEVEL, CONTROL, INCREASE, ANIMAL, SERUM, INSULIN, DECREASE, EFFECT
21	REACTION, COMPOUND, COMPLEX, STUDY, STRUCTURE, MOLECULE, MOLECULAR, BOND, ENERGY, RESULT
22	SURFACE, WATER, PROPERTY, PH, TEMPERATURE, STUDY, RESULT, INCREASE, FILM, SOLUTION
23	VIRUS, INFECTION, ANTIBODY, CRUZI, DISEASE, VACCINE, VIRAL, PARASITE, STRAIN, ANTIGEN
24	CELL, STUDY, TISSUE, MEMBRANE, MICROSCOPY, PROCESS, DEVELOPMENT, STAGE, OBSERVE, CHANGE
25	PATIENT, DISEASE, CLINICAL, DIAGNOSIS, SYNDROME, TREATMENT, LESION, REPORT, LIVER, CAUSE
26	STRAIN, ISOLATE, INFECTION, RESISTANCE, COLI, ANTIBIOTIC, STUDY, BACTERIUM, BACTERIAL, AUREUS
27	MODEL, RESULT, DATUM, TIME, USE, STUDY, METHOD, DYNAMICS, PARAMETER, OBTAIN
28	EFFECT, INCREASE, RAT, CA2, CHANNEL, RECEPTOR, DECREASE, RENAL, CALCIUM, RESPONSE
29	CELL, RESPONSE, IMMUNE, MOUSE, INCREASE, EXPRESSION, INFLAMMATORY, CYTOKINE, INDUCE, MACROPHAGE
30	HEALTH, COUNTRY, REVIEW, CARE, RESEARCH, STUDY, INCLUDE, PROVIDE, DISEASE, DATUM

Mediante aprendizaje automático, a partir del texto de los resúmenes, se detectaron temas y sus palabras pertinentes. Se indicó al sistema que agrupara toda la información del corpus en 30 temas, con 10 palabras características en cada uno de ellos. Se respeta el idioma inglés del corpus

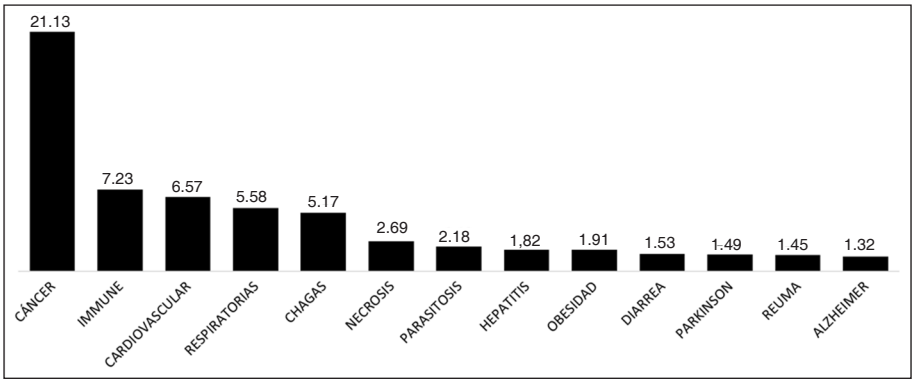
i) por sitio de acción; ii) por mecanismo de acción. Debe tenerse en cuenta que estas agrupaciones tienen limitaciones, ya que un mismo fármaco puede tener más de un sitio y/o mecanismo de acción. Las primeras 10 categorías de drogas por sitio de acción (n = 364; 83,5%) y por mecanismo de acción (n = 139; 31,9%) aparecen en las Figura 7 A y B, respectivamente.

Al compararse las Figuras 5 y 7, se observa que no hay una correlación directa entre los porcentajes de enfermedades estudiadas y de drogas aprobadas para su tratamiento. Podría concluirse que los estudios científicos sobre afecciones incluyen no solo el trabajo con estas drogas sino también con otros enfoques, como nuevas técnicas o testeos con sustancias todavía no aprobadas para su uso medicinal.

Discusión

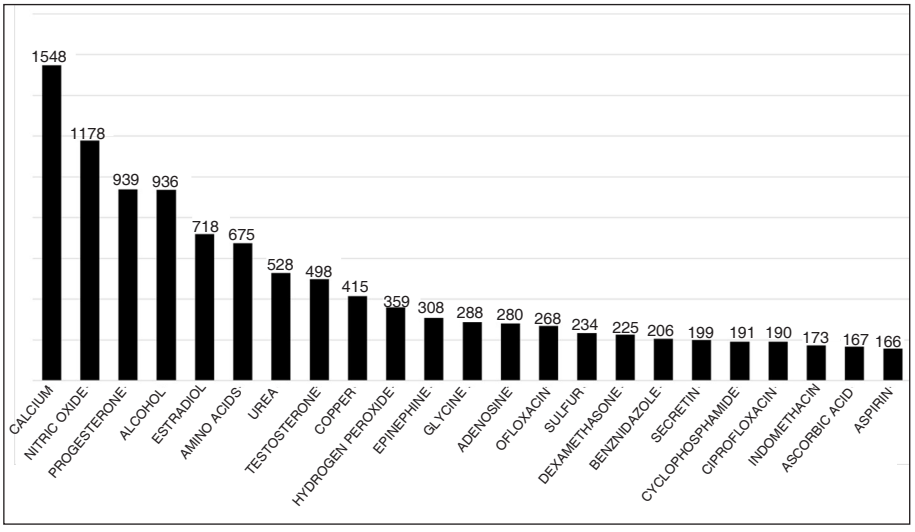
Si bien existen antecedentes de la minería de texto sobre temas relacionados con la salud[13], entendemos que es la primera vez que se presenta un estudio exhaustivo con esta herramienta sobre una masa temporal y cuantitativamente extensa de trabajos con participación argentina en el campo de las ciencias de la vida. La metodología aplicada permitió relacionar datos previamente dispersos, no estructurados, y presentarlos de forma compacta y estructurada.

Fig. 5.– Porcentaje de enfermedades mencionadas en los resúmenes de los artículos



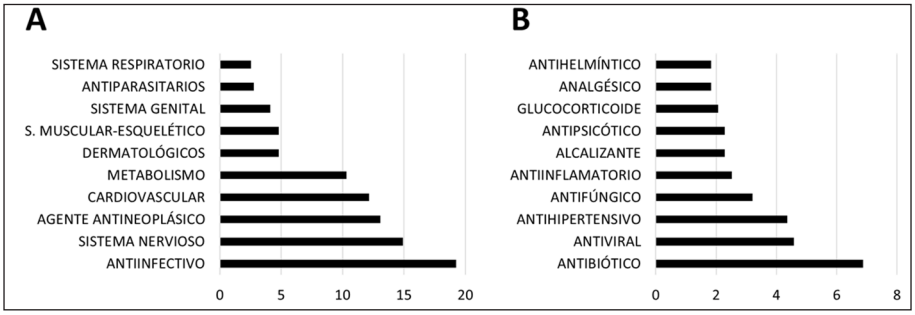
Se presenta un recorte correspondiente al 60% del total de enfermedades mencionadas. El porcentaje se indica arriba de cada columna

Fig. 6.– Sustancias aprobadas por *Food and Drug Administration* (FDA) mencionadas en los resúmenes



Se presentan las mencionadas en más de 150 oportunidades

Fig. 7.– Agrupaciones de drogas aprobadas por *Food and Drug Administration* (FDA) mencionadas en los resúmenes



A: Categorías de drogas por sitio de acción. Los 10 grupos representados corresponden al 83.5% de 436 mencionadas. B: Categorías de drogas por mecanismo de acción. Los 10 grupos representados corresponden al 31.9% de las mencionadas. Los números corresponden a los porcentajes obtenidos

Los datos fueron analizados sin ningún sesgo, como el factor de impacto de los sitios de publicación, el grupo de trabajo o los países involucrados, dando objetividad al análisis realizado.

No todos los artículos indexados en ePMC mencionan la filiación de sus autores, especialmente los más antiguos. Como ejemplo, si se hace una búsqueda de las publicaciones que firma Luis Federico Leloir (Leloir LF), resultan 85 trabajos; si a la misma búsqueda se agrega Aff=Argentina, el resultado disminuye a solo 2 trabajos. Teniendo en cuenta esta carencia, es probable que el número de publicaciones obtenido con nuestra búsqueda subestime el real; de todas formas, sigue siendo un número significativo para la estadística que presentamos.

Este trabajo nos permitió concluir que i) el número de publicaciones indexadas con participación argentina aumentó sostenidamente en los últimos 35 años, con un primer cambio en la tasa de publicaciones en 1986 y un segundo en 2006 (Figura 1); ii) la comunidad científica argentina publicó artículos relacionados con enfermedades de impacto en el país, aunque los temas tratados no se limitan exclusivamente a ellas (Tabla 3 y Figura 5); iii) existe una relación entre las enfermedades que se estudian y la búsqueda de tratamientos, más allá del uso de drogas ya aprobadas (Tabla 2, Fig. 5 y 7); iv) hay un trabajo colaborativo importante de científicos en la Argentina con investigadores extranjeros de 161 países.

Considerando los medios de publicación elegidos por los autores, la revista argentina *Medicina (Buenos Aires)* encabeza la lista de medios detectados. La revista es leída y consultada principalmente por profesionales argentinos y de países de habla hispana. Citando, sus artículos son "una retribución a la inversión que el país (la Argentina) ha hecho en su sistema científico"[14]. En relación al sistema científico vigente, es innegable que las revistas con mayor factor de impacto se publican en inglés, y tienen mayor peso en evaluaciones y promociones de sus autores. Sin embargo, marcando un déficit, "la hipercentralidad del inglés en ese sistema de publicaciones contribuyó al abandono paulatino de las lenguas locales, con el empobrecimiento cultural que conlleva y el efecto negativo que tiene en las posibilidades de vinculación entre la sociedad y la producción de conocimientos", como señalan Beigel y Gallardo[15].

En el análisis de palabras en los resúmenes se observa el uso de nuevos términos por parte de los autores a medida que pasan los años (Tabla 1), probablemente debido al desarrollo de técnicas novedosas y al descubrimiento de estructuras que conllevan la utilización de nuevos sustantivos y verbos.

En los resúmenes aparecen también términos utilizados conspicuamente en los últimos años pero que no tuvieron la misma preponderancia en otras décadas. Uno de ellos es *model* (en relación con modelos animales, matemáticos, informáticos, etc.). Esto podría deberse a una tendencia en la construcción de modelos a partir de la información científica obtenida en décadas anteriores, que más tarde pudieron integrarse.

El *corpus* general y fraccionado en décadas muestra el énfasis puesto en el paciente (*patient*) en todas las épocas analizadas, así como un enfoque de trabajo dirigido al estudio a nivel de la célula (*cell*).

La posición que ocupa la palabra *treatment* (tratamiento), a lo largo de las décadas analizadas, puede asociarse, más allá de su polisemia, con la propuesta de tratamientos para las enfermedades abarcadas en los estudios publicados; su aparición en los resúmenes permitiría descartar su uso con significado de metodología en el procesamiento de muestras.

En cuanto a la detección no supervisada de temas, consideramos que la metodología aplicada permite una agrupación que es difícil de obtener sin herramientas digitales, sobre todo cuando se trata de bases de datos como la utilizada, con decenas de miles de publicaciones. El análisis de agrupamientos y palabras clave abren la posibilidad de un seguimiento objetivo de intereses científicos que de otra forma permanecen ocultos. Los resultados pueden implementarse como forma de analizar las decisiones de los autores para elegir temas de investigación, y así detectar posibles avances y déficits en el marco de una política científica organizadora.

Con respecto a las predicciones, entendemos que este estudio debe tomarse con recaudo, ya que no puede aseverarse la dirección futura de un tema de investigación, especialmente debido a la posibilidad de aparición de eventos singulares como la enfermedad por coronavirus (COVID-19) declarada pandemia en 2020, que está teniendo fuertes implicancias en la reconversión de temas y el redireccionamiento del financiamiento. Sin embargo, junto con la mencionada identificación de temas de estudio, las predicciones y proyecciones pueden ayudar a la toma de decisiones por parte del Estado, las instituciones y los equipos de trabajo, aumentando la eficiencia en el uso de los recursos otorgados a la investigación.

Tenemos conocimiento de la publicación de artículos que aplican minería de textos sobre un *corpus* compuesto por el texto completo de las publicaciones y no solo sobre los resúmenes[16]. El mayor volumen de texto, y por ende la mayor cantidad de datos, favorecería en principio la riqueza del análisis. Sin embargo, consideramos que la información acotada que los autores vuelcan en el título, y sobre todo en los resúmenes de sus trabajos, contiene las ideas principales a ser estudiadas, lo cual es consistente con los fines que nos propusimos. Más aun, el cuerpo entero de la publicación puede contener información redundante y distractora del centro del análisis, además de implicar un alto costo en tiempo computacional, no siempre disponible en las instituciones. El flujo de trabajo que presentamos y la metodología que aplicamos permiten un análisis riguroso y fiable mediante el uso de

computadoras presentes en casi todos los laboratorios de investigación, demostrando que el uso de minería de texto puede aplicarse con relativa facilidad al estudio de otros temas similares.

Adicionalmente a la obtención de los resultados mencionados, la utilización de la minería de textos sobre publicaciones científicas puede considerarse una contribución para imaginar formas innovadoras de abordar los datos científicos, en un esfuerzo por mejorar campos como la prevención, la investigación y el tratamiento de muchas enfermedades y otros aspectos pertinentes en el ámbito de la salud humana mundial.

Agradecimientos: Este trabajo fue realizado gracias a los aportes del subsidio "Proyecto de Investigación de Unidades Ejecutoras (P-UE 2017) # 22920170100041CO" del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina, y del subsidio "UBACyT N° 20020170100733BA" de la Universidad de Buenos Aires (UBA), Argentina. Agradecemos el apoyo de NVIDIA Corporation por la donación de una GPU Titan Xp utilizada para nuestra investigación, y al Dr. Jorge Aliaga por la desinteresada provisión de sus análisis de la evolución del Presupuesto en Ciencia y Técnica en la Argentina.

Conflicto de intereses: Ninguno para declarar

Bibliografía

1. Renganathan V. Text mining in biomedical domain with emphasis on document clustering. *Health Inform Res* 2017; 23: 141-6.
2. Tshitoyan V, Dagdelen J, Weston L, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* 2019; 571, 95-8.
3. Europe PMC. En: *https://europepmc.org/*; consultado abril de 2020.
4. Lemma list 5. En: *https://lexically.net/downloads/BNC_lemmafile5.txt*; consultado abril de 2020.
5. Newman D, Asuncion A, Smyth P, Welling M. Distributed algorithms for topic models. *Journal of Machine Learning Research* 2009; 10: 1801-28.
6. Yao L, Mimno D, McCallum A. Efficient methods for topic model inference on streaming document collections, 2009. *En: https://mimno.infosci.cornell.edu/papers/fast-topic-model.pdf*; consultado octubre 2020.
7. Aad G, Abbott B, Abdallah J, et al. Combined measurement of the Higgs Boson Mass in pp collisions at sqrt[s]=7 and 8 TeV with the ATLAS and CMS experiments. *Phys Rev Lett* 2015; 114: 191803.
8. Naciones Unidas. En: *https://www.un.org/es/about-un/index.html;* consultado abril de 2020.
9. Houssay BA, Foglia VG, Smyth FS. Endocrine function of the surgically reduced pancreas. *J Exp Med* 1941; 74: 283-95.
10. Carvalho-Silva D, Pierleoni A, Pignatelli M, et al. Open Targets Platform: new developments and updates two years on. *Nucleic Acids Research* 2019; 47(D1): D1056–65. En: *https://www.targetvalidation.org/downloads/data*; consultado mayo 2020.
11. U.S. Food & Drug. En: *https://www.fda.gov/drugs/development-approval-process-drugs/drug-approvals-and-databases*; consultado mayo 2020.
12. Psychogios N, Hau DD, Peng J, et al. The human serum metabolome. *PLoS One* 2011; 6:e16957.
13. Luque C, Luna JM, Luque M, Ventura S. An advanced review on text mining in medicine. WIREs Data Mining and Knowledge Discovery 2018. En: *http://www.uco.es/grupos/kdis/wp-content/uploads/draft_advanced-review-Text-Mining-in-Medicine.pdf*; consultado octubre 2020.
14. Kotsias B. Las publicaciones científicas argentinas. *Medicina (B Aires)* 2013; 73: 597-600.
15. Beigel F, Gallardo O. Productividad, bibliodiversidad y bilingüismo en un corpus completo de producciones científicas. *Revista Iberoamericana de Ciencia, Tecnología y Sociedad* 2021; 46: en prensa.
16. Westergaard D, Stærfeldt HH, Tønsberg C, Jensen LJ, Brunak S. A comprehensive and quantitative comparison of text mining in 15 million full-text articles versus their corresponding abstracts. *PLoS Comput Biol* 2018; 14:e1005962.