Una publicación del Grupo de Estudios sobre Comunicación, Política y Seguridad
del Instituto de Investigaciones Gino Germani (UBA)

Aportes desde el procesamiento de lenguaje natural para incrementar la escalabilidad en los estudios sobre tópicos de noticias digitales securitarias

Contributions from natural language processing to increase scalability in studies on topics of digital security news

---

Florencia Nathalia Piñeyrúa
Escuela Interdisciplinaria de Altos Estudios Sociales (IDAES). Universidad Nacional de San Martín (Argentina).
Correo: pinieyrua@gmail.com

Fecha de recepción: 15 de junio de 2021
Fecha de aceptación: 17 de diciembre de 2021

Resumen:

Este trabajo explora la aplicación de técnicas de procesamiento de lenguaje natural y *web scraping* para el estudio de contenido de noticias digitales a gran escala. Para ello, trabajamos con datos primarios construidos a partir de la técnica de *web scraping* utilizando como soporte empírico las noticias publicadas desde julio a septiembre 2019 en los portales *Clarín*, *La Nación, Infobae, Página 12, Télam, Perfil*, *Crónica* y *Minuto Uno.* En el procesamiento del corpus empleamos la técnica de procesamiento de lenguaje natural para la detección de tópicos con la implementación del método *Latent Dirichlet Allocation* (LDA). Los resultados muestran que los principales tópicos de la agenda mediática digital durante el contexto de las elecciones Primarias Abiertas Simultáneas y Obligatorias son las elecciones, los espectáculos, el deporte, la seguridad y la política

exterior. El caso securitario es un tópico estable y relevante de la agenda mediática digital, aunque su prevalencia no aumenta durante el mes electoral. La conclusión principal es que la combinación de las técnicas *web scraping* y procesamiento de lenguaje natural pueden ser útiles para incrementar la escalabilidad (aumentar la captura de información y reducir los tiempos de selección y análisis de tópicos) en los estudios de contenido de noticias.

Palabras clave: tópicos, procesamiento de lenguaje natural, *web scraping* y noticias digitales.

Abstract

This paper explores the application of natural language processing and web scraping techniques for the study of large-scale digital news content. For this purpose, we work with primary data constructed from the web scraping technique using as empirical support the news published from July to September 2019 in the portals Clarín, La Nación, Infobae, Página 12, Télam, Perfil, Crónica and Minuto Uno. In the corpus processing we employed the natural language processing technique for topic detection with the implementation of the Latent Dirichlet Allocation (LDA) method. The results show that the main topics of the digital media agenda during the context of the Simultaneous and Mandatory Open Primary Elections are elections, entertainment, sports, security and foreign policy. The security case is a stable and relevant topic of the digital media agenda, although its prevalence does not increase during the electoral month. The main conclusion is that the combination of web scraping and natural language processing techniques can be useful to increase scalability (increase information capture and reduce topic selection and analysis times) in news content studies

Keywords: topics, natural language processing, web scraping

techniques and digital news.

1. Introducción

Este artículo explora las potencialidades y limitaciones del uso de técnicas de procesamiento de lenguaje natural (topic modelling con la implementación del método *Latent Dirichlet Allocation* LDA) y *web scraping*[1] para el análisis de medios desde la propuesta teórica de la Agenda *Setting*. Al interior de este campo de estudio, que aborda los contenidos de noticias, existen ciertas dificultades metodológicas: problemas para relevar volúmenes grandes de información y los altos costos en tiempo, recursos, etc. que implica la detección manual de los tópicos (Orozco Gómez y González, 2012). Sin embargo, el uso de técnicas de procesamiento de lenguaje natural y *web scraping*, aunque poco exploradas por las Ciencias de la Comunicación, pueden ser una herramienta que permita sortear alguna de tales dificultades.

El presente trabajo se centra en la dimensión metodológica del problema, nos proponemos mostrar que es posible ampliar la cobertura y reducir los tiempos de detección y análisis de los tópicos de noticias digitales combinando la recolección automatizada de piezas periodísticas y la técnica de modelado de tópicos. Nuestro objetivo es explorar la aplicación de algunas técnicas de análisis vinculadas al campo del procesamiento de lenguaje natural, ya que el esfuerzo por automatizar tareas que normalmente se realizan de forma manual podría permitir sortear algunas limitaciones metodológicas -como ser la escalabilidad y replicabilidad- presentes en los análisis de medios (Orozco Gómez y González, 2012). En efecto, la relevancia del abordaje del problema de investigación reside en la existencia de dificultades en el tratamiento metodológico de los tópicos de noticias relacionado a las grandes cantidades de tiempo que conllevan la selección y el análisis de tópicos.

A nivel empírico, analizamos de forma exploratoria los cambios y continuidades que experimentaron los tópicos de la agenda mediática digital desde

[1] Técnica que permite descargar y formatear la información disponible en sitios web, la cual en general no se encuentra en condiciones de ser trabajada de forma cuantitativa (Mitchell, 2015 en Rosati, 2021).

julio a septiembre 2019, período en el cual se desarrollaron las elecciones Primarias Abiertas Simultáneas y Obligatorias (PASO)[2]. Estudios recientes (Koziner et al., 2018) estiman que el nivel de consumo de los medios digitales asciende a prácticamente la mitad de la población (SINCA - Encuesta Nacional de Consumos Culturales del Ministerio de Cultura de la Nación Argentina, 2017). Por lo anterior, optamos abordar los objetivos de este trabajo utilizando como soporte empírico las piezas periodísticas publicadas en los principales medios digitales de comunicación de Argentina: *Clarín*, *La Nación, Infobae, Página 12, Télam, Perfil*, *Crónica* y *Minuto Uno*. A su vez, el periodo seleccionado entre julio a septiembre de 2019, contexto electoral de las PASO 2019, no se ha estudiado aún desde la perspectiva de la agenda mediática securitaria, optamos por incluir un mes de campaña electoral -julio-, un mes donde se desarrollan los comicios -agosto- y un mes posterior a los resultados –septiembre-.

Con base a lo relatado, formulamos una hipótesis que explora la dimensión metodológica del problema y el proceso de cobertura mediática. La aplicación de técnicas de procesamiento de lenguaje natural y *web scraping* permite aumentar la cobertura de noticias y reducir los tiempos de detección y análisis de tópicos relevantes y caracterizar la agenda mediática digital desde julio a septiembre de 2019, donde la prevalencia de noticias sobre delito y seguridad aumenta durante el mes de las elecciones Primarias Abiertas Simultáneas y Obligatorias.

Nos proponemos explorar las principales fortalezas y problemas de emplear técnicas de análisis textual computacional en el estudio de contenido de noticias. Para ello, el artículo se organiza de la siguiente manera, primero, revisamos los antecedentes teórico-metodológicos de investigaciones recientes sobre tópicos de noticias en Argentina. Segundo, desarrollaremos la perspectiva metodológica y el flujo de trabajo para el análisis textual computacional. Tercero, presentamos los resultados del análisis del modelado de tópicos a partir de tablas y gráficos. En un cuarto momento especificamos los aportes y problemas de la aplicación de técnicas de procesamiento de lenguaje natural empleadas, estableciendo una comparación con las técnicas de análisis de contenido cuantitativo que caracterizan las

[2] Las PASO se realizaron el domingo 11 de agosto de 2019. En estas elecciones los espacios políticos dirimen sus candidaturas de cara a las elecciones generales, tanto a cargos nacionales y provinciales del Poder Ejecutivo como del Poder Legislativo.

estrategias metodologías de las investigaciones sobre tópicos de noticias securitarias en Argentina. Por último, finalizamos con una reflexión sobre las fortalezas de combinar los enfoques metodológicos computacionales y cuantitativos en la investigación social.

2. Los estudios de contenido: una metodología para analizar los tópicos de noticias

Las investigaciones en Argentina que indagan sobre el contenido mediático, en general, comparten las propuestas teóricas de la Agenda *Setting* y del *Framing* (Ariza y Beccaria, 2019; Aruguete, 2015; Zunino y Grilli Fox, 2019). Dentro de las teorías que se han dedicado al análisis de contenido de medios, la Agenda *Setting* postula que el público es consciente o ignora, presta atención o descuida, enfatiza o pasa por alto elementos específicos de los temas públicos en relación con lo que muestran los medios de comunicación masivos[3]. La agenda mediática es definida como el patrón de cobertura de noticias durante un tiempo determinado (McCombs, 2015), es decir, un conjunto de cuestiones comunicadas en función de una determinada jerarquía (Aruguete, 2009). La cobertura mediática hace referencia a un dispositivo que tiene la capacidad de incluir y descartar ciertos acontecimientos y omitir otros u otorgar diferentes niveles de jerarquías informativas (Aruguete, 2015). De este modo, los temas que aparecen en la agenda tienen preferencia sobre aquellos que no están. Siguiendo con las elaboraciones, la importancia de estudiar la cobertura mediática de ciertos temas radica en que su mera presencia marca la prioridad de intereses (Sádaba, 2008 en Aruguete, 2009). En las Ciencias Sociales existen varios enfoques para el análisis de contenido, en este apartado abordamos la técnica de análisis de contenido cuantitativo porque es una de las más empleadas en las investigaciones argentinas que abordan la cobertura y el tratamiento mediáticos de un asunto tanto en la prensa gráfica y *online* (Aruguete, 2009).

La técnica análisis de contenido cuantitativo es introducida en los Estados Unidos en la década de 1930 junto al nacimiento de las escuelas de periodismo y

[3] Para más detalles, véase: Arguete, N. (2015) *El poder de la agenda. Política, medios y público.* Editorial Biblos/Cuadernos de comunicación.

fue concebida como "una técnica de investigación destinada a formular, a partir de ciertos datos, inferencias reproducibles y válidas que puedan aplicarse a su contexto" (Krippendorff, 1990: 28 en Zunino y Grilli Fox, 2019: 403). A partir de la revisión de bibliografía especializada, Zunino y Grilli Fox (2019) destacan tres de sus características centrales. La primera es la sistematicidad pues está sometida a reglas explícitas replicables por otros investigadores. Si bien existen reglas, la codificación sigue siendo manual, con lo cual es muy probable que cada codificador tenga cierto margen de interpretación de tales reglas. La segunda es que es una técnica cuantitativa porque su aplicación permite medir diversas variables al transformar un documento en resultados numéricos. En tercer lugar, Zunino y Grilli Fox (2019) afirman que la técnica de análisis de contenido cuantitativo es objetiva ya que a partir de técnicas específicas se intenta reducir al máximo los sesgos del investigador en los hallazgos del estudio. A continuación, comparamos las estrategias metodológicas de diversos estudios sobre tópicos de noticias en Argentina contemporánea.

Los trabajos que emplean esta técnica para estudiar la cobertura y el tratamiento mediáticos en la prensa argentina tienen la potencialidad de emplear sistemas de categorías extensos que relacionan múltiples variables en el tratamiento mediático del delito. Particularmente, Focás y Zunino (2017) y Zunino y Focás (2019a y 2019b) indagan sobre la frecuencia de temas y tópicos, las fuentes de información externas a la redacción -fuentes oficiales o familiares de la víctima, entre otras-, la localización de los ilícitos -por provincia o grandes regiones-, la jerarquía de la información -si aparece en tapa, si abre sección, si está en página impar, en mitad superior, si tiene gran tamaño, firma o títulos grandes, etc.-. También, analizan el tratamiento de víctimas y victimarios, la edad, la clase social y los actores sociales según el rol. Además, estudian las causas atribuidas por los periódicos a la problemática securitaria como ser inseguridad, género, DDHH, protesta social, abuso de menores u otros. Asimismo, analizan las soluciones promovidas por los diarios a dichas problemáticas; estas pueden ser salidas punitivistas, políticas de contención o que no haya salida posible. Se interrogan además sobre la evaluación moral que ejercen los medios sobre los hechos que relatan. Por su parte, Zunino y Grilli Fox (2019) indagan en los medios *online* el promedio de la extensión, la autoría de las piezas, el género del autor y

la frecuencia de fotografías, audios y videos por noticia.

En relación con los abordajes metodológicos, los estudios en Argentina que emplean el análisis de contenido cuantitativo presentan algunas características metodológicas comunes (Orozco Gómez y González, 2012). Suelen recolectar el *corpus* de noticias de manera manual e incluir las piezas periodísticas siguiendo criterios específicos definidos por los investigadores previamente. En general, los estudios sobre la prensa online seleccionan los periódicos *Clarín, La Nación* e *Infobae* por su relevancia en términos de preferencias masivas en el consumo de noticias digitales (Koziner, 2019). A su vez, como los *corpus* recolectados en la prensa *online* suelen ser extensos los investigadores los reducen a una escala abordable. Para ello, existe una amplia variedad de formas de realizar este proceso. A continuación, centraremos nuestra atención en dos estrategias que emplean las investigaciones sobre tratamiento informativo en la Argentina.

Por un lado, se realiza un muestreo simple al azar donde todas las piezas periodísticas tienen la misma probabilidad de ser seleccionadas. Este tipo de muestreo "garantiza una distribución representativa de los casos en relación con los períodos y diarios relevados" (Zunino y Focás, 2019a: 88). Por ejemplo, en el estudio de Focás y Zunino (2017) se recolectó un *corpus* compuesto de 1328 piezas periodísticas y que luego se redujo al 22,5% a través de una muestra aleatoria simple. Para establecer la fiabilidad de los datos, los cientistas sociales recurren a una serie de estrategias: seleccionan aleatoriamente el 10% de la muestra para calcular el Coeficiente de Correlación Rho de Spearman (Focás y Zunino, 2017 y 2019a; Zunino y Grilli Fox, 2019) o el valor alfa de Krippendorff y Fleiss (Dammert y Erlandsen, 2020). En otras palabras, los investigadores recodifican una parte del *corpus* y ensayan pruebas estadísticas de correlación para comparar entre la codificación original y la prueba de fiabilidad y observar si existen (o no) diferencias significativas. Si no las hay, se admite que el trabajo empírico está bien hecho.

Por otro lado, Observatorio de Medios de la UNCuyo realiza la recolección del *corpus* con el método de una semana construida aleatoriamente para cada mes. La estrategia metodológica se basa en un análisis de contenido sobre 1470 piezas periodísticas recolectadas entre abril y noviembre de 2019. El universo de estudio se delimitó de la siguiente manera: las cinco primeras notas publicadas en las *homepage* de los diarios y en dos cortes horarios (9 y 19 horas). A partir de los

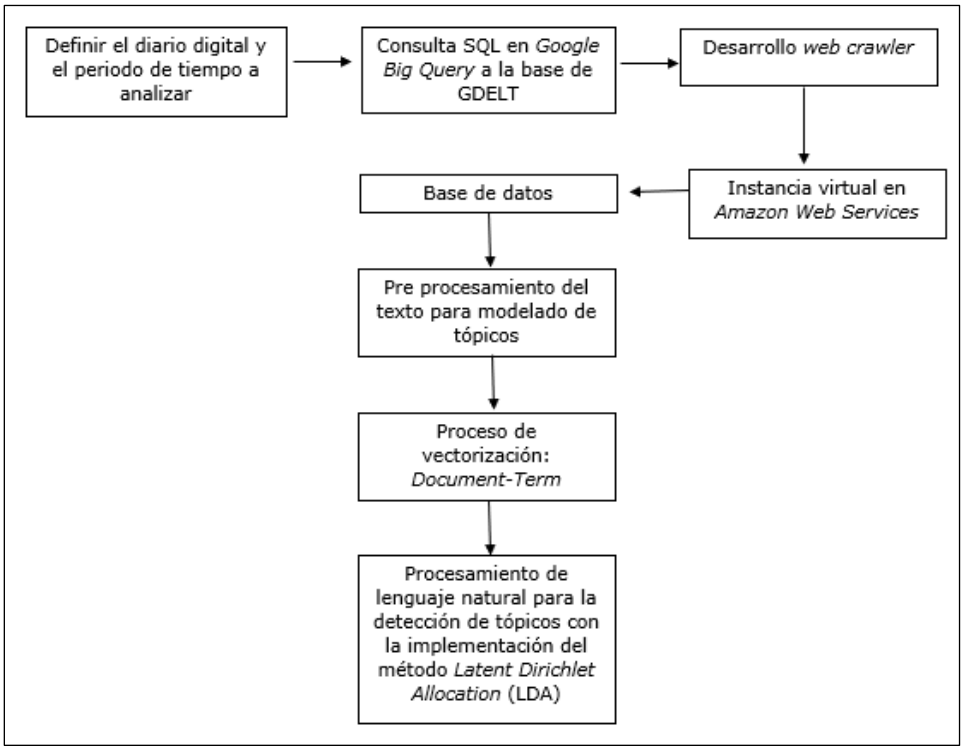
datos proporcionados por el Observatorio, Koziner (2019) realiza un estudio sobre los principales medios digitales del país en el año 2019 centrando su atención en la frecuencia y relevancia de diversos tópicos de las noticias *online.*

En resumen, las principales fortalezas del análisis de contenido cuantitativo para abordar el estudio de tópicos de noticias se relacionan a la multiplicidad de aspectos que es posible abordar sobre las características de la cobertura mediática de dichas noticias. También aportan un sistema categorial con el cual abordar las diversas relaciones entre las variables. A la par de las fortalezas existen algunas limitaciones de este enfoque. El abordaje metodológico de los *corpus* de noticias que puede calificarse como poco escalable (Orozco Gómez y González, 2012). A su vez, los criterios de clasificación definidos por los investigadores y el carácter manual de las codificaciones pueden introducir sesgos.

3. Notas metodológicas: construcción del *corpus*, pre procesamiento del texto y detección de tópicos

En la búsqueda de moderar alguna de las limitaciones presentes en el análisis textual de los estudios de contenido cuantitativos, optamos por emplear una metodología computacional que nos permite incrementar la cobertura de noticias y reducir los tiempos de detección y análisis de los tópicos relevantes. Por esta razón aumentamos la cantidad de medios de comunicación y la cobertura de noticias analizadas, en relación con los estudios de medios abordados en el apartado anterior. En este artículo trabajamos sobre un *corpus* compuesto por 52154 noticias publicadas desde julio a septiembre de 2019 en los medios digitales *Clarín*, *La Nación, Infobae, Página 12, Télam, Perfil*, *Crónica* y *Minuto Uno*. A continuación, explicamos los procedimientos para recibir, almacenar y procesar la información, es decir, cómo construimos el *corpus* de textos, qué métodos empleamos para armar la matriz de datos y para la detección de tópicos latentes.

Esquema del flujo de trabajo para el análisis textual computacional.



En la construcción del *corpus* usamos la técnica de *web scraping*. Para la recolección de las noticias partimos de realizar una consulta SQL en *Google Big Query*[4] a la base de datos de *Global Database of Events, Language and Tone* (GDELT)[5] y exportamos un archivo que contiene los 52154 *links* de las noticias. Luego, armamos un *web crawler* que detecta la información asociada a las noticias disponible en los sitios web de los medios de comunicación digitales para guardarla en una base de datos. En otras palabras, desarrollamos un programa (*web crawler)* que recorre automáticamente los links y detecta la información asociada a las noticias digitales (título, fecha, medio, texto, link, entre otros) para guardarla en una base de datos. A continuación, generamos una instancia virtual en *Amazon Web Services* (un proveedor de servicios de cómputo en la nube) que contiene al

[4] Una consulta Google Big Query es tipo de consulta a una base de datos empleando lenguaje similar a la sintaxis Structured Query Language (SQL).
[5] GDELT es un grafo global de datos abiertos en tiempo real (se actualiza cada 15 minutos) sobre la sociedad humana tal como se la ve a través de los medios de comunicación del mundo. Este proyecto incluye en su base de datos los links de las noticias que son publicadas en los portales de los diversos medios digitales.

scraper y a la base de datos, donde se ejecuta y se almacena la información recolectada.

Tabla 1. Ejemplo de base de datos utilizada.

	DATE	link	titulo	texto	Medio
0	2019-09-29 23:45:00	https://www.clarin.com/fama/soledad-pastorutti...	Soledad Pastorutti contó el detrás de escena d...	Recién llegada de España, donde participó junt...	Clarín
1	2019-09-29 23:45:00	https://www.clarin.com/politica/alberto-fernan...	Alberto Fernández le hará un homenaje al Procu...	Amado Boudou está cumpliendo 5 años y 10 meses...	Clarín
2	2019-09-29 23:45:00	https://www.clarin.com/sociedad/muerte-argenti...	Muerte de una argentina en Punta Cana: la auto...	"Hacemos constar que le fue practicada la necr...	Clarín
3	2019-09-29 23:45:00	https://www.clarin.com/politica/salarios-vs-in...	Salarios vs. inflación: 17 derrotas, 1 empate ...	Marcos Peña suele mostrar su celular cuando qu...	Clarín
4	2019-09-29 23:45:00	https://www.clarin.com/politica/cerro-votacion...	Cambiemos se impuso con comodidad en Mendoza y...	Había alerta de viento Zonda, pero al final no...	Clarín

Fuente: elaboración propia.

Especificaciones sobre el pre procesamiento del texto

Partimos de realizar un pre-procesamiento del *corpus* con la finalidad de producir datos limpios. El primer paso es normalizar el texto convirtiendo todo el *corpus* en minúscula. En segundo lugar, eliminamos sitios web, signos de puntuación y números. Luego realizamos el proceso de *tokenization* que refiere a "dividir" a cada documento del *corpus* en sus "unidades mínimas", en nuestro caso en palabras. Por último, removimos las palabras más comunes de la lengua[6], proceso denominado *stop words* tanto por lista (artículos, etc.) como por frecuencia. Es decir, se eliminaron tanto los términos muy frecuentes (que aportan poca información sobre el contenido del texto) como los muy poco frecuentes que probablemente produzcan alguna forma de *overfitting*[7].

Para que el texto no estructurado sea procesado por un algoritmo o técnica de aprendizaje automático es necesario representar el documento (las noticias) en una forma de espacio vectorial. En otras palabras, dado que la información, es en su mayoría texto libre es necesario darle un formato que permita hacerla operable. El objetivo de este procesamiento es representar el texto de forma "vectorial", es decir, representar cada documento del *corpus* como un vector en un cierto espacio. Existe una amplia variedad de formas de realizar el proceso de vectorización, en

[6] Un ejemplo son las preposiciones que aparecen en todos los documentos y no aportan información valiosa para distinguirlos.
[7] En aprendizaje automático el término *overfitting* o sobreajuste hace referencia al efecto de sobreentrenar un algoritmo con ciertos datos para los que se conoce el resultado deseado a predecir.

este caso optamos por emplear la llamada *Document-Term Frequiency Matrix* (DTM).

El diseño de la matriz de frecuencia término-documento implica tabular el *corpus* (texto libre) de forma tal que respete la estructura tripartita del dato (Galtung, 1970). El *corpus* de textos se dispone en una representación vectorial de la siguiente manera: en la columna, los documentos; en la fila, cada *token* (en este caso, cada palabra del vocabulario); y en la intersección entre ambos está la frecuencia de aparición (una forma de conteo crudo o ponderado mediante una métrica llamada TF-IDF que veremos enseguida) de cada *token* en cada documento. Al construir esta matriz se pierde la información sobre el orden de las palabras, el orden de las columnas es ahora arbitrario y no respeta la estructura secuencial de las palabras en un texto. Por eso, suele llamarse "bolsa de palabras" (*bag of words*) a esta forma de representar el texto.

Siguiendo las elaboraciones de Rosati (2021), existen dos dimensiones de las frecuencias de los términos de un *corpus*:

1. un término *t* es importante si es muy frecuente en un documento *d* del *corpus c* analizado,
2. un término *t* es más informativo del contenido de un documento *d* si el *t* está presente en pocos documentos y no en todos los documentos del *corpus*.

Por consiguiente, hay que observar la frecuencia del término a lo largo de todo el *corpus* y al interior del documento determinado. A modo de ejemplo mostramos la tercera noticia del *dataset* y la forma en que se dispone la información en la matriz de frecuencia.

Imagen 1. Ejemplo de la forma de disponer la información de las piezas periodísticas en la matriz de frecuencia.

Fuente: elaboración propia.

Tabla 2. Ejemplo de la construcción de matriz término-documento en frecuencia absoluta.

	documento 1	documento 2	documento 3	documento 4	documento 5
muerte	1		1		
autopsia			1		1
fiscal		2	1		

Fuente: elaboración propia en base a Rosati (2021).

Uno de los principales problemas que trae aparejada esta forma de representar la información es que en los textos más extensos hay una sobrerrepresentación de palabras pues se realiza un conteo absoluto. Para sortear dicha dificultad se suele normalizar el conteo como una proporción sobre el total de palabras en el documento. La tabla 3, que presentamos a continuación, permite observar cómo se normaliza el texto por fila.

Tabla 3. Ejemplo de la construcción de matriz término-documento en proporciones.

	documento 1	documento 2	documento 3	documento 4	documento 5
muerte	0,25		0,5		
autopsia			0,25		0,5
fiscal		0,5	0,5		

Fuente: elaboración propia en base a Rosati (2021).

De este modo, para medir la importancia de una palabra en un documento empleamos la métrica *Term Frequency* (TF): el conteo crudo normalizado por la

extensión (el total de términos) del documento.

Por lo que respecta a la informatividad de un término a lo largo del *corpus*

$$TF(t,d) = \frac{c(t,d)}{\sum_{t \in d} c(t,d)}$$

c, empleamos la métrica *Inverse Document Frequency* (IDF) que calcula la proporción de documentos del *corpus* que contienen el término *t*. Asimismo, es importante aclarar que utilizamos la inversa de la métrica *Document Frequency* (DF) porque permite realizar una lectura más intuitiva: cuanto mayor es DF(t) menos informativo es el término *t.* Así, IDF(t) es mayor cuanto más imformativo

$$IDF(t) = \log \frac{|C|}{df(t)}$$

es *t*, es decir, cuanto menor sea la frecuencia de *t* en el *corpus c*.

Las métricas *Term Frequency* (TF) y *Inverse Document Frequency* (IDF) se agrupan en la matriz *Term Frequency-Inverse Document Frequency* (TF-IDF) que mejora el conteo crudo de la aparición de cada palabra en los documentos y permite medir la importancia y la informatividad de cada término a lo largo del *corpus* de textos analizado, en nuestro caso de estudio son las noticias digitales de los principales diarios online de Argentina desde julio a septiembre de 2019. De esta manera, generamos el *input* para la detección de tópicos.

Detección de tópicos

Una vez obtenida la matriz de frecuencia término-documento empleamos técnica de procesamiento de lenguaje natural para la detección de tópicos (*topic modelling*) y, en particular, con la implementación del método *Latent Dirichlet Allocation* (LDA) que permite estimar los principales temas dentro de un *corpus* y clasificar documentos individuales en esas categorías. Al ser una técnica de aprendizaje automático no supervisado, donde no hay variable dependiente y los datos a modelar no ofrecen información acerca del resultado a predecir, existen complejidades para estimar la evaluación del modelo. La estructura de tópicos -la composición de tópicos por documento y la probabilidad de pertenencia de cada

palabra a un tópico- puede ser considerada como un conjunto de variables no observadas que se tratan de estimar (Rosati, 2021). De esta forma, permite agrupar las noticias según su tema prevalente sin recurrir a una codificación *a priori* del investigador. No obstante, la participación y el conocimiento del investigador es necesario al momento de interpretar los tópicos detectados. En particular, el modelado de tópicos cambia la interpretación sustantiva del cientista social a un paso posterior en el proceso analítico (Mützel, 2015). Según Nelson (2017), este pasaje de interpretación de la creación de categorías a la interpretación de categorías estimadas aleja a los investigadores de los datos y de los sesgos culturales e históricos que los acompañan.

Uno de los instrumentos más difundidos en la actualidad para realizar una modelización de tópicos es *Latent Dirichlet Allocation Models* (LDA). El algoritmo LDA identifica tópicos latentes (grupos de palabras) a partir de un modelo que se basa en la coocurrencia (repetición) de palabras y en el significado contextual para realizar la detección de tópicos (Mützel, 2015). A grandes rasgos, este es un modelo inferencial *bayesiano* que propone un proceso generativo de los documentos del *corpus*. Cada palabra es el resultado de un encadenamiento de distribuciones sobre las que luego se realiza inferencia hacia atrás para calcular la distribución más probable dada las palabras y los documentos (Koslowski, 2019). Siguiendo las elaboraciones, el modelo LDA selecciona aleatoriamente una distribución a lo largo de la cantidad de tópicos (hiperparámetros definido por el investigador), luego para cada término del documento elige un tópico de la distribución de tópicos del documento y, por último, escoge de manera aleatoria una palabra del tópico correspondiente.

En este marco, es importante subrayar cuatro supuestos del modelo LDA. Primero, se considera a un texto como una secuencia de palabras y a una palabra como una secuencia significativa de caracteres (Rosati, 2021). En segundo lugar, los tópicos son preexistentes a los documentos. El tópico es definido como una distribución de probabilidad sobre el vocabulario a lo largo del *corpus* de texto. El conjunto de palabras tiene una determinada probabilidad de pertenecer a un tópico. Así, el modelo LDA estima cuáles son los términos que tienen la mayor probabilidad de pertenecer a un determinado tópico. El tercer supuesto es que cada documento es una mixtura de tópicos, en otras palabras, contiene palabras que

pertenecen a una pluralidad de tópicos en proporciones particulares. Por ejemplo, al realizar la modelización de tópicos observamos que la tercera noticia de nuestro *corpus* tiene un 56% de probabilidad de pertenecer al tópico 4 (seguridad) y un 6% al tópico 5 (política exterior). El cuarto supuesto del modelo LDA es que cada tópico es una mixtura de palabras. A modo de ejemplo, en el tópico "seguridad" las palabras como "policía", "víctima", "hombre", "joven", "mujer" tienen altas probabilidades de pertenencia. En cambio, términos como "primarias", "kirchnerismo", "oposición", "oficialismo", "votos" están más asociadas a un tópico que hable sobre "elecciones". Es importante subrayar que las palabras pueden ser compartidas entre tópicos, por ejemplo, la palabra "argentina" tiene una alta probabilidad de pertenencia a los tópicos sobre deportes, economía y obra pública/interés general. Así, el método LDA permite que los tópicos se "solapen" en un documento particular del corpus, en lugar de tratar a los tópicos como categorías excluyentes.

El resultado del algoritmo LDA es, por un lado, una distribución de palabras por tópico y, por otro, una distribución de los tópicos por documento. Estos datos se encuentran organizados en dos matrices ordenadas en una estructura de filas y columnas. Para analizar la evolución en el tiempo de la relevancia de los tópicos de las noticias digital desde julio a septiembre de 2019 realizamos la visualización de los resultados del modelo LDA a partir de la librería *ggplot* del lenguaje R.

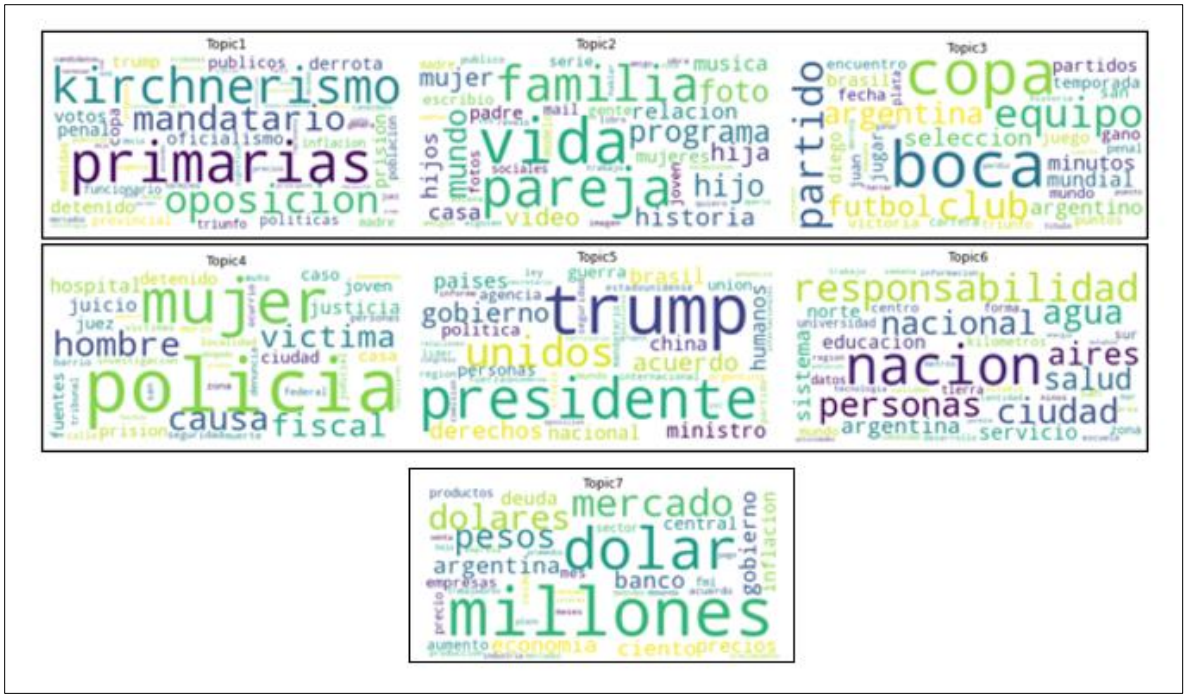
4. Composición y evolución temporal de los tópicos de las noticias digitales

Utilizando el modelo LDA buscamos responder a nuestra pregunta de investigación: ¿cuáles son los principales temas de la agenda mediática digital entre julio y septiembre de 2019? Como mencionamos anteriormente, para identificar los tópicos más relevantes con la implementación de este algoritmo se debe determinar la cantidad de tópicos que se busca detectar. Al analizar nuestro *corpus* es importante destacar que independientemente de la cantidad de tópicos seleccionada, existen algunos temas que se mantienen en la agenda: elecciones, seguridad, economía, espectáculos y deportes. A su vez, es importante generar los tópicos que tengan sentido semántico, es decir, que sean interpretables.

Recapitulando, el *dataset está compuesto* por 52154 noticias de los medios digitales de comunicación *Clarín*, *La Nación, Infobae, Página 12, Télam, Perfil*,

Crónica y *Minuto Uno* que fueron publicadas en sus portales desde julio a septiembre de 2019. En nuestro caso seleccionamos el modelo que muestra un total de 7 tópicos. En el gráfico 1 mostramos las 30 palabras con mayor probabilidad de pertenencia a cada tópico donde el tamaño de cada término del vocabulario es proporcional a dicha probabilidad.

Gráfico 1. Nubes de palabras con mayor probabilidad de partencia a cada tópico noticioso.



Fuente: elaboración propia.

En el primer tópico las palabras con mayor probabilidad de pertenencia son "primarias", "kirchnerismo", "oposición", "mandatario", "oficialismo", "votos" que están asociadas a las elecciones PASO 2019. Una pieza periodística que pertenece al mismo es:

Imagen 2. Ejemplo de noticia agrupada en el tópico Elecciones.

Fuente: elaboración propia.

Las palabras con mayor probabilidad de pertenecer al segundo tópico son "vida", "pareja", "familia", "programa", "hijo", "foto" y "mundo", términos vinculados al espectáculo. A modo de ejemplo presentamos una noticia agrupada al tópico:

Imagen 3. Ejemplo de noticia agrupada en el tópico Espectáculos.

Fuente: elaboración propia.

El tópico 3 contiene piezas periodísticas sobre deportes, los términos con mayor probabilidad de coocurrencia son "boca", "copa", "equipo", "partido", "club", "argentina" y "fútbol", "selección".

Imagen 4. Ejemplo de noticia agrupada en el tópico Deportes.

Fuente: elaboración propia.

El tópico 4 capta noticias securitarias. Los términos "policía", "mujer", "hombre", "causa", "víctima", "fiscal", "hospital", "justicia", "juicio", "detenido", "joven", "prisión", "ciudad", "juez" tienen mayores probabilidades de pertenencia. A modo de ejemplo, en este tópico aparece la siguiente noticia.

Imagen 5. Ejemplo de noticia agrupada en el tópico Seguridad.

Fuente: elaboración propia.

El quinto tópico se vincula a noticias sobre Política exterior, los términos con mayor probabilidad de coocurrencia son "trump", "presidente", "unidos", "gobierno", "acuerdo", "derechos", "países", "brasil", "ministro", "humanos" y "china".

Imagen 6. Ejemplo de noticia agrupada en el tópico Política exterior.

Fuente: elaboración propia.

El sexto tópico compuesto por palabras como "nación", "responsabilidad", "personas", "ciudad", "agua" y "nacional" muestra noticias vinculadas a la sección Obra pública/Interés General.

Imagen 7. Ejemplo de noticia agrupada en el tópico Obra pública/Interés General.

Fuente: elaboración propia.

Por último, en el séptimo tópico se logran evidenciar noticias sobre economía y menciona términos como "millones", "dólar", "mercado", "dólares", "pesos", "argentina", "banco", "gobierno", "economía", "precios", "deuda", "inflación".

Imagen 8. Ejemplo de noticia agrupada en el tópico Economía.

Fuente: elaboración propia.

Por lo hasta aquí relevado, podemos establecer que el modelado de tópicos permite clasificar las piezas periodísticas según su tópico prevalente sin recurrir a una codificación *a priori* del investigador/a. Las técnicas de análisis textual computacional cambian la interpretación sustantiva del científico social a una etapa posterior en el proceso analítico (Mützel, 2015). En este caso, la interpretación se pone de manifiesto en la etiqueta de tópicos.

Tabla 4. Etiquetado de tópicos noticiosos.

Tópico	**Etiqueta**
1	Elecciones
2	Espectáculos
3	Deportes
4	Seguridad
5	Política exterior
6	Obra pública/Interés General
7	Economía

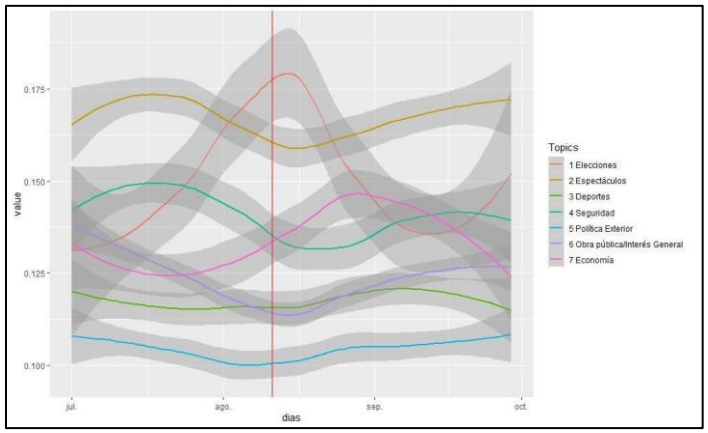
Fuente: elaboración propia.

Los resultados del modelo LDA muestran al caso de las elecciones nacionales como el principal tópico de la agenda mediática *online* desde julio a septiembre de 2019. Los espectáculos, los deportes, la seguridad, la política exterior, la obra pública y la economía también fueron prioridad en las agendas mediáticas digitales entre

julio y septiembre de 2019. En la misma dirección, el Observatorio de Medios de la UNCuyo establece el tópico electoral como el principal de la agenda mediática *online* desde abril a septiembre de 2019 (Koziner, 2019). Sin embargo, al comparar los hallazgos de ambas investigaciones observamos diferencias en el orden de relevancia de estos tópicos: las elecciones nacionales, la economía, la seguridad, los deportes, la política, la corrupción y los espectáculos. Las diferencias en los resultados pueden deberse a los criterios de recolección del *corpus* (el método de una semana construida aleatoriamente para cada mes), a las definiciones del universo de estudio (las cinco primeras noticias publicadas en las *homepage* de los diarios y en dos franjas horarios) y/o a la diferencia entre los *corpus* de noticias.

En este marco nos interesa indagar en la evolución durante julio a septiembre de 2019 de la relevancia de los tópicos de la agenda mediática digitales. Para ello calculamos la evolución de la media de tópicos por día y aplicamos un suavizado GAM (*generalized additive models*)[8]. Tal como desarrollamos en el apartado anterior, el algoritmo LDA en su versión básica asume supuestos fuertes para el análisis temporal: los tópicos preexisten a los textos y son constantes en el tiempo[9].

Gráfico 2 Argentina: evolución de la media de los tópicos de noticias por día, julio a septiembre 2019 (suavizado GAM).



Fuente: elaboración propia.

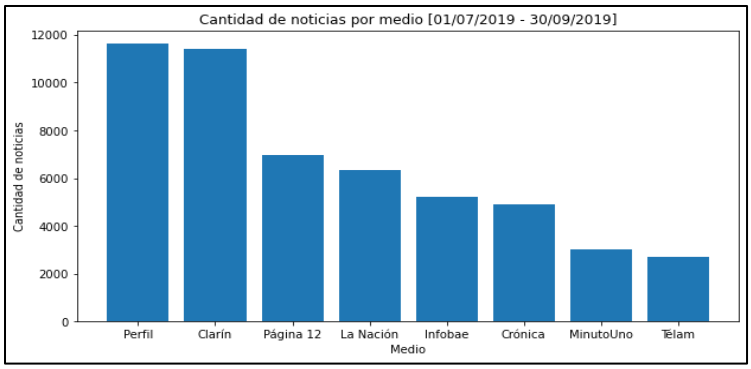
[8] Los GAMs son una clase de modelos lineales en los que se reemplaza la función lineal por un set de funciones aditivas. Suelen ser usados para realizar filtrados y suavizados en datos ruidosos (Hastie y Tibshirani, 1986).

[9] El modelo LDA en su versión básica tiene supuestos con fuertes implicancias para el análisis temporal. Es por ello que existen versiones del modelado de tópicos que permiten flexibilizar estos supuestos, por ejemplo, *Dynamic topic modeling*. Para más detalles, véase: Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, 55 (4).

La línea vertical indica el día de las elecciones Primarias Abiertas Simultáneas y Obligatorias, el domingo 11 de agosto de 2019. El gráfico 3 nos muestra al caso de las elecciones como el tópico con mayores variaciones en su evolución temporal: aumenta sostenidamente desde los primeros días de julio, alcanzando su punto máximo días posteriores a la elección cuando comienza a decrecer y vuelve aumentar en los últimos días de septiembre de cara a las elecciones generales realizadas el 27 de octubre de 2019. El tópico sobre economía también presenta oscilaciones: parece caer levemente en julio e incrementarse sostenidamente en agosto, cuando alcanza su punto máximo en los últimos días del mes y vuelve a caer en septiembre. Mientras que los tópicos sobre deportes, seguridad y política exterior se muestran estables como temas relevantes de la agenda mediática *online*. En contraposición parcial a nuestra hipótesis preliminar no observamos que aumente la prevalencia de las noticias securitarias durante el mes de la elección. Más bien, la relevancia del tópico seguridad desciende levemente durante agosto y vuelve a incrementarse en septiembre de 2019 sin llegar a los valores previos a la elección. Lo anterior guarda sintonía como la tendencia observada en las elecciones generales de 2015, donde no se incrementó la frecuencia de piezas securitarias (Zunino y Focás, 2019b).

Previo a analizar la evolución temporal de los tópicos desagregados por el medio *online* es importante observar la cantidad de noticias publicadas según el medio de comunicación. El gráfico 2 muestra a *Perfil* (11611) y *Clarín* (11399) como los medios de comunicación con mayor producción de noticias *online*. En el extremo opuesto se encuentran *MinutoUno* (3010) y *Telam* (2723). Con valores intermedios de producción de piezas periodísticas digitales se encuentran *Página 12 (*6953), seguido por *La Nación* (6338), *Infobae* (5234) y *Crónica* (4886).

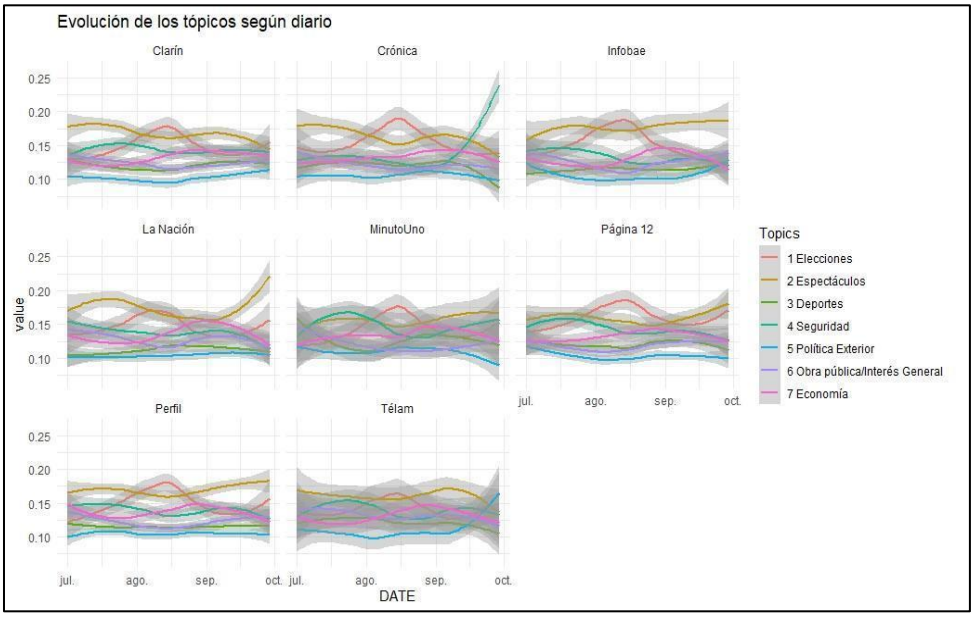
Gráfico 3 Argentina: cantidad de noticias según medio de comunicación *online*, julio a septiembre 2019.



Fuente: elaboración propia.

Ahora bien, al analizar la evolución de la media de los tópicos de noticias desagregados según el medio de comunicación evidenciamos que los tópicos electoral y securitario son estables y relevantes en la agenda mediática digital de los medios *Clarín*, *La Nación, Infobae, Página 12, Télam, Perfil*, *Crónica* y *Minuto Uno*. En todos los medios analizados el caso electoral asume el mayor nivel de relevancia en los días posteriores a las PASO 2019. En la misma dirección, el caso de la economía incrementa relevancia luego de las elecciones. Estos resultados permiten inferir que, en general, los medios digitales comparten una misma agenda mediática, aunque con algunas diferencias leves. No obstante, el dato más elocuente que muestra el análisis del *corpus* refiere al incremento exponencial de la relevancia del tópico securitario en el medio *Crónica* durante septiembre de 2019.

Gráfico 4 Argentina: evolución de la media de los tópicos de noticias según diarios online por día, julio a septiembre 2019 (suavizado GAM).



Fuente: elaboración propia.

A partir de los resultados de la modelización podemos establecer que la hipótesis preliminar se corrobora de forma parcial. A nivel metodológico observamos que aplicar técnicas de procesamiento de lenguaje natural y *web scraping* nos permite incrementar la cobertura de noticias y reducir los tiempos de detección y análisis de tópicos relevantes y caracterizar la agenda mediática digital desde julio a septiembre de 2019. En este sentido, el tamaño total del *corpus* construido en este artículo es sensiblemente mayor que los estudios reseñados en el segundo apartado. A nivel del estudio empírico, la prevalencia de noticias securitarias no aumenta durante el mes de las elecciones Primarias Abiertas Simultáneas y Obligatorias.

5. Abriendo la caja de herramientas metodológicas

¿Cuáles son las potencialidades que las técnicas de análisis textual computacionales tienen para aportar al análisis de medios? En esta sección buscamos responder este interrogante. A partir del ejercicio de modelización de tópicos, presentado en el apartado anterior, identificamos que este conjunto de

técnicas computacionales permite aumentar sensiblemente la escalabilidad al incrementar la cobertura de noticias y reducir los tiempos de detección y análisis de tópicos relevantes. Los resultados de la modelización permiten establecer que esta herramienta es útil para el análisis textual de *corpus* amplios y de forma sistemática la evolución temporal de los tópicos. Por lo anterior, nos parece que las técnicas de *web scraping* y procesamiento de lenguaje natural abordadas en este artículo pueden aportar a las investigaciones de Agenda *Setting* que tienen pretensión de análisis de contenido a gran escala.

En segundo lugar, la implementación de algoritmos de modelización de tópicos puede ser de utilidad para robustecer la replicabilidad de la codificación de piezas periodísticas, dado que empleando técnicas procesamiento de lenguaje natural y *web scraping* es posible sistematizar y alcanzar cierto grado de automatización de las diferentes etapas de pre-procesamiento de un texto (Rosati, 2021). En la actualidad, las investigaciones de Agenda *Setting* que utilizan la técnica de análisis de contenido cuantitativo recolectan el *corpus* de noticias de manera manual e incluyen las noticias siguiendo criterios específicos definidos por los investigadores previamente. De manera similar a la codificación tradicional, el modelado de tópicos clasifica documentos de un *corpus* en categorías. Si bien esta etapa implica una serie de decisiones (cantidad de tópicos, interpretación de los tópicos estimados, etc.) estos juicios subjetivos se escriben directamente en el proceso de codificación asistido por computadora, por lo tanto, la salida del texto codificado computacionalmente es total e inmediatamente reproducible. A diferencia del texto codificado manualmente donde el análisis de contenido no es fácilmente reproducible debido a que es difícil lograr que la misma persona codifique la misma pieza periodística dos veces del mismo modo, y mucho más complejo es entrenar a un equipo completamente nuevo para codificar un *corpus* de manera similar que un anterior equipo (Nelson, 2017).

En nuestro caso empleamos el modelo LDA que clasifica los textos de la misma manera cada vez, haciendo que la etapa de clasificación sea reproducible. Si se emplean las mismas técnicas y se le atribuyen los mismos valores a los hiperparámetros del modelo estos mismos resultados pueden ser replicados por otros investigadores. En concreto, el código y *scripts* desarrollados permiten replicar el trabajo. En este marco, es importante aclarar que el flujo de trabajo

debe ser revisado para cada problema de investigación en particular (Grimmer & Stewart, 2013 en Rosati, 2021). El flujo de trabajo implantado en este artículo es uno de los posibles, pero no el único ni necesariamente el "mejor" en términos absolutos.

En tercer lugar, el análisis textual computacional puede ser una herramienta que permita encontrar tópicos noticiosos de manera inductiva o no establecidos *a priori* por el investigador. Al incorporar un modo de inducción en el diseño de la investigación pueden sugerir categorías relevantes para el texto que los científicos sociales no habían considerado previamente debido a sus prenociones o la complejidad del *corpus* (Grimmer & Stewart, 2011 en Nelson, 2017) y puede ayudar a evitar los sesgos de los investigadores y la volatilidad natural que conlleva la lectura de grandes volúmenes de texto. Sin embargo, el rol del investigador no queda totalmente difuminado de esta etapa del flujo de trabajo. El proceso de análisis, interpretación y etiquetado de los tópicos continúa siendo de carácter manual. En nuestro caso surgió un tópico (Política exterior) que no buscábamos al principio de la investigación. De esta manera, el modelado de tópicos abre la posibilidad de clasificar o agrupar las piezas periodísticas según su tópico prevalente sin recurrir a una clasificación *a priori* del investigador.

A la par de estas ventajas existen limitaciones a tener en cuenta. En primer lugar, existe una fuente de sesgos en la recolección de las noticias que conforman el *corpus,* que pese a su amplitud ha sido confeccionado a partir del grafo GDELT. Por lo anterior, no es posible afirmar que se analizaron todas las piezas periodísticas publicadas en los ocho medios digitales analizados en este trabajo. En segundo lugar, algunos metadatos recabados (título y texto) presentan diversos grados de calidad. En algunos casos las piezas periodísticas habían sido eliminadas de los diarios, por esta razón decidimos eliminar 300 unidades de análisis que no estaban en condiciones óptimas de ser incluidas en la base de datos. En tercer lugar, la técnica empleada en este trabajo se limita a estimar los principales tópicos dentro de un *corpus* y clasificar cada pieza periodística en esas categorías. Esta herramienta no es eficiente para abordar la multiplicidad de aspectos sobre las características de la cobertura mediática securitaria abordados desde la teoría de Agenda *Setting*. Estudios recientes desde esta perspectiva (Focás y Zunino, 2017; Zunino y Focás, 2019a y 2019b) emplean sistemas de categorías extensos que

relacionan múltiples variables en el tratamiento mediático del delito en la prensa *online*. En concreto, analizan las fuentes de información externas a la redacción -fuentes oficiales o familiares de la víctima, etc.-, la localización de los ilícitos, el tratamiento de víctimas y victimarios, la edad, la clase social y los actores sociales según el rol, las causas atribuidas por los periódicos a la problemática securitaria y las soluciones promovidas y la evaluación moral que ejercen los diarios a dichas problemáticas.

Los resultados del ejercicio propuesto están acotados a mostrar la aplicación de una técnica de procesamiento de lenguaje natural más que agotar determinaciones del objeto en cuestión. Para acotar el alcance de este artículo, el mismo se centró en la detección, frecuencia y evolución que adquieren en la prensa *online* los tópicos relativos a la cuestión securitaria. Sin embargo, esta perspectiva metodológica puede ser útil para abordar otros elementos propios de la perspectiva teórica de la Agenda *Setting*. En nuestro caso, el *corpus* de noticias recolectadas en las páginas web de medios de comunicación presenta características de multimedialidad -imagen, sonido, hipervínculo. El soporte noticioso objeto de análisis presenta elementos propios del escenario digital que pueden ser incluidos en el *scraper* a partir del análisis más específico de los elementos no textuales de los sitios *web*. La teoría de la Agenda *Setting* postula que la relevancia mediática se mide a partir de dos criterios básicos de noticiabilidad: la frecuencia de publicación y la jerarquía de la información (Aruguete, 2015). En la segunda sección mostramos como investigaciones recientes desde esta perspectiva teórica (Focás y Zunino, 2017; Koziner, 2019; Zunino y Focás, 2019a y 2019b; Zunino y Grilli Fox, 2019) operacionalizan el concepto de jerarquía de la información desde las plataformas digitales a partir de diferentes atributos de la noticia: si aparece en tapa, si abre sección, si está en página impar, en mitad superior, si tiene gran tamaño, firma o títulos grandes, el promedio de la extensión, la autoría de las piezas, el género del autor y la frecuencia de fotografías, audios y videos por noticia, etc. Si bien las técnicas de modelado de tópicos presentadas en este trabajo sirven para estudiar la frecuencia de publicación de los diversos tópicos, es posible tomar la operacionalización del concepto de jerarquía de la información e incluirlos como criterios en el *scraper*. En este caso queda abierta una futura línea de investigación que explore las técnicas y los modelos computacionales de análisis

de los elementos no textuales de los sitios *web* y evaluar, a partir de la decisión teóricamente sustentada, la manera más plausible de abordar soportes noticiosos con características de multimedialidad.

Con base a lo relatado, consideramos que la perspectiva de la Agenda *Setting* podría enriquecer los estudios de contenido de noticias, agenda y relevancia mediática al emplear una estrategia metodológica mixta: métodos computacionales (procesamiento de lenguaje natural y *web scraping*) y análisis de contenido cuantitativo. Por un lado, el análisis textual computacional habilita la aplicación de métodos cuantitativos para una amplia variedad de tareas como la detección, la frecuencia y la evolución de los tópicos noticiosos y la clasificación de las piezas periodísticas. Por otro lado, las principales fortalezas del análisis de contenido cuantitativo se relacionan con el amplio sistema de categorías que emplean para estudiar la jerarquía mediática. Estas investigaciones abordan una multiplicidad de aspectos sobre las características de la cobertura mediática securitaria (Focás y Zunino, 2017; Zunino y Focás, 2019a y 2019b). La combinación de metodologías podría ser útil para que las investigaciones empíricas de la Agenda *Setting* que abordan una multiplicidad de aspectos de la cobertura mediática moderen los problemas metodológicos de la escalabilidad y replicabilidad. Este panorama deja abierta una futura línea de investigación para la cual proponemos el siguiente flujo de trabajo. Primero, con el modelo LDA podemos identificar las noticias securitarias sobre un *corpus* de noticias digitales y seleccionar las piezas con alta prevalencia de este tópico, para posteriormente realizar un análisis de contenido cuantitativo (e incluso un análisis cualitativo profundo) sobre este recorte específico.

6. Conclusiones

Este artículo se ubica y limita a exponer un abordaje metodológico posible de análisis computacional de texto a los estudios sobre tratamientos informativos. A diferencia de la prensa en papel, el *corpus* analizado en este trabajo presenta características de multimedialidad que en la actualidad la perspectiva de Agenda *Setting* la aborda desde las plataformas digitales. Sin embargo, para acotar el alcance del artículo, optamos por centrarnos en la detección, frecuencia y evolución que adquieren los tópicos securitarios en la prensa *online* durante el contexto

electoral de las PASO 2019.

La pregunta central que movilizó el artículo es ¿cuáles son los principales temas de la agenda mediática digital entre julio y septiembre de 2019? A partir de emplear la herramienta de modelados de tópicos con la implementación del modelo LDA pudimos estimar los principales tópicos noticiosos dentro del *corpus* y clasificar cada pieza periodística en esas categorías sin recurrir a una codificación *a priori* del investigador. De esta forma, identificamos que el caso de las elecciones Primarias Abiertas Simultáneas y Obligatorias 2019 ocupó un lugar relevante en las agendas de los principales medios digitales nacionales. En menor nivel de relevancia los tópicos vinculados a espectáculos, deporte, seguridad, política exterior, obra pública y economía fueron prioridad en la agenda de los medios de comunicación *online.* Al analizar la evolución temporal de los tópicos de la agenda mediática digital en los medios digitales de *Clarín*, *La Nación, Infobae, Página 12, Télam, Perfil*, *Crónica* y *Minuto Uno*, observamos que el caso electoral y el asunto securitario se constituyen como tópicos estables y relevantes en las noticias digitales en el contexto electoral de las PASO 2019. Los resultados arribados con la implementación del algoritmo LDA permiten refutar de manera parcial nuestra hipótesis de investigación. En nuestro caso, no se incrementó la frecuencia de piezas securitarias durante agosto de 2019, mes de las elecciones PASO 2019. Este dato guarda relación con la tendencia general observada en las elecciones generales de 2015, donde la prevalencia de noticias de seguridad no aumentó (Zunino y Focás, 2019b). En relación con el objeto de estudio (las noticias digitales securitarias), el dato más elocuente que se desprende del análisis del *corpus* hace referencia al aumento significativo de la frecuencia de publicación de las piezas securitarias en el diario digital *Crónica* durante septiembre de 2019.

Este artículo persiguió el objetivo de explorar la aplicación de algunas técnicas de análisis vinculadas al campo del procesamiento de lenguaje natural al análisis de medios desde la perspectiva de Agenda *Setting*. Con base a lo relatado, concluimos que estas técnicas permiten escalar el trabajo de forma eficiente y podrían ser útiles para moderar algunas de las limitaciones propias de la técnica de análisis de contenido cuantitativo, como ser la escalabilidad debido a las grandes cantidades de tiempo que estos estudios emplean para la codificación y el análisis de las piezas periodísticas (Orozco Gómez y González, 2012). Por un lado, la

herramienta de web scraping puede ser de utilidad para incrementar la cobertura de noticias. Por otro lado, las técnicas de procesamiento de lenguaje natural abordadas en este artículo –y muchas otras que no mencionamos- abren la posibilidad de escalar el análisis de forma eficiente (Rosati, 2021).

En comparación con los estudios de contenido cuantitativo donde el investigador tiene que leer cada una de las noticias de un *corpus* -tarea que se vuelve imposible de realizar en cantidades extensas por lo que suelen reducir la población a una dimensión abordable (Zunino y Focás, 2019a)- las técnicas de procesamiento de lenguaje permiten analizar de forma automática *corpus* textuales a escalas notablemente más grandes. Además, este conjunto de herramientas abre la posibilidad de aplicar métodos cuantitativos de análisis textual (por ejemplo, clasificación de textos y detección de temas y tópicos). De esta forma, consideramos que el conjunto de técnicas computacionales abordadas en este trabajo puede aportar a las investigaciones de Agenda *Setting* que tienen pretensión de análisis de contenido a gran escala.

La perspectiva de Agenda *Setting* podría enriquecer sus investigaciones empíricas al incorporar a su repertorio metodológico técnicas de análisis textual computacional (*web scraping* y procesamiento de lenguaje natural). Esta perspectiva metodológica puede ser útil para incrementar la replicabilidad en las investigaciones sobre contenido de noticias. Las técnicas abordadas en este artículo permiten alcanzar una sistematización de las diversas etapas de pre-procesamiento de piezas periodísticas y alcanzar -eventualmente- cierto grado de automatización. El presente abordaje del problema puede ser útil para el desarrollo de futuras investigaciones en torno a la Comunicación, pudiendo utilizar la base de datos que hemos producido e incorporar el presente enfoque a sus estrategias metodológicas. A su vez, podría ampliarse el período de análisis de forma relativamente simple, dado que existen *scripts* y rutinas que permiten replicar los procedimientos de captura y procesamiento de la información.

Es importante destacar algunos problemas del análisis textual computacional aplicado al análisis de contenido mediático. En primer lugar, la fuerte dependencia de la base de datos GDELT donde se recolectaron los links de las noticias para la construcción del *corpus*. En este sentido, encontramos sesgos en la recolección de la información que no nos permiten afirmar que se abordaron la totalidad de las

piezas publicadas en los medios digitales bajo estudio. En segundo lugar, la calidad de los metadatos (título, texto, fecha) no siempre es óptima, por lo cual optamos por eliminar 300 piezas periodísticas del *dataset.*

Ahora bien, las investigaciones desde la perspectiva de la Agenda *Setting* tienen la fortaleza de estudiar una multiplicidad de elementos que caracterizan la cobertura mediática securitaria a partir de un robusto sistema de categorías. Los estudios teóricos sobre tratamientos informativos relevados en este trabajo abordan la jerarquía de la información y las plataformas digitales. Por lo anterior, consideramos que la combinación de metodologías computacionales y cuantitativas permitiría moderar algunas dificultades metodológicas de escalabilidad y replicabilidad que implica la detección manual de los tópicos, sin por ello limitar el análisis profuso de contenido mediático que caracteriza a las investigaciones de Agenda *Setting* en Argentina. En concreto, proponemos para futuras investigaciones emplear una estrategia metodológica mixta. A partir del modelo LDA podemos identificar las piezas digitales securitarias sobre un *corpus* y posteriormente seleccionar las noticias con alta prevalencia de este tópico para un análisis de contenido cuantitativo profundo sobre el caso securitario en la agenda mediática digital.

Cómo citar este artículo:
Piñeyrúa, F. N. (2021). Aportes desde el procesamiento de lenguaje natural para incrementar la escalabilidad en los estudios sobre tópicos de noticias digitales securitarias. *Revista Comunicación, Política y Seguridad*, 3, 111-142. Recuperado de https://publicaciones.sociales.uba.ar/index.php/revistacomunicacion/article/view/6627

Bibliografía

Ariza, L. y Beccaria, L. (2019). Víctimas y victimarios: niñez y adolescencia en las noticias televisivas. *Comunicación, Política y Seguridad, 1(1),* 63-87.

Aruguete, N. (2009). Estableciendo la agenda. Los orígenes y la evolución de la teoría de la Agenda Setting. *Ecos de la comunicación, 2(2).*

Aruguete, N. (2015). *El poder de la agenda. Política, medios y público.* Editorial Biblos/Cuadernos de comunicación.

Barriola, J. M. y Gncchi, L. (2020). COVID-19: Medios a Contramano de la

Pandemia - Parte I. *Medium*.

Carrillo, F. (4 de noviembre de 2019). *Análisis computacional de texto aplicado a las ciencias sociales. Aprendizaje automático y Big Data*. Workshop de Factor DATA, Universidad Nacional de San Martín, Escuela de Altos Estudios Sociales, Argentina.

Dammert, L. y Erlandsen, M. (2020). Migración, miedos y medios en la elección presidencial en Chile (2017). *Revista CS, 31*, 43-76.

Focás, B. M. y Zunino, E. (2017). El tratamiento informativo de la "inseguridad" en la Argentina: víctimas, victimarios y demandas punitivas. *Communication & Society, 31(3)*, 189-209.

Koslowski, D. (4 de noviembre de 2019). *Análisis computacional de texto aplicado a las ciencias sociales. Aprendizaje automático y Big Data*. Workshop de Factor DATA, Universidad Nacional de San Martín, Escuela de Altos Estudios Sociales, Argentina.

Koziner, N. (2019). Temas y fuentes en medios digitales argentinos. Un estudio en contexto electoral. *Más Poder Local*, 40, 46-56.

Mützel, S. (2015). Facing Big Data: Making sociology relevant. *Big Data & Society*. https//doi:10.1177/2053951715599179

Nelson, L. K. (2017). Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*, *49*(1), 3–42.

Pallavicini, C. (4 de noviembre de 2019). *Análisis computacional de texto aplicado a las ciencias sociales. Aprendizaje automático y Big Data*. Workshop de Factor DATA, Universidad Nacional de San Martín, Escuela de Altos Estudios Sociales, Argentina.

Rosati, G. (2021). Procesamiento de Lenguaje Natural aplicado a las ciencias sociales. Detección de tópicos en letras de tango. Revista Latinoamericana de Metodología de la Investigación Social (RELMIS), en prensa.

Orozco Gómez, G. y González, R. (2012) *Una coartada metodológica. Abordajes cualitativos en la investigación en comunicación, medios y audiencias*. Kindle Edition.

Zunino, E. y Focás, B. M. (2019a) Territorios, tópicos y fuentes de la inseguridad. Un estudio sobre la prensa argentina. *Cuadernos info, 45*, 73-93.

Zunino, E. y Focás, B. M. (2019b) Revisitando la agenda de la seguridad en los medios: un análisis exploratorio de los contenidos de las noticias policiales y de inseguridad durante el gobierno de Cambiemos (2015-2019). *Cuestiones criminales, 2(4)*, 78-104.

Zunino, E. y Grilli Fox, A. (2019). Medios digitales en la Argentina: posibilidades y límites en tensión. *Estudios sobre el Mensaje Periodístico*, 401-413.