DOSSIER

Una aproximación a los temas acerca de la COVID-19

Aplicación de técnicas de procesamiento de lenguaje natural sobre comentarios de lectores de noticias digitales

Germán Rosati,[1] Adriana Chazarreta,[2] Laia Domenech,[3] Tomas Maguire[4]

Resumen

El presente trabajo intenta un primer acercamiento a los discursos construidos alrededor de la pandemia desatada en el año 2020. ¿Qué temas se encuentran asociados a la pandemia y a la COVID-19? ¿Cuáles son los más relevantes? ¿Cuál es su evolución temporal en los primeros momentos de la misma? Con este objetivo, se utilizará una fuente poco utilizada en ciencias sociales: los comentarios que los lectores de noticias digitales producen en los foros de los medios. A su vez, el trabajo se propone aportar elementos que permitan ponderar la utilidad de las técnicas de procesamiento de lenguaje natural para abordar este tipo de problemas en ciencias sociales.

Para la construcción de los datos se utilizaron técnicas de *web scraping* y para su análisis se aplicaron algoritmos de procesamiento de lenguaje natural. Un hallazgo relevante se

1 Investigador Asistente del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Investigador en Programa de Investigaciones sobre el Movimiento de la Sociedad Argentina. german.rosati@gmail.com. ORCID 0000-0002-9775-0435

2 Investigador Asistente del CONICET. adchazarreta@gmail.com. ORCID 0000-0002-4737-9578

3 Asistente de investigación en Escuela Interdisciplinaria de Estudios Sociales - UNSAM . laiadomenechburin@gmail.com. ORCID 0000-0003-4576-3143

4 Asistente de investigación en Escuela Interdisciplinaria de Estudios Sociales - UNSAM . tomasmaguire@gmail.com. ORCID 0000-0001-6511-4728

vincula a la aparente estabilidad en la evolución en el tiempo de los to icos, independiente de la métrica utilizada y del diario analizado.
Palabras clave: comentarios en foros; *topic modeling;* discursos; pandemia; COVID-19; procesamiento de lenguaje natural

Abstract

This work attempts a first approach to the discourses built around the pandemic unleashed in 2020. What issues are associated with the pandemic and COVID-19? Which are the most relevant? How do they change over time? To this end, the comments that digital news readers produce in media forums are used. Besides, the paper aims to provide elements about the usefulness of natural language processing techniques to address this type of problem in social sciences.

For the construction of the data, web scraping techniques were used and natural language processing algorithms were applied. A relevant finding seems to be the apparent stability in the evolution of the papers over time, regardless of the metric used and the newspaper analyzed.
Key words: *forum comments; topic modeling; discourses; pandemic; COVID-19; natural language processing*

Introducción[5]

El presente trabajo se propone realizar un primer acercamiento a los temas y discursos construidos alrededor de la pandemia desatada en el año 2020 a causa del virus COVID-19. ¿Qué aspectos temáticos se encuentran asociados a la pandemia y a la COVID-19? ¿Cuáles son los más relevantes? ¿Cuál es su evolución temporal en los primeros momentos de la misma? Para ello, utilizaremos una fuente relativamente poco utilizada: los comentarios que los lectores de noticias digitales producen en los foros de los medios. Se hace necesaria una aclaración preliminar: si bien se trabajará con comentarios de noticias sobre o vinculadas al COVID-19, el objeto del trabajo no será las noticias, es decir, no se tratará de realizar un análisis de *agenda setting* (McCombs, 2015) o de *news salience* (McCombs y Guo, 2014). Este trabajo no busca avanzar en el conocimiento de los mecanismos por los cuales los diferentes medios definen la importancia de sus temas, tampoco el rol que los medios tienen en la definición de la agenda pública,[6] ni de la relación entre la "saliencia" de una noticia o un hecho en los medios y su relevancia como tema relevante *(issue)* para el público general.[7] Tampoco se trata de un estudio acerca de las audiencias en sí mismas. En todo

5 Todos los autores contribuyeron en igual proporción a la producción de este trabajo.

6 Estos temas han sido abordados desde la perspectiva de la *agenda setting* y con técnicas de modelado de tópicos en Maguire (2021).

7 En términos generales, la "saliencia" de un tema puede definirse como la importancia de ese tema para el público general o para

caso, se evaluará una parte de las audiencias (la que se corresponde con aquellos lectores que producen comentarios) como dispositivo metodológico para explorar dos aspectos: 1) la utilidad de las técnicas de procesamiento de lenguaje natural para abordar este tipo de problemas en ciencias sociales y 2) la viabilidad de este tipo de fuentes como una opción para el estudio de cierto tipo de discursos que se producen alrededor de un tema dado (en este caso, la pandemia desatada en 2020).

Este tipo de cuestiones se conectan con un campo vasto de problemas ampliamente trabajado: los procesos de génesis y desarrollo de las diversas formas del conocimiento social. Las aproximaciones conceptuales, entre otras, a la formación y evolución de la ideología (Marx y Engels, 2004; Gramsci, 1972) y del *sentido común* (Gramsci, 1971), a la idea durkheimiana de "representaciones colectivas" (Durkheim, 1968), retomada por Moscovici (1979), o los estudios sobre las llamadas "ideas inherentes" en distintas protestas pre-industriales (Rudé, 1980) son diferentes maneras de abordar múltiples dimensiones de las formas de producción de conocimiento sobre lo social.

Existe un amplio campo de debate teórico acerca de la forma en que estos discursos asumen, su función, su estructura interna y su significado. Independientemente de los diferentes fundamentos conceptuales y metodológicos del que provengan, los estudios sobre representaciones y discursos sociales parecen coincidir en que los mismos no conforman un sistema lógico totalmente consistente: pueden estar configurados por fragmentos de ideas no vinculadas de forma coherente. Tampoco resultan un sistema completamente estático: los mismos se encuentran abiertos al cambio. Exploraciones conceptuales tan diversas como las de Gramsci (1971) sobre el sentido común, las de Schütz y Luckmann (2009) sobre los acervos de conocimiento y las vinculadas a la equilibración de los sistemas cognitivos de Piaget (2010) coinciden en estos puntos. A su vez, estudios más recientes desde el campo de la psicología social han postulado la existencia de una estructura y una diferenciación de funciones en los elementos de estas representaciones: un núcleo central y una periferia (LoMonaco *et al*., 2017, Piermatteo *et al*., 2018).

Métodos y datos

Enfoques metodológicos habituales

Los trabajos que abordan las diferentes formas de conocimiento social comparten, más allá del posicionamiento teórico, algunas herramientas metodológicas básicas. Así, el uso de métodos y técnicas cualitativas y cuantitativas (incluso en abordajes mixtos) es habitual.

una fracción del público. A su vez, la news salience (la importancia que un medio le da a una noticia) tiene un rol fundamental, en tanto "the salience of these issues on the media's agenda influenced the salience of the same issues on the public's agenda" (McCombs y Guo, 2014: 252). En este sentido, en el caso propuesto resulta difícil (y sobre todo en el período en estudio, es decir, los primeros meses de la pandemia) sostener que el issue en cuestión (la pandemia) sea únicamente efecto de la construcción mediática de la agenda pública.

Particularmente relevante es el uso de métodos cuantitativos, los cuales parecen centrarse en la utilización de diversas metodologías de encuestas. A continuación, reseñaremos algunas de tales aproximaciones. No obstante, debe tenerse en cuenta que el objetivo de esta reseña busca solamente ilustrar algunos procedimientos metodológicos habituales en el abordaje de estas preguntas y no espera constituir una descripción exhaustiva de los mismos.

El trabajo de Moscovici (1979) es uno de los primeros en abordar de forma sistemática el problema de las representaciones y los discursos construidos alrededor del psicoanálisis en la sociedad francesa de la década de 1960: a partir de una encuesta a población general y del estudio de contenido de noticias de medios franceses al respecto, se analizan las diversas etapas de construcción, difusión y divulgación de la disciplina del psicoanálisis. El estudio acerca de las "disposiciones psicológicas" de la personalidad autoritaria (Adorno et al., 1950) comparte el uso de encuestas y escalas de opinión.

José Nun (2015) abordó en un trabajo escrito en los años 80 el estudio de los "significados alrededor del peronismo" a través de una encuesta a trabajadores despedidos del sector automotriz con una batería de preguntas acerca de las percepciones sobre antagonismos sociales y homogeneidad de intereses. Los trabajos de Edna Muleras (2008) buscan analizar, a través de encuestas, las representaciones sobre el orden social y los comportamientos sacralizados de diversas fracciones de las clases trabajadores metropolitanas.

En estos primeros ejemplos, las encuestas consisten principalmente en diferentes baterías de preguntas clásicas de opinión en distintos formatos (abiertas, cerradas, escalas de acuerdo y opinión). A su vez, desde el campo del análisis de las representaciones sociales (LoMonaco et al., 2017) se ha desarrollado una serie de herramientas metodológicas que también se basan en encuestas pero difieren de las anteriores en que uno de los objetivos principales es la detección de palabras, frases y términos que estén asociados al objeto del cual se busca reconstruir las representaciones. Métodos como la asociación libre de palabras o las llamadas evocaciones jerárquicas (Dany et al., 2015; Piermattéo et al., 2018) piden a los entrevistados que nombren una serie de palabras o frases que consideren asociadas y, eventualmente, las ordenen en términos de su importancia. A continuación se analizan las frecuencias de aparición y la importancia atribuida por los sujetos entrevistados: las palabras más frecuentes e importantes constituirán aproximaciones a los elementos centrales de las representaciones sociales sobre algún objeto o tema. Si bien ambas aproximaciones metodológicas tienen fortalezas y debilidades (muchas de las cuales provienen de la dificultad de conocer la totalidad del sentido atribuido por los sujetos al responder) en este trabajo nos centraremos en una de ellas.

Una característica común en los enfoques anteriores, tanto en los análisis de entrevistas (etnográficas, en profundidad, etc.) como en las encuestas de carácter cuantitativo, es la existencia de una situación de entrevista. En este sentido, independientemente del uso cuantitativo o cualitativo que se dé la misma, lo cierto es que cualquier situación de entrevista supone una interacción entre al menos dos sujetos, entrevistador y entrevistado,

en la que el primero guía una conversación (de forma más o menos estructurada, según el método en cuestión) a partir de ciertas hipótesis o preguntas de investigación.[8] De esta forma, el entrevistador obliga al entrevistado a elaborar ciertos discursos retrospectivos o proyectivos. La situación de entrevista dista mucho de una conversación espontánea y se encuentra cruzada por diversos factores que inciden sobre la validez y confiabilidad de la misma que deben ser considerados: problemas de recordación, racionalizaciones ex-post, vergüenza, etc. Además, los métodos basados en entrevistas se caracterizan por su "alta reactividad" (Webb et al., 1966), ya que pueden tender a generar sesgos o cambios en las respuestas, producto, por ejemplo, de la "adaptación" del entrevistado a lo que considera serían respuestas aceptables para el entrevistador.[9]

Ventajas y limitaciones del presente abordaje

Este trabajo propone aplicar un enfoque diferente a los anteriores. Se trabajará sobre los comentarios de lectores a noticias que hablan sobre la COVID-19 en cinco medios nacionales. Al igual que en cualquier foro, los usuarios realizan intervenciones de forma directa y en respuesta o bien al estímulo inicial (el texto de la noticia original) o bien a otros comentarios. En ese sentido, se produce una dinámica de conversaciones de carácter relativamente espontáneo y sin intervención ex-ante de la figura de un entrevistador.[10] Una de las principales ventajas de este tipo de enfoque es que trabaja con una fuente de baja reactividad (Salganik, 2018).

Son habituales los estudios que monitorean uno o varios temas a lo largo de las redes sociales como Twitter para medir su prevalencia y/o los posicionamientos alrededor de los mismos (Calvo y Aruguete, 2018 y 2020). Efectivamente, una de las principales ventajas de trabajar con Twitter es la relativa accesibilidad de sus datos a través de la API provista por la empresa. Los foros de lectores (o los foros en general) parecen no haber sido explotados con la misma intensidad, probablemente porque la extracción de datos de estas fuentes supone un trabajo de scraping notablemente más dificultoso[11]. No obstante, la existencia de un "estímulo" al cual los lectores responden, como mencionamos más arriba (la noticia)

8 "A diferencia de una conversación cotidiana, la entrevista se sustenta siempre en una hipótesis y será guiada por objetivos establecidos en función de nuestros intereses cognitivos." (Cortazzo y Trindade, 2014)

9 Así, por ejemplo, Wu (2019) analizó la problemática del sesgo de género analizando la forma en que las mujeres (estudiantes o profesionales) vinculadas a la economía eran descriptas en un foro de ofertas laborales de la disciplina. El "anonimato" que proveen los foros ayudó a la autora a recuperar expresiones y discursos que no hubieran sido posibles de visibilizar con las herramientas de entrevistas tradicionales.

10 No es el objetivo de este trabajo entrar en una taxonomía de métodos de recolección de datos "espontáneos" pero el campo disciplinar de la "etnografía digital" o "virtual" presenta ciertos puntos de contacto metodológicos con el enfoque propuesto aquí -con una diferencia central en la escala de la información recogida y analizada.

11 Es posible llegar a un esquema similar en Twitter a partir de reconstruir hilos y conversaciones enmarcadas en un tema. El mismo también es sumamente complejo de construir.

permite suponer que (en términos generales) el *framing* de los comentarios se encuentra más acotado en este tipo de foros que en Twitter.

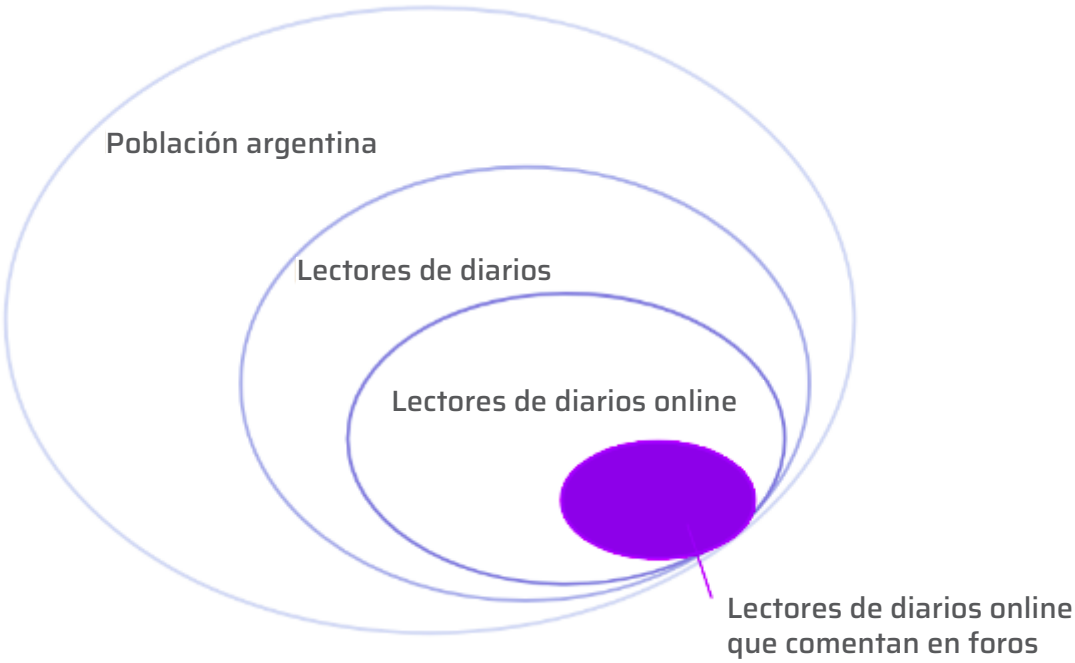
Sin embargo, estas formas de discurso espontáneo no han sido un foco de interés para abordar este tipo de problema. Schuth et al. (2007) es de las pocas investigaciones que hay al respecto, en la cual los autores extrajeron la estructura discursiva de los comentarios de lectores en varios diarios holandeses.

A su vez, dado que los comentarios de noticias (al igual, por ejemplo, que posteos en Facebook o Twitter) son producidos constantemente, habilitan la posibilidad de construir información de forma continua en el tiempo: de hecho, en el caso que presentamos haremos un análisis diario de tales comentarios.

Sin embargo, al igual que todas las fuentes provenientes del llamado *"big data"* (Salganik, 2018), los datos utilizados en este trabajo adolecen de diversos sesgos y problemas que es necesario explicitar.

Por un lado, y quizás sea el sesgo más importante de todos, es importante remarcar que existe una limitación de cobertura. No se está recolectando información sobre toda la población argentina. De hecho, ni siquiera puede afirmarse que estemos llegando a la totalidad de lectores de diarios.

FIGURA 1. Esquema de la cobertura de los datos utilizados



Solamente estamos aproximándonos a una parte de los sujetos que leen alguno de los cinco diarios relevados de forma digital (es decir, a través de su sitio web) y que, además, tienen incentivos para realizar algún comentario en el foro (Fig. 1). Este último hecho tiene

influencia en el tono de los comentarios que son posteados: son predominantemente negativos. Este es un problema ampliamente abordado en los estudios que trabajan con este tipo de información (por ejemplo, Calvo y Aruguete, 2020). Al mismo tiempo, es habitual encontrar la presencia (al igual que en redes sociales) de trolls y spammers que afectan el tono y los temas de la discusión. En este caso, como veremos más adelante fue posible filtrar parcialmente la influencia de este tipo de actores.

A su vez, los medios con los que se han trabajado no cubren la totalidad de los existentes en la Argentina (ni tampoco una muestra representativa de los mismos). No se han abordado medios provinciales o locales, sino solamente cinco grandes medios de circulación nacional.

Al mismo tiempo, existe un factor que afecta el grado de reactividad que mencionamos antes como una ventaja. En efecto, los comentarios analizados son reacciones a una noticia publicada en un medio digital (o bien respuestas a otros comentarios de la misma noticia). Esto supone que el comentario observado puede ser dependiente de varios factores exógenos vinculados a las diferentes líneas editoriales: la definición de una agenda de temas y la visibilidad o "*salience*" de la noticia, el encuadre que el medio o periodista le da a la nota, etc. Analizaremos esta cuestión más adelante en el corpus de comentarios sobre COVID-19.

Por último, otro factor importante que afecta, no tanto la calidad de los datos sino más bien la posibilidad de ampliar los análisis, es la ausencia de la información sobre los usuarios que comentan. A diferencia del caso de otras redes sociales, en los foros de lectores no se accede a más que el nombre del usuario. Esto hace que no se disponga de ningún tipo de información acerca de los emisores de los mensajes.

Estos sesgos y limitaciones en la información analizada hace necesaria la precaución al momento de hacer extrapolaciones poco reflexivas de los resultados alcanzados. Bajo ningún punto de vista deben considerarse los resultados de este trabajo como "representativos" de lo que la "población" o la "opinión pública" entiende al respecto. Tampoco permiten afirmar la importancia que tales temas tienen. Los temas detectados aquí esperan servir como una primera aproximación a ciertas ideas asociadas a la pandemia y como una ilustración de algunas potencialidades que las fuentes y las técnicas presentadas tienen para el abordaje del problema.

Scrapeo[12] de noticias y comentarios

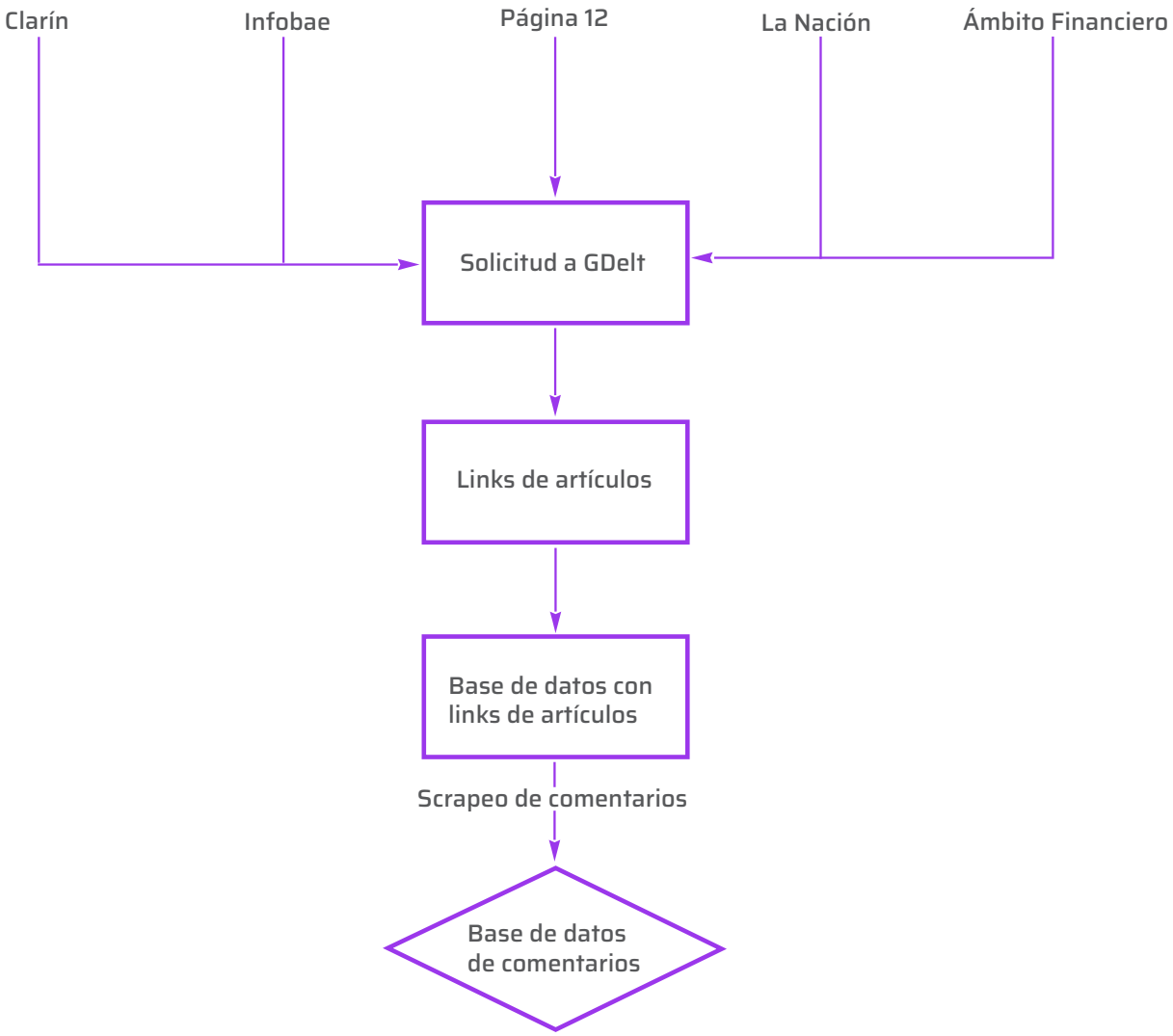
Las unidades de análisis fueron comentarios hechos en noticias online disponibles en cinco diarios de circulación nacional: *Página 12*, *La Nación*, *Clarín*, *Infobae* y *Ámbito Financiero*. Para conseguir estos comentarios fue necesario, en primer lugar, hacer un *scrapeo* de

12 El *scrapeo*, preprocesamiento del texto y el modelado de tópicos fueron realizados utilizando herramientas y librerías del stack científico de Python (BeautifulSoup, Selenium, scikit-learn, etc.). Las visualizaciones fueron producidas utilizando la librería de R ggplot.

las noticias que hablasen de COVID-19. Esto lo logramos a partir de una serie de pedidos GDELT: una base de datos de acceso abierto que monitorea y sube noticias de múltiples países, etiquetándolas según tópicos, ubicaciones, emociones e idiomas. Esto nos dio como resultado todos los links a los artículos de diarios argentinos que hiciesen mención a la COVID-19.

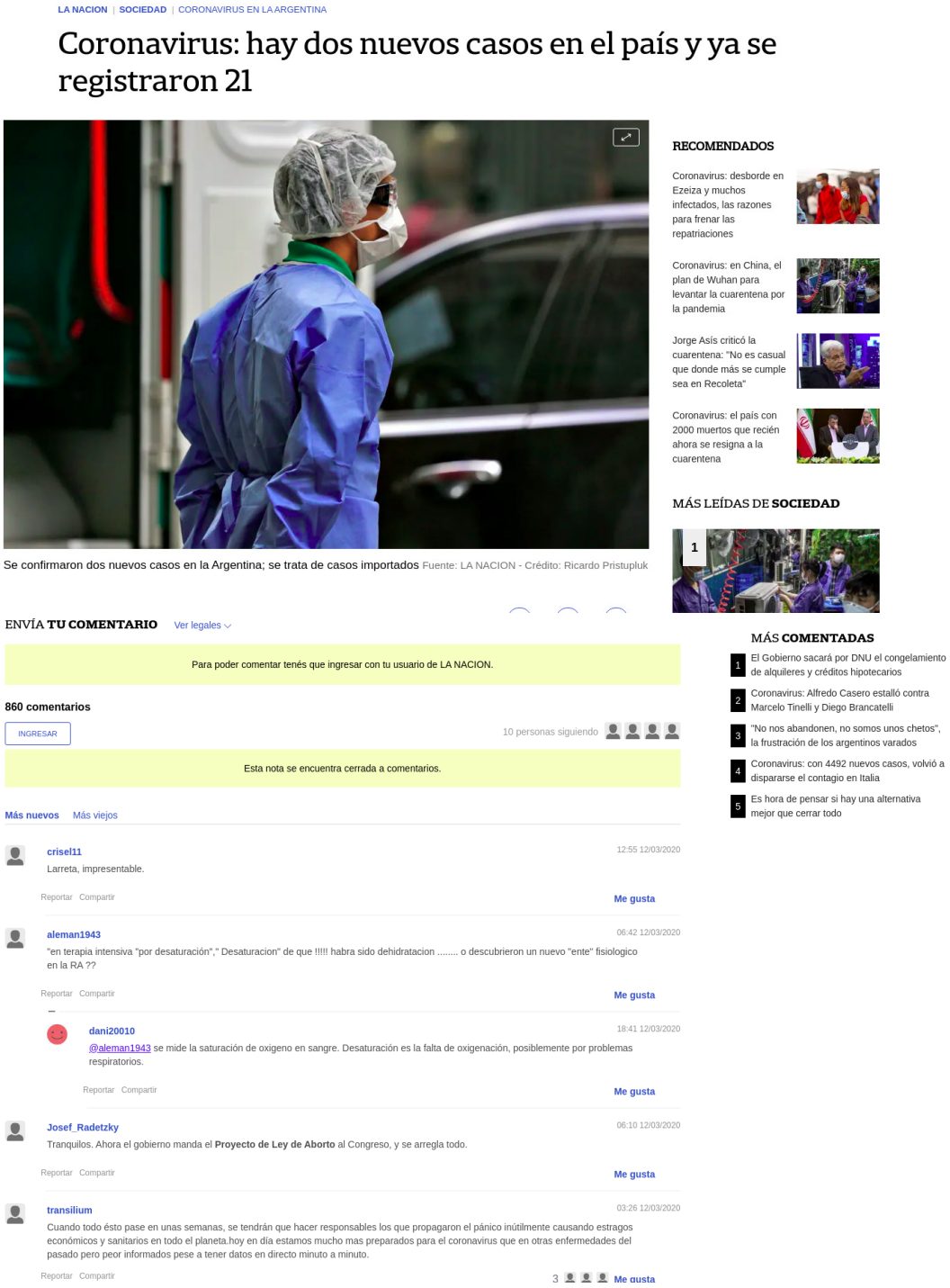
Luego, se aplicó un *scraper* previamente desarrollado sobre cada set de artículos. Esta herramienta entró a cada enlace del *input*, descargó todos los comentarios disponibles y los insertó en una base de datos. La base de datos final contiene 385.255 comentarios producidos entre el 13/03/2020 y el 01/06/2020.

FIGURA 2. Diagrama de flujo del proceso de adquisición de los datos



De esta forma, los datos a utilizar se componen básicamente de una fecha (en que se produce el comentario), un medio, un nombre de usuario y un texto del comentario.

FIGURA 3. Ejemplo de la información recolectada: texto de la noticia (arriba), comentarios de lectores (abajo)



Topic modeling con Latent Dirichlet Allocation

El segundo paso fue avanzar en la detección de estos temas relevantes en los comentarios de lectores a noticias sobre COVID-19. Existen varias técnicas para la detección automática de tópicos en corpus textuales, algunas de las cuales están basadas en alguna forma de

descomposición de la TFM.[13] Para el presente trabajo se utilizará una de las más conocidas: *Latent Dirichlet Allocation* o LDA.

La intuición detrás de LDA (Blei, 2012) es que cada documento del corpus puede exhibir varios tópicos, es decir, puede hablar de varios temas simultáneamente. El objetivo es poder operacionalizar esta intuición a través de un modelo generativo, es decir, asume la existencia de un "proceso generador de textos".

Más formalmente, un tópico se define como una distribución de probabilidad a lo largo de un vocabulario V fijo. Por ejemplo, si existiera un tópico sobre los problemas sanitarios vinculados a la pandemia sería esperable que palabras como "hospitales", "camas", "vacunas" tuvieran mayores probabilidades de inclusión en este tópico. En cambio, palabras como "kirchnerismo", "macrismo" estarán más asociadas a un tópico que discuta sobre la llamada "grieta".

Para cada documento d en el corpus C se generan las palabras w que lo componen en un proceso de dos etapas:

1. Se selecciona de forma aleatoria una distribución de tópicos para d
2. Para cada palabra (w_i) en d
   1. se selecciona aleatoriamente un tópico de la distribución general de tópicos
   2. se selecciona aleatoriamente una palabra correspondiente a la distribución de todo el vocabulario V

De esta forma, cada documento d exhibe ciertos tópicos t en diferente proporción (paso 1), cada palabra w es extraída de uno de los tópicos (paso 2.2), donde el tópico seleccionado es elegido de la distribución de tópicos de ese documento d particular (paso 2.1).

El modelado de tópicos busca descubrir de forma automática los temas vinculados a un determinado conjunto de documentos. Lo único observado es el conjunto de documentos (preprocesado como una TFM). La estructura de tópicos (es decir, la composición de tópicos por documento y la asignación de palabras a un documento) puede ser considerada como un conjunto de variables no observadas (justamente lo que se trata de estimar).

A su vez, LDA tiene algunos supuestos que conviene explicitar:

- Cada documento d se compone de varios tópicos
- Un tópico, a su vez, se compone de palabras; más precisamente, un tópico es una distribución de probabilidad sobre la totalidad de palabras del vocabulario V
- Los tópicos "preexisten" a los documentos y la distribución de probabilidad sobre V es constante; el orden de los documentos no es relevante (lo cual puede volverse un

13 Una TFM (matriz de frecuencia de términos, por su sigla en inglés) es una tabla en la que cada fila representa un documento (en este caso, un comentario) y cada columna consistirá en un término *t* del vocabulario general *V* del corpus *C*. Cada celda estará constituida por el conteo crudo de ocurrencias de cada palabra (columna) en cada documento (fila). Esta representación es la que se denomina Bag of Words o "bolsa de palabras". A su vez, existen diferentes formas de ponderar dichos conteos (*term-frequency*, tf-idf, etc.).

problema al momento de analizar la evolución de un tópico)
- Se asume que las palabras no tienen orden.

Experimentos y resultados

Aplicamos varias operaciones de preprocesamiento de texto para tener un corpus limpio y estandarizado: borramos las *stopwords*, los signos de puntuación, dígitos, links, y pasamos todo a minúscula. Luego, construimos la matriz de frecuencia de términos (TFM) y el ponderador TF-IDF[14].

Luego de iterar el modelo, encontramos diez tópicos conceptualmente relevantes.[15]

TABLA 1. Identificación de tópicos y etiquetado

ORDEN	DISTRIBUCIÓN DE PALABRAS	ETIQUETA
Topic 01	[alberto peronismo impuesto chile acuerdo razon excelente paso tipo ministro brasil ignorante test gato casos]	Gobierno, casos y testeos a nivel regional
Topic 02	[vos sos gamurra tenes porota cuba troll venezuela foto matanza mano barbijo decis paga quiero]	Insultos entre comentaristas
Topic 03	[gente pandemia china argentina salir cuarentena argentinos muertos gusta gobierno paises virus digo infectados hablar]	Testeos, casos y muertes a nivel local
Topic 04	[globo pobre vieja jajajaja arroyo grande fideos mujer cosa sueldo albertitere miedo patria queda dictadura]	Misceláneo
Topic 05	[larreta peronista inutil seguro deja viejo cabeza vergüenza kk ojo peronistas veo basura pasa idea]	Peronismo-Macrismo
Topic 06	[che cuarentena ja anos muertos fernandez titere pena argentina vas mira favor hijo economía delincuentes]	Gobierno y política económica
Topic 07	[gobierno alverso dólar virus científicos coronavirus presidente inútiles covid leer default bolsonaro meses inflación pobres]	Insultos al gobierno, crisis económica
Topic 08	[macri jajaja nota comentario ladrones chorros cree voto diario clarín jaja sale alguien cerebro llama]	Peronismo-Macrismo
Topic 09	[cara médicos votaron cubanos falta políticos deuda aca provincia mundo comer odio conurbano espero perdón]	Insultos al gobierno, crisis económica
Topic 10	[millones gente pagar cuarentena impuestos casa gobierno personas presos trabajar anda anos ojala trabajo plan]	Administración pública

FUENTE: Elaboración propia sobre datos recolectados de GDELT y sitios de los medios

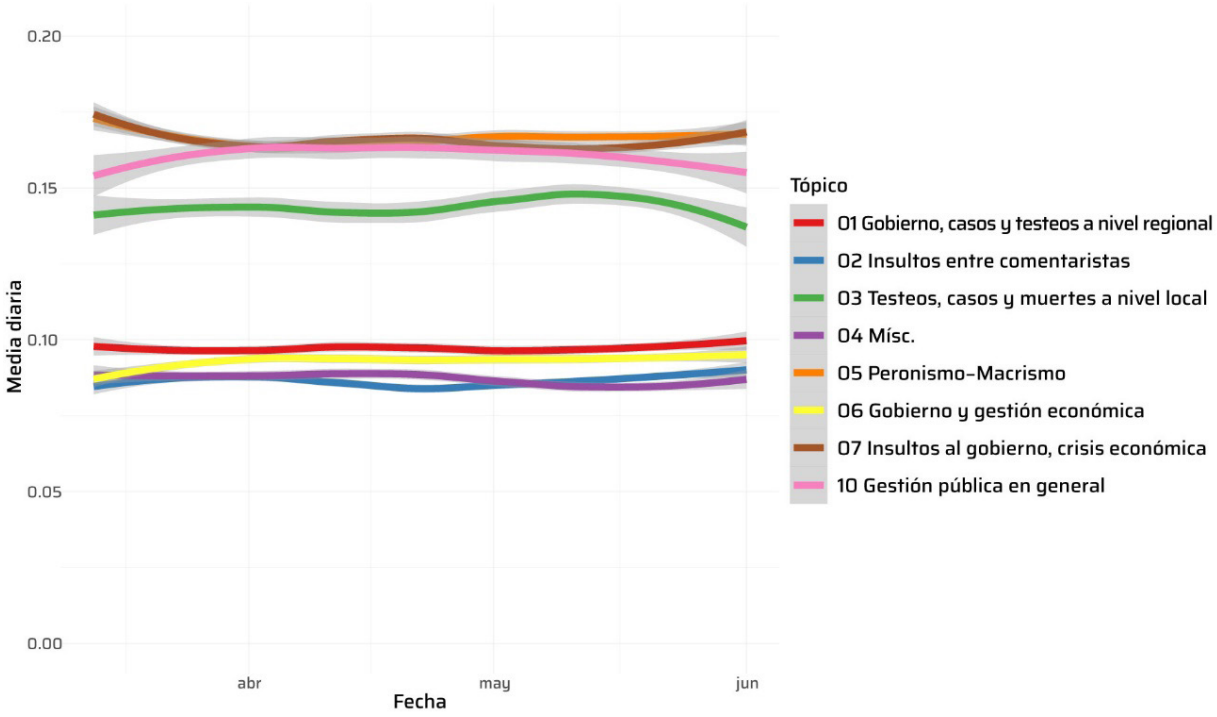
14 Se trata de una métrica que combina la frecuencia de cada término de la TFM en cada documento, y la importancia que tiene el término a lo largo de todo el corpus. Puede verse al respecto Wiedemann (2016).

15 Como se desprende del apartado anterior, uno de los problemas principales es determinar la cantidad de tópicos a estimar. Este problema es análogo al problema de determinación de la cantidad de *clusters* al aplicar algoritmos de *clustering* tales como *K-means*.

Los tópicos 07 y 09 se fusionaron en "Insultos al gobierno y crisis económica", y los tópicos 05 y 08 se fusionaron como "Peronismo-Macrismo". Finalmente, obtuvimos 8 tópicos lo suficientemente interpretables.[16]

Ahora bien, la técnica LDA permite estimar para cada documento del corpus la prevalencia que cada tópico tiene. De esta forma, podemos calcular diferentes métricas para analizar su evolución temporal.[17] El primer gráfico muestra la media que cada tópico alcanza en cada día de análisis. Para construir el segundo gráfico, calculamos el tópico de mayor prevalencia que cada comentario tiene y, luego, para cada día, construimos la frecuencia relativa de cada tópico mayoritario.

GRÁFICO 1. Evolución de la media de composición de los tópicos en comentarios de lectores de noticias sobre COVID-19.

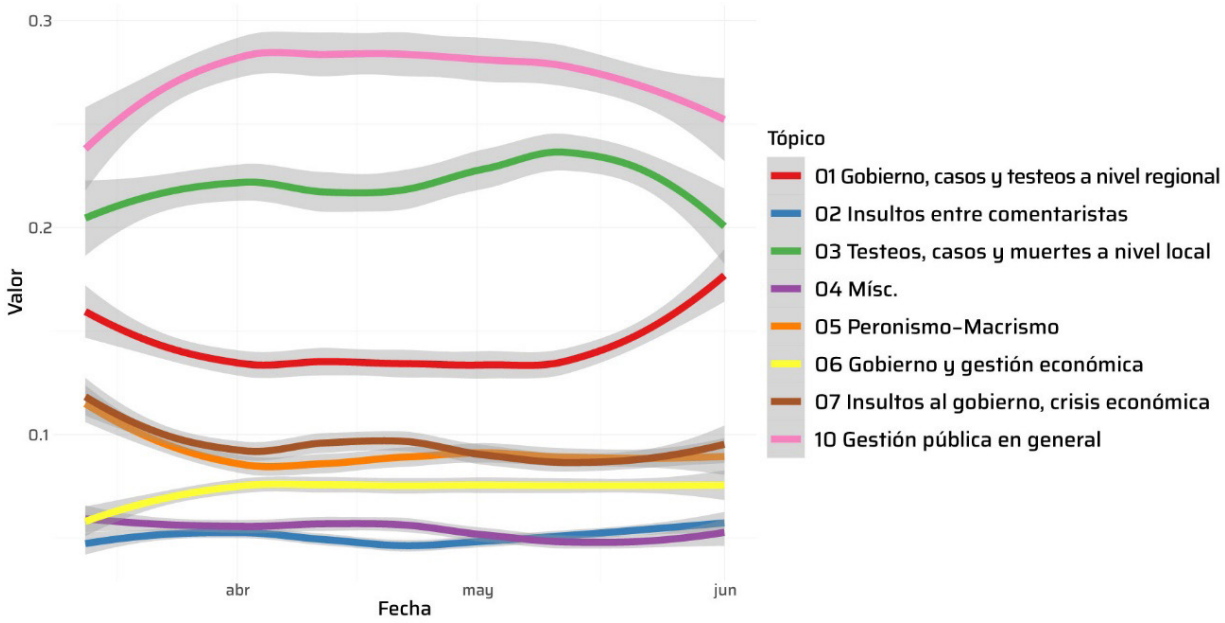


FUENTE: Elaboración propia sobre datos recolectados de GDELT y sitios de los medios

16 Existen diversas métricas que permiten cuantificar el ajuste (es decir, qué tan "bueno" es) del número de tópicos definido en términos cuantitativos (log-likelihood, perplexity, etc.). En general el uso de estas métricas conduce a modelos que logran buena performance estadística pero que no necesariamente generan tópicos interpretables. En términos generales, mayor cantidad de tópicos generan mejores valores de estas métricas. Aún así, el incremento en el número de tópicos tiende a generar una degradación en la interpretabilidad de los mismos. Al igual que en muchos otros problemas, complejidad del modelo e interpretabilidad tienden a ir en direcciones contrarias (Mimno et al., 2011; Chang et al., 2009).

17 Como hemos mencionado previamente, uno de los supuestos de LDA es la preexistencia y estabilidad temporal de los tópicos: los mismos no cambian en el tiempo. Otras técnicas de modelo de tópicos flexibilizan estos supuestos: *dynamic topic modeling*, por ejemplo (Blei, 2012).

GRÁFICO 2. Evolución de la proporción de comentarios de lectores considerando solo el tópico predominante en noticias sobre COVID-19



FUENTE: Elaboración propia sobre datos recolectados de GDELT y sitios de los medios

En ambas métricas, se nota que los tópicos en general son estables a lo largo del período. Los debates acerca de la administración pública (Tópico 10) junto con la discusión sobre los testeos y casos a nivel local (Tópico 03) parecen ser los temas más predominantes en ambas métricas.

Es particularmente importante la evolución de la media de los tópicos (Gráfico 1) en tanto nos muestra que hay un componente fuerte a lo largo del período: el debate macrismo-peronismo (Tópico 05) junto con los insultos al gobierno (Tópicos 07) mantienen una relevancia que queda oscurecida si solamente observamos los temas mayoritarios de cada tópico.

En el movimiento de los tópicos 10, 03 y 01 pareciera evidenciarse que la discusión sobre la cuestión local (tópicos 10 y 03) pierden peso hacia el final del período -junio del 2020- y gana importancia la discusión sobre la cuestión internacional.

En este punto, resulta importante remarcar un aspecto que se vincula con una de las limitaciones de esta fuente mencionada más arriba: el peso que los llamados *trolls* y otros actores tienen en las discusiones dentro de este tipo de espacios y foros. En un modelo preliminar, se mantuvieron en la base de datos los comentarios duplicados. El tópico mayoritario resultó ser las críticas al gobierno, y teníamos un tópico general que representaba los *trolls*, *bots* y *spammers*. En una segunda iteración se eliminaron dichos comentarios repetidos y eso provocó que este último tópico desapareciera. Lo cual podría indicar que se trata de una forma (aún rudimentaria) de lograr un mínimo control de la influencia de ciertos usuarios que caen en la categoría de *spammers* (que suelen repetir el mismo mensaje varias veces).

TABLA 2. Comentarios ilustrativos (muestra aleatoria de comentarios con alta prevalencia de los tópicos 05 y 08)

ORDEN	TEXTO DE COMENTARIO MUESTREADO
Topic 05 - Macrismo-Peronismo	LA CUARENTENA SIGUE HASTA EL 2023... PARA ALBERTITO, EL CORONAVIRUS ES SU TABLA DE SALVACIÓN... EL LAVADO DE K ULO, ES PARA LA FASE 3 !! OK? El verdadero virus que destruyó Argentina es el peronismo. Cristina, Alverso, el General, Menem,Duhalde, Eva. Todos lo mismo. El día que logremos una vacuna efectiva contra el peronismo, ese dia va a arrancar el pais Cómo es posible que la oposición se prestara a este bochorno¡ lamentable R Larreta!!! los gobiernos no..... los peronistas.... no generalices....
Topic 08 - Macrismo-Peronismo	SI ESTUVIERA MAURICIO....LAS INSTITUCIONES FUNCIONARIAN !!!!! Mauricio, ¿Quién es ese individuo? Naaaa ya estábamos quebrados, Macri lo hizo
	NO SOLO ORKO CEREBRO DE AMEBA,..SINO QUE UN REVERENDO PE-LO-TU-2,. ESTAMOS INFECTADOS DE ESTE VIRUS KAKA
	Son K el ADN de chorros no se los quita nadie

FUENTE: Elaboración propia sobre datos recolectados de GDELT y sitios de los medios

Como se ve en la tabla, ambos tópicos capturan la "grieta". Contienen imágenes positivas y negativas en relación a la administración actual, el peronismo en general y el gobierno anterior.

Se analiza, ahora, la misma evolución para cada uno de los diarios a partir de dos métricas: la evolución del promedio diario de tópicos para cada medio relevado y la evolución de la tasa de variación de cada medio contra la media total.

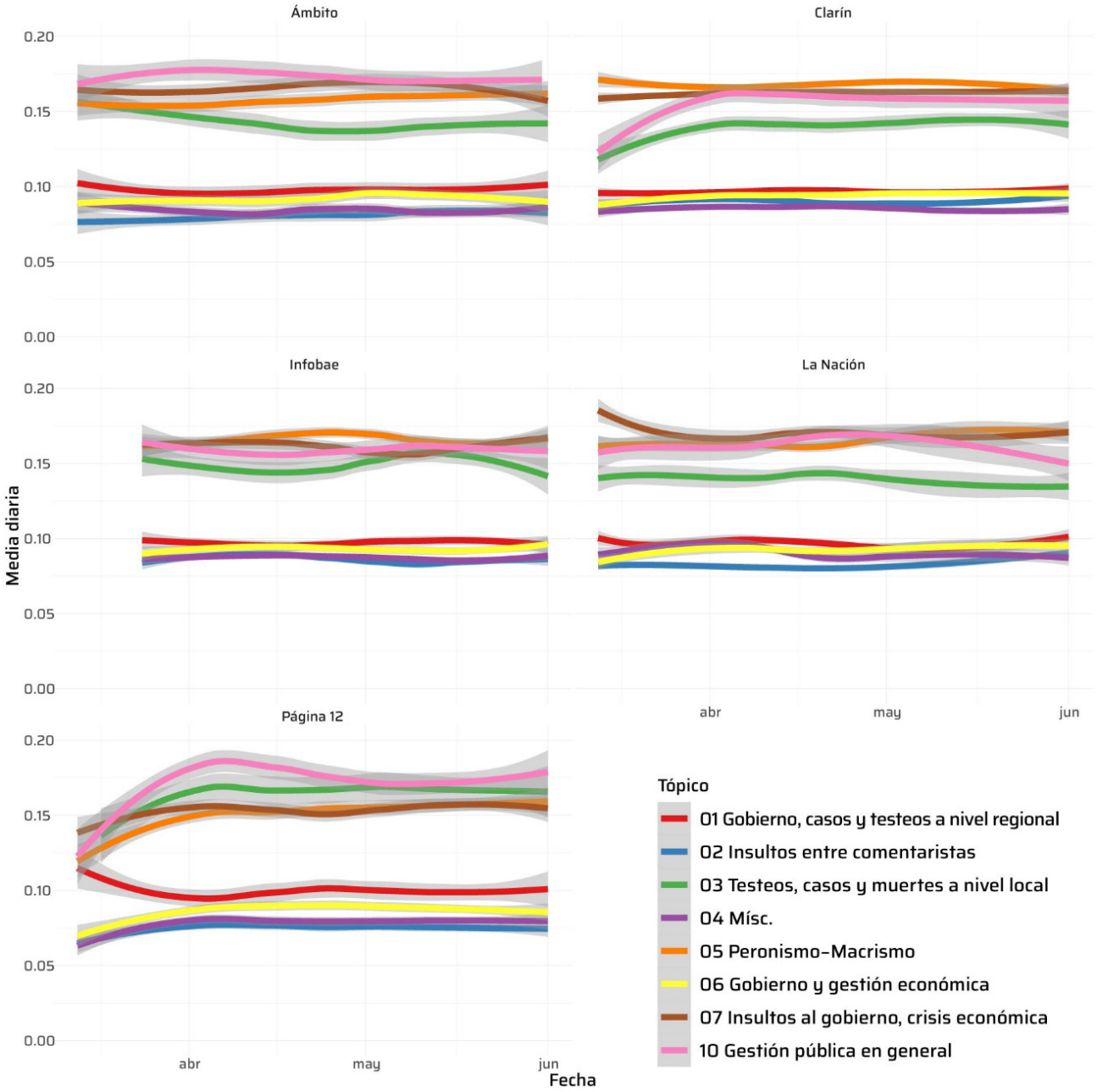
En ambas se observa que en líneas generales no hay muchas diferencias entre lo que los lectores de cada diario hablan. En la mayoría de los casos, el tópico que habla sobre la administración pública es el más alto, excepto en Clarín donde el eje antiperonismo-macrismo es ligeramente más importante en algunos momentos.

Ahora bien, podría ser razonable suponer que las diferencias en los tópicos de los comentarios (que es lo que observamos en el trabajo) son resultado de al menos dos efectos: por un lado, el tema y el encuadre de la noticia (que se vincula a la agenda del medio o del periodista en cuestión);[18] por otro, de los intereses propios de los comentaristas.

Si bien no es el objetivo de este artículo avanzar en la separación de ambos efectos, sí podemos plantear una primera aproximación al problema. ¿Cómo responden los comentaristas a diferentes "estímulos"? ¿Se observan diferencias entre los tópicos de los comentarios en las diferentes noticias de acuerdo a su tema?

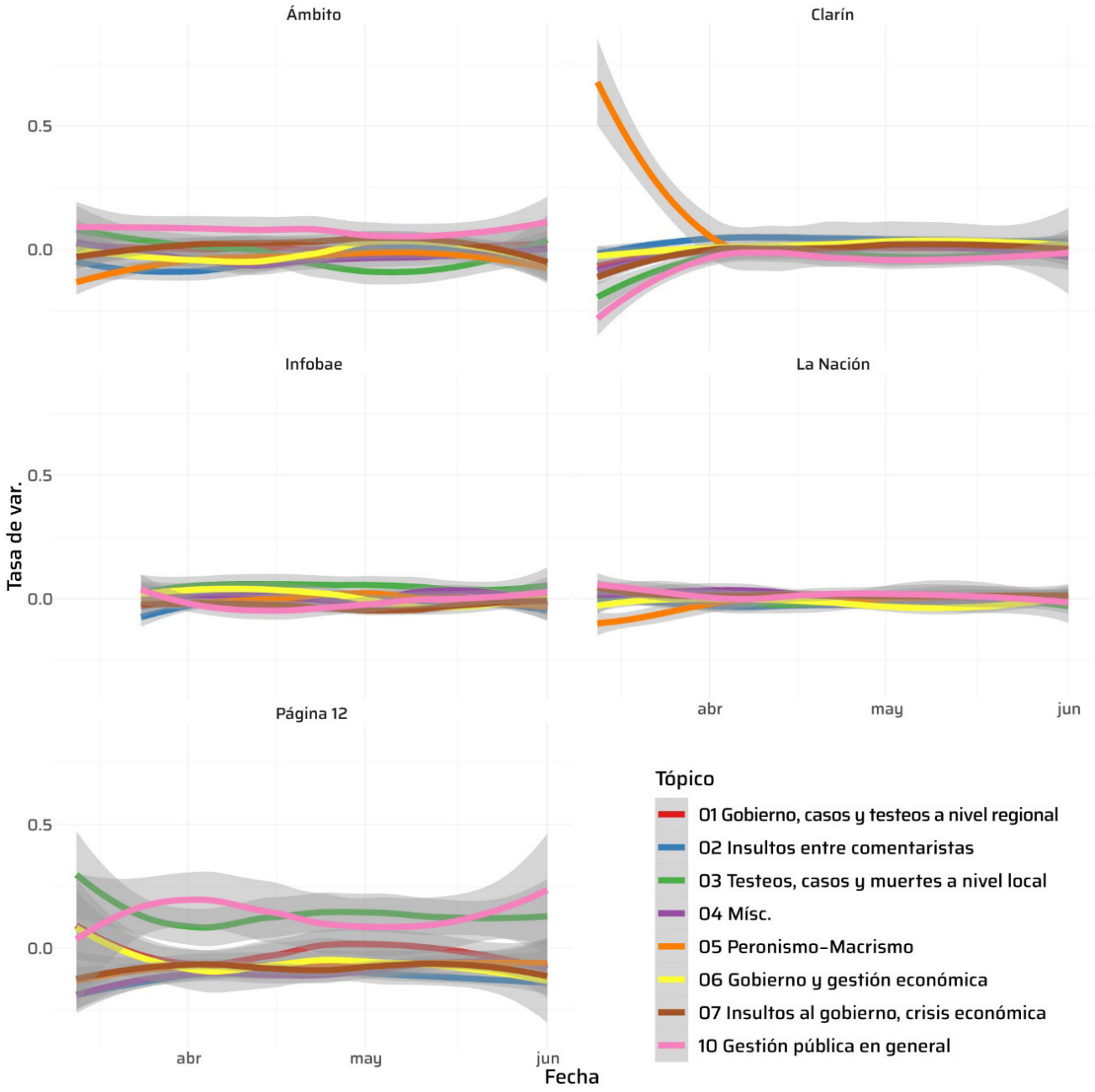
18 Si bien, como hemos mencionado antes, el tema y el encuadre general de las noticias analizadas se vincula al COVID-19, sí es probable que existan diferentes matices en el abordaje que cada medio o periodista realice en las diferentes notas.

GRÁFICO 3. Evolución temporal de la media de composición de tópicos en comentarios sobre noticias de COVID-19 separado por diario.



FUENTE: Elaboración propia sobre datos recolectados de GDELT y sitios de los medios

GRÁFICO 4. Evolución temporal tasa de variación de la media de composición de tópicos en comentarios sobre noticias de COVID-19 por diario

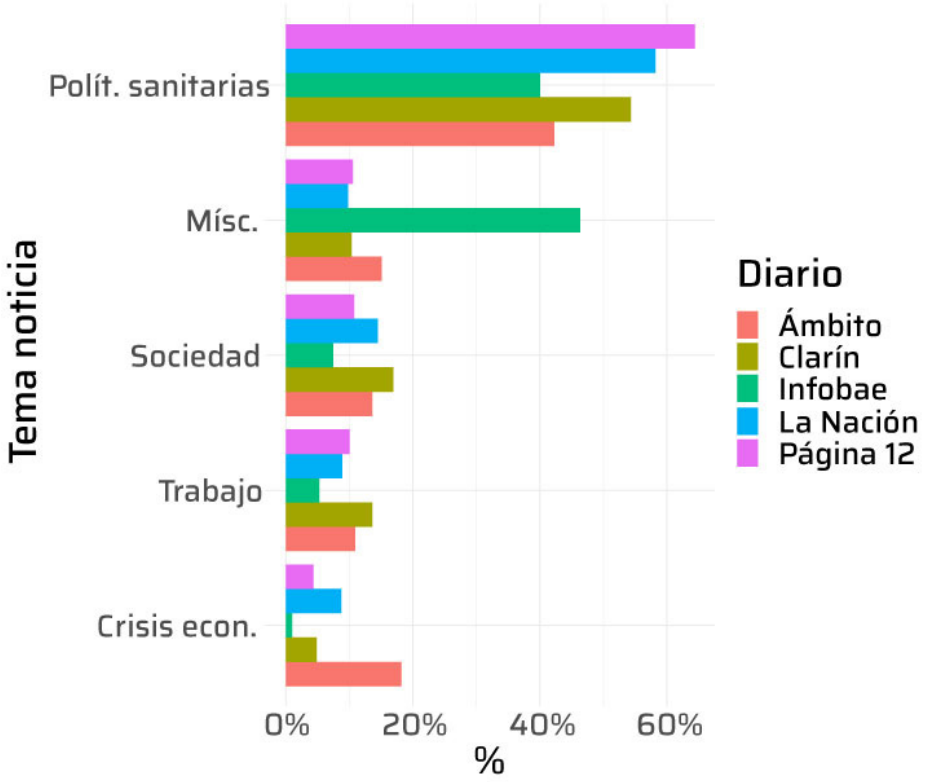


FUENTE: Elaboración propia sobre datos recolectados de GDELT y sitios de los medios

Dado que en este diseño metodológico el tema de las noticias constituye un análogo a dicho estímulo, cabe preguntarse si noticias con diferentes temas (es decir que hablan de diferentes aspectos de la pandemia) tienen algún efecto en los temas de discusión y su evolución.

Para realizar una aproximación a esta pregunta se utilizó el sistema de etiquetado de GDelt, clasificando manualmente las primeras 300 etiquetas de noticias con las frecuencias más altas en cuatro categorías: crisis económica, políticas de salud, noticias sobre la situación de empleo, y sociedad.[19]

GRÁFICO 5. Frecuencias relativas de tipo de artículo según portal de noticias.

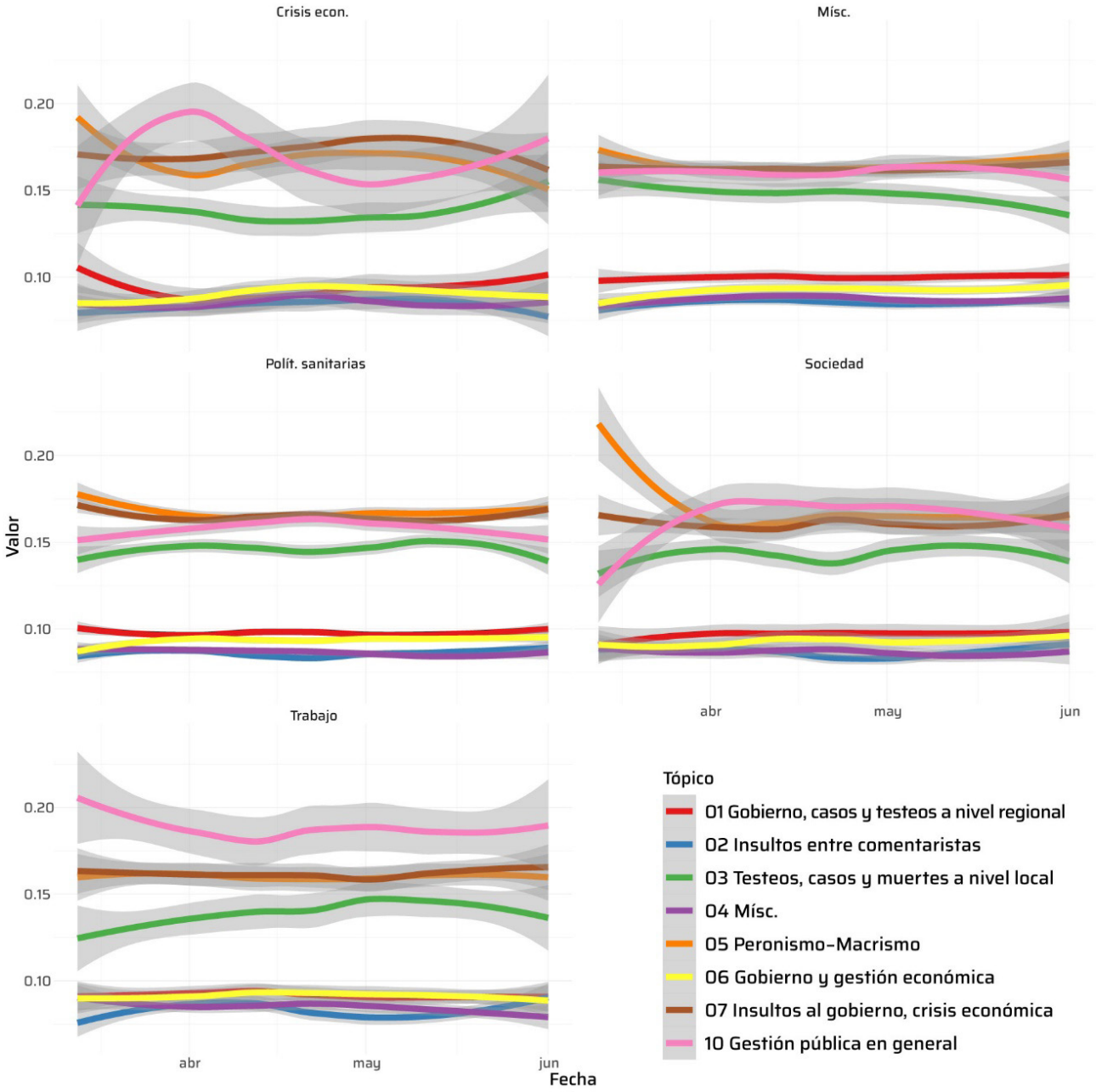


FUENTE: Elaboración propia sobre datos recolectados de GDELT y sitios de los medios

Las noticias sobre políticas de salud son las principales en todos los diarios, seguidas por los artículos sobre la sociedad en general. En *Clarín*, *La Nación* y *Página 12* las primeras representan más de la mitad de los artículos que publican. *Ámbito Financiero* es el diario con mayor cantidad de noticias sobre la crisis económica, e *Infobae* tiene la distribución más atípica, donde los artículos misceláneos tienen la frecuencia más alta.

19 GDELT realiza un etiquetado automático de las temáticas de las noticias. Una de estas etiquetas es la que permite filtrar las noticias que hablan de COVID-19 pero dado que se trata de un sistema de etiquetas múltiples es posible encontrar "subtemas" en las noticias. Pueden verse los criterios de agrupamiento de las categorías de GDELT en el Anexo metodológico.

GRÁFICO 6. Evolución de la media de composición de tópicos en comentarios sobre noticias de COVID-19 según tema de noticia



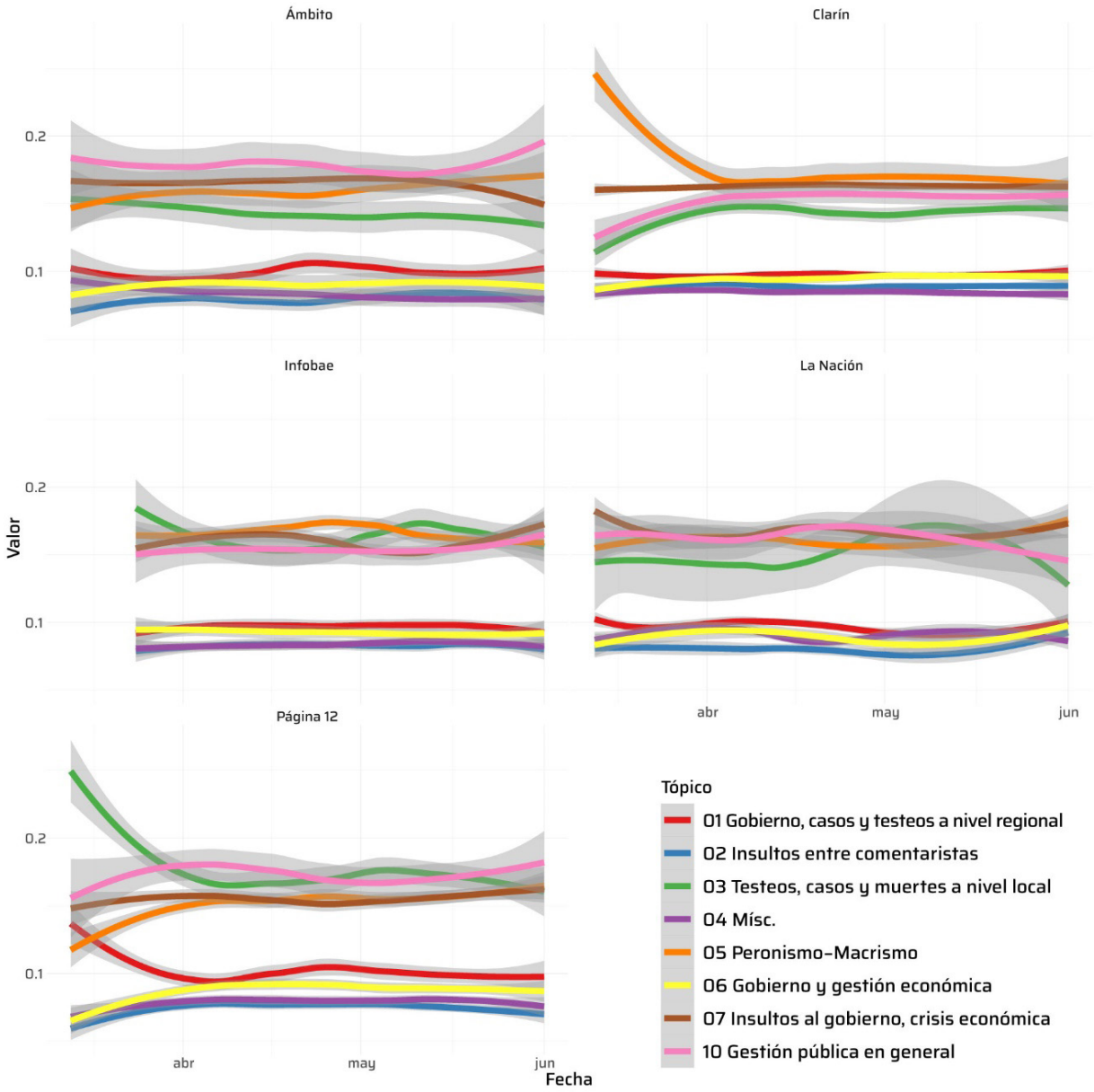
FUENTE: Elaboración propia sobre datos recolectados de GDELT y sitios de los medios

En noticias que no están relacionadas a cuestiones económicas, tenemos una distribución de tópicos similar a la distribución general. Hay una distribución estable de los tres tópicos principales: administración pública, antiperonismo-macrismo e insultos al gobierno y crisis económica. Sólo en las noticias vinculadas a "sociedad" hay una tendencia decreciente del eje antiperonismo-macrismo y una tendencia creciente de las políticas públicas hasta abril.

En la distribución de tópicos en noticias vinculadas a la crisis económica y la situación del mercado laboral se encuentran ligeras variaciones en la distribución de tópicos. En el primer

caso, la administración pública tiene un comportamiento irregular: alcanza un pico en abril superando los otros dos tópicos mayoritarios, decrece hasta mediados de mayo y luego empieza una tendencia creciente de nuevo. En noticias relacionadas al mercado laboral, este tópico se mantiene estable pero constantemente por encima de los otros dos tópicos generales. En líneas generales, parece notarse que las noticias vinculadas al campo de lo económico presentan mayor relevancia de comentarios sobre la administración pública.

GRÁFICO 7. Evolución de la composición de tópicos por diario en noticias que hablan sobre las políticas sanitarias



FUENTE: Elaboración propia sobre datos recolectados de GDELT y sitios de los medios

Si se analiza la evolución de los comentarios en noticias de política sanitaria para cada uno de los medios relevados se manifiesta nuevamente una distribución muy similar a la

general. Aunque *Página 12* y *Clarín* comienzan con una prevalencia muy alta de los tópicos de testeos y de peronismo/antiperonismo respectivamente, ambos decrecen hasta estabilizarse al nivel del resto de los tópicos mayoritarios.

Ahora bien, un punto que resulta llamativo tiene que ver con la relativa homogeneidad que podemos ver en los tópicos de los comentarios en los diferentes medios que se analizan. Surge, casi de inmediato, la pregunta acerca de las causas de dicha homogeneidad. Los tópicos detectados constituyen una primera capa relacionada a los conceptos vinculados a la COVID en tales comentarios. Como tal resulta importante marcar que análisis posteriores deberían poder realizar clasificaciones más detalladas acerca de las diversas características de estos comentarios.[20]

Más allá de estas razones técnicas, parece plausible plantear un interrogante: ¿es posible que exista algún grado de relación entre la homogeneidad de temas que se observan entre los diferentes medios y una posible homogeneidad entre los comentarias/lectores de cada medio? ¿Cómo es la composición social de los comentaristas en los diferentes medios relevados? ¿Existe algún grado de similitud entre algunas características sociodemográficas de los comentaristas de los diferentes medios pese a las evidentes y conocidas diferencias en las líneas editoriales?

¿Homogeneidad en las audiencias de los medios digitales?

Si bien no resulta posible realizar una caracterización exhaustiva de los comentaristas (por las razones expuestas en el apartado sobre las limitaciones y alcances de este trabajo), sí podemos acotar la población en estudio a los lectores de medios digitales. Es razonable asumir que los comentaristas de cada diario constituyen alguna subpoblación de los lectores del mismo. Por ello, intentamos reconstruir el perfil de los mismos.

Para ello, trabajamos con Digital News Report[21] que presenta datos recopilados de 80.000 personas en 40 países (entre ellos Argentina) y registra el uso de los medios entre enero y febrero y en abril de 2020.

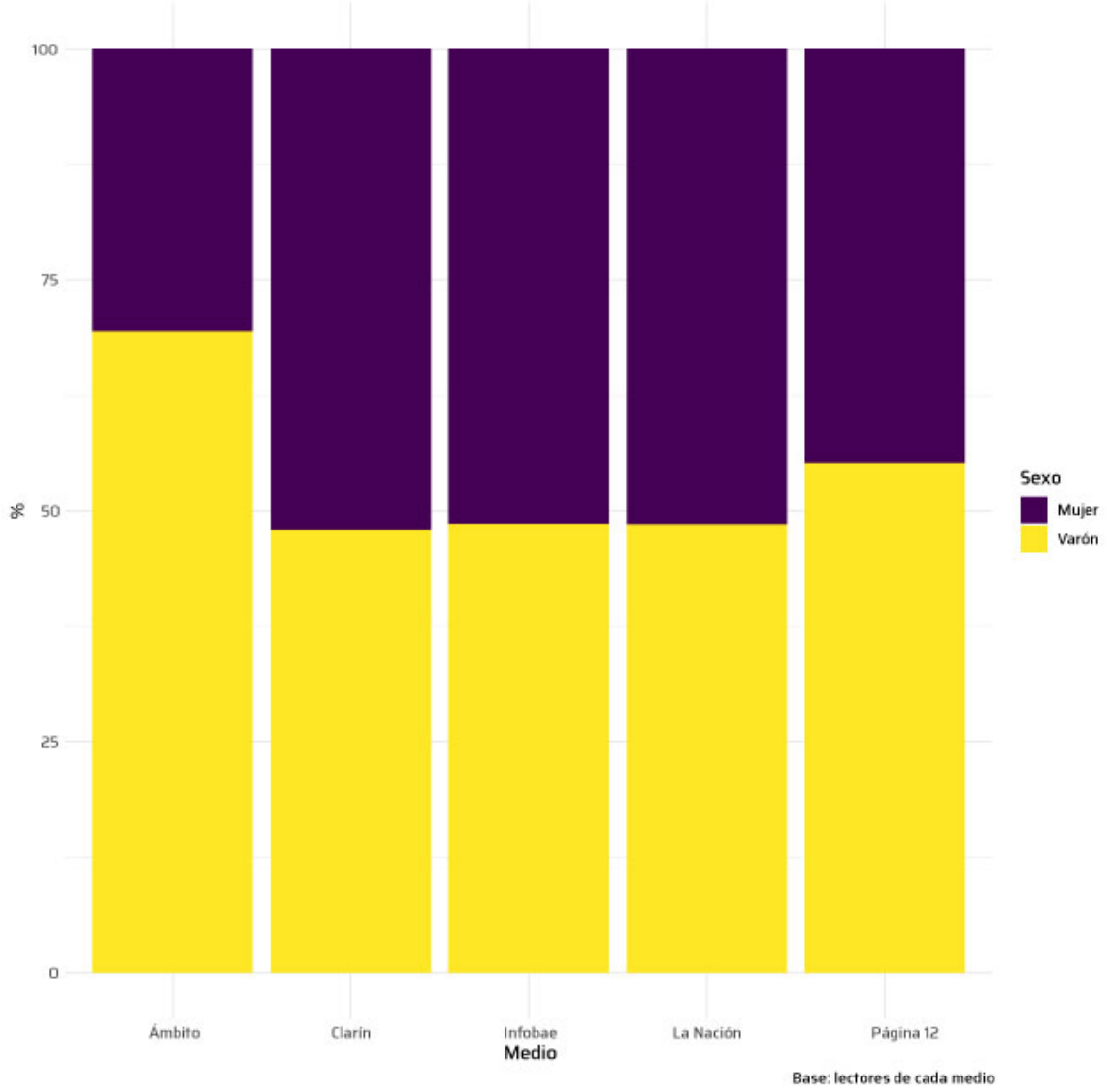
Los lectores de *Clarín*, I*nfobae* y *La Nación* se distribuyen equilibradamente según sexo, esto implica que mujeres y varones concentran, cada uno, alrededor del 50%. En cambio, entre los

20 Por ejemplo, podría intentar realizarse un "sentiment analysis" de estos comentarios: se trata de un conjunto de técnicas de NLP que buscan clasificar un determinado texto en función del tono y las emociones predominantes.

21 Puede encontrarse información al respecto en https://www.digitalnewsreport.org/survey/2020/foreword-2020/. Se trata de una encuesta autoadministrada vía web; es decir, una encuesta que a los efectos prácticos es sumamente similar a un relevamiento coincidental. No se trata de un muestreo probabilístico, lo cual hace que los resultados analizados aquí deban ser tomados con cautela. Resulta llamativa, no obstante, la escasez de estudios exhaustivos (publicados, al menos) sobre las características de las audiencias de los medios digitales en Argentina. Una excepción la constituye el relevamiento de la Encuesta de Satisfacción Política y Opinión Pública (ESPOP) de la Universidad de San Andrés (UDESA) que a mediados del 2020 realizó una onda especial sobre medios de comunicación junto con el Centro de Estudios de Medios y Sociedad en Argentina (MESO-UDESA). Lamentablemente, los microdatos no se encontraban disponibles al momento de la redacción de este trabajo.

lectores de *Ámbito* se destacan los varones, ya que estos significan cerca del 70% del total de personas que leen este diario. Por su parte, entre los lectores de *Página 12*, también se destacan los varones (55%), pero la diferencia con las mujeres (45%) no es tan profunda.

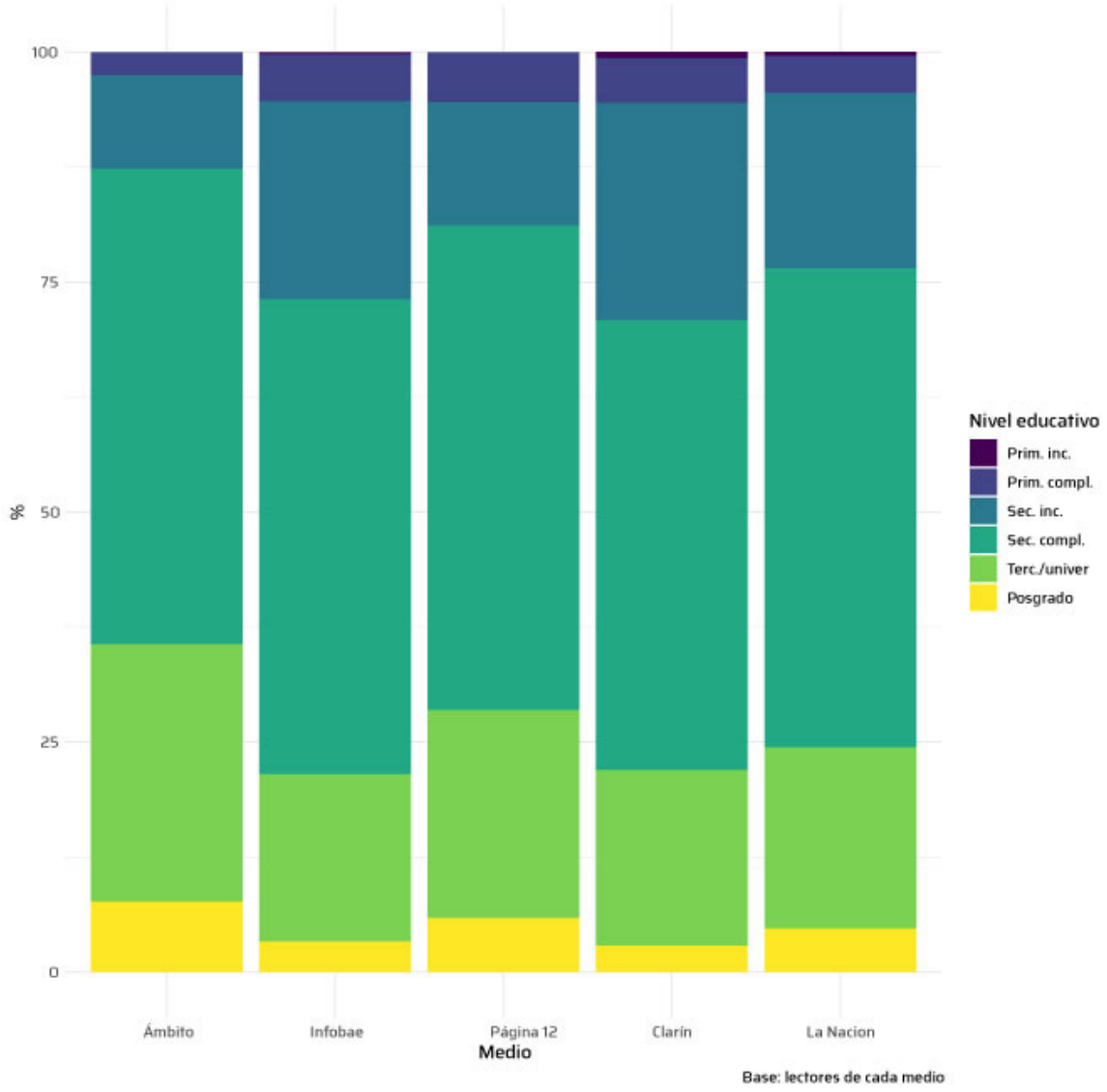
GRÁFICO 8. Personas que leen diarios en formato digital por medio según sexo



FUENTE: Elaboración propia sobre microdatos de Reuters Institute Digital News Report 2020

Los niveles educativos más altos se observan entre los lectores de *Ámbito*, alrededor del 38% corresponde a universitarios/técnicos superiores o personas con posgrados. A su vez, los lectores de *Página 12* con esos niveles altos de educación concentran cerca del 30%. Por el contrario, los diarios cuyos lectores tienen los niveles educativos más bajos son *Infobae* y *Clarín*. En estos periódicos cerca del 25% de los lectores alcanza como máximo nivel de instrucción hasta primaria completa, guarismo que en el caso de *Ámbito* corresponde al 10% y en *Página 12* no supera el 15%.

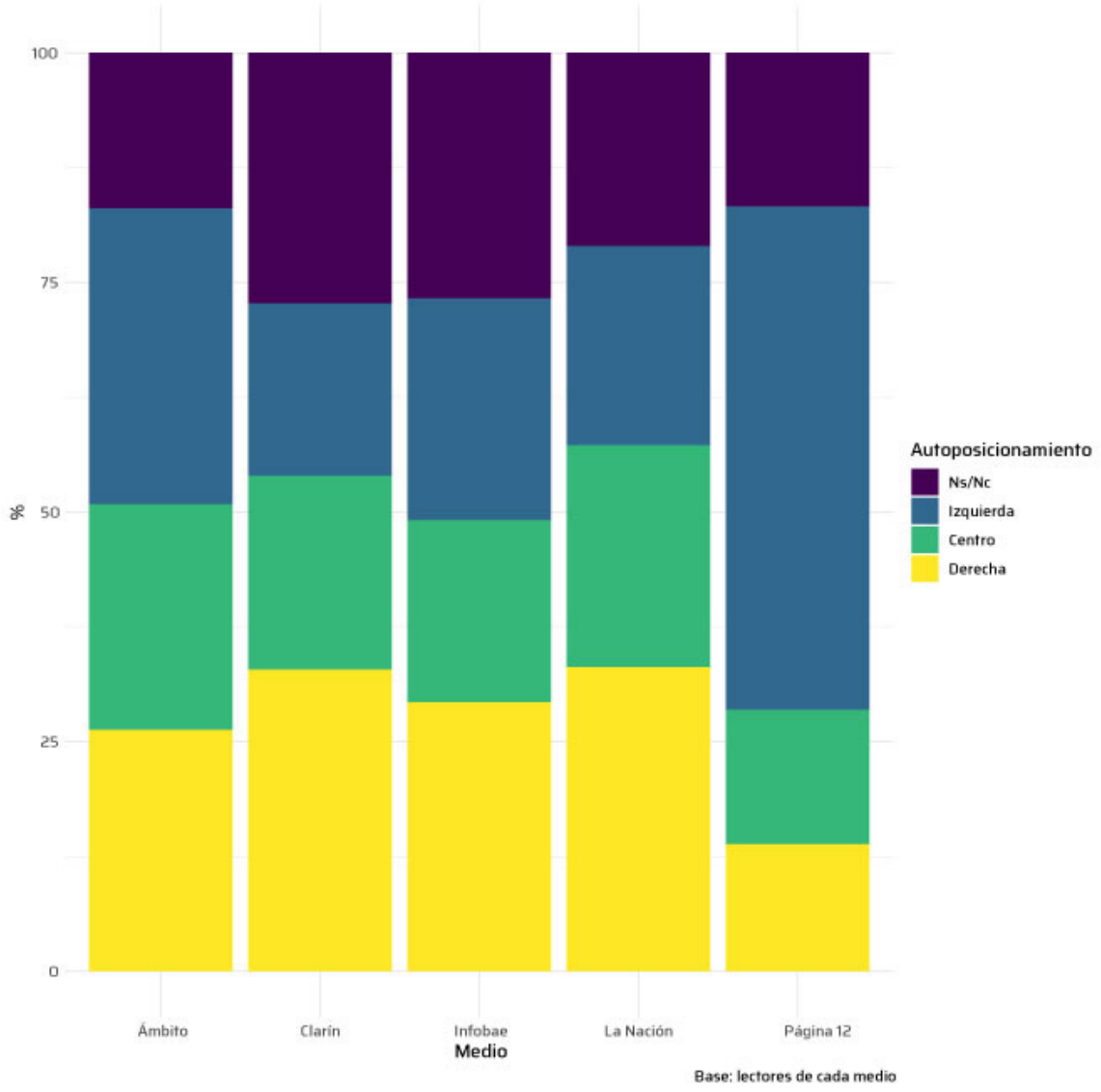
GRÁFICO 9. Personas que leen diarios en formato digital por medio según por nivel educativo



FUENTE: Elaboración propia sobre microdatos de Reuters Institute Digital News Report 2020

Otra característica que analizamos en los lectores de los diferentes diarios, es el autoposicionamiento ideológico. El diario donde se registra con mayor claridad el autoposicionamiento respecto a una sola ideología, es *Página 12*: alrededor del 60% de los lectores se consideran de izquierda. En el caso de *Ámbito* también se destaca entre sus lectores, aunque en menor medida, el autoposicionamiento de izquierda, ya que alcanza valores cercanos al 30%. Entre los lectores de *Clarín*, *Infobae* y *La Nación*, su distribución según el autoposicionamiento ideológico es bastante equitativa entre las tres posiciones y los que respondieron ns/nc. Sin embargo, se puede destacar que los lectores que se consideran de derecha se distinguen principalmente entre quienes leen *Clarín* y *La Nación*, alcanzando cerca del 30% de los mismos. Un último aspecto que se destaca entre los lectores de todos los diarios son los porcentajes altos de quienes no se identifican con ninguna ideología y respondieron Ns/Nc: en *Clarín* e *Infobae* esta característica la registran más del 25% de los lectores y en *La Nación* corresponde a un valor cercano al 20%.

GRÁFICO 10. Personas que leen diarios en formato digital por medio según autoposicionamiento ideológico



FUENTE: Elaboración propia sobre microdatos de Reuters Institute Digital News Report 2020

Así, el análisis de estas características básicas parece aportar algunos elementos a la hipótesis de homogeneidad de lectores. Si bien existen algunos diarios con ciertas características diferenciales (probablemente, *Página 12* con su alta proporción de autoposicionamiento ideológico de izquierda o Ámbito financiero que presenta niveles educativos ligeramente más elevados) las distribuciones no son tan divergentes respecto a sexo, nivel educativo y posicionmiento ideológico. Es decir, los lectores de cada medio parecen tener características sociodemográficas similares. En qué medida esta homogeneidad de los lectores se relaciona con la homogeneidad de la evolución de los tópicos detectados, es un problema que queda planteado para futuras aproximaciones.

Discusión de resultados y nuevos problemas

Este trabajo presentó algunas aproximaciones preliminares y de carácter exploratorio acerca de las posibilidades que abren el uso de técnicas vinculadas al *web scraping* y al procesamiento de lenguaje natural para el estudio de los temas y discursos que se producen alrededor de cierto tópico. Se utilizó la COVID-19 como una primera prueba para evaluar la viabilidad metodológica de estas herramientas.

Para ello, se descargaron los (aproximadamente) 360.000 comentarios en foros de lectores de cinco diarios de circulación nacional entre marzo y junio del 2020 y se aplicó una técnica de modelado de tópicos (LDA) para detectar los principales temas de las conversaciones.

Así, las conversaciones sobre testeos y casos a nivel local, los debates sobre la administración pública, el debate peronismo-antiperonismo y los insultos al gobierno aparecen como temas relevantes.

La existencia de estos últimos tópicos vinculados a la contradicción entre peronismo-antiperonismo reactualiza el debate sobre la llamada "grieta" y la "polarización" en las redes sociales. De hecho, resulta interesante que este tópico sea visible al calcular los promedios de prevalencia de cada tópico y no al analizar la evolución de los tópicos predominantes de los comentarios. Este hecho hace posible plantear como hipótesis que ese componente de contradicción se exprese de forma menos manifiesta que otros temas (hay menos comentarios con ese tópico predominante) aunque de manera más latente (muchos comentarios tiene alguna prevalencia de este tópico que va articulada con algún otro tema). ¿Qué tan estable es la existencia del peronismo-antiperonismo en los tópicos? ¿Es posible hallar este eje en otros debates? ¿Asume la misma forma "latente"? Quedan planteadas estas preguntas para futuras aproximaciones.

Un punto relevante parece ser la aparente estabilidad en la evolución de los tópicos en este primer período. Independientemente de la métrica que se utilice (y más allá de algunas diferencias menores) la pauta que parece observarse es la de pocos cambios relativos. Solamente al final del período, pareciera comenzar a alterarse levemente la composición de los tópicos detectados. También entre los diferentes medios y entre muchos de los temas de las noticias (que funcionarían como el estímulo que dispara la discusión en cada foro) se observa esta pauta de homogeneidad.

La pregunta sobre los motivos de esta homogeneidad llevó a evaluar la posibilidad de que existieran perfiles similares de lectores entre los diferentes medios. Es decir: independientemente de la línea editorial de los medios, existe un conjunto de lectores que presentan características similares (e incluso, podría haber lectores que se solapen, es decir, que lean y participen en los foros de más de un medio). No obstante, esta aproximación adolece de una serie de limitaciones. La primera tiene que ver con la fuente utilizada (una encuesta coincidental sobre consumo de medios a nivel mundial). La segunda, más conceptual y que se vincula con un problema de cobertura: el universo de lectores de medio digitales es, seguramente, más amplio que el de comentaristas de foros.

Este último punto, lleva a remarcar algunos aspectos metodológicos del trabajo que resultan relevantes. Una primera característica de este tipo de estudios se vincula con su "baja reactividad". A diferencia de los métodos basados en entrevistas (sean cualitativas o cuantitativas) la información generada y analizada en este trabajo no se encuentra mediatizada por una situación de entrevista, generalmente caracterizadas como de "alta reactividad". Más bien (y de forma análoga a otras redes sociales y foros) las conversaciones se dan de manera espontánea con algunos disparadores o estímulos muchas veces endógenos, como puede ser la noticia o las intervenciones de otros comentaristas.

Este carácter espontáneo tiene su contracara en el escaso control que el investigador tiene del proceso de captura de los datos. Esto se evidencia en varios aspectos pero uno de los más relevantes es el que se produce como consecuencia de la intervención de agentes como trolls o spammers en las conversaciones de los foros. Como se mencionó, a partir del control de los comentarios duplicados se logró morigerar su influencia.

En este sentido, teniendo en cuenta estas características metodológicas es importante recordar que los tópicos detectados apuntan a constituir una primera aproximación que logre habilitar más preguntas de investigación al respecto y no buscan ser "representativos" de las opiniones de ninguna población.

Al mismo tiempo, el uso de estas herramientas permite la generación de información en una escala temporal continua: dado que los comentarios se producen de forma constante sería posible (en teoría) generar información y análisis con periodicidad muy cercana al "tiempo real".

También sería posible escalar este prototipo metodológico tanto a nivel geográfico, incorporando diarios de menor tirada y/o diarios locales (GDELT cuenta con información de medios gráficos locales); como a nivel temático, abarcando otros tipos de noticias (y comentarios) de interés.

Finalmente, quedan como futuros aspectos técnicos pendientes de evaluación ponderar mejores aproximaciones a la estimación de los modelos (mejores estrategias de tuneo de los hiper parámetros) y testear la aplicación de modelos de NLP más avanzados (como *word embeddings* contextuales).

Anexo metodológico

TABLA A1. Etiquetas usadas para clasificar el tipo de artículo.

CLASIFICACIÓN	ETIQUETAS
Crisis económica	-WB_1104_MACROECONOMIC_VULNERABILITY_AND_DEBT, WB_695_POVERTY, WB_450_DEBT, ECON_DEBT, EPU_POLICY_CENTRAL_BANK, ECON_STOCKMARKET, ECON_DEVELOPMENTORGS, WB_1973_FINANCIAL_RISK_REDUCTION,
-Situación de empleo	-WB_2689_JOBS_DIAGNOSTICS, WB_697_SOCIAL_PROTECTION_AND_LABOR, WB_2690_CATEGORIES_OF_EMPLOYMENT, WB_855_LABOR_MARKETS, WB_2745_JOB_QUALITY_AND_LABOR_MARKET_PERFORMANCE, RETIREMENT, UNGP_JOB_OPPORTUNITIES_EMPLOYMENT, UNEMPLOYMENT, WB_2747_UNEMPLOYMENT, WB_1654_ACTIVE_LABOR_MARKET_POLICIES, WB_856_WAGES, WB_2748_EMPLOYMENT
-Políticas de salud	-WB_635_PUBLIC_HEALTH, UNGP_HEALTHCARE, WB_2165_HEALTH_EMERGENCIES, WB_2166_HEALTH_EMERGENCY_PREPAREDNESS_AND_DISASTER_RESPONSE, WB_1305_HEALTH_SERVICES_DELIVERY, EPU_CATS_HEALTHCARE, WB_1466_SOCIAL_ASSISTANCE, HEALTH_VACCINATION, WB_1462_WATER_SANITATION_AND_HYGIENE
-Sociedad	-SOC_QUARANTINE, EPU_ECONOMY_HISTORIC, SOC_GENERALCRIME, TAX_FNCACT_CITIZENS, PROTEST

Referencias Bibliográficas

Adorno, Theodore, Frenkel-Brunswik, Else Levinson, Daniel, Sanford, Nevitt (1950). "Studies in authoritarian personality". *New York, Harper & Row*, pp. 976.

Blei, David, Ng Andrew, Jordan, Michel I. (2003). "Latent Dirichlet Allocation". *The Journal of Machine Learning* (3), pp. 993-1022.

Calvo, Ernesto, Aruguete, Natalia (2018). "#Tarifazo. Medios tradicionales y fusión de agenda en redes sociales". *Inmediaciones de la Comunicación* (13), pp. 189-213.

Calvo, Ernesto y Aruguete, Natalia (2020). *Fake news, trolls y otros encantos. Cómo funcionan (para bien y para mal) las redes sociales*. Buenos Aires, Siglo XXI.

Chang, Jonathan, Gerrish, Sean, Wang, Chong, Boyd-graber, Jordan, Blei David (2009) "Reading tea leaves: How humans interpret topic models". *Neural information processing systems*. Disponible en: https://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models

Dany, Lionel, Urdapilleta, Isabel, y Lo Monaco, Gregory (2015). "Free associations and social representations: some reflections on rank-frequency and importance-frequency methods". *Quality & Quantity*, 49(2), pp. 489–507.

Durkheim, Emile (1968). Las formas elementales de la vida religiosa. El sistema totémico en Australia. Buenos Aires, Schapire.

Gramsci, Antonio (1971). *El materialismo histórico y la filosofía de Benedetto Croce.* Buenos Aires, Nueva Visión.

Gramsci, Antonio (1972). *Notas sobre Maquiavelo, sobre la política y sobre el estado moderno*. Buenos Aires, Nueva Visión.

Lo Monaco, Gregory, Piermattéo, Anthony, Rateau, Patrick, Tavani, Jean Louis (2017). "Methods for studying the structure of social representations: A critical review and agenda for future research". *Journal for the Theory of Social Behaviour*, 47(3), pp. 306-331.

Maguire, Tomás (2021). *Aprendizaje automático y modelización de tópicos: un estudio de caso sobre la agenda mediática en contexto de las elecciones Argentina 2015*, Tesina de Licenciatura en Sociología, Escuela Interdisciplinaria de Estudios Sociales, Universidad Nacional de San Martín.

Marx, Karl, Engels, Friederich (2004). *La ideología alemana*. México, Pueblos Unidos.

McCombs, Maxwell. (2006). *Estableciendo la agenda*, Buenos Aires, Paidós Comunicación.

McCombs, Maxwell y Guo, Lei (2014). "Agenda-setting Influence of the Media in the Public Sphere". Fortner, Robert y Fackler, Mark (eds.). The Handbook of Media and Mass Communication Theory, New York, John Wiley & Sons, pp. 251-268.

Mimno, David, Wallach, Hannah, Talley, Edmund, Leenders, Miriam, y Mccallum, Andrew (2011) "Optimizing semantic coherence in topic models". *Empirical methods on natural language processing*. Disponible en: https://mimno.infosci.cornell.edu/papers/mimno-semantic-emnlp.pdf

Moscovici, Sergei (1979). *El psicoanálisis, su imagen y su público*. Buenos Aires: Huemul.

Muleras, Edna (2008). *Sacralización y desencantamiento. Las formas primarias del conocimiento en el orden social*. Buenos Aires, Miño y Dávila.

Nun, José (2015). "Averiguación sobre algunos significados del peronismo". En Nun, Jose, *El sentido común y la política. Escritos teóricos y prácticos*. Buenos Aires, Fondo de Cultura Económica, pp. 221-278.

Piaget, Jean (2010). *La equilibración de las estructuras cognitivas. Problema central del desarrollo*. México, Siglo XXI.

Piermattéo, Anthony, Tavani, Jean Lois, y Lo Monaco, Gregory. (2018). "Improving the study of social representations through word associations: Validation of semantic contextualization". *Field Methods*, 30(4), pp. 329–344.

Rudé, George (1980). *Ideology and Popular Protest*. Gran Bretaña, Lawrence & Wishart.

Salganik, Matthew (2018). *Bit by bit. Social research in the digital age*. Princeton, Princeton University Press.

Schütz, Alfred y Luckmann, Thomas (2009). *Las estructuras elementales del mundo de la vida*. Buenos Aires, Amorrortu.

Schuth, Anne., Marx, Maarten, Rijke, Maarten (2007). "Extracting the discussion structure in comments on news-articles". *9th ACM International Workshop on Web Information and Data Management*, pp. 97-104. Lisbon, Portugal (2007).

Wagner, Wolfgang, Duveen, Gerard, Farr, Robert, Jovchelovitch, Sandra, Lorenzi-Cioldi, Fabio, Markova, Ivana, Rose, Diana. (1999). "Theory and method of social representations".

Asian Journal of Social Psychology (2), pp. 95–125.

Webb, Eugene, Campbell, Donald, Schwartz, Richard y Sechrest, Lee (1966). *Unobtrusive Measures: Nonreactive Research in the Social Science*. Chicago, Rand McNally.

Wiedemann, Gregor (2016). *Text mining for qualitative data analysis in the social sciences. a study on democratic discourse in Germany*. Berlín, Springer.

Wu, Alice (2020). "Gender bias in rumors among professionals: an identity-based interpretation". *The Review of Economics and Statistics*, 102(5), pp. 867–880.