Procesamiento de lenguaje natural aplicado al estudio de tópicos de noticias de seguridad en Argentina: julio a septiembre 2019

Tesina para obtener el título de Licenciada en Sociología
Carrera de Sociología

Escuela Interdisciplinaria de Altos Estudios Sociales
Universidad Nacional de San Martín

Estudiante: Florencia Piñeyrúa
Director: Dr. Germán Rosati
Co-directora: Dra. Brenda Focás
Evaluador: Dr. Gabriel Kessler
Fecha: mayo 2021

Procesamiento de lenguaje natural aplicado al estudio de tópicos de noticias de seguridad en Argentina: julio a septiembre 2019

Autora: Florencia Nathalia Piñeyrúa

Director: Dr. Germán Rosati

Co-directora: Dra. Brenda Focás

Evaluador: Dr. Gabriel Kessler

Resumen:

La tesina es un trabajo exploratorio donde mostramos la aplicación de técnicas de procesamiento de lenguaje natural y *web scraping* sobre un corpus de noticias digitales. El objetivo general es explorar la aplicación de una técnica de procesamiento de lenguaje natural (modelado de tópicos) para estudiar el contenido de noticias digitales. Para ello, utilizamos como soporte empírico las piezas periodísticas publicadas desde julio a septiembre 2019 en los portales *Clarín*, *La Nación, Infobae, Página 12*, *Télam*, *Perfil*, *Crónica* y *Minuto Uno*. Se utilizan datos primarios construidos a partir de la técnica de *web scraping.* La metodología aplicada combina el análisis descriptivo y una 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). Los resultados de la modelización de tópicos muestran que los principales temas de la agenda mediática digital son las elecciones, los espectáculos, el deporte, la seguridad, la política exterior, la obra pública y la economía. En el contexto de las elecciones Primarias Abiertas Simultáneas y Obligatorias la frecuencia de publicación de las noticias securitarias fue de casi 2 de cada 10 piezas periodísticas. La principal conclusión de este trabajo es que la combinación de la técnica *web scraping* y procesamiento de lenguaje natural pueden ser de utilidad 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 claves: tópicos, procesamiento de lenguaje natural, *web scraping* y noticias de seguridad.

Agradecimientos

Quiero agradecer, en primer lugar, a mis directores, Germán Rosati y Brenda Focás, quienes me acompañaron y aconsejaron incansablemente. Sus aportes, contribuciones y enseñanzas han sido imprescindibles y valiosísimos en el desarrollo de este trabajo y para ampliar mi mirada sobre la temática en general. También quiero agradecer a Marina Dossi, profesora del Taller de redacción de tesina II, por sus aportes para la construcción del objetivo de investigación. A mis compañeros de los talleres de redacción por los intercambios y sugerencias y al equipo de trabajo de FACTOR DATA por la transferencia de conocimientos metodológicos.

Un agradecimiento muy especial es para los profesores Leandro López y Pablo Dalle que a lo largo de la carrera estuvieron siempre para guiarme, en particular, por su excelente predisposición a escuchar todas mis dudas e inquietudes.

Al equipo Migrantas en Reconquista que lograron que todos los días quiera ir a trabajar con una sonrisa, y a Lucila Nejamkis y Natalia Gavazzo por iniciarme en el oficio de la investigación y constantemente proponerme desafíos.

A los integrantes del Núcleo de Estudios sobre Violencia y Muerte que en un diálogo horizontal han ampliado y enriquecido mi visión sobre la temática del delito y la seguridad; y, en especial, a Violeta Dikenstein por leer el borrador de la tesina.

Un agradecimiento muy afectuoso merece mi familia. A mi mamá, Lorena, por ser mi incansable lectora. A Iván y a mis abuelos, Eduardo y Norma, por escuchar mis reflexiones sociológicas y calmar mis ansias. A mi prima, Avril, y a mis tías, Eugenia y Analía, por su cariño incondicional.

Índice de la tesina

Introducción

Esta tesina trata sobre una técnica de aprendizaje automático de procesamiento de lenguaje natural (modelado de tópicos, particularmente mediante el método LDA) aplicada al estudio de los tópicos de noticias de delito y seguridad durante el periodo de julio a septiembre 2019 en Argentina. Al interior de los campos de estudio de la Sociología de la Comunicación y la Ciencias de la Comunicación, que abordan 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, la utilización de técnicas de procesamiento de lenguaje natural y *web scraping*[1], aunque poco exploradas por ambas disciplinas, 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 securitarias combinando la recolección automatizada de noticias y la técnica de modelado de tópicos. Como soporte empírico utilizamos las noticias digitales desde julio a septiembre de 2019, período en el cual se desarrollaron las elecciones Primarias Abiertas Simultáneas y Obligatorias (PASO)[2], y analizamos de forma exploratoria los cambios y continuidades que experimentó los tópicos de la agenda mediática digital.

La presente tesina emerge del diálogo entre las Ciencias Sociales y las Ciencias Computacionales y se enmarca en la subdisciplina Ciencias Sociales Computacionales, donde las Ciencias Sociales proporcionan temáticas y problemáticas para ser abordadas con el desarrollo y el uso de técnicas computacionales y estadísticas. El estudio de procesos sociales a través de métodos computacionales no es una novedad. Desde la mitad de la década de 1960 se han realizado investigaciones de este estilo en disciplinas como la Sociología, la Antropología o las Ciencias Políticas (Rodriguez Zoya y Roggero, 2015). A su vez, el cambio de milenio trajo aparejada la acumulación de grandes volúmenes de información estadística, longitudinal, multivariada y no estructurada sobre tendencias económicas y sociales y del comportamiento humano, comúnmente denominada *Big Data*.

Esta tesina tiene como objetivo general explorar la aplicación de una técnica de procesamiento de lenguaje natural (modelado de tópicos) para estudiar el contenido de noticias

[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).
[2] Las PASO se realizaron el domingo 11 de agosto de 2019.

digitales. Para ello, se emplean como soporte empírico las noticias digitales en el contexto electoral de las PASO 2019 en Argentina. A su vez, la tesina surge en el marco de reflexión del laboratorio Factor DATA (IDAES-UNSAM) que propone nuevas herramientas para el estudio de lo social. La pregunta central que moviliza el presente trabajo es cuál es la prevalencia de las noticias sobre delito y qué relevancia tienen en comparación con otros temas de la agenda mediática entre julio y septiembre de 2019. La metodología que aplicamos combina el análisis descriptivo y técnicas de aprendizaje automático empleando un algoritmo de modelado de tópicos. El universo de estudio, es decir, las fuentes que utilizamos son las noticias publicadas en los portales de los principales medios digitales de comunicación de Argentina: *Clarín*, *La Nación, Infobae, Página 12, Télam, Perfil*, *Crónica* y *Minuto Uno* desde julio a septiembre de 2019. La información la procesamos a partir del instrumento de modelado de tópicos con la implementación del método *Latent Dirichlet Allocation Models* (LDA).

En base a lo relatado, se formularon dos hipótesis, una que explora la dimensión metodológica del problema y otra sobre el proceso de cobertura mediática.

Primera hipótesis: la aplicación de técnicas de aprendizaje automático 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.

Segunda hipótesis: estas técnicas computacionales hacen posible incrementar sensiblemente la escala del análisis 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.

Es importante tener en cuenta que nuestro objetivo central 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. En lo que respecta al soporte de los medios, optamos por trabajar con diarios online ya que estudios recientes (Koziner et al., 2018) estiman que el nivel de consumo de estos medios asciende a prácticamente la mitad de la población (SINCA - Encuesta Nacional de Consumos Culturales del Ministerio de Cultura de

la Nación Argentina, 2017). A su vez, el delito en sus distintas variantes ha sido una de las principales preocupaciones tanto de la opinión pública como de las Ciencias Sociales a lo largo del siglo XX, y en particular de la Sociología. En América Latina se trata de uno de los principales asuntos de preocupación ciudadana configurándose desde el 2008 como el principal problema de la región (Kessler, 2009), ayudada entre otros factores por el crecimiento cuantitativo de las noticias de inseguridad (Focás y Kessler, 2015). Como consecuencia, la batalla contra la inseguridad se transformó en un eje central de las campañas electorales que se centraron, con frecuencia, en discursos sobre el miedo al crimen (Calzado et al., 2014).

Independientemente de las estadísticas delictivas, la llamada "lucha contra el crimen" se consolida como elemento privilegiado en las agendas electorales y la inseguridad se presenta como una temática recurrente de las disputas políticas. Autores como Calzado, Lio y Gómez (2019) consideran a los contextos electorales como espacios conflictivos donde se plasman los imaginarios políticos y las preocupaciones sociales contemporáneas. Los actores pertenecientes a la administración pública y al gobierno juegan un rol decisivo en la definición de los temas políticos que se debaten en un momento dado e influyen en las formas que cobran los mensajes en los medios (Retegui et al., 2019). En las campañas electorales la problemática securitaria se expresa en forma de demandas hacia los funcionarios públicos (Zunino y Focás, 2019a). Por todo lo anterior, optamos por estudiar la prevalencia y relevancia de los tópicos de noticias de delito y seguridad en la agenda mediática digital durante el proceso electoral de las elecciones Primarias Abiertas Simultáneas y Obligatorias en Argentina. 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. De manera general, el escenario político argentino tuvo dos partidos como protagonistas: Juntos por el Cambio y la colación Frente de Todos. Cambiemos es una alianza de tendencia liberal conservadora surgida en 2005 bajo el liderazgo del empresario local Mauricio Macri, que llegó al Poder Ejecutivo Nacional en diciembre de 2015. El Frente de Todos es la colación gobernante en Argentina (2019 – 2023), la fórmula presidencial encabezada por Alberto Fernández y Cristina Fernández de Kirchner se impuso con el 49,49% de los votos sobre el entonces presidente Mauricio Macri y Miguel Ángel Pichetto (32,93%)[3].

Asimismo, el periodo seleccionado (julio a septiembre de 2019, contexto electoral de las PASO 2019) no se ha estudiado aún desde la perspectiva de la agenda mediática securitaria,

[3] Los porcentajes corresponden al escrutinio final publicado el 3 de septiembre de 2019 por la Cámara Nacional Electoral.

optamos por incluir un mes (julio) de campaña electoral, un mes donde se desarrollan los comicios (agosto) y un mes (septiembre) posterior a los resultados. El presente abordaje del problema puede ser útil para el desarrollo de futuras investigaciones tanto en el campo de la Sociología de la comunicación como en la Ciencias de 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. Asimismo, el enfoque metodológico empleado puede servir para abrir nuevas preguntas de investigación sobre el problema. Así, como parte de las conclusiones formulamos un conjunto de preguntas-problemas que queda abierta: ¿cuáles son los principales tópicos de la agenda mediática securitaria en el contexto electoral de las PASO 2019? y ¿cuál es la dinámica temporal de la agenda mediática securitaria durante dicho contexto?

Los principales resultados de la investigación se presentan en tres capítulos. En el primero, desarrollamos el abordaje de los antecedentes teórico-metodológicos de las investigaciones sociológicas que aplican técnicas de aprendizaje automático a nivel internacional y de los estudios de tópicos de noticias en Argentina, con el propósito de analizar las principales fortalezas y problemas de ambos enfoques. Este capítulo culmina con la explicación de la estrategia metodológica y el flujo de trabajo para el análisis textual computacional. En el segundo capítulo, presentamos los resultados del análisis del modelado de tópicos a partir de tablas y gráficos. El tercer capítulo dialoga con la literatura del campo de análisis de medios: especificamos los aportes de la aplicación de técnicas de procesamiento de lenguaje natural empleadas en esta tesina, estableciendo una comparación con las técnicas de análisis de contenido cuantitativo que caracterizan las estrategias metodologías de las investigaciones sobre tópicos de noticias securitarias en Argentina. Por último, finalizamos el último capítulo con una reflexión sobre las fortalezas de combinar los enfoques metodológicos computacionales y cuantitativos en la investigación social.

Las principales conclusiones que se desprenden de esta tesina se relacionan a las potencialidades de abordar el estudio de tópicos de la agenda mediática digital a partir de técnicas de aprendizaje automático que destacan características del texto. En este sentido, las técnicas de *web sacraping* y procesamiento de lenguaje natural permiten sistematizar las diversas etapas del proceso de investigación (la recolección de datos, la construcción del corpus, el pre-procesamiento de un texto y su análisis), incrementar sensiblemente la escala de captura y análisis de información y abordar de forma sistemática la evolución temporal.

A nivel de estudio del caso empírico, nos propusimos explorar la prevalencia y relevancia de los tópicos de las noticias de delito y seguridad comparadas con otros temas de distintos diarios online de Argentina– *Clarín, La Nación, Infobae, Página 12, Télam, Perfil, Crónica* y *Minuto Uno*– entre julio a septiembre de 2019. En este marco es importante aclarar que el ejercicio propuesto tiene como objetivo mostrar un caso de uso de la herramienta metodológica más que agotar determinaciones del objeto de estudio. Durante el contexto de las elecciones Primarias Abiertas Simultáneas y Obligatorias las elecciones, los espectáculos, el deporte, la seguridad, la política exterior, la obra pública y la economía fueron prioridad en las agendas mediáticas digitales. La frecuencia de publicación de las noticias online de delito y seguridad fue de casi 2 de cada 10 piezas periodísticas. En comparación con el resto de los tópicos detectados en el corpus, el caso de seguridad es el cuarto en orden de relevancia. La evolución temporal del tópico securitario muestra que se mantiene como un tema relevante y estable en la agenda mediática.

La estrategia metodológica que empleamos en esta tesina podría formar parte de una investigación más amplia sobre contenidos de noticias. En este caso queda abierta una futura línea de investigación basada en una metodología mixta: análisis computacional de tópicos y análisis de contenido cuantitativo. La técnica de modelado de tópicos permite identificar en un corpus de noticias los tópicos relevantes y seleccionar las piezas periodísticas con alta prevalencia de estos tópicos, para luego realizar un análisis de contenido cuantitativo sobre estos textos seleccionados. De esta forma se pueden combinar las metodologías cuantitativas y computacionales en una misma investigación.

Capítulo 1

Panorama teórico-metodológico del problema

Tres interrogantes guían el desarrollo del capítulo. ¿Cuáles son las dificultades y potencialidades de aplicar técnicas de aprendizaje automático al estudio de lo social? ¿Qué clases de temas, preguntas-problemas, objetos de estudios y objetivos de investigación ameritan la utilización de técnicas de aprendizaje automático en la investigación sociológica? ¿Cómo son las estrategias metodológicas empleadas en las investigaciones en Argentina sobre tópicos de noticias? El capítulo se organiza en tres apartados. En el primero, realizamos un breve recorrido sobre los principios epistemológicos y teórico-metodológicos que subyacen a la utilización de las técnicas de aprendizaje automático en relación con los estudios sociológicos a nivel mundial, y abordamos el modelado de tópicos. En el segundo, analizamos las estrategias metodológicas que utilizan las investigaciones en Argentina sobre tópicos de noticias, en especial, las que indagan sobre la cuestión securitaria. Por último, el capítulo finaliza con la explicación de la estrategia metodológica y del flujo de trabajo para el análisis textual computacional.

Antiguos problemas, nuevas herramientas: las técnicas de aprendizaje automático aplicadas a la investigación sociológica

El ingreso al siglo XXI trajo aparejados cambios técnicos que están impulsando los avances en el campo del aprendizaje automático. Este campo de estudio se guía por hallazgos experimentales de modo que los avances algorítmicos sólo son posibles cuando los conjuntos de datos y el hardware apropiados están disponibles (Challote, 2018). Según diversos autores, el aumento de disponibilidad de grandes fuentes de información observacionales que refieren a una gran variedad de fenómenos sociales representa un momento decisivo para las Ciencias Sociales (Kitchin, 2014; McFarland et al., 2015; Uman, 2018). Las tecnologías móviles (celulares, apps, etc.), la "internet de las cosas" (IoT), la información producida por los usuarios de redes sociales son algunos de los vectores de este incremento en la disponibilidad de información. En esta misma dirección, Wallach (2014) sostiene que se trata de conjuntos de datos sociales que documentan los comportamientos de las personas en su vida cotidiana; son "huellas digitales" de la actividad e interacción humana y no información expresada por un individuo colocado en una relación delimitada con un investigador (McFarland et al., 2015).

A su vez, estos datos han creado la demanda de nuevos métodos que reducen/simplifican su dimensionalidad, identifican nuevos patrones y relaciones y predicen

resultados. En la búsqueda de modos de reducirlos a dimensiones abordables y significativas existen diversas variedades de técnicas computacionales (por ejemplo, los experimentos in situ) y disciplinas como la lingüística computacional. En esta tesina abordamos la modelización de tópicos que es parte del conjunto de técnicas de aprendizaje automático. De manera general, las técnicas de aprendizaje automático hacen referencia a algoritmos que emplean características para predecir varios resultados. Siguiendo las elaboraciones, estas consisten en un conjunto de algoritmos que pueden detectar automáticamente patrones/asociaciones en los datos o "aprender" de los datos.

En la década de 1950 surgen las primeras técnicas de aprendizaje automático como subárea en el campo de la inteligencia artificial. El objetivo central de estas técnicas es predecir correctamente datos que están fuera de la muestra, para lo cual es fundamental una estrategia de construcción de modelos. Por su parte, el supuesto clave en el aprendizaje automático sobre el mecanismo de generación de datos es que los mismos se extraen independiente e idénticamente distribuidos de una distribución desconocida (Mazzocchi, 2015), es decir, "[that] uses algorithmic models and treats the data mechanism as unknown"[4] (Breiman, 2001: 199). También una parte sustancial se dedica a elegir una complejidad óptima: un modelo es más complejo si contiene una mayor cantidad de variables o una mayor cantidad de interacciones y no linealidades en los mismos. El aumento de la complejidad de un modelo puede redundar en modelos menos sesgados pero más "erráticos" lo que en la jerga produce el llamado "*overfitting*" (Sosa Escudero, 2019). Los parámetros del modelo son estimados con los datos y luego se realizan las predicciones.

En la Sociología el concepto de modelo y la práctica de modelización no conforman el *habitus* metodológico[5]. En la actualidad de las Ciencias Sociales latinoamericanas, este enfoque es poco conocido y empleado, aunque sí se destacan incipientes contribuciones en Argentina (Calvo y Aruguete, 2020; Kessler et al., 2021; Sosa Escudero, 2019). En general, en las Ciencias Sociales se trabaja en mayor medida con modelos discursivos que con modelos formales (Rodriguez Zoya y Roggero, 2015); por supuesto que en la Sociología los modelos formales más habituales son los estadísticos centrados en la relación entre variables y en los cálculos de probabilidad que conforman el núcleo de la metodología cuantitativa. Por su parte, los modelos computacionales constituyen otro tipo de modelo formal expresado como

[4] "[que] usa modelos algorítmicos y trata el mecanismo de datos como desconocido" (Breiman, 2001: 199).
[5] La sociología pragmática francesa representa una excepción a la afirmación anterior. No obstante, la concepción de modelo (gramática) y el proceso de modelización difiere radicalmente de la propuesta por las Ciencias Computacionales.

programas informáticos que pueden ser ejecutados en una computadora. Tal como exponen Rodriguez Zoya y Roggero (2015) estos modelos permiten expandir el horizonte de la investigación social al repensar algunos problemas clásicos de la disciplina a partir de herramientas novedosas.

Para el cientista social la aplicación de técnicas de aprendizaje automático a grandes volúmenes de información no estructurada trae aparejada ciertas ventajas y desventajas. A partir de la revisión de la bibliografía especializada destacamos las principales potencialidades y problemas del empleo de dichas técnicas y de grandes volúmenes de información multivariada, longitudinal y no estructurada. A su vez, partimos de la premisa de que los métodos computacionales no anulan los métodos clásicos de las Ciencias Sociales (métodos cuantitativo y cualitativo), más bien afirmamos que guardan complementariedad entre sí (Garg et al., 2018; Mazzocchi, 2015; Mützel, 2015; Rodriguez Zoya y Roggero, 2015). Las técnicas de aprendizaje automático nos muestran la presencia de patrones, pero muchas veces existen dificultades de interpretación y no llegamos a comprender qué procesos sociales e individuales subyacen a dichas secuencias (Sautu, 2019; Uman, 2018).

Un problema clásico en la investigación social proviene de la utilización de datos secundarios: no siempre existe una interacción entre la teoría y la operacionalización empírica en el proceso de creación de los mismos (Mazzocchi, 2015; Sautu, 2019). Este problema reaparece (incluso de forma ampliada) en el uso de los "grandes datos". Aunque, cuando se construyen estas bases de datos con fines cognoscitivos –como en el caso de esta tesina- es posible recolectar los datos empíricos en diálogo con la teoría. A su vez, estos datos tienden a ser propensos a errores, presentar sesgos y poca representatividad (McFarland et al., 2015; Salganik, 2017; Sautu, 2019; Uman, 2018). Si bien el gran tamaño de los datos reduce la necesidad de preocuparse por errores aleatorios, en realidad aumenta la necesidad de prestar atención a los errores sistemáticos que surgen de sesgos en la forma en que se crean o recolectan los datos (Salganik, 2017). Ahora bien, la poca representatividad de los datos no permite hacer generalizaciones fuera de la muestra[6], pero puede ser bastante útil para comparaciones dentro de la muestra; siempre que los investigadores tengan claro las características de su muestra y sustente las afirmaciones sobre la transportablilidad de los patrones encontrados con evidencia teórica o empírica. En relación con los márgenes de error y los niveles de significación estadísticamente aceptados en el análisis de estos datos no siempre es posible determinarlos.

[6] En general, estos datos masivos, multivariados, longitudinales y no estructurados no provienen de una muestra aleatoria probabilística de una población definida, por lo tanto, no suele ser posible establecer generalizaciones al universo de estudio.

Por ejemplo, los datos extraídos de redes sociales o páginas web tienen diversas dimensiones asociativas (en Facebook, amistades; en diarios online, noticias digitales; en *jornal*, artículos científicos). Además, estos conjuntos de datos con frecuencia consisten en selecciones sesgadas de los individuos al captar personas que utilizan más las redes sociales o sujetos que generan tipos particulares de registros (por ejemplo, investigadores académicos que escriben artículos y no libros).

Por lo que respecta a la potencialidad, proveen información complementaria a las fuentes de datos tradicionales permitiendo responder interrogantes desde diferentes perspectivas. Estas grandes fuentes de información pueden complementar los datos censales -que tiene fines de conocimiento demográfico de la población- y a los datos de las encuestas permanentes -que permiten conocer dimensiones socioeconómicas. Otra de las grandes ventajas de las fuentes masivas de información es que recopilan datos a lo largo del tiempo. Más aún, algunos datos geolocalizados tiene una resolución espacial[7] y temporal[8] muy alta, es decir, los sistemas de datos están recopilando información constantemente lo que posibilita la producción de estimaciones cercanas al tiempo real. Esta característica implica la disponibilidad de datos longitudinales para la investigación social sobre eventos inesperados, estimaciones de heterogeneidad y detección de pequeñas diferencias (Salganik, 2017). Sin embargo, estas fuentes de datos tienen limitaciones para el estudio de cambios durante periodos largos de tiempo debido a que, para medir el cambio de manera confiable, la medición del propio sistema debe ser estable. En particular, estos sistemas de datos tienen asociadas tres formas principales de cambio: variación de la población (cambio en quién los usa), variación conductual (cambio en la forma en que las personas los usan) y variación del sistema (cambio en el propio sistema). Por consiguiente, la modificación significativa observada en un patrón podría deberse a un cambio importante en el mundo empírico o a alguna de estos tres cambios.

Por su parte, las técnicas de aprendizaje automático constituyen una fuente potente para la investigación social en la predicción de cursos futuros (por ejemplo, procesos de movilización colectiva), el diagnóstico de problemáticas sociales (la incidencia de ciertos tipos de delito o la propagación de una pandemia) y el diseño e implementación de políticas públicas (Sautu, 2019). A su vez, permiten el análisis a lo largo del tiempo de grandes volúmenes de información, aunque es necesario tener en cuenta las limitaciones desarrolladas en el párrafo anterior y que las conclusiones arribadas con estas técnicas carecen del detalle propio de los

[7] El dato está localizado por sus coordenadas geográficas, no agregado espacialmente.

[8] El dato se almacena en el momento en que se genera (año, mes, día, hora, minuto y segundo), por lo cual se dispone de información siempre actual y se pueden realizar estudios longitudinales de los procesos.

estudios cualitativos. También las diversas investigaciones que aplican métodos computacionales se encuentran más cercanas al ideal de replicabilidad. Además, por su carácter exploratorio pueden representar un gran potencial en las investigaciones sociológicas pues permiten encontrar asociaciones o patrones de manera inductiva o no establecidos por la teoría. De este modo, las técnicas de aprendizaje automático posibilitan a los sociólogos generar, evaluar y priorizar sus hipótesis de trabajo (Mazzocchi, 2015; Nelson, 2017). Sin embargo, incorporar un modo de inducción en el diseño de la investigación, tal como se realiza en la Ciencia de Datos, no implica abandonar el método científico y con él la formulación de hipótesis que guían la investigación y articulan la teoría con el mundo empírico (Kitchin, 2014). Parafraseando a Sautu (2019), la búsqueda de regularidades y asociaciones es solo una parte de la investigación científica. Para comprender por qué suceden de esa manera y no de otra y adentrarnos en los procesos subyacentes de las regularidades, debemos recurrir a las teorías o a modelos explicativos de tales procesos.

A partir de los nuevos datos y las técnicas de aprendizaje automático es posible estudiar los problemas tradicionales de la Sociología desde nuevas perspectivas. Hay disponibles novedosas herramientas para el abordaje de objetos de estudios y preguntas-problemas que se relacionan a dimensiones macro sociales, a las dinámicas temporales de procesos sociales y a la integración e interacciones de los individuos, pues permiten estudiar de forma longitudinal los conjuntos societales, las características o atributos propios de estos agregados. En especial, las grandes fuentes de información representan un gran insumo para los estudios comparativos entre ciudades o países porque los mismos tienen cobertura global. Además, investigaciones recientes sobre polarización política en situaciones de interacción (Calvo y Aruguete, 2020; Kessler et al., 2021) y en redes sociales (Calvo, 2015) han empleado metodologías computacionales.

También para los antiguos interrogantes sobre la desigualdad y la pobreza se encuentran nuevas herramientas, como observar la forma de movilidad cotidiana a partir de datos móviles (conexión a torres de teléfonos celulares) o registros de computadoras (McFarland et al, 2015). Por ejemplo, Blumenstock et al. (2015) realizaron un estudio para medir la pobreza sobre la base de la intensidad del uso de teléfonos móviles en Ruanda, donde por las características geo-políticas e institucionales del país no es posible implantar una estrategia similar a la muestra EPH[9] (Sosa Escudero, 2019). Por su parte, Sarraute et al. (2014) a partir de un estudio

[9] La Encuesta Permanente de Hogares (EPH) indaga sobre las características sociodemográficas y socioeconómicas de la población de Argentina, y es realizada sistemática y permanentemente por el Instituto Nacional de Estadística y Censos (INDEC).

observacional sobre datos del uso de teléfonos móviles detectaron diferencias significativas según género y edad en los patrones de llamadas individuales y proporcionaron una nueva metodología para predecir las características demográficas (género y edad) de usuarios no etiquetados, es decir, para conocer el género y la edad de los usuarios. También se han realizado investigaciones aplicadas, con datos de geolocalización de los teléfonos celulares, con el objetivo desarrollar un mapa de prevalencia potencial de la enfermedad de Chagas (ECh) de alta desagregación espacial, para posteriormente desarrollar políticas públicas para atemperar este acuciante problema social (Vazquez Brust et al., 2018). En la misma dirección, Rosati et al. (2020) han realizado una medición empírica de la vulnerabilidad sanitaria de la población argentina con el objetivo de construir un instrumento que sirva como complemento y para contextualizar el análisis de la prevalencia de ciertas patologías en determinadas zonas. Particularmente, la generación de información con alto nivel de desagregación a partir de datos abiertos es potencialmente útil para la toma de decisiones costo-efectivas en la locación de recursos para la política pública. Otros ejemplos de investigaciones en Ciencias Sociales que emplean técnicas de aprendizaje automático son la detección de villas y asentamientos informales en el partido de La Matanza a partir de clasificar imágenes satelitales (Baylé, 2016), identificar las posiciones ideológicas de los políticos a partir de los textos de proyectos de ley a lo largo de 12 años (Gerrish y Blei, 2012), la clasificación de textos en Twitter según su sentimiento positivo o negativo (Wang et al., 2012), la imputación de datos perdidos y sin respuesta para variables de ingreso en la EPH (Rosati, 2017), el análisis sobre estereotipos de género en las revistas *Brando* y *OhLaLá* (Koslowski, 2019), el reconocimiento de cepas de cannabis analizando relatos de usuarios (Pallavicini, 2019) y la identificación de pacientes con esquizofrenia a partir del análisis del discurso (Carrillo, 2019).

Entre los usos típicos que los cientistas sociales hacen de las técnicas de aprendizaje automático se destaca medir ciertas características o cantidades latentes en un objeto de estudio determinado; en nuestro caso estudio buscamos medir tópicos prevalentes (que es una característica o una cantidad latente) en noticias digitales. Existen una amplia gama de técnicas computacionales que permiten el análisis de contenido en grandes corpus de texto. A partir de la decisión teóricamente sustentada sobre la mejor manera de abordar el conjunto de datos de manera que revele información de interés para la investigación (Kitchin, 2014), optamos por trabajar con una técnica de procesamiento de lenguaje natural: el modelado de tópicos con la implementación del método *Latent Dirichlet Allocation*. El procesamiento de lenguaje natural es una subdisciplina del campo de estudio del aprendizaje automático y la lingüística computacional que trata de emular la interpretación humana de textos y utiliza conjuntos de

algoritmos que se usan sobre datos de textos no estructurados (Rosati, 2021). Esta área de investigación –de manera general– hace referencia al proceso de extracción y el pre-procesamiento de los datos, la construcción de la matriz, la elección del modelo, las consideraciones de inferencia, las métricas, la visualización, entre otras. A lo largo de la tesina continuaremos desarrollando las anteriores elaboraciones.

El modelado de tópicos se usa cada vez más en las Ciencias Sociales para simplificar y dar sentido a grandes cuerpos de textos (Koslowski, 2019; McFarland et al., 2015). Esta técnica permite estimar los principales temas dentro de un corpus y clasificar documentos individuales en esas categorías. A grandes rasgos, es un proceso de aprendizaje automático el cual asume la existencia de un cierto proceso generador del texto que permite estimar qué composición de temas muestra cada noticia. En las técnicas de modelado de tópicos, en especial en el modelo *Latent Dirichlet Allocation* (LDA), la detección de tópicos está basada en la coocurrencia (repetición) de palabras en un mismo documento. El resultado de la modelización de tópicos son listas de palabras ponderadas, donde cada lista es un tópico y donde las palabras con mayor ponderación en una lista son más indicativas de ese tema, y representa cada documento como una distribución sobre temas, que se puede utilizar para detectar patrones temáticos en los documentos.

Es 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, lo cual conlleva mayores complejidades para estimar la evaluación del modelo. En otras palabras, 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.

A modo de reflexión parcial, esta introducción a las técnicas de aprendizaje automático, al procesamiento de lenguaje natural y al modelado de tópicos nos permite abordar las especificidades de la estrategia metodológica de la tesina, presentada en el último apartado de este capítulo.

Los estudios de contenido: una metodología para analizar los tópicos de noticias de delito y seguridad en Argentina

Tanto el modelado de tópicos como la codificación tradicional clasifican documentos de un corpus en categorías**.** Los sociólogos han intentado desarrollar técnicas de análisis de contenido que cumplan con tres requisitos para el análisis científico. El primero es que debe ser confiable, es decir, el análisis producirá los mismos resultados en todo momento. El segundo requisito es que tiene que ser intersubjetivamente válido: dos analistas informados interpretarán los resultados de manera similar. Tercero, debe ser totalmente reproducible, en otras palabras, al proporcionar una descripción detallada de los pasos del procesamiento de datos y la estrategia analítica, así como los datos en sí, cualquier investigador podrá reproducir de forma independiente el análisis completo. En Sociología 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 el contenido de noticias.

El análisis de contenido cuantitativo es utilizado habitualmente en las Ciencias Sociales argentinas, y en especial en las Ciencias de la Comunicación y la Sociología de la comunicación (Ariza y Beccaria, 2019; Koziner, 2019; Zunino y Grilli Fox, 2019)[10] para estudiar la cobertura y el tratamiento mediáticos de un asunto tanto en la prensa gráfica y online como en los noticieros televisivos (Aruguete, 2009). En general los estudios sobre análisis de contenido cuantitativo emplean grandes cantidades de tiempo en la construcción de los corpus y en la detección y análisis de los tópicos de noticias, por lo cual presentan limitaciones metodológicas relacionadas a la escalabilidad (Orozco Gómez y González, 2012).

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

[10] Las diferentes investigaciones abordadas en este apartado comparten la propuesta teórica de la Agenda Setting y del Framing.

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. La objetividad siempre se construye con ciertos matices en las Ciencias Sociales[11]. A continuación, se compara las estrategias metodológicas de diversos estudios sobre tópicos de noticias en Argentina contemporánea.

Los estudios en Argentina que emplean el análisis de contenido cuantitativo a la cobertura mediática de la inseguridad presentan algunas características metodológicas comunes. 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 gráfica seleccionan los periódicos *Clarín, La Nación* y *Página 12*. Los criterios de selección de los mismos son la relevancia en términos de circulación y la capacidad de influir en las agendas públicas y políticas (Martini, 2007). A su vez, como los corpus recolectados en la prensa gráfica y online suelen ser extensos los investigadores los reducen a una escala abordable. Para ello, emplean 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. Asimismo, 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.

Los trabajos mencionados tienen la potencialidad de emplear sistemas de categorías extensos que relacionan múltiples variables en el tratamiento mediático del delito en la prensa gráfica y online argentina. 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

[11] Marradi et al. (2018) plantean que para sortear las dificultades metodológicas de alcanzar la objetividad (cuando se parte de conceptos y categorías construidas por el investigador) es necesaria la construcción intersubjetiva de un sistema de categorial aplicable al objeto de estudio y la aplicación uniforme del esquema de codificación.

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. También se interrogan sobre la evaluación moral que ejercen los medios sobre los hechos que relatan. Por otro lado, 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.

Por su parte, el Observatorio de Medios de la UNCuyo realizó un estudio sobre los principales medios digitales del país en 2019. Los medios *Clarín, La Nación* e *Infobae* fueron seleccionados por ser lo más importantes del país en términos de preferencias masivas en el consumo de noticias digitales (Koziner, 2019). En sintonía con los estudios mencionados anteriormente, la estrategia metodológica se basa en un análisis de contenido sobre 1470 piezas periodísticas recolectadas entre abril a 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). La recolección del corpus se realizó con el método de una semana construida aleatoriamente para cada mes. A partir de los datos proporcionados por el Observatorio, Koziner (2019) indaga sobre la frecuencia y relevancia de diversos tópicos de las noticias online.

En sintonía con las investigaciones sobre medios gráficos y online, los estudios de contenido de noticieros televisivos utilizan metodologías similares de recolección de datos. Por este motivo, analizamos la metodología empleada en los informes de monitoreo llevados a cabo por la Defensoría del Público de Servicios de Comunicación Audiovisual (DPSCA) de Argentina realizados por el Programa de monitoreo de noticias de canales de aire de la Ciudad Autónoma de Buenos Aires (CABA). Desde el año 2013 se realiza este estudio con carácter sistemático y estandarizado a través del cual se recopila información sobre la tematización de los programas noticiosos de la televisión abierta de gestión pública y privada. En semejanza con el estudio de Focás y Zunino (2019b), en los informes de DPSCA se emplean análisis de tipo cuantitativo y cualitativo. La recolección de datos cuantitativos permite conocer aspectos como la cantidad y duración de las noticias presentadas por canal, programa y franja horaria

de emisión, los tópicos a través de los cuales la noticia es tematizada, la localización geográfica, los columnistas y la cantidad y tipos de actores y fuentes presentados en las noticias. El eje analítico central del monitoreo es el concepto tópico, con el cual se "busca evidenciar la producción discursiva mediática que convierte un hecho en noticia" (Ariza y Beccaria, 2019: 64). En efecto, la pieza periodística se clasifica dentro del campo temático propuesto por el noticiero, que es el resultado de un "proceso de tematización (la inclusión de la noticia dentro de un campo semántico" (Ariza y Beccaria, 2019: 70), y no por la mirada del investigador. También con la categoría tópico se hace referencia a una pluralidad de líneas de tematización posibles, con diferentes grados, en cada una de las noticias. Por lo tanto, la categoría no es excluyente con opciones de respuestas dicotómicas (DPSCA, 2017).

Por otro lado, el muestreo sistemático se compone por todos los programas televisivos emitidos durante la primera semana completa de los meses pares de cada año: febrero, abril, junio, agosto, octubre y diciembre. Además, se monitorea las cuatro franjas horarias y se incluyen todas las noticias emitidas en ella. Los intercambios entre conductores que no hacen referencia a una noticia determinada y los cortes publicitarios se excluyen de la recolección de información. Tal como se expuso en la introducción de la tesina, estos estudios conllevan grandes cantidades de tiempo para realizar la codificación de las noticias. En este sentido, en 2017 se relevaron 19.160 noticias a partir de casi 591 horas de observación de TV. Los monitores son aquellos que recolectan los datos para lo cual visionan y atribuyen valores a una cantidad de variables (o aspectos) alojadas en una matriz de datos registrada en una planilla de carga en formato Excel. A cada pieza periodística (unidad de análisis) se le atribuye un único valor en cada variable. De este modo, la codificación se realiza de forma manual. Posteriormente a esta etapa de recolección de datos se realizan diversos procesamientos estadísticos.

En resumen, las principales fortalezas del análisis de contenido cuantitativo para abordar el estudio de tópicos de noticias de delito e inseguridad 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. Por otra parte, las debilidades de este enfoque se encuentran en 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.

Notas metodológicas: proceso de recolección de datos

En el presente trabajo empleamos una estrategia metodológica que combina el análisis descriptivo y técnicas de aprendizaje automático. El objetivo general es explorar la aplicación de una técnica de procesamiento de lenguaje natural para estudiar el contenido de noticias digitales con la finalidad de sortear algunas dificultades metodológicas presentes en los estudios de contenido de cuantitativos, los cuales se desarrollaron en el apartado anterior. Por esta razón incrementamos la cantidad de medios de comunicación y la cobertura de noticias analizadas, en relación a los estudios de análisis de medios abordados en el apartado anterior. El universo de estudio son las noticias digitales publicadas en los portales de los medios *Clarín, La Nación, Infobae, Página 12, Télam, Perfil, Crónica* y *Minuto Uno* desde julio a septiembre de 2019. En la tesina trabajamos con un corpus compuesto por 52.154 noticias. La unidad de análisis es la pieza periodística. Conforme a la advertencia de Ruth Sautu sobre que en "la investigación científica es clave tener en cuenta cómo han sido producidos originalmente los datos" (Sautu, 2019: 107) en este apartado abordamos cómo han sido recogidos y registrados los mismos.

La fuente de la cual extraemos la unidad de análisis es la base de datos de *Global Database of Events, Language and Tone* (GDELT)[12]. Por lo que respecta, el proyecto 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. 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.

El corpus lo construimos a partir de un proceso de *web scraping*. Para realizar la recolección de las noticias partimos por realizar una consulta SQL en Google Big Query[13] a la base de datos de GDELT y exportamos un archivo que contiene los 52.154 links de las noticias publicadas desde julio a septiembre de 2019 en algunos de los principales diarios online de Argentina. A continuación, generamos una instancia virtual en Amazon Web Services (un proveedor de servicios de cómputo en la nube) que contiene al *scraper* y a la base de datos, donde se ejecuta y se almacena la información recolectada. Desarrollamos un *web crawler*, es decir, un programa que recorre automáticamente los links y detecta la información asociada a

[12] https://www.gdeltproject.org

[13] Una consulta Google Big Query es tipo de consulta a una base de datos empleando lenguaje similar a la sintaxis Structured Query Language (SQL).

las noticias digitales (título, fecha, medio, texto, link, entre otros) para guardarla en una base de datos.

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

Dado que la información es en su mayoría texto libre es necesario darle un formato que permita hacerla operable, realizamos un pre-procesamiento del corpus con la finalidad de producir datos limpios. 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. En este caso, se utilizó una representación vectorial llamada *Document-Term Frequency Matrix* y su lógica es simple: se busca representar cada documento como una columna de la matriz y cada *token* (en este caso, cada palabra del vocabulario) como una fila; el valor que intersecta cada fila y columna es alguna forma de conteo (crudo o ponderado mediante una métrica llamada TF-IDF que veremos enseguida) de cada *token* en cada documento, volveremos sobre esto. Primero normalizamos 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[14] 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*[15].

[14] Un ejemplo son las preposiciones que aparecen en todos los documentos y no aportan información valiosa para distinguirlos.

[15] 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.

Especificaciones sobre el pre-procesamiento

El análisis del contenido de las noticias digitales lo realizamos con una 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). Para que 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. Existe una amplia variedad de formas de realizar el proceso de vectorización, en esta tesina optamos por emplear la llamada *Document-Term Frequiency Matrix* (DTM). La confección 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 palabra o *token*; y en la intersección entre ambos está la frecuencia de aparición (cuántas veces aparece el término en cada uno de los documentos). 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 a esta forma de representar el texto suele llamarse "bolsa de palabras" (*bag of words*).

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.

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 suele normalizarse 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.

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

Por lo que respecta a la informatividad de un término a lo largo del corpus *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 es *t*, es decir, cuanto menor sea la frecuencia de *t* en el corpus *c*.

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

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.

Posteriormente, empleamos la técnica de modelado de tópicos -que trata de captar los tópicos subyacentes de un corpus de textos- para alcanzar los objetivos específicos a) relevar y graficar los tópicos de las noticias digitales y b) determinar la frecuencia de publicación de las noticias securitarias digitales. Uno de los instrumentos más difundidos en la actualidad 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[16] y, por otro, una distribución de los tópicos por documento[17]. Estos datos se encuentran organizados en dos matrices ordenadas en una estructura de filas y columnas. De este modo, abordamos el objetivo general de la tesina: explorar la aplicación de una técnica de procesamiento de lenguaje natural (modelado de tópicos) para estudiar el contenido de noticias digitales. Para alcanzar el objetivo específico c) analizar la evolución en el tiempo de la relevancia de los tópicos securitarios en las noticias digitales desde julio a septiembre de 2019

[16] Podemos caracterizar cada tópico por sus palabras más importantes.
[17] Un documento se puede caracterizar por sus tópicos más importantes.

realizamos la visualización de los resultados del modelo LDA a partir del paquete de Python *pyLDAvis* y de la librería ggplot del lenguaje R.

En este capítulo abordamos, desde una lectura metodológica, diversos estudios sobre tópicos de noticias de delito y seguridad en Argentina contemporánea. En el próximo capítulo presentamos los resultados del modelo LDA en diálogo con los hallazgos empíricos de dichas investigaciones. A su vez, en el capítulo 3 recuperamos el análisis de las principales fortalezas y problemas de ambos enfoques metodológicos desarrollados en este capítulo y los articulamos con los hallazgos empíricos, con la finalidad de especificar los aportes y limitaciones de emplear técnicas de procesamiento de lenguaje natural y *web scraping* en el presente estudio sobre tópicos de noticias securitarias.

Capítulo 2

El caso de la seguridad: un tópico estable y relevante en la agenda mediática digital

El propósito de este capítulo es presentar los resultados de la modelización de tópicos y abordar las preguntas centrales de la tesina: ¿cuál es la prevalencia de las noticias sobre delito? y ¿qué relevancia tienen en comparación con otros temas de la agenda mediática entre julio y septiembre de 2019? Decidimos emplear una estrategia metodológica centrada en técnicas de aprendizaje automático de procesamiento de lenguaje natural (modelado de tópicos) y *web scraping* con el objetivo de probar una nueva técnica metodológica en el análisis de contenido de noticias. Retomando la hipótesis de trabajo argumentamos que 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.

A su vez, es posible aplicar técnicas de modelado de tópicos al estudio de diversos temas pero en este trabajo optamos por realizar un recorte temático vinculado al delito y la seguridad, pues "la batalla contra el crimen" se consolida desde 2008 como elemento privilegiado en las agendas electorales y como una temática recurrente de las disputas políticas (Calzado et al., 2014). Las elecciones son un mecanismo para elegir gobernantes, pero también una vía para canalizar y moderar las angustias sociales (Natanson, 2019). En la misma dirección, Calzado et al. (2019) definen a los contextos electorales como espacios de conflicto donde se plasman los imaginarios políticos y las preocupaciones sociales contemporáneas. En los procesos electorales un tema relevante de la agenda política y mediática es la inseguridad, que además se ubica entre las principales preocupaciones ciudadanas. Desde principios de siglo se observa un incremento sostenido en la representación mediática del delito en contextos electorales (Calzado, 2013; Martini, 2007). En los momentos electorales la problemática securitaria se expresa en forma de demandas hacia los funcionarios públicos y candidatos políticos, quienes suelen ubicarla entre sus principales temas de campaña (Zunino y Focás 2019a). Por lo anterior, escogimos trabajar con el recorte temporal de julio a septiembre de 2019 durante el cual se desarrollaron las elecciones Primarias Abiertas Simultáneas y Obligatorias (PASO).

Argentina ingresó al año electoral sumida en una profunda crisis económica (fuerte endeudamiento, devaluación del peso, altos niveles de inflación) y social (altas tasas de desempleo y el consecuente crecimiento de los índices de pobreza), y con el Fondo Monetario Internacional monitoreando las cuentas públicas. En las PASO los espacios políticos dirimieron sus candidaturas a cargos nacionales y provinciales del Poder Ejecutivo y Legislativo de cara

a las elecciones generales de octubre. El escenario político nacional tuvo dos principales protagonistas: Juntos por el Cambio y la colación Frente de Todos. Como explicitamos en la introducción, Cambiemos es una alianza de tendencia liberal conservadora que surgió en 2005 bajo el liderazgo del empresario local Mauricio Macri. El Frente de todos es la colación gobernante en Argentina (2019 – 2023) en la que convergen cuatro grandes sectores políticos: el Partido Justicialista, los sectores peronistas y no peronistas del kirchnerismo liderado por la expresidenta Cristina Fernández de Kirchner, la mayoría de los gobernadores peronistas y el Frente Renovador liderado por Sergio Massa. El domingo 11 de agosto de 2019 el Frente de Todos, encabezado por Alberto Fernández y Cristina Fernández de Kirchner, se impuso a nivel nacional con el 49,49% de los votos sobre el expresidente Mauricio Macri y Miguel Ángel Pichetto que obtuvo el 32,93%.

Ahora bien, ¿qué entendemos por delito? En esta tesina optamos por una definición amplia del delito pensándolo como un producto socio-histórico, constructo de la sociedad. En el cuarto tópico están agrupadas noticias sobre seguridad, inseguridad, casos de corrupción en el ámbito público, entre otros. Entre los relatos securitarios, que son muchos y diversos, las noticias de los medios resultan los de mayor alcance masivo (Martini, 2019). Las noticias digitales sobre delito y seguridad se vuelven relevantes, puesto que operan como caja de resonancia y amplificación del fenómeno de la "inseguridad" (Arce y Zapata, 2019), sobre todo porque –como señala Koziner et al. (2018)- el nivel de consumo de los medios digitales alcanza a la mitad de la población. La noción de inseguridad se define como una sensación de indefensión de los individuos contra una amenaza aleatoria, que opera con autonomía relativa respecto de los hechos delictivos (Kessler, 2009). Según un estudio diacrónico de la consultora Latinobarómetro, la "inseguridad" aparece como principal preocupación ciudadana configurándose como el principal problema de importancia en América Latina que asciende, con vaivenes, desde 2004 en adelante (Focás y Kessler, 2015).

Por lo que respecta a las investigaciones en Argentina que indagan sobre el análisis de medios, en general se abordan desde de la teoría 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[18].

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

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). Para estudiar la agenda mediática de la cuestión securitaria en esta tesina se analiza la composición de los tópicos de las noticias digitales.

Este panorama nos habilita a preguntarnos cuál es la frecuencia de publicación de las noticias securitarias en la prensa online y cómo es su evolución temporal. En este capítulo nos proponemos responder estos interrogantes y alcanzar los tres objetivos específicos de la tesina: a) identificar los tópicos de las noticias digitales, b) determinar la frecuencia de publicación de las noticias online de delito y seguridad y c) analizar su evolución en el tiempo. También sometemos a prueba la segunda hipótesis preliminar: la prevalencia de noticias sobre delito y seguridad aumenta durante el mes de las elecciones Primarias Abiertas Simultáneas y Obligatorias.

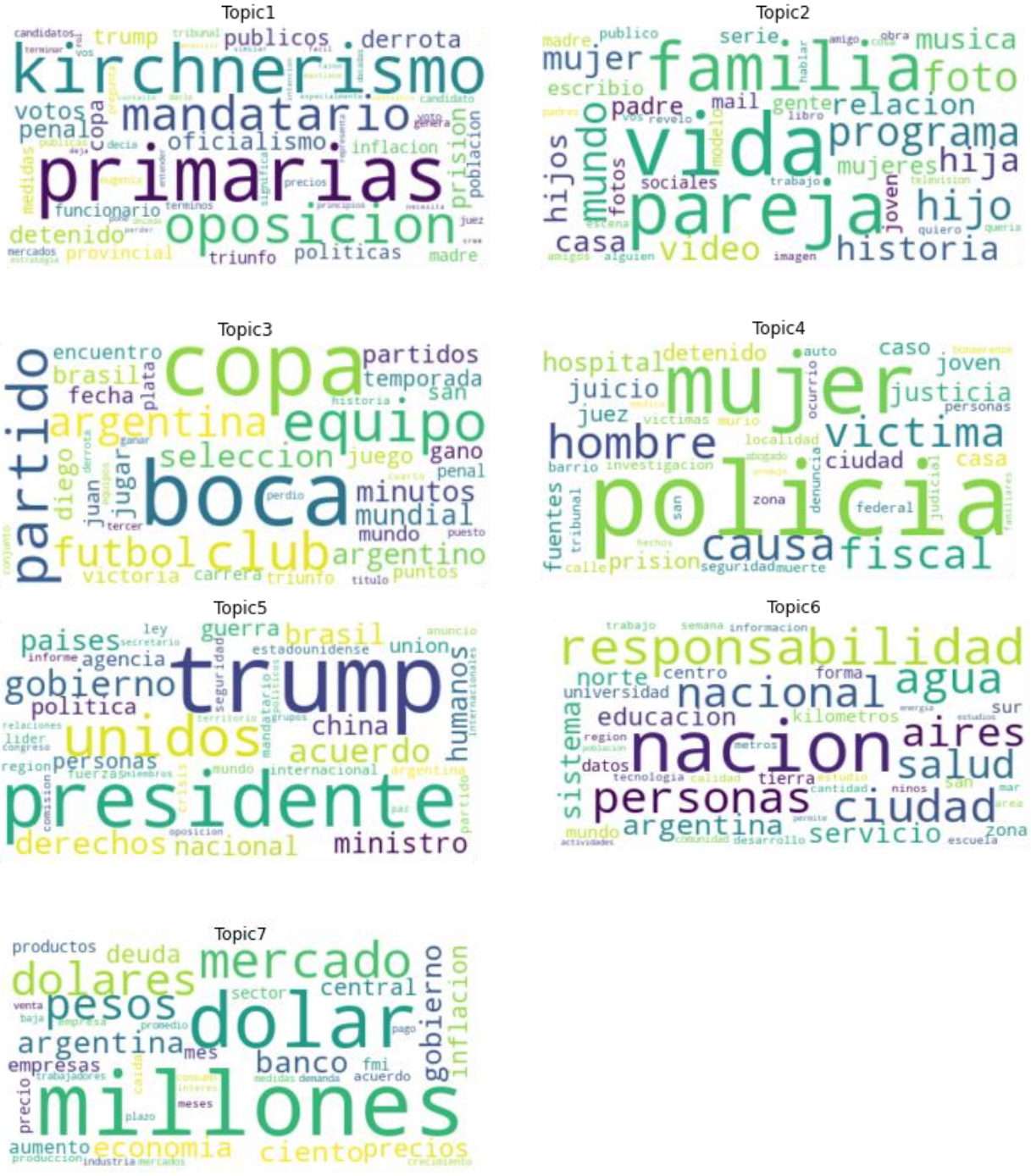
Composición de los tópicos de las noticias digitales

Utilizando el modelo LDA buscamos detectar los tópicos más relevantes en el *dataset* compuesto por 52.154 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. Tal como se desprende del capítulo anterior, una de las principales decisiones que se deben tomar al emplear este modelo es determinar la cantidad de tópicos que se busca detectar. En nuestro caso de estudio es importante observar que independientemente de la cantidad de tópicos seleccionada, existen algunos temas que se mantienen vigentes en la agenda: elecciones, seguridad, economía, espectáculos y deportes. Al emplear estas técnicas de aprendizaje automático es necesario generar tópicos que tengan sentido semántico, es decir, que sean interpretables. En nuestro caso seleccionamos el modelo que muestra un total de 7 tópicos.

Recordamos que el tópico se define como una distribución de probabilidad sobre las palabras del vocabulario, por lo cual una misma palabra tiene una determinada probabilidad de pertenencia a todos los tópicos. Por ende, la diferencia relativa entre tópicos es que ciertos

términos pertenecen a ciertos tópicos con mayor probabilidad que otros, por ejemplo, la palabra "mujer" pertenece con mayor probabilidad al tópico "seguridad" y en menor medida al tópico "espectáculos". A su vez, el modelo LDA tiene como supuesto que el conjunto de palabras que pertenecen a un mismo tópico tienen una probabilidad de coocurrencia (repetición) a lo largo del corpus (Rosati, 2021). A continuación, observamos las 30 primeras palabras de cada uno de los tópicos, es decir, las palabras con mayor probabilidad de pertenencia. En el gráfico 1 el tamaño de las palabras es proporcional a dicha probabilidad.

Gráfico 1. Composición de las 30 palabras de cada tópico



El primer tópico detectado tiene palabras asociadas a las elecciones PASO 2019, como ser “primarias”, “kirchnerismo”, “oposición”, “mandatario”, “oficialismo”, “votos”, entre otras. A modo de ejemplo, mostramos una noticia agrupada en el presente tema.

El segundo tópico contiene términos vinculados al espectáculo: “vida”, “pareja”, “familia”, “programa”, “hijo”, “foto” y “mundo”. Una pieza periodística que pertenece al mismo es:

El tópico 3 capta noticias sobre deportes pues las palabras con mayor probabilidad de coocurrencia son “boca”, “copa”, “equipo”, “partido”, “club”, “argentina” y “futbol”, “selección”.

El tópico 4 contiene los términos “policía”, “mujer”, “hombre”, “causa”, “victima”, “fiscal”, “hospital”, “justicia”, “juicio”, “detenido”, “joven”, “prisión”, “ciudad”, “juez”, entre otras. Es decir, nos habla de noticias sobre delito y seguridad. A modo de ejemplo, en este tópico aparece la siguiente noticia.


MENÚ Martes 19 de enero crónica Firme junto al pueblo

POLICIALES 30-08-2019 08:24

Impactante triple robo en simultáneo en el subte

El hecho ocurrió en la estación San Juan de la línea C.Las cámaras de seguridad muestran cómo tres sujetos esperaron a que la formación cerrara sus puertas para meter el cuerpo por la ventana y arrebatar celulares. *Mirá el increíble video.*

El quinto tópico muestra noticias vinculadas a la sección Política exterior y está compuesto por términos como “trump”, “presidente”, “unidos”, “gobierno”, “acuerdo”, “derechos”, “países”, “brasil”, “ministro”, “humanos” y “china”.


Política Economía Sociedad Policiales Provincias Deportes Internacional Espectáculos Cultura DDHH Portugués Ambiente Entrevistas

INTERNACIONAL

29/09/2019 RUSIA

La oposición rusa volvió a protestar en las calles de Moscú

La manifestación fue organizada por el Partido Libertario, tres semanas después de haber obtenido el mejor resultado de su historia en las elecciones municipales de Moscú y San Petersburgo.

El sexto tópico se vincula a noticias sobre Obra pública/Interés General y las palabras con mayor probabilidad de pertenencia son “nación”, “responsabilidad”, “personas”, “ciudad”, “agua”, “nacional”, “aires”, “salud”, “argentina”, “sistema”, “servicio” y “educación”.


Página12

Defendé

19 de enero de 2021 | EDICIÓN IMPRESA | PDF

AM 750 INICIAR SE

SECCIONES Y SUPLEMENTOS El país | Economía | Sociedad | Cultura y Espectáculos | Deportes | Ciencia | El mundo Hoy: Verano12 | NO

SOCIEDAD CIENCIA Y TECNOLOGÍA ALBERTO KORNBLIHTT

13 de septiembre de 2019

Charla con lectores de Página/12

Kornblihtt: "Para este gobierno la ciencia y la tecnología no son importantes"

El reconocido científico y director del Conicet estuvo en la Sala Caras y Caretas, donde opinó sobre política científica, la actualidad del organismo y el futuro que se viene. Dijo que al gobierno de Macri "le molesta tanto el pensamiento crítico que intenta reemplazarlo con falso optimismo y entusiasmo".

Por último, se observan el séptimo tópico logra evidenciar noticias sobre economía y menciona términos como “millones”, “dólar”, “mercado”, “dólares”, “pesos”, “argentina”, “banco”, “gobierno”, “economía”, “precios”, “deuda”, “inflación”.

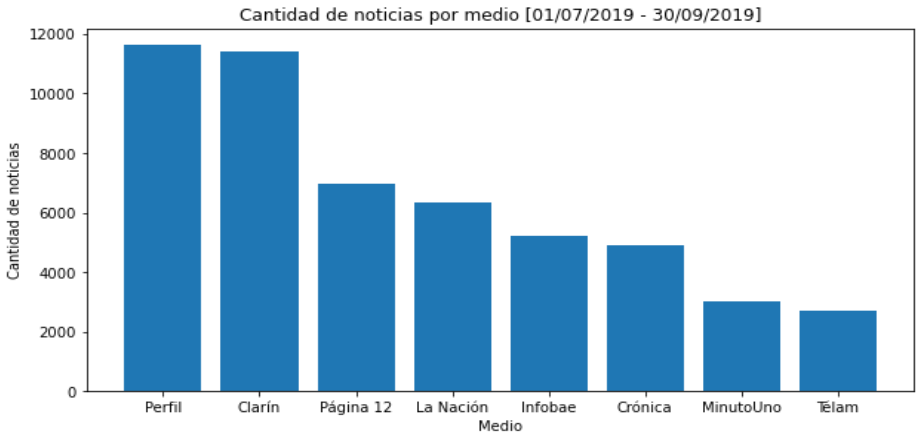
De esta forma, el modelado de tópicos permite agrupar las noticias según su tópico prevalente de forma automática sin recurrir a una codificación a priori del cientista social. Estas técnicas cambian la interpretación sustantiva del investigador a una etapa posterior en el proceso analítico (Mützel, 2015). En nuestro caso de estudio, la interpretación del investigador se pone de manifiesto en la etiqueta de tópicos.

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

Una vez identificados e interpretados los tópicos relevantes y previo a establecer la frecuencia de publicación que adquieren en la prensa online del tópico relativo a la cuestión securitaria es importante observar la cantidad de noticias publicadas según el medio de comunicación. El gráfico 2, presentado a continuación, permite comparar la cantidad de noticias online publicadas en los portales desagregadas por medio digital. Durante los meses de julio a septiembre de 2019 se observa que los medios de comunicación que publicaron una mayor cantidad de noticias online son *Perfil* (11.611) y *Clarín* (11.399). En el extremo opuesto se encuentran los medios *MinutoUno* (3.010) y *Telam* (2.723). Por su parte, el diario *Página 12* también ostenta valores altos de producción de noticias online con una cantidad de 6.953,

seguido por *La Nación* con 6.338 noticias. A su vez, *Infobae* (5.234) y *Crónica* (4.886) muestran valores intermedios de publicación de noticias.

Gráfico 2 Argentina: cantidad de noticias según medio de comunicación online, julio a septiembre 2019



Fuente: elaboración propia

Resulta de interés cotejar estos datos con aquellos que se vinculan al consumo de medios digitales. Siguiendo las elaboraciones, el medio de comunicación más consumido en internet es *Clarín* que es también uno de los mayores productores de noticias online. El segundo medio más consumido es *La Nación* y el tercero, *Infobae*. Ambos diarios online ostentan valores intermedios de producción de noticias digitales (Becerra, 2019).

Los resultados de la modelización de tópicos permiten establecer que las elecciones (21%), los espectáculos (17%), los deportes (16,3%), la seguridad (15,9%), la política exterior (11,7%), la obra pública (9,7%) y la economía (8,4%) fueron prioridad en las agendas mediáticas digitales entre julio y septiembre de 2019. En sintonía con nuestros hallazgos, el Observatorio de Medios de la UNCuyo muestra al caso de las elecciones nacionales como el principal tópico de la agenda mediática online desde abril a septiembre de 2019. Si comparamos los resultados arribados en esta tesina con los datos provistos por el Observatorio observamos diferencias en el orden de relevancia de los tópicos. El Observatorio muestra que entre abril y septiembre de 2019 las elecciones nacionales (27,7%), la economía (22,6%), la seguridad (11,6%), los deportes (9,4%), la política (8,6%), la corrupción (6,3%) y los espectáculos (2,8%) ocuparon un lugar relevante en la agenda de los portales *Clarín*, *La Nación* e *Infobae*.

En base a lo relatado abordamos el segundo objetivo específico: determinar la frecuencia de publicación de las noticias online de delito y seguridad. En el contexto de las elecciones PASO de 2019 se observa que el caso de la seguridad es el cuarto tópico más prevalente en las noticias digitales publicadas en los portales de los medios digitales *Clarín, La Nación, Infobae, Página 12, Télam, Perfil, Crónica* y *Minuto Uno*. Las noticias securitarias representan el 16% del corpus compuesto por 52.154 noticias, en términos absolutos son 8272 piezas periodísticas. La frecuencia de publicación de las noticias securitarias en la prensa digital fue de casi 2 de cada 10 piezas periodísticas. Según estudios recientes "el caso de la inseguridad se trata de un tópico familiarizado para los medios de comunicación, que mantiene una omnipresencia en el espacio mediático, tanto en el tiempo como el espacio" (Focás, 2018: 13). En las últimas décadas la agenda noticiosa estuvo protagonizada por el tópico del delito y la violencia urbana, incluso en ciertos momentos más que la información general y los deportes (Calzado et al., 2019).

La evolución temporal de la relevancia del tópico securitario

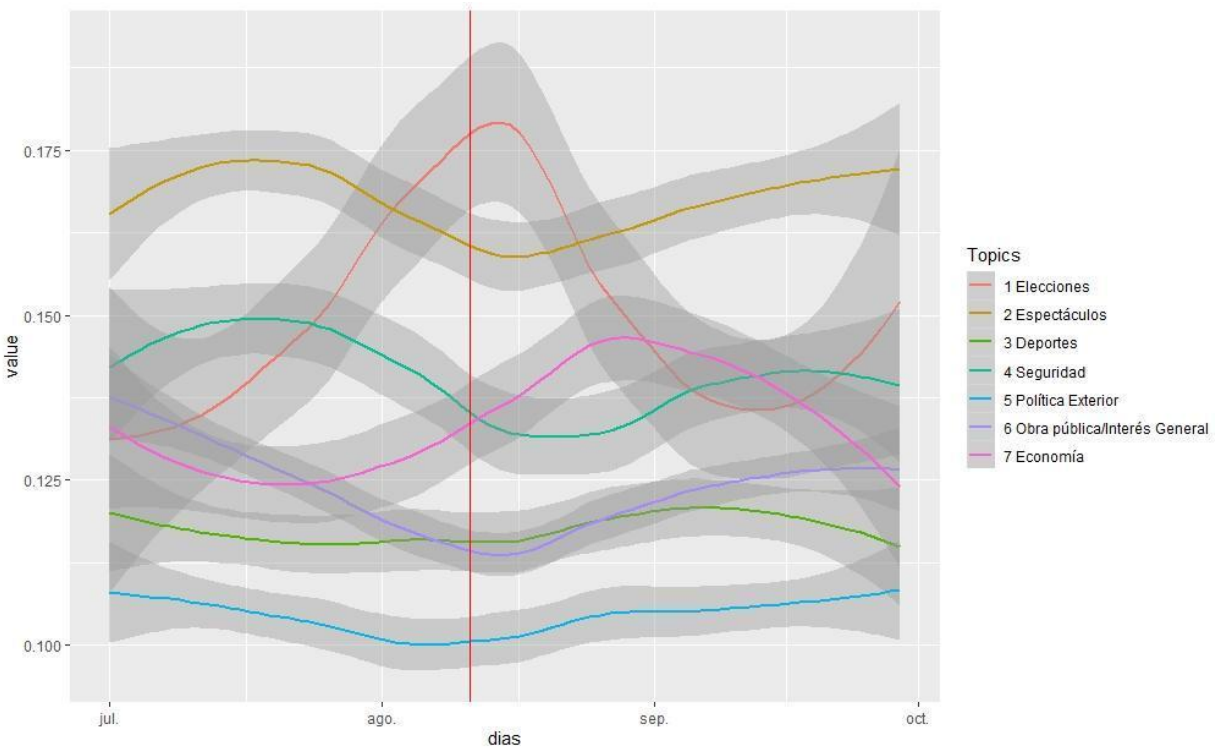
Ahora bien, el tercer objetivo específico apunta a conocer la evolución de la relevancia del tópico securitario en las noticias digitales desde julio a septiembre de 2019. Para ello, calculamos la evolución de la media de tópicos por día y aplicamos un suavizado GAM (*generalized additive models*)[19]. Es importante recordar que el modelo LDA en su versión básica asume supuestos fuertes para el análisis temporal: los tópicos preexisten a los textos y son constantes en el tiempo[20]. De manera general, a lo largo de los tres meses analizados la frecuencia de la cobertura mediática securitaria se mantiene estable como un tema relevante de la agenda mediática online, aunque es posible visualizar oscilaciones leves al comparar cada mes entre sí. Desde los primeros días de julio el tópico securitario gana relevancia y comienza a descender levemente hacia fin de mes. En los días posteriores a las elecciones Primarias Abiertas Simultáneas y Obligatorias la relevancia de las noticias securitarias alcanzan su nivel más bajo, pero vuelve a incrementarse hacia el mes de septiembre aunque sin llegar a los valores previos a la elección. Este dato coincide con la tendencia observada en las elecciones

[19] 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).

[20] 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).

Generales de 2015, donde no se incrementó la frecuencia de piezas de este tipo (Zunino y Focás, 2019b).

Gráfico 3 Argentina: evolución de la media de los tópicos de noticias por día, julio a septiembre 2019 (suavizado GAM)

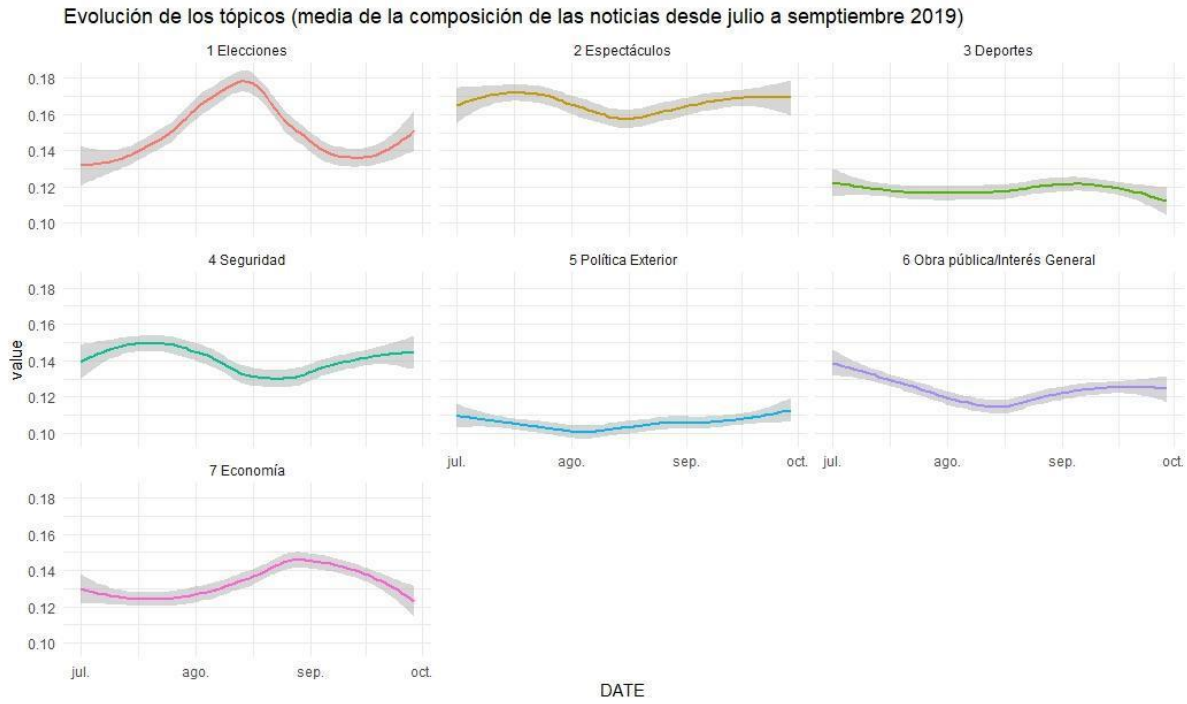


Fuente: elaboración propia

El gráfico 3 permite observar cómo se modifica la relevancia del tópico securitario comparado con otros temas presentes en los medios digitales *Clarín, La Nación, Infobae, Página 12, Télam, Perfil, Crónica* y *Minuto Uno* desde julio a septiembre de 2019. La línea vertical indica el día de las elecciones Primarias Abiertas Simultáneas y Obligatorias, el domingo 11 de agosto de 2019. En primer lugar, el tópico 1 sobre elecciones es el que presenta 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. También resultan interesantes las oscilaciones que presenta el tópico 7 sobre economía. 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. En cambio, los tópicos 3 (deportes) y 5 (política exterior) se mantienen estables durante todo el periodo analizado.

El gráfico que mostramos a continuación permite observar la evolución de cada uno de los 7 tópicos por separado. En comparación con el resto de los tópicos detectados en el corpus el caso de seguridad muestra variaciones en su relevancia similares al tópico 2, aunque el tópico sobre espectáculos ostenta una media de la composición de noticias más alta en comparación al cuarto tópico.

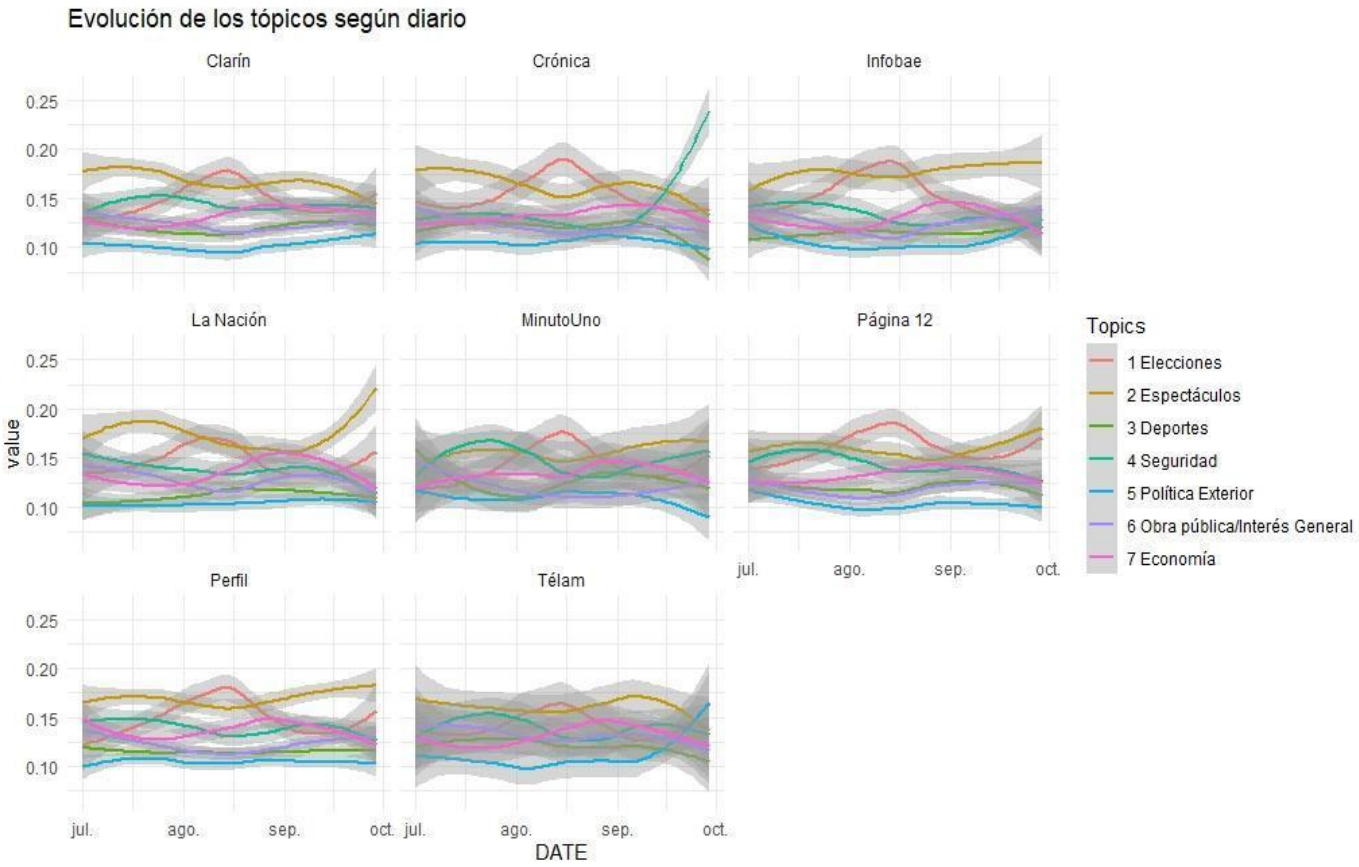
Gráfico 4 Argentina: evolución de la media de cada tópico de noticias por día, julio a septiembre 2019 (suavizado GAM)



Fuente: elaboración propia

Finalmente, al analizar la evolución temporal de los tópicos diferenciando según el medio online se evidencia al caso de la seguridad como un tema estable y relevante en la agenda mediática digital de los diarios los medios digitales *Clarín*, *La Nación*, *Infobae*, *Página 12*, *Télam*, *Perfil*, *Crónica* y *Minuto Uno*. Sin embargo, el dato más elocuente que muestra el análisis del corpus de esta tesina se relaciona con la evolución temporal de tópico seguridad en el diario *Crónica* que crece exponencialmente la relevancia durante septiembre de 2019.

Gráfico 5 Argentina: evolución de la media de los tópicos de noticias según diarios online por día, julio a septiembre 2019 (suavizado GAM)



Fuente: elaboración propia

También, el caso de las elecciones aparece como un tópico estable y relevante de la agenda mediática digital en el contexto electoral de las Primarias Abiertas Simultáneas y Obligatorias. Resulta de interés evidenciar que en todos los medios digitales el tema de las elecciones alcanza su valor de relevancia más alto en los días posteriores al sufragio del 11 de agosto de 2019. En sintonía, el tópico sobre economía aumenta la relevancia luego de las PASO 2019. Estos hallazgos permiten arriesgar que, en general, los medios digitales comparten una misma agenda mediática, aunque con algunas diferencias leves.

A partir de la visualización de los resultados de la modelización de tópicos sometemos a prueba la segunda hipótesis preliminar: las técnicas computacionales empleadas en esta tesina permiten escalar el análisis 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. Es posible afirmar que la hipótesis se corrobora de forma parcial, pues, por un lado, aplicar técnicas de procesamiento de lenguaje natural y *web scraping* nos permite incrementar la cobertura de noticias. En este

sentido, el tamaño total del corpus construido en esta tesina es sensiblemente mayor que los estudios reseñados en el primer capítulo. Por otro lado, no observamos que aumente la prevalencia de las noticias securitarias durante agosto de 2019. Más bien, la relevancia del el tópico seguridad desciende levemente durante el mes del escrutinio y vuelve a incrementarse en septiembre.

A lo largo de este capítulo presentamos los hallazgos empíricos alcanzados a partir de la técnica de modelado de tópico y abordamos los objetivos y preguntas específicas junto a la segunda hipótesis preliminar de la tesina. En base a lo anterior y en diálogo con el panorama teórico-metodológico presentado en la primera parte, en el próximo capítulo analizamos los aportes metodológicos de la implementación de técnicas de aprendizaje automático de procesamiento de lenguaje natural en el estudio del caso empírico y en las posibilidades que se abren para los estudios de Agenda Setting.

Capítulo 3

Abriendo la caja de herramientas metodológicas

En esta tesina nos propusimos explorar la aplicación de una técnica de aprendizaje automático de procesamiento de lenguaje natural, el modelado de tópicos, para estudiar el contenido de noticias digitales. Para ello, tomamos como soporte empírico las noticias digitales publicadas en ocho medios de comunicación desde julio a septiembre de 2019. Así, en el capítulo anterior mostramos los principales tópicos de las noticias digitales y analizamos su evolución temporal durante el contexto electoral. En este capítulo trabajaremos sobre algunas de las posibilidades que se abren para los análisis de contenidos mediáticos con este tipo de técnicas.

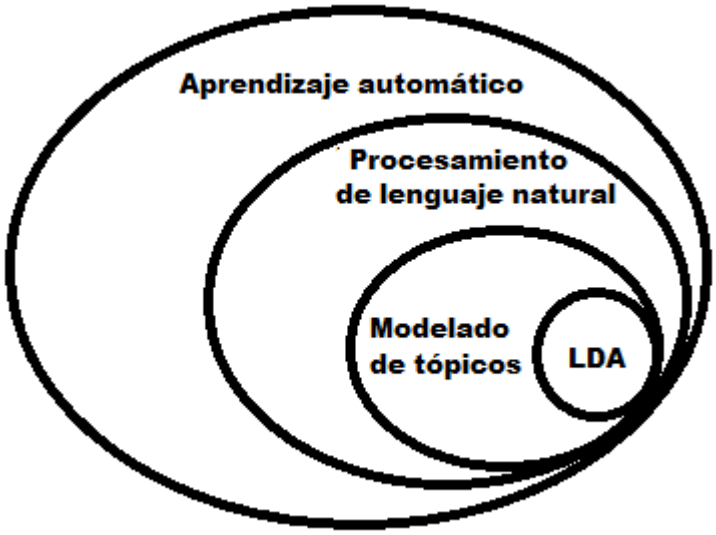
El capítulo se organiza de la siguiente manera. En el primer apartado realizamos una aproximación metodológica al análisis textual computacional y presentamos algunas de las potencialidades de incluir una técnica especifica de detección de tópicos (*Latent Dirichlet Allocation*) al trabajo cotidiano de las Ciencias Sociales, a partir de su aplicación a un caso de estudio concreto: los tópicos de las noticias digitales en Argentina desde julio a septiembre de 2019. El empleo de técnicas de procesamiento de lenguaje natural busca morigerar algunas limitaciones (replicabilidad y estabilidad) presentes en el análisis textual, una herramienta ampliamente utilizada en las Ciencias Sociales. En el segundo apartado establecemos algunas de las potencialidades de incluir técnicas de análisis computacional de textos al repertorio metodológico de las investigaciones empíricas de Agenda Setting.

Aportes del modelado de tópicos aplicado al estudio del caso empírico

De manera general, el procesamiento de lenguaje natural es un área de investigación del campo de estudio del aprendizaje automático que ofrece la posibilidad de explorar grandes corpus de textos, organizarlo en forma de datos, establecer asociaciones y extraer información útil (Botta-Ferret y Cabrera-Gato, 2007). El objetivo de la técnica de modelado de tópicos y, en particular, del algoritmo LDA es descubrir de forma automática los tópicos a los que alude un determinado conjunto de documentos, en nuestro caso de estudio: las noticias digitales. Las técnicas de procesamiento de lenguaje natural abordadas en esta tesis –y muchas otras que no mencionamos- abren la posibilidad de escalar el trabajo 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. Más aún, el análisis textual computacional permite reducir los tiempos de detección y análisis de los tópicos en comparación a los estudios cuantitativos de análisis de medios abordados en el primer capítulo. El gráfico 6, que presentamos a continuación, permite observar cómo se organizan las técnicas de aprendizaje automático abordadas en esta tesina.

Gráfico 6 Técnicas de aprendizaje automático



En la búsqueda de morigerar alguna de las limitaciones presentes en el análisis textual de los estudios de contenido cuantitativos, desarrollados en el capítulo 1, 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. En esta tesina trabajamos sobre un corpus compuesto por 52.154 noticias digitales con la finalidad de aumentar la escalabilidad. Para armar el corpus empleamos la técnica de *web scraping* y 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. Previo a aplicar la técnica de *web scraping* es necesario recolectar los links de las noticias digitales para lo cual hay que definir cuál es el universo de estudios, es decir, cuál es el recorte temporal y qué diarios online utilizamos como fuente. Como mencionamos anteriormente, la recolección de los links de las noticias se realiza a partir de una consulta SQL en Google Big Query a la base de datos de GDELT. En nuestro caso, exportamos un archivo que contiene los links de las noticias publicadas desde julio a septiembre de 2019 en los portales de los principales medios digitales de Argentina: *Clarín, La Nación, Infobae, Página 12, Télam, Perfil, Crónica* y *Minuto Uno*.

Siguiendo las elaboraciones, la técnica de *web scraping* posibilita aumentar la escalabilidad en el análisis textual porque permite confeccionar corpus de manera automática a partir de un programa (*web crawler*) que recorre los links de las noticias. En nuestro caso descargamos y formateamos el título, la fecha, el medio y el texto de las noticias en una base de datos. De esta manera es posible aumentar la cobertura de noticias, es decir, la técnica de *web scraping* permite escalar y abordar la casi totalidad las noticias digitales publicadas en los portales los principales diarios online de Argentina.

Recuperando el panorama teórico-metodológico presentado en la primera parte, los resultados del modelo LDA nos muestran la presencia de patrones, es decir, tópicos relevantes en nuestro corpus. Para interpretar y comprender los procesos sociales que subyacen a dichas secuencias debemos recurrir a las teorías (Sautu, 2019; Uman, 2018). De este modo, utilizando la técnica de modelado de tópicos, en particular el modelo LDA, y en diálogo con la literatura especializada sobre análisis de medios y agenda mediática securitaria se respondió la pregunta de investigación ¿cuál es la prevalencia de las noticias sobre delito y qué relevancia tienen en comparación con otros temas de la agenda mediática entre julio y septiembre de 2019?, se abordaron los tres objetivos específicos y se sometió a prueba la segunda hipótesis (las técnicas de procesamiento de lenguaje natural y *web scraping* permiten escalar el análisis 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).

Tal como se desprende del segundo capítulo, los resultados de la modelización de tópicos muestran que las elecciones, los espectáculos, el deporte, la seguridad, la política exterior, la obra pública y la economía fueron prioridad en las agendas mediáticas digitales de los medios *Clarín*, *La Nación*, *Infobae*, *Página 12*, *Télam*, *Perfil*, *Crónica* y *Minuto Uno* desde julio a septiembre de 2019. En particular, la cuestión securitaria tuvo una frecuencia de publicación de casi 2 de cada 10 piezas periodísticas. En el contexto electoral de las PASO 2019 la evolución temporal del tópico seguridad se mantiene estable, aunque al comparar los tres meses entre sí observamos algunas oscilaciones leves en su nivel de relevancia. Un mes previo a la elección el caso de la seguridad alcanzó su punto de mayor relevancia, disminuyó levemente en agosto y volvió a incrementarse en septiembre. En oposición a la segunda hipótesis preliminar, la prevalencia de noticias sobre delito y seguridad disminuye en el mes de las elecciones Primarias Abiertas Simultáneas y Obligatorias. Este hallazgo coincide con la tendencia observada en las elecciones generales de 2015 cuando la frecuencia de publicación de las noticias securitarias no aumentó (Zunino y Focás, 2019b). A partir de los resultados de

la modelización de tópicos, concluimos que la seguridad es un tópico estable y relevante en la agenda mediática digital en el contexto de las PASO 2019.

En base a lo relatado argumentamos que la utilización de técnicas de aprendizaje automático de procesamiento de lenguaje natural y *web scraping*, aunque son poco exploradas por las Ciencias Sociales, pueden ser una herramienta que permita sortear algunas de las dificultades metodológicas presentes en los estudios de análisis de contenido de noticias: la imposibilidad de abordar (casi) la totalidad de noticias y los grandes tiempos que conlleva la detección y el análisis de tópicos (Orozco Gómez y González, 2012). Los resultados del ejercicio propuesto en esta tesina, desarrollados en el capítulo anterior, tienen como objetivo mostrar un caso de usos de la herramienta metodológica más abordar el campo problemático de la agenda mediática securitaria en la totalidad de sus determinaciones. A partir del análisis de los resultados del modelo LDA es posible corroborar la primera hipótesis de investigación: aplicar técnicas de aprendizaje automático de procesamiento de lenguaje natural permite aumentar la cobertura de noticias, sistematizar las diversas etapas de codificación de un texto y reducir los tiempos de detección y análisis de tópicos relevantes.

Ahora bien, a la par de estas ventajas existen algunas limitaciones a tener en cuenta. Por un lado, hay una fuente de sesgos en la recolección de las piezas periodísticas que conforman el corpus –pese a su amplitud- ha sido confeccionado a partir del grafo GDELT, que abordamos en el primer capítulo. En este sentido, no se puede afirmar que se analizaron todas las noticias publicadas en los 8 diarios analizados. Por otro lado, 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.

El procesamiento de lenguaje natural y los estudios de contenido de noticias

En esta sección buscamos mostrar las potencialidades que este tipo de técnicas computacionales tienen para el análisis de contenidos, en particular para el análisis de medios. Las técnicas de aprendizaje automático que destacan características del texto son herramientas que permiten analizar corpus amplios y de forma sistemática la evolución temporal de los tópicos. En nuestro caso abordamos la prevalencia del tópico securitario en noticias digitales y su evolución en un contexto electoral pero también puede ser utilizado para otras aplicaciones: servir para otros tópicos, como ser la cobertura mediática digital de la evolución del COVID-19 (Barriola y Gncchi, 2020), el análisis sobre estereotipos de género en las

revistas *Brando* y *OhLaLá* (Koslowski, 2019). También puede aplicarse fuera del estudio de la comunicación mediática los temas en las letras de tango (Rosati, 2021), el reconocimiento de cepas de cannabis analizando relatos de usuarios (Pallavicini, 2019) y la identificación de pacientes con esquizofrenia a partir del análisis del discurso (Carrillo, 2019).

La implementación del análisis automático de textos tiene potencialidades para el trabajo cotidiano de las Ciencias Sociales argentinas que utilizan la técnica de análisis de contenido cuantitativo. Emplear en la investigación social técnicas de procesamiento de lenguaje natural como las abordadas en este trabajo tiene tres potencialidades. En primer lugar, permiten alcanzar una sistematización de las diversas etapas de pre-procesamiento de un texto. En segundo lugar, abre la posibilidad de aplicar métodos cuantitativos de análisis textual (por ejemplo, clasificación de textos y detección de temas y tópicos). En tercer lugar, habilita la posibilidad de analizar de forma automática corpus notablemente grandes, es decir, permite escalar el trabajo de forma eficiente. En el campo de la Sociología de la Comunicación y análisis de medios la aplicación de algunas de las técnicas de análisis textual computacional que abordamos en esta tesina (y muchas otras que no son mencionadas) pueden ser útiles para morigerar algunas dificultades metodológicas del análisis textual. Nos parece que está técnica puede aportar a las investigaciones de Agenda Setting que tienen pretensión de análisis de contenido a gran escala. En particular, las técnicas de modelado de tópicos permiten sortear 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).

Los estudios en Argentina que emplean la técnica de análisis de contenido cuantitativo a la cobertura mediática securitaria recolectan el corpus de noticias de manera manual e incluyen las piezas periodísticas siguiendo criterios específicos definidos por los investigadores previamente. En este sentido, las técnicas procesamiento de lenguaje natural y *web scraping* pueden ser de utilidad dado que permiten realizar una sistematización y, eventualmente, lograr un cierto grado de automatización de los diversos pasos de pre-procesamiento de un texto (Rosati, 2021). En concreto, todo el flujo de trabajo implementado en esta tesina y las operaciones contenidas en el mismo (que se describe en el primer capítulo) también son replicables, es decir, que si se emplean las mismas técnicas y se le atribuyen los mismos valores a los hiperparámetros del modelo estos mismos resultados pueden ser alcanzado por otros investigadores. El código y *scripts* desarrollados para la tesina permiten replicar el trabajo. No obstante, es menester aclarar que el flujo de trabajo implantado en este trabajo es uno de los posibles, pero no el único ni necesariamente el "mejor" en términos absolutos. Siguiendo las

elaboraciones, el flujo de trabajo debe ser revisado para cada problema de investigación en particular (Grimmer & Stewart, 2013 en Rosati, 2021).

De manera similar a la codificación tradicional, el modelado de tópico clasifica documentos de un corpus en categorías. Si bien este paso todavía implica una serie de juicios subjetivos (cantidad de tópicos, interpretación de los tópicos estimados, etc.) estas decisiones se escriben directamente en el proceso de codificación asistido por computadora, por lo que, a diferencia del texto codificado manualmente, la salida del texto codificado computacionalmente es total e inmediatamente reproducible. En comparación, el análisis de contenido no es fácilmente reproducible ya que es difícil lograr que la misma persona codifique la misma noticia de la misma manera dos veces, y mucho más entrenar a un equipo completamente nuevo para codificar un corpus de la misma manera que un equipo anterior (Nelson, 2017). En general, estos algoritmos como LDA clasifican los textos de la misma manera cada vez, haciendo que el paso de clasificación sea completamente reproducible.

Asimismo, trabajar con la herramienta de modelado de tópicos permite corregir algunas limitaciones del método manual de análisis de tópicos. En relación a los estudios que analizan textos y tienden a buscar temas específicos, las técnicas de aprendizaje automático permiten encontrar una multiplicidad de tópicos en los corpus diferentes a los hallados en las investigaciones de análisis de contenido cuantitativo. La modelización de tópicos abre la posibilidad de agrupar o clasificar los documentos según su tópico o tema prevalente sin recurrir a una clasificación a priori del investigador. Más aún, el carácter exploratorio del análisis textual computacional puede sugerir categorías relevantes para el texto que los investigadores no habían considerado previamente debido a sus nociones preconcebidas o la complejidad del texto (Grimmer & Stewart, 2011 en Nelson, 2017) y puede ayudar a los investigadores a evitar sus sesgos y la volatilidad natural que conlleva la lectura de grandes volúmenes de texto. En nuestro caso de estudio surgió un tópico que no buscábamos al principio de la investigación (Política exterior). No obstante, el rol del investigador no queda totalmente difuminado de este proceso. En efecto, el proceso de análisis, interpretación y etiquetado de los tópicos continúa siendo de carácter manual.

Retomando los desarrollos argumentativos del primer capítulo se postula que la potencialidad de aplicar estas técnicas computacionales en las investigaciones sociales se debe a su carácter exploratorio que permite encontrar relaciones, asociaciones y/o patrones de manera más inductiva. De este modo, la utilización de técnicas de aprendizaje automático de procesamiento de lenguaje natural permite a los sociólogos generar, evaluar y priorizar sus hipótesis preliminares de investigación (Mazzocchi, 2015; Nelson, 2017). No obstante, es

importante subrayar que incorporar técnicas con carácter más inductivo en el diseño de la investigación no implica abandonar la formulación de hipótesis preliminares que articulan la teoría con el mundo empírico (Kitchin, 2014). En la misma dirección, Sautu (2019) postula que la búsqueda de regularidades y asociaciones es solo una parte de la investigación científica, es necesario recurrir a las teorías sociales para comprender por qué suceden de esa manera (y no de otra) y adentrarnos en los procesos subyacentes de las regularidades.

A partir de los resultados que arribamos en este trabajo realizamos una comparación con los hallazgos de una investigación reciente sobre la agenda mediática digital en 2019 y con estudios que abordan la agenda mediática securitaria entre 2015 y 2019. Tal como definimos en el segundo capítulo, la agenda mediática es el patrón de cobertura de noticias durante un tiempo determinado (McCombs, 2015). A lo largo de la tesina mostramos que es posible identificar patrones en los contenidos de noticias durante un periodo de tiempo utilizando la técnica de modelado de tópicos, en particular, con el método LDA. En nuestro caso de estudio, la cobertura mediática online en el contexto electoral de las PASO 2019 estuvo compuesta por los tópicos elecciones, espectáculos, deportes, seguridad, política exterior, obra pública y economía. De esta forma, los temas que aparecen en la agenda mediática tienen preferencia sobre aquellos que no están. En sintonía con el estudio exploratorio de Koziner (2019) que aborda la agenda mediática digital entre abril y noviembre de 2019, en esta tesina evidenciamos que las elecciones, la seguridad, la economía, el deporte y los espectáculos fueron prioridad en la agenda mediática digital en el contexto de las PASO 2019. Aunque observamos diferencias en el orden de relevancia de estos tópicos. Estos hallazgos diferentes pueden deberse a las diferencias en 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 a la diferencia entre los corpus de noticias. En nuestro caso de estudio relevamos 52.154 de noticias de los portales de ocho medios, mientras de Koziner (2019) analizó el contenido de 1470 piezas periodísticas recolectadas de la *homepage* de tres medios. A partir de un análisis textual computacional en esta tesina concluimos que el principal tópico de la agenda mediática digital desde agosto a septiembre de 2019 fue las elecciones, relegando a las noticias sobre delito y seguridad a un cuarto lugar.

Investigaciones recientes sobre la agenda mediática securitaria han señalado que durante el gobierno de Mauricio Macri (2015-2019), y en particular con la designación de Patricia Bullrich como ministra de Seguridad, la narrativa mediática sobre la inseguridad se modificó (Calzado et al., 2019; Retegui et al., 2019; Zunino y Focás, 2019b). En la literatura

especializada se destacan cuatro movimientos en el tratamiento de la información securitaria, donde los actores pertenecientes al gobierno de Cambiemos jugaron un papel decisivo en la definición de los temas políticos que se debatieron en un momento dado y en las formas en que los medios tomaron estas temáticas. El primer movimiento fue la relevancia del tópico narcotráfico que se efectuó desde la agenda política. La "lucha contra el narcotráfico" constituyó un eje central de la batalla de la discursiva de Cambiemos, que fue recogida por los medios tradicionales. El segundo movimiento fue la mayor relevancia de la corrupción, en especial la atribuida a los gobiernos de Néstor Kirchner (2003-2007) y Cristina Fernández (2007-2015), como un tipo de delito que tiene relevancia en las agendas mediática televisivas y digitales. En los meses previos a las elecciones Primarias Abiertas Simultáneas y Obligatorias, Mauricio Macri dejó de lado el discurso económico para enfatizar sus supuestos logros en la lucha contra la inseguridad y el narcotráfico, al tiempo que exhibía temas relacionados con las instituciones y la corrupción (Natanson, 2019). El tercer movimiento fue el tratamiento o la entrada en la agenda securitaria de temas de "violencia de género", el cual estuvo entre los *issues* más relevantes del periodo. A partir del 2015, con el surgimiento del movimiento social y político Ni Una Menos, la matriz sobre inseguridad-seguridad se modificó y la violencia hacia las mujeres comenzó a ser una temática que conforma la agenda securitaria. En nuestro caso de estudio, las noticias sobre violencia hacia las mujeres se agrupan en el tópico 4 sobre "seguridad", tal es así que la palabra "mujer" es uno de los principales términos que conforman este tópico. A modo de ejemplo, mostramos una noticia del tópico "seguridad" que aborda la cuestión de género.


PERFIL ÚLTIMAS NOTICIAS POLÍTICA OPINIÓN SOCIEDAD

RADIO PERFIL FM 101.9 INGRESAR SUSCRIBITE

Temas del día · Mauro Viale · Segunda ola · Martín Guzmán · Cristina Kirchner · Alejandro Slokar · Remes Lenicov · Daniel Gollan · Médicos · HIJOS · Créditos UVA · Banco Central

POLICIA | MENDOZA 29-09-2019 15:59

Eran buscados por abuso y violencia de género: los atraparon cuando fueron a votar

Uno de los casos sucedió en el departamento de Guaymallén y el otro en San Martín. Los detenidos quedaron a disposición de la Justicia.

Finalmente, Zunino y Focás (2019b) destacan la entrada del tópico protesta social dentro de las secciones policiales e inseguridad. La alianza Cambiemos logró trasladar el eje de discusión pública sobre la criminalización de protesta a los medios de comunicación. Si bien los tópicos de la agenda mediática de seguridad se modificaron, la inseguridad se mantuvo

entre los temas más relevantes y es manifiesta su constitución como un tópico estable en la agenda mediática (Galar y Focás, 2019). Este panorama deja abierta una futura línea de investigación que estudie la composición de los tópicos de la agenda mediática securitaria digital en Argentina durante el contexto electoral de las PASO 2019 (julio a septiembre) e indague sobre una transición en torno al abordaje de los tópicos corrupción y narcotráfico.

Por su parte, 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). Las técnicas de modelado de tópicos presentadas en este trabajo sirven para estudiar la frecuencia de publicación de los diversos tópicos. En el primer capítulo mostramos cómo Focás y Zunino (2017) y Zunio y Focás (2019a y 2019b) operacionalizan el concepto de jerarquía de la información 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, etc. Por su parte, Koziner (2019) indaga sobre la relevancia de diversos tópicos de las noticias online en función de si la pieza periodística aparece en página impar, en mitad superior y si tiene gran tamaño, firma o títulos grandes. Los criterios mencionados pueden ser incluidos en el *scraper* a partir del análisis más específico de los elementos no textuales de los sitios web, es decir, es posible tomar la operacionalización del concepto de jerarquía de la información e incluirlos como criterios en el *scraper*. Para acotar el alcance esta tesina, la misma se centró en la detección, frecuencia y evolución que adquieren en la prensa online los tópicos relativos a la cuestión securitaria.

Por último, este capítulo finaliza con una reflexión sobre las potencialidades de combinar metodologías computacionales con los métodos tradicionales de las Ciencias Sociales y, en particular, con las técnicas de análisis de contenido cuantitativo en las investigaciones empíricas de Agenda Setting que estudian el contenido de noticias, la agenda y la relevancia mediática. La presente tesina es un paso en esta dirección. A continuación, tomamos dos investigaciones recientes para reflexionar de manera general sobre las potencialidades de emplear técnicas procesamiento de lenguaje natural en las Ciencias Sociales.

Un buen ejemplo es el trabajo de Nelson (2017), en el que se postula un marco general para incorporar métodos computacionales en el análisis de contenido sociológico. En la Sociología se ha utilizado durante mucho tiempo la teoría fundamentada (Glaser y Strauss 1999 en Nelson, 2017) para realizar investigaciones rigurosas que produzcan teorías. La teoría fundamentada es un método diseñado para permitir que las categorías y los temas emerjan de forma inductiva a partir de los datos, culminando en una comprensión teórica abstracta, basada

en los datos, del mundo social subyacente. La propuesta de Nelson (2017) denominada *Computational Grounded Theory* actualiza la teoría fundamentada para la investigación contemporánea agregando técnicas computacionales que brindan la capacidad de incorporar cantidades masivas de datos en la investigación generadora de teoría de una manera rigurosa y confiable, mitigando las deficiencias de la investigación puramente cualitativa. El enfoque propuesto se basa en tres etapas para medir el significado textual. La primera etapa, la detección de patrones, implica el uso de técnicas computacionales para reducir el texto libre a grupos de palabras interpretables para revelar patrones dentro del texto de una manera imparcial y reproducible. La segunda etapa, refinamiento de patrones, implica una lectura profunda del texto guiada computacionalmente e incorpora una interpretación holística. A través del modelado de tópicos, el investigador puede identificar matemáticamente textos que son representativos de un tema en particular y calcular la prevalencia relativa de ese tópico. En otras palabras, el cientista social puede leer fácilmente los documentos principales de cada tema, sabiendo que su lectura está dirigida a ese tópico. Estas etapas ayudan a los investigadores a explorar el texto de forma inductiva para descubrir patrones significativos basados en datos. En la tercera etapa, confirmación del patrón, se emplea el uso de técnicas computacionales adicionales para evaluar la validez de los patrones identificados inductivamente en el texto. De esta forma, combina el conocimiento especializado y las habilidades de interpretación con el poder de procesamiento y el reconocimiento de patrones que aportan las computadoras. A su vez, la autora postula que, al combinar los enfoques interpretativo y computacional, su marco propuesto supera las deficiencias de cada método individual.

Para reflexionar sobre las posibilidades que se abren para las Ciencias Sociales argentinas tomamos como ejemplo un estudio exploratorio reciente (Kessler et al., 2021) basado en una metodología mixta: grupos focales sobre consumo de noticias de delito en televisión y análisis computacional de polarización semántica. Este método computacional permite cuantificar la polarización sobre los distintos tópicos (corrupción, femicidios e inseguridad urbana) discutidos en los grupos focales. En este caso específico, la combinación de métodos cualitativos y computacionales permitió a los investigadores "seguir" a los individuos en situaciones de interacción durante las cuales se configura y reconfiguran discusiones que tienden (o no) a la polarización. El método computacional al estar basado en el lenguaje permite explorar diversas aristas de la controversia: qué individuos expresan los discursos más polarizados y cuáles se ubican más en las "fronteras" de cada comunidad (oficialistas u opositores). En este sentido, primero se genera un modelo de procesamiento del

lenguaje natural que clasifica textos como oficialistas u opositores en Argentina y, en segundo lugar, se mapea el texto a espacios vectoriales según ese mismo criterio. De esta forma, es posible estimar las distancias semánticas de las intervenciones y la polaridad de los tópicos. Los hallazgos de este estudio muestran que el único tópico no polarizado es femicidios mientras que para los sí polarizados el más controvertido es corrupción, seguido por inseguridad. Asimismo, es importante destacar que estos métodos computacionales se pueden aplicar tanto a contexto de redes sociales y medios digitales como también para interacciones entre grupos de individuos, en la medida que los intercambios verbales hayan sido grabados y trascriptos. De este modo, las potencialidades de este método para el estudio de polarización y de situaciones de interacción son muy amplias.

En el campo de la Sociología de la Comunicación y análisis de medios la combinación de métodos computacionales (procesamiento de lenguaje natural y *web scraping*) con técnicas de análisis de contenido cuantitativo puede ser fructífero para enriquecer los estudios de contenido de noticias, agenda y relevancia mediática. Por un lado, las técnicas de procesamiento de lenguaje natural habilitan la aplicación de métodos cuantitativos de análisis para una amplia variedad de tareas, como ser la clasificación de textos, la detección de temas y tópicos, entre otros. 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. Como mencionamos en el primer capítulo, estos estudios abordan una multiplicidad de aspectos sobre las características de la cobertura mediática securitaria. Por lo anterior, la combinación de métodos computacionales y técnicas de análisis de contenido cuantitativas puede enriquecer los estudios sobre relevancia mediática. En concreto, el modelo LDA permite identificar los tópicos relevantes en un corpus de noticias y seleccionar las piezas periodísticas con alta prevalencia del tópico específico que se quiera estudiar, para posteriormente realizar un análisis de contenido cuantitativo (e incluso un análisis cualitativo profundo) sobre estos documentos específicos. De esta manera es posible realizar una triangulación metodológica que combine métodos cuantitativos, computacionales y cualitativos en una misma investigación.

Recapitulando, en este capítulo abordamos la primera hipótesis de investigación y, luego del análisis de los resultados del modelo LDA, argumentamos que el empleo métodos computacionales posibilita escalar en la construcción y el análisis de grandes corpus de texto. También, analizamos las potencialidades que se abren para el trabajo cotidiano de las Ciencias Sociales y, en particular, para las investigaciones empíricas de Agenda Setting al incorporar la propuesta de técnicas mixtas a su repertorio metodológico. En concreto, aplicar técnicas de

procesamiento de lenguaje natural y *web scraping* permite aumentar sensiblemente la captura de información, sistematizar el proceso de pre-procesamiento de texto y reducir los tiempos de detección y análisis de tópicos relevantes.

Conclusiones

Esta tesina pretendió mostrar una aproximación metodológica posible para el análisis textual computacional a partir de la aplicación de una técnica de detección de tópicos sobre un corpus de noticias. Para ello tomamos como corpus de análisis, las piezas periodísticas publicadas durante julio a septiembre de 2019 en los medios online *Clarín*, *La Nación*, *Infobae*, *Página 12*, *Télam*, *Perfil*, *Crónica* y *Minuto Uno*. Con esto pretendemos ilustrar un flujo de trabajo para la utilización una técnica de procesamiento de lenguaje natural y, en particular, del modelo *Latent Dirichlet Allocation* (LDA) para la detección de tópicos. También, discutimos algunas herramientas para el pre-procesamiento del texto (eliminación de stopwords y otros signos, normalización de conteos, construcción de una TFM con el modelo *Bag of Words*). El objetivo fue analizar, de modo exploratorio, la aplicación de una técnica de procesamiento de lenguaje natural (modelado de tópicos) al estudio de contenido de noticias digitales con la finalidad de sortear algunas limitaciones metodológicas -como ser la escalabilidad y replicabilidad- presentes en los análisis de medios (Orozco Gómez y González, 2012).

De esta forma, identificamos los tópicos de la agenda mediática digital de los principales diarios online de Argentina durante el contexto de las elecciones Primarias Abiertas Simultáneas y Obligatorias. Encontramos, en primer lugar, que el asunto de las elecciones PASO 2019 ocupó un lugar relevante en las agendas de los principales medios digitales nacionales. También, aunque en menor nivel, los espectáculos, el deporte, la seguridad, la política exterior, la obra pública y la economía fueron prioridad en la agenda de los medios de comunicación online. En segundo lugar, nos focalizamos en las noticias de delito y seguridad, para comprobar, si como muestra la literatura local, este tópico había aumentado durante las elecciones. Observamos que la frecuencia de publicación de las noticias online de seguridad, en el contexto electoral de las PASO 2019, fue casi de 2 de cada 10 piezas periodísticas. Entre julio y septiembre de 2019, dado al contexto nacional de elecciones presidenciales la prevalencia fue de las noticias políticas relegando a las noticias securitarias a un cuarto lugar.

A partir de la visualización de la evolución en el tiempo de los tópicos de la agenda mediática digital observamos que el caso de la seguridad se constituye como un tópico estable y relevante en las noticias digitales en el contexto electoral. La tendencia general del tópico securitario muestra que el nivel de relevancia se mantiene estable en la agenda mediática digital. En contraposición a la segunda hipótesis preliminar, la prevalencia de las noticias securitarias no aumentó durante el mes de las elecciones Primarias Abiertas Simultáneas y Obligatorias (agosto).

La revisión de literatura que conformó esta articulación teórica no pretendió ser exhaustiva, sino que recuperó aportes específicos sobre la metodología de los estudios sobre contenido de noticias de delito y seguridad, y estuvo circunscrita mayormente a las producciones académicas argentinas que trabajaron en el tema. La cuestión securitaria se ubica entre las principales preocupaciones ciudadanas, influenciada entre otros factores por el crecimiento en la cantidad de noticias de inseguridad (Focás y Kessler, 2015). En sintonía con nuestros hallazgos, las distintas investigaciones sobre contenido de noticias muestran a la seguridad como un tema relevante de la agenda mediática durante el gobierno de Cambiemos (2015-2019) y en elecciones Primarias Abiertas Simultáneas y Obligatorias 2019. Los estudios locales muestran que la cobertura de noticias delictivas se incrementa en momentos electorales, en los que se plasman y canalizan las preocupaciones sociales contemporáneas. En nuestro caso de estudio empírico el aumento en la publicación de las noticias securitarias tiene lugar en el mes previo a las elecciones PASO 2019, lo cual plantea algunos interrogantes. ¿Se incrementa la publicación de piezas periodísticas securitarias durante el contexto electoral de las PASO 2019? Si esto sucede, ¿cuál es su dinámica temporal? Para responder a esta pregunta es necesario ampliar el periodo de análisis.

Ahora bien, además de los resultados del ejercicio propuesto acotado a mostrar la aplicación de una técnica de procesamiento de lenguaje natural más que agotar determinaciones del objeto en cuestión (las noticias securitarias), la tesina busca explorar las potencialidades que el análisis textual computacional tiene para las investigaciones sobre análisis de contenido. Una de sus principales ventajas radica en su escalabilidad ya que permite incrementar la capacidad de captura de información: el tamaño total del corpus (52.154 noticias) construido en este trabajo es de una escala sensiblemente mayor que los utilizados en los estudios reseñados en el primero y en el tercer capítulo. Los antecedentes mencionados tenían una escala más pequeña, lo que omite la mayoría del texto disponible que se puede usar como datos: alrededor de 1328 y 170 piezas periodísticas. Además, permite reducir el tiempo de la detección y análisis de los tópicos. Entre las fortalezas de emplear la técnica de modelización de tópicos al análisis de contenido de noticias se destaca la replicabilidad. El flujo de trabajo implementado en esta tesina y las operaciones contenidas en el mismo (el proceso de extracción y el pre-procesamiento de los datos, la construcción de la matriz, la elección del modelo y la visualización) son replicables. El código y *scripts* desarrollados permiten que los mismos resultados arribados en esta tesina pueden ser alcanzado por otros investigadores. También, brinda la posibilidad de morigerar ciertas limitaciones del análisis textual tradicional de las Ciencias Sociales causados porque el carácter manual de las codificaciones puede introducir

sesgos y los criterios de clasificación (codificación, etc.) son definidos por los investigadores. No obstante, los estudios sobre contenido de noticias reseñados en la tesina logran gran profundidad analítica sobre corpus medianos que no es posible alcanzar con técnicas de análisis textual computacional. Por ello, es importante subrayar que el uso de técnicas de análisis textual computacional no implica un desplazamiento de los enfoques basados en la interpretación manual de los documentos.

En este tipo de enfoques existen algunas limitaciones que es importante remarcar. En relación con la construcción del corpus, la dependencia de la fuente de datos, en este caso GDELT que utilizamos resulta una primera desventaja que no permite afirmar que se analizar la totalidad de las piezas periodísticas publicadas en los medios digitales *Clarín*, *La Nación*, *Infobae*, *Página 12*, *Télam*, *Perfil*, *Crónica* y *Minuto Uno* durante julio a septiembre de 2019. Otra limitación para considerar surge de los metadatos recabados (título y texto), algunos presentan diversos grados de calidad y no siempre están condiciones óptimas de ser incluidas en la base de datos, por esta razón eliminamos 300 unidades de análisis. También se presenta como problemas potenciales a resolver en la etapa de análisis la correcta determinación de la cantidad de tópicos y el etiquetado de los mismos. En este punto, la complementación con análisis de contenido cuantitativo de las noticias del corpus resulta de suma utilidad.

Por lo hasta aquí relevado, consideramos que la combinación de técnicas de *web scraping*, y procesamiento de lenguaje natural brinda a las Ciencias Sociales la posibilidad de sortear algunas dificultades metodológicas vinculadas a la escalabilidad y la replicabilidad. Estas herramientas de análisis textual computacional permiten realizar una sistematización de las diversas etapas del proceso de investigación (la recolección de datos, la construcción del corpus, el pre-procesamiento de un texto, su análisis y su evolución temporal). En este sentido, consideramos que estas técnicas pueden aportar herramientas metodológicas a las investigaciones de Agenda Setting que tienen pretensión de análisis de contenido a gran escala.

Este es un estudio exploratorio y deja planteados interrogantes académicos para ulteriores trabajos. El primero, que ya planteamos, es cuáles son los principales tópicos de la agenda mediática securitaria en el contexto electoral de las PASO 2019. En otras palabras, ¿de qué hablan las noticias securitarias?, ¿sobre qué tipo de delitos informan? La revisión de la literatura evidencia cambios en la composición de los tópicos de las noticias securitarias en los meses previos a las elecciones Primarias Abiertas Simultáneas y Obligatorias. Estos estudios muestran una transición en torno al abordaje de los tópicos corrupción, narcotráfico y criminalización de la protesta social. La segunda línea de investigación que queda abierta para futuros trabajos, se basa en una metodología mixta que combine el análisis computacional de

tópicos y el análisis de contenido cuantitativo (Nelson, 2017). A partir de la detección de tópicos automática podemos identificar en un corpus de noticias los tópicos securitarios y seleccionar las piezas periodísticas con alta prevalencia de estos temas. Para posteriormente emplear un análisis de contenido cuantitativo sobre estos textos seleccionados.

Referencias bibliográficas

Bibliografía citada:

Arce, V. V. y Zapata, N. (2019) Stella Martini y María Eugenia Contursi: Comunicación pública del crimen y gestión del control social. *Cuestiones criminales, 2(4),* 323-331.

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).*

- (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*. Recuperado el 10 de septiembre de 2020 de https://medium.com/dataholics/covid-19-medios-a-contramano-de-la-pandemia-parte-i-58ca5de2c192

Baylé, F. (2016). *Detección de villas y asentamientos informales en el partido de la matanza mediante teledetección y sistemas de información geográfica*. Tesis de Maestría. Universidad de Buenos Aires, Argentina.

Becerra, M. (2019): *Medios digitales en Argentina: la película y la foto*. Recuperado el 15 de febrero de 2021, de: https://www.letrap.com.ar/nota/2018-9-20-16-3-0-medios-digitales-en-argentina-la-pelicula-y-la-foto

Blumenstock, J., Cadamuro, G. y On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science, 350(6264),* 1073-1076.

Botta-Ferret, E. y Cabrera-Gato, J. E. (2007). Minería de textos: una herramienta útil para mejorar la gestión del bibliotecario en el entorno digital. *ACIMED, 16(4).*

Breiman, L. (2001). Statistical Modeling: The Two Culture. *Statistical Science*, 16(3), 199-215.

Calvo, E. (2015). *Anatomía política de Twitter en Argentina: tuiteando #Nisman*. Capital Intelectual.

Calvo, E. y Aruguete, N. (2020). *Fake news, trolls y otros encantos. Cómo funcionan (para bien y para mal) las redes sociales*. Siglo XXI editores.

Calzado, M. (2013). Ciudad segura. Vecindad, víctimas y gubernamentalidad. Notas sobre la campaña electoral del PRO en la Ciudad de Buenos Aires (2011) (Safe city. Neighborhood, victims and government. Notes on the electoral PRO campaign in Buenos Aires (2011)). Confuenze. *Revista di Studi Iberoamericani*, 5(1), 249-263.

Calzado, M.; Lio, V. y Fernández, M. (2014). El concepto de inseguridad en las campañas electorales latinoamericanas. El caso del PRO en la Ciudad de Buenos Aires (2007-2011). *Mediaciones Sociales*, nº 13, 211-237.

Calzado M., Lio V., y Gómez Y. (2019). Noticias policiales y nuevos modos de narrar la "inseguridad" en la televisión Argentina de aire. *Ámbitos, Revista Internacional de Comunicación*, n° 44, 217-243.

Carrillo, F. (2019). Análisis computacional de texto aplicado a las ciencias sociales. Aprendizaje automático y Big Data. Simposio llevado a cabo en el workshop de Factor DATA, Universidad Nacional de San Martín, Escuela de Altos Estudios Sociales.

Challot, F. (2018) *Deep learning with Python*, Shelter Island, NY: Manning Publications.

Dammert, L. y Erlandsen, M. (2020). Migración, miedos y medios en la elección presidencial en Chile (2017). *Revista CS, n° 31*, 43-76. https://doi.org/10.18046/recs.i31.3730

Defensoría del Público de Servicios de Comunicación Audiovisual, Dirección de Análisis, Investigación y Monitoreo. (2017). Monitoreos de Noticieros Televisivos de Canales de Aire de la Ciudad de Buenos Aires. Resumen Ejecutivo 6 Monitoreos (Febrero / Abril / Junio / Agosto / Octubre / Diciembre 2017).

Focás, B. M. (2018) *(In)seguridad, medios y miedos. Una mirada desde las experiencias y las prácticas cotidianas en América Latina*, Buenos Aires: Imago Mundi.

Focás, B. M. y Kessler, G. (2015). Inseguridad y opinión pública: debates y líneas de investigación sobre el impacto de los medios, *Revista Mexicana de Opinión Pública*, n° 19, 41–59.

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)*, ISSN 2386-7876, 189-209.

Galar, S. y Focás, B. M. (2019) El regreso de las víctimas. Reconfiguraciones en el procesamiento público de la inseguridad en la actual coyuntura política nacional (2016-2017). *Revista Austral*, vol. 8, n°1, 131-150.

Galtung, J. (1970) *Teoría y métodos de la investigación social*, Buenos Aires: Eudeba

Garg, N.; Schiebingerb, L.; Jurafskyc, D. y Zoue, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS, 115(16)*, 3635–3644. www.pnas.org/cgi/doi/10.1073/pnas.1720347115

Gerrish, S. y Blei, D. M. (2012). How they vote: Issue-adjusted models of legislative behavior. *NIPS*. https://papers.nips.cc/paper/4715-how-they-vote-issue-adjusted-models-of-legislative-behavior.pdf

Grimmes, J. (2015). We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together. *Cambridge University Press: 31*, 80-83. https://doi.org/10.1017/S1049096514001784

Hastie, R. y Tibshirani, R. (1986). Generalized Additive Models. *Statiscial Science 1(3)*, 297-318.

Kessler, G.; Focás, B. M.; Ortiz de Zárate, J. M. y Feuerstein, E. (2020). Los divergentes en un escenario de polarización. Un estudio exploratorio sobre los "no polarizados" en situaciones de interacción. *Revista SAAP* (en prensa).

Kessler, G. (2009). *El sentimiento de inseguridad. Sociología del temor al delito*, Buenos Aires: Siglo XXI.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, April–June 2014, 1–12.

Koslowski, D. (2019). Análisis computacional de texto aplicado a las ciencias sociales. Aprendizaje automático y Big Data. Simposio llevado a cabo en el workshop de Factor DATA, Universidad Nacional de San Martín, Escuela de Altos Estudios Sociales. https://diegokoz.github.io/workshop_text_mining/1_explicacion.nb.html

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

Koziner, N.; Zunino, E. y Aruguete, N. (2018) Las fuentes de la corrupción, *Voces del Fénix*, 76–81.

Marradi, A.; Archenti, N. y Piovani, J. I. (2018) *Manual de metodología de las ciencias sociales*, Siglo XXI.

Martini, S. (2007). Argentina, prensa gráfica, delito e inseguridad en G. Rey (Ed.), *Los relatos periodísticos del crimen* (pp. 21–54). Ebert-Stiftung

- (2019). Delincuentes, crímenes y monstruosidades: la noticia sobre el delito en los medios masivos, *Cuestiones criminales, 2(4)*, 268-278.

Mazzocchi, F. (2015). Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. *EMBO reports*, *16*(10), 1250–1255. https://doi.org/10.15252/embr.201541001

McCombs, M. F. (2006). *Estableciendo la agenda*, Paidós Comunicación.

McFarland, D.; Lewis, K. y Goldberg, A. (2015). Sociology in the Era of Big Data: The Ascent of Forensic Social Science. *American Sociologist*. 10.1007/s12108-015-9291-8

Mützel, S. (2015). Facing Big Data: Making sociology relevant. Big Data & Society. Disponible en: https://journals.sagepub.com/doi/full/10.1177/2053951715599179

Natanson, J. (2019). Argentina: elecciones en tiempos de grieta. *Nueva Sociedad*, n° 281, 4-11.

Nelson, L. K. (2020). Computational Grounded Theory: A Methodological Framework. *Sociological Methods & Research*, *49*(1), 3–42.

Pallavicini, C. (2019). Análisis computacional de texto aplicado a las ciencias sociales. Aprendizaje automático y Big Data. Simposio llevado a cabo en el workshop de Factor DATA, Universidad Nacional de San Martín, Escuela de Altos Estudios Sociales.

Retegui, L.; Carboni, O.; Koziner, N. y Aruguete, N. (2019). Fuentes periodísticas, standing y rutinas de trabajo en las noticias de delito, inseguridad y violencia en los noticieros de AMBA, *Cuestiones criminales, 2(4),* 236-265.

Rodríguez Zoya, L. y Roggero, P. (2015). La modelización y simulación computacional como metodología de investigación social. *Polis, 13(39).* http://polis.revues.org/10568

Rosati, G. (2017). Construcción de un modelo de imputación para variables de ingreso con valores perdidos a partir de Ensamble Learning. Aplicación a la Encuesta Permanente de Hogares (EPH). *SaberES. Revista de Ciencias Económicas y Estadística 9(1).* http://dx.doi.org/10.35305/s.v9i1.132

- (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.

Rosati, G.; Olego, T. y Vazquez Brust, A, (2020). Vulnerabilidad Sanitaria en Argentina Construyendo un mapa de vulnerabilidad sanitaria a partir de datos abiertos. *Revista internacional para la equidad en salud, 19(204).*

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

Salganik, M. J. (2017) *BIT BY BIT: Social research in the digital age*, Princeton University Press.

Sarraute, C.; Blanc, P. y Burroni, J. (2014). A study of age and gender seen through mobile phone usage patterns in Mexico. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* 836-843. 10.1109/ASONAM.2014.6921683

Sautu, R. (2019). *Estrategias teórico metodológicas en el diseño de la Investigación en Ciencias Sociales*, Lumiere.

Sosa Escudero, W. (2019) Big Data y Aprendizaje Automático: Ideas y desafíos para Economistas en Ahumada, H., Gabrielli, M., Herrera, M. y Sosa Escudero, W. (Ed.), *Una nueva econometría. Automatización, big data, econometría espacial y estructural* (157-201). Editorial Universidad Nacional del Sur.

Uman, I. (2018). Big Data y memoria digital. Claves para su exploración e investigación desde las ciencias sociales. *AVATARES de la comunicación y la cultura, nº 15,* ISSN 1853-5925.

Vazquez Brust, A.; Olego, T.; Rosati, G.; Lang, C.; Bozzoli, G.; Weinberg, D.; Chuit, R.; Minnoni, M. A. y Sarraute, C. (2018). Detección de Zonas de Alta Prevalencia Potencial de Chagas en Argentina. https://3b6a36a6-378a-4d09-b2fd-9e3fa9574447.filesusr.com/ugd/2aae47_a893b943bb7f4a93a3474157ed6855ab.pdf?index=true

Wallach, H. (2014). Big Data, Machine Learning, and the Social Sciences: Fairness, Accountability, and Transparency. https://medium.com/@hannawallach/big-data-machine-learning-and-the-social-sciences-927a8e20460d

Wang, H.; Can, D.; Kazemzadeh, A.; Bar, F. y Narayanan, S. (2012). A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle. *Association for Computational Linguistics*, 115–120. https://www.aclweb.org/anthology/P12-3020.pdf

Zunino, E. y Focás, B. M. (2019a) Territorios, tópicos y fuentes de la inseguridad. Un estudio sobre la prensa argentina. *Cuadernos info, Nº 45* ISSN 0719-3661, 73-93.

- (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* ISSN-e: 1988-2696, 401-413.

Bibliografía consultada:

Anderson, C. (2008). The end of theory. *Wired, 16(7).*

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, 55 (4).

Calvo, E. y Aruguete, N. (2018). #Tarifazo. Medios tradicionales y fusión de agenda en redes sociales. *InMediaciones de la Comunicación, 13(1),* 189-213.

Calzado, M. (2010). Miedo y sensación térmica. Hacia un análisis de los protagonistas de lo inseguro. Oficios Terrestres, año XVI, nº 25, p. 107-116. https://revistas.ort.edu.uy/inmediaciones-de-la-comunicacion/article/view/2831/2841

Defensoría del Público de Servicios de Comunicación Audiovisual, Dirección de Análisis, Investigación y Monitoreo (2019). Manual para el Monitoreo de Programas Noticiosos de Canales de Aire de la Ciudad de Buenos Aires.

Lemieux, C. (2017 [2009]) *Gramáticas de la acción social. Refundar las ciencias sociales para recuperar su dimensión crítica*, Siglo XXI.

Meraz, S. (2009). Is there an elite hold? Traditional media to social agenda setting influence in blog networks. *Journal of Computer-Mediated Communications, 14 (3),* 682-707. 10.1111 / j.1083-6101.2009.01458.x

- (2011). Using time series analysis to measure intermedia agenda setting influence in traditional media and political blog networks. *Journalism & Mass Communication Quartely, 88 (1).*