Tomás Montemagno

Analisis de agenda setting en medios digitales argentinos

Este trabajo intentará dar cuenta de algunos interrogantes sobre un corpus de noticias publicadas en los sitios webs de los principales medios de noticias argentinos. Se trata de un corpus elaborado por Florencia Piñeyrúa para su tesina de grado “Procesamiento del lenguaje natural aplicado al estudio de tópicos de noticias de seguridad en Argentina: julio a septiembre 2019” . Consta de siete mil noticias scrapeadas entre julio y septiembre de 2019 de los siguientes medios de circulación nacional:

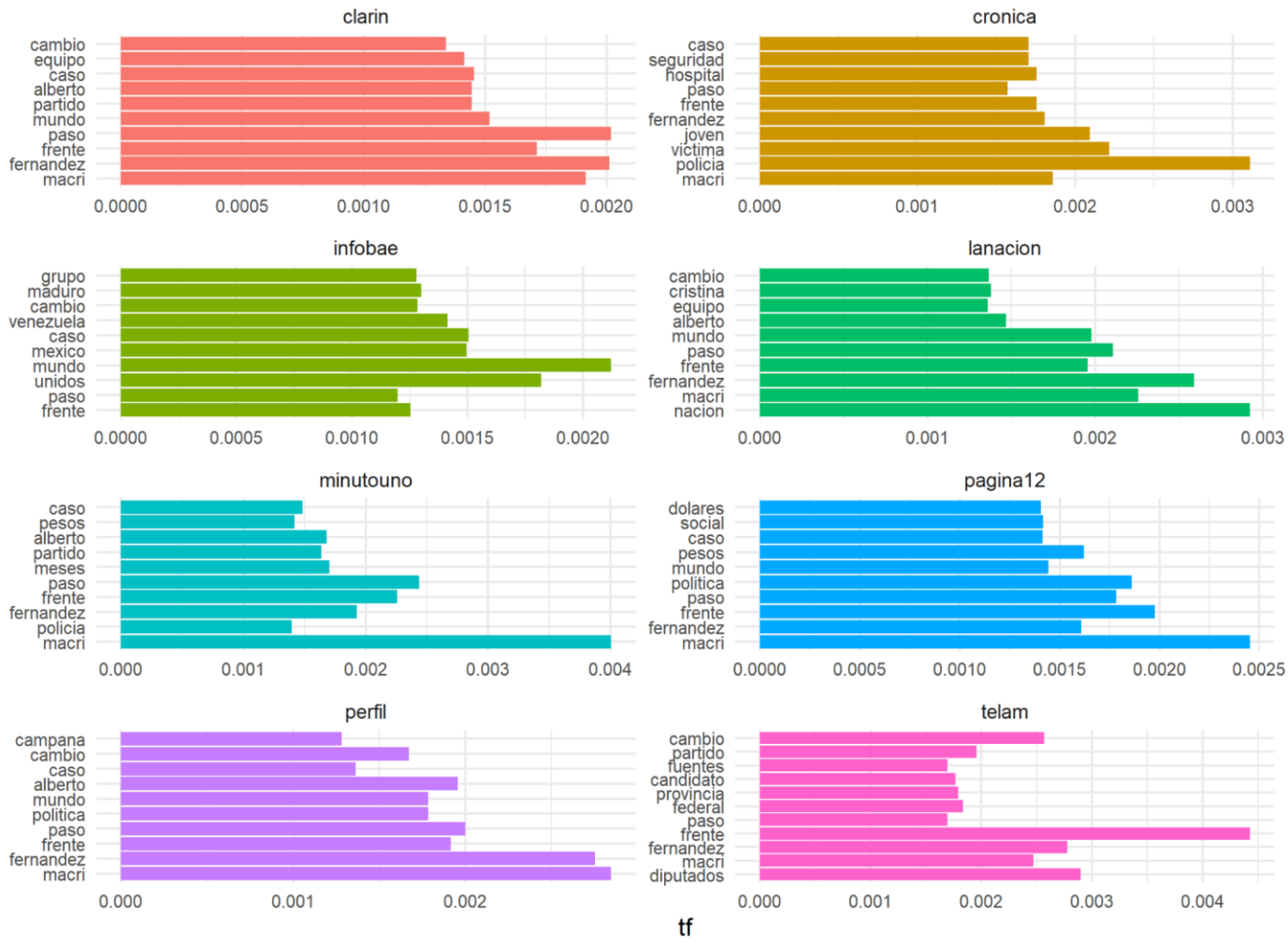
- Télam
- La Nación
- Clarín
- Perfil
- Infobae
- MinutoUno
- Página 12
- Perfil
- Crónica

Para el procesamiento de dicho corpus utilizamos algunas técnicas de procesamiento de lenguaje natural. Dichas técnicas nos facilitaron ahondar en algunos interrogantes que requieren mucho procesamiento de información. Pero antes, tuvimos que hacer un preprocesamiento del texto: pasarlo a minúscula, eliminar números, tildes y marcadores HTML. También lo *tokenizamos* , es decir, dividimos cada documento del corpus (artículos) en unidades mínimas (palabras). Una vez realizada la *tokenización* se le aplicó un diccionario de *stopwords* para remover las palabras más frecuentes de la lengua que aportan poco información sobre el texto.

Luego de la *tokenización* realizamos una matriz de palabras por medio y le aplicamos las métricas TF (term frequency) , IDF (inverse document frequency) yTFIDF (term frequency inverse document frequency). TF es el conteo de las palabras normalizado por la extensión (el total de términos) del documento. IDF calcula la “informatividad” de un término de acuerdo a la proporción de documentos que lo contienen. Las métricas Term Frequency (TF) e 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.

Finalmente nos quedamos con la métrica term frequency y generamos un gráfico de barras horizontal con la frecuencia de términos (TF) por cada medio (Gráfico 1). Podemos observar como la mayoría de los medios se concentraron en palabras asociadas a las elecciones PASO (primarias, abiertas, simultáneas y obligatorias). Casi todos frecuentan las palabras “Macri”, “Fernandez”, “Alberto”, “paso”, “frente”. Las excepciones aparecen en los sitios de Crónica e Infobae. En crónica las palabras de mayor frecuencia parecerían indicar que se enfocan en policiales con palabras como “jóven”, “víctima”, “policía”. En Infobae observamos que frecuentan

palabras como “Unidos”, “Venezuela”, “México”, “mundo” con lo que podríamos inferir que tiende a tratar temáticas asociadas a la política internacional.



Gràfico 1

Para profundizar un poco más en el gráfico de barras decidimos construir una visualización de nube de palabras por cada medio y observar así específicamente su predominio.

Medio	Nube de palabras	Predominio

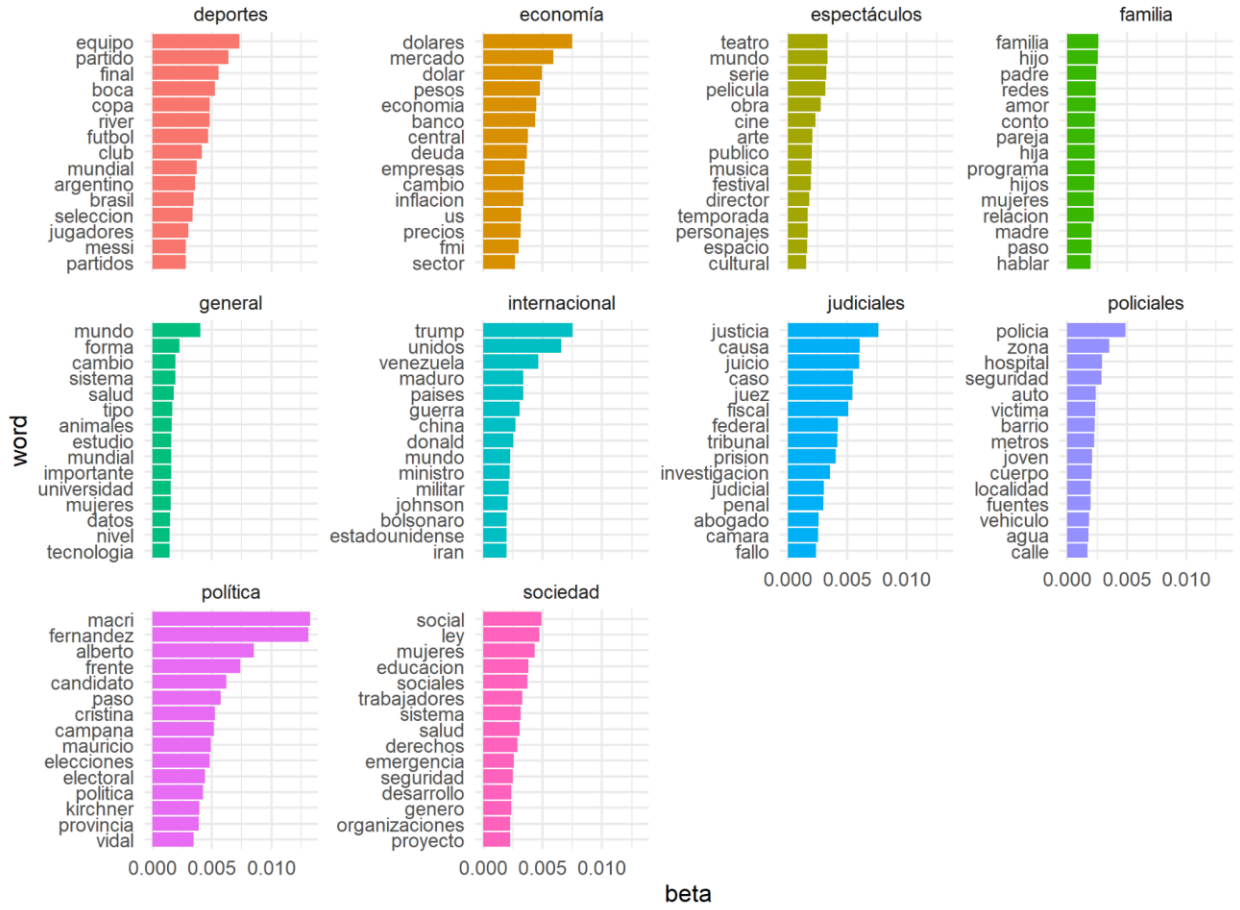
Tomás Montemagno

Página 12		**Elecciones PASO**
Clarín		**Elecciones PASO**
La Nación		**Elecciones PASO**
Perfil		**Elecciones PASO**
Telam		**Elecciones PASO**

Infobae		**Política Internacional**
Crónica		**Policial**

A su vez, empleamos otra técnica de procesamiento de lenguaje natural que nos ayudó a detectar tópicos entre las noticias. Este método se conoce como Latent Dirichlet Allocation (LDA). Es un método de aprendizaje no supervisado utilizado para *topic modeling* donde el algoritmo 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 (Piñeyrúa, 2021). El LDA supone que un texto es una secuencia de palabras y una palabra una secuencia de caracteres. También entiende a los tópicos como distribuciones de probabilidad sobre el vocabulario a lo largo del corpus del texto. El modelo LDA estima cuáles son los términos que tienen mayor probabilidad de pertenecer a un determinado tópico. A su vez, cada documento es una *mixtura* de tópicos: una noticia, por ejemplo, tiene una probabilidad del 60% de ser una noticia policial y una probabilidad del 13% del tópico judicial. Y cada tópico es una *mixtura* de palabras: los tópicos se componen de palabras con altas probabilidades de generar pertenencia (ejemplo: justicia, causa, juicio, juez) y al mismo tiempo las palabras pueden repetirse en varios tópicos (la palabra provincia puede estar en los tópicos de política, economía y sociedad, por ejemplo). El resultado del algoritmo LDA es, por un lado, una distribución de palabras por tópico y, por otro, una distribución de los tópicos por documento.

En nuestro caso le pedimos al modelo LDA diez tópicos (a partir del hiper parámetro K) que los agrupó en palabras por pertenencia. Luego interpretamos la relación entre esas palabras y les asignamos un nombre a cada tópico. Los resultados los podemos observar en el gráfico (gráfico 2) debajo. El gráfico muestra grupos de palabras, categorizadas en un tópico, las cuales tienen una probabilidad (beta) de ser parte de ese tópico. En este caso seleccionamos las 15 palabras con más probabilidades de ser parte de un mismo tópico, en grupos de 10 tópicos los cuales clasificamos a partir de la interpretación de la relación entre esas palabras.



Gràfico 2

Una vez clasificados los tópicos nos preguntamos, considerando todos los medios, cuáles fueron los tópicos con probabilidades más altas de ser parte de una noticia. Generamos un gráfico de barras (gráfico 3) y lo ordenamos de mayor a menor. Observamos como las temáticas más fuertes son política, luego policiales y en tercer lugar familia.

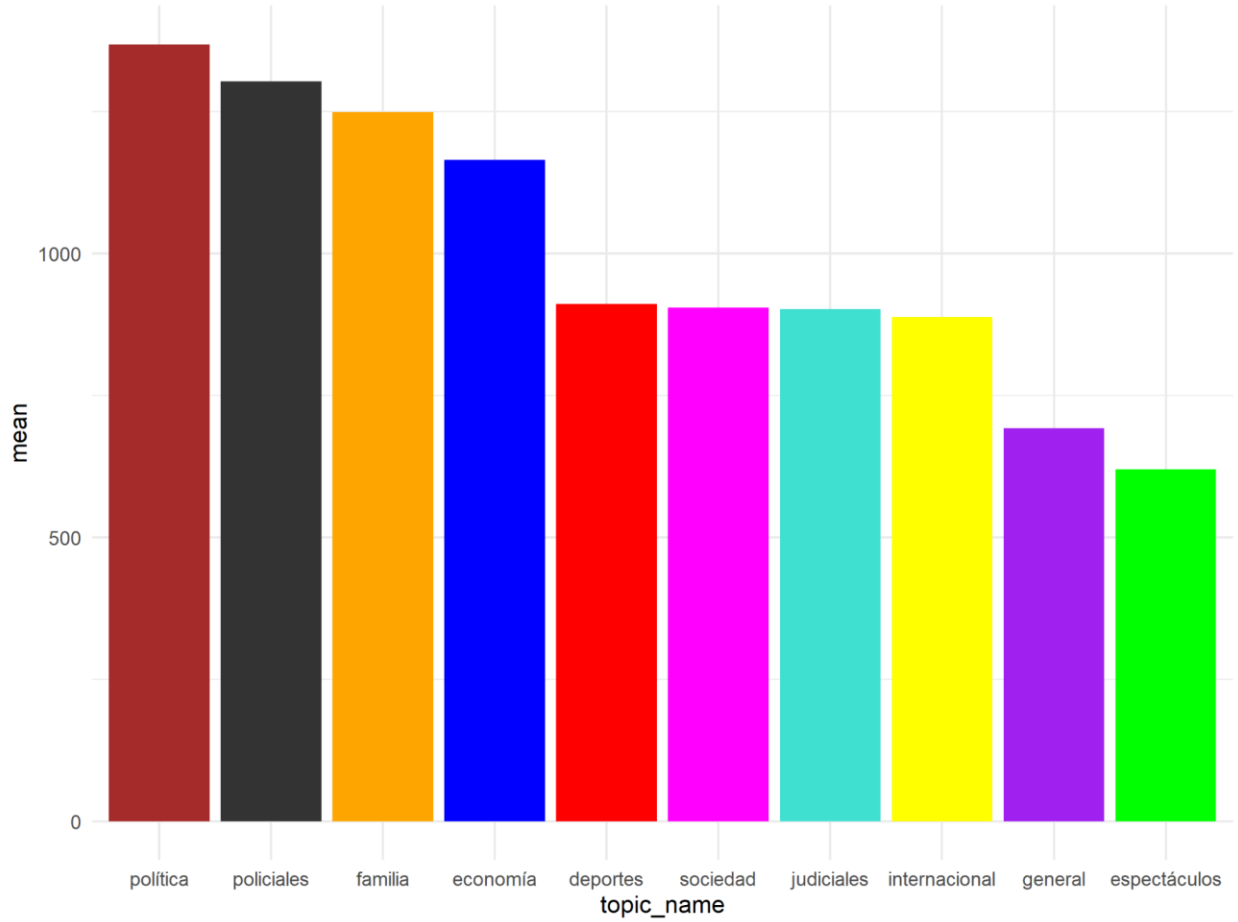


Gráfico 3

Si bien el gráfico anterior nos indica cierta tendencia sobre el predominio del tipo de noticias lo hace a un nivel generalista sin considerar la particularidad de cada medio. Es decir que predomine “política” como tópico en términos generales no implica que “política” sea el tópico predominante para, por ejemplo, Crónica o Minuto Uno. Además un medio puede contar con más noticias sobre otro generando más peso sobre un tópico determinado.
Nos preguntamos, entonces, cómo se dio la distribución de tópicos por medio y generamos un gráfico de barras por medio (gráfico 4) para conocerla.

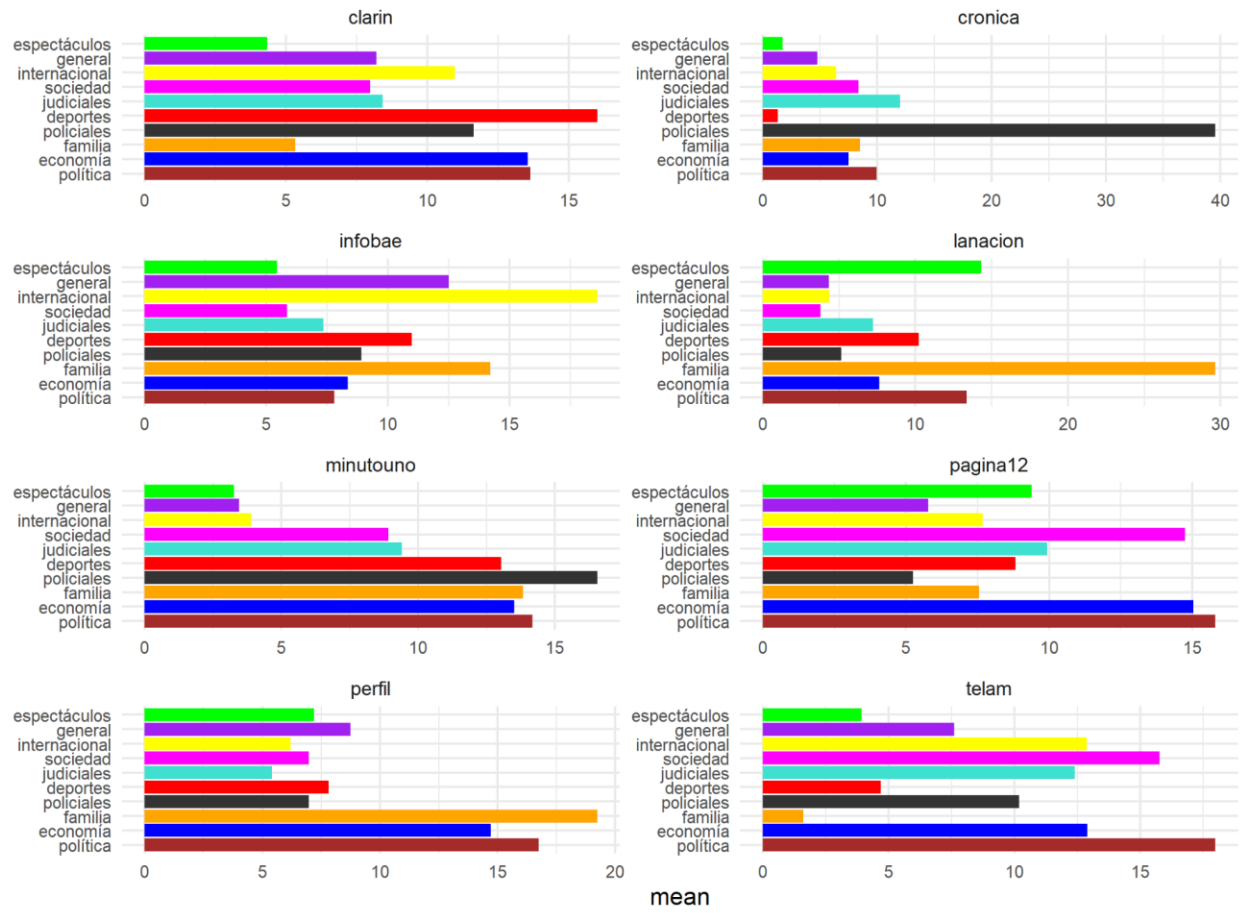


Gráfico 4

A partir de este gráfico podemos observar diferencias entre los medios. Por ejemplo, para el caso de Clarín, observamos que el tópico que más aparece es deportes seguido de política y economía (ambos casi con la misma media). En el caso de Crónica observamos como se destacan las noticias policiales sobre el resto de los tópicos con una diferencia de 27 pts de media sobre el segundo tópico que es judiciales, muy por debajo y claramente en consonancia con el tópico policial. En Crónica, el tópico deportes casi que ni aparece. En Infobae predomina el tópico internacional, seguido de familia y general. En La Nación observamos que el tópico predominante es familia, seguido de espectáculos y política. En Minuto Uno también predomina el policial pero lo acompañan muy de cerca política, economía, deportes y familia. En Página 12 predomina política, economía y sociedad, mientras que en perfil familia y luego política. Telam tiene un predominio similar a Página 12 con predominio de política, seguido de sociedad y luego economía.

Si bien hay diferencias centrales de entre cada medio, política supera la media de 10 casi en todos los medios (salvo en Infobae) por lo que podemos asumir que en este período y para este corpus de noticias fue un tópico de fuerte relevancia. Considerando el contexto de las elecciones PASO, nuestra hipótesis es que casi todos los medios coordinaron una misma agenda sobre esta temática dándole mayor cobertura que al resto de los tópicos durante las primeras semanas de

agosto (el 11 de agosto de 2019 se celebraron las elecciones). Para comprobar esta hipótesis generamos un gráfico de líneas evolutivo (gráfico 5), desde julio a octubre, con los cambios de la media móvil por tópico[1].

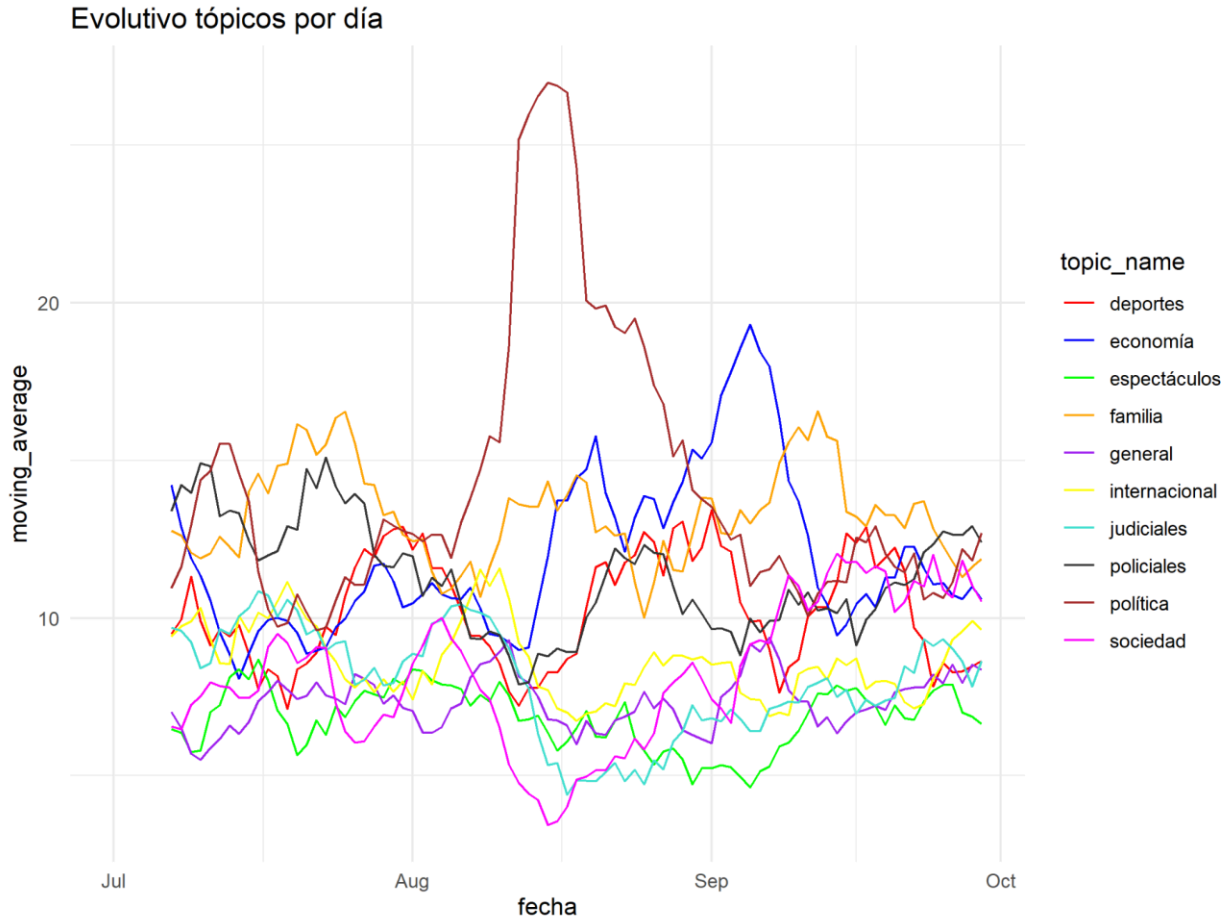


Gráfico 5

Observamos como, considerando todos los medios, hay un pico en el tópico política durante las dos primeras semanas de agosto. Luego decrece y emerge el pico de economía las primeras semanas de septiembre. Como planteamos en nuestra hipótesis existe una predominancia general del tópico política en las fechas aledañas a las elecciones PASO, lo que implica que este tópico se jerarquiza sobre el resto de los tópicos en todos los medios de comunicación.

[1] Elegimos la media móvil por sobre la media ya que es una técnica estadística que nos permite suavizar las fluctuaciones aleatorias y mostrar la tendencia subyacente. Esta técnica también nos permite atenuar otro supuesto del LDA básico: los tópicos preexisten a los textos y son constantes en el tiempo.

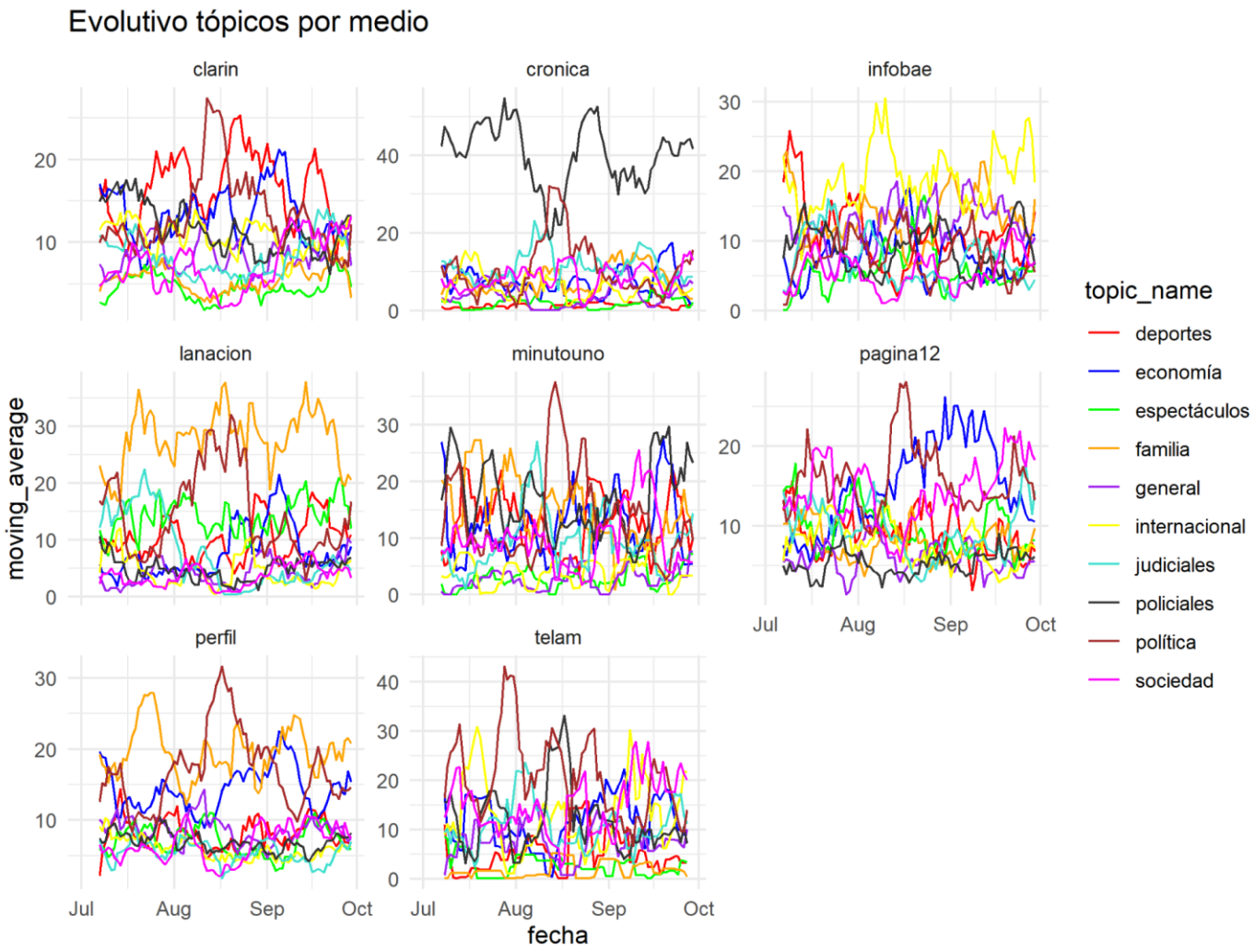


Gráfico 6

Al observar la evolución de la media móvil por tópico por medio (gráfico 6) podemos corroborar nuestra hipótesis: en todos los medios hay picos del tópico política en las primeras semanas de agosto salvo por Infobae. En Infobae se observa un pico de internacional en esa semana, con lo cual habría que explorar si las notas en ese período son referencias a las PASO desde una mirada de política internacional. A su vez, observamos que el tópico economía también cobra relevancia las últimas semanas de agosto y principios de septiembre en varios medios.

A través de este análisis computacional de tipo cuantitativo, podemos esbozar como estos medios digitales comparten una misma agenda mediática sobre algunas temáticas que son consideradas de interés general para luego continuar con su línea jerárquica de tópicos. Por ejemplo, en Clarín predomina deportes que cae un poco cuando son más relevantes las temáticas de política y economía (por agenda compartida). En Crónica sucede lo mismo pero su predominancia es policiales. En Perfil y La Nación lo mismo pero con la temática de familia y en Página 12 y Télam lo mismo pero con la temática sociedad.

De este modo pudimos dar respuesta a algunos interrogantes gracias al procesamiento de lenguaje natural y las técnicas computacionales empleadas para el procesamiento de datos. Gracias a la técnica de tokenización pudimos observar las palabras más utilizadas por los medios y observar algunas diferencias entre ellos. Con el algoritmo LDA pudimos identificar los principales tópicos de las noticias para luego trazar una evolución diaria y así observar su variación en el tiempo y las diferencias entre los medios. Podemos concluir que, luego del análisis de este corpus de noticias para estos medios digitales, los medios mantienen constante una determinada jerarquía de tópicos a lo largo del tiempo, según su elección editorial, pero que la jerarquía puede ser modificada con la aparición de temáticas de interés general (como las elecciones PASO de agosto de 2019) consideradas como parte de una agenda compartida por todos los medios.