Martín Tapia Serrano, Eugenia Peiretti, Alejandra Fauquié

**Trabajo final integrador**

Diplomatura en Ciencias Sociales Computacionales y Humanidades Digitales

# Consigna 1: Comparativas TF-IDF

*¿Cuáles son las palabras más utilizadas en cada uno de los medios? ¿Pueden verse diferencias? (Tener en cuenta las diferentes métricas trabajadas en el curso: tf, tf-idf, etc.) Generar las visualizaciones que considere más pertinentes para responder la pregunta*

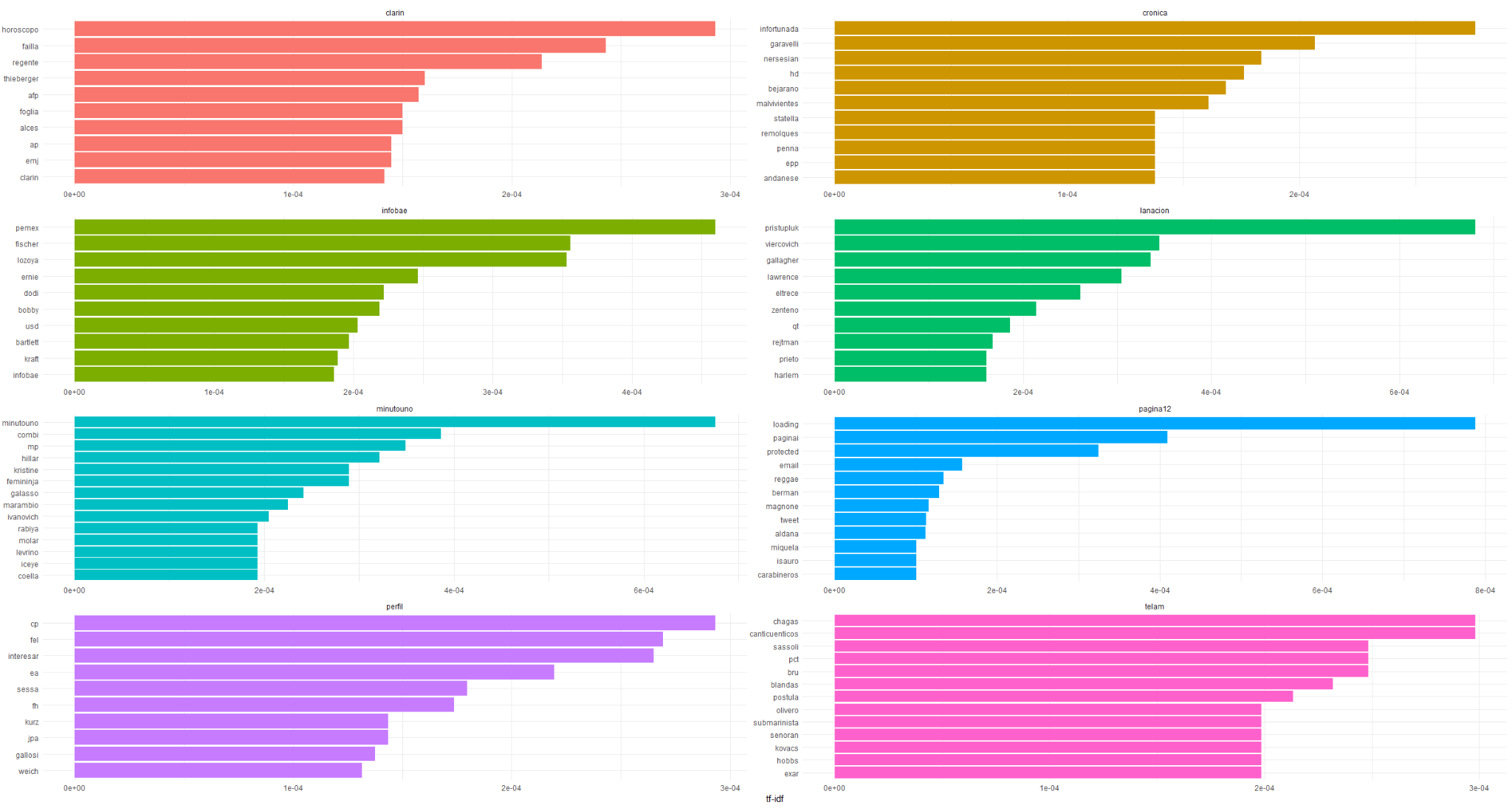
A partir del corpus base conformado por 7.000 noticias publicadas entre entre julio y septiembre de 2019 de 8 medios nacionales (Télam, La Nación, Clarín, Perfil, Infobae, MinutoUno, Página 12 y Crónica)[[1]](#footnote-0), se realizó un preprocesamiento de datos para luego poder hacer una primera exploración de palabras contenidas en el corpus.

La primera parte del preprocesamiento consistió en una limpieza centrada en detectar observaciones que por diferentes motivos resultó conveniente eliminar de la base: porque estaban duplicadas, porque el texto de la noticia se ha perdido o por algún problema durante el scraping. Esta primera limpieza resultó útil, a su vez, para conocer con mayor profundidad la estructura de la base de datos y sus características. Luego de este proceso, la base de datos quedó reducida a 6847 artículos.

Sobre esta nueva base, se avanzó con la normalización y limpieza de textos. Para esto se eliminaron las stopwords, caracteres especiales y números. A su vez, se cambiaron mayúsculas por minúsculas.

Con esto, se procedió a observar distribuciones de frecuencias de las palabras y a realizar el cálculo de métricas Term Frequency-Inverse Document Frequency (tf-idf). Estas métricas reflejan la relevancia y la informatividad de las palabras dado que permiten detectar aquellas palabras con alta frecuencia en los documentos y con baja frecuencia en el corpus, lo que ayuda a resaltar términos distintivos en cada texto. A continuación, se visualizan los resultados obtenidos de las métricas tf-idf para los distintos medios realizada a partir de la primera base limpia global.

**Gráfico 1. Palabras más relevantes por medio de comunicación - métricas tf-idf**

****

Fuente: elaboración propia

Antes de pasar a la descripción de las palabras más relevantes por medio, es importante destacar que tras esta primera serie de exploraciones de la base de datos, encontramos dos problemas extras: el primero, que hay un conjunto de observaciones que no contienen noticias en sí, sino referencias a videos en redes sociales o tweets; el segundo, que hay artículos que contienen errores en el scraping y en el campo texto encontramos la oración \_article download failed\_.

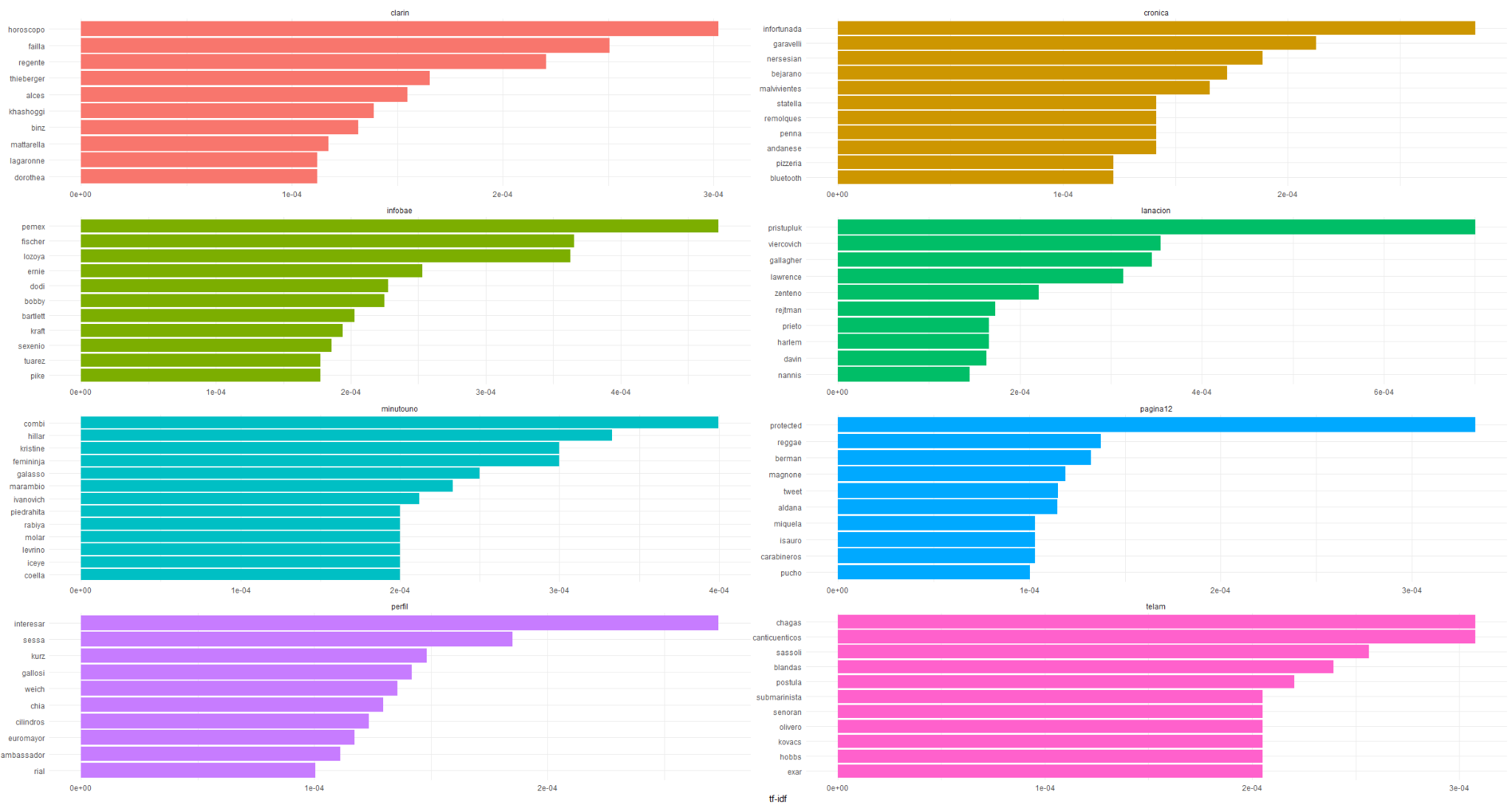
Luego de algunas pruebas, se concluye que es posible solucionarlos eliminando aquellas entradas cuyo texto no contenga más de 250 caracteres. Si bien esto no nos asegura eliminar todos ellos, sí permite limpiar la mayoría.

A su vez, a través de las exploraciones se detectó que, para algunos medios, algunas palabras que cobraban una relevancia significativa no tenían relación con las noticias en sí mismas, sino con segmentos de los portales vinculados al uso de redes sociales (compartir, facebook, whatsapp, etc.). También se encontraron palabras calificadas como relevantes por las métricas que hacían alusión a los nombres de los medios o nombres de periodistas.

Dada esta circunstancia, se realizó una exploración por medio y se eliminaron específicamente las palabras correspondientes a cada uno que se consideraron irrelevantes.

La siguiente visualización presenta los datos luego de haber realizado este filtrado extra de tokens (eliminando las palabras de 3 caracteres o menos y un set de términos considerados irrelevantes). Luego, se describen los resultados de la primera y la segunda prueba de métricas de forma comparativa para cada medio.

**Gráfico 2. Palabras más relevantes por medio de comunicación - métricas tf-idf en Base nueva**

****

Fuente: elaboración propia

***Clarín:*** el resultado obtenido en la primera prueba tf-idf nos arroja una gran cantidad de términos cortos. Al explorar los textos nos encontramos con que en su amplia mayoría son firmas que aparecen al final de algunas notas. También nos encontramos con algunos nombres propios (failla y tesone), el nombre del propio medio (clarin) y el nombre de una agencia de noticias (afp).

Por su parte, al ejecutar un filtrado extra no encontramos con que siguen predominando los nombres propios y aparecen los términos corresponsal y corresponsalía. Esto nos indica que los términos más relevantes para este medio que resultan de la prueba tf-idf no refieren necesariamente a temáticas o estilos específicos, sino a referencias sobre la autoría u origen de la noticia.

***Crónica****:* tanto en la primera como en la segunda prueba nos encontramos con que en este medio parecen emerger una temática específica y un estilo particular. Varias de las palabras parecen referir a hechos policiales y coincidir con el estilo amarillista que se le atribuye al medio: infortunada, malvivientes, sátiro, salvajemente, asaltantes, malheridos, acuchillado, son las palabras que están en el top.

Tras realizar el filtrado nos encontramos con que un conjunto de varias palabras pasan a tener el mismo peso.

***Infobae:*** lo llamativo en ambas pruebas son las menciones de palabras vinculadas a México: pemex, amlo, tabasco. También la mención a EEUU. Esto puede deberse a la vinculación de dicho medio con México y al tratamiento de noticias respecto a la política internacional.

***La Nación****:* en ambas pruebas destacan términos que podrían vincularse al ámbito del espectáculo: eltrece, pachano, archivo, fidalgo, polino, pampita. En la primera prueba también nos encontramos con términos cortos que pueden relacionarse con los orígenes de las noticias: ap, afp.

***Minuto*** *Uno:* en ambas pruebas destacan términos vinculados al deporte (o que podrían asociarse a dicha temática): riverplate, igualan, fifacom, fcbarcelona\_es. Luego nos encontramos con varios términos que, en primera instancia, no podrían asociarse directamente a una temática o un estilo.

***Página 12****:* en la primera prueba, el listado de términos está encabezado por palabras que no parecen formar parte de las noticias en sí. Por ejemplo, los términos email y protected, parecen surgir de direcciones de correo que no se han podido raspar como tales y que se encuentran al final de algunas notas.

Más allá de dichas palabras, en este caso parece emerger una línea editorial, dado que nos encontramos con términos con carga ideológica como: disciplinamiento, antiderechos, violencias (en plural), precarización, ultraderechistas, punitivistas.

***Perfil****:* en la primera prueba nos arroja términos cortos sin mucha relevancia y palabras que parecen ser publicitarias.

Tras ejecutar el filtrado tampoco encontramos resultados claros. Llama la atención la aparición de términos como morocha y rubia[[2]](#footnote-1).

***Télam****:* lo que destaca a primera vista es que este medio tiene una gran cantidad de términos que tienen el mismo valor (o muy similar) para la prueba tf-idf. Al revisar las tablas nos encontramos con que se trata con términos de poca frecuencia de aparición, pero alta informatividad dentro del set completo de medios.

Por su parte, vemos que Clarín, Infobae, La Nación y Télam, tienen términos con coeficientes tf-idf más elevados que el resto de medios.

Hemos decidido realizar una prueba más, filtrando aquellas palabras que cuenten con una frecuencia de aparición mínima de 5 dentro de cada medio. El resultado para la prueba tf-idf ha sido el siguiente.

**Gráfico 3. Palabras más relevantes por medio de comunicación - métricas tf-idf en Base filtros de frecuencia mínima.**



Fuente: elaboración propia

No se registran grandes diferencias en los resultados.

Podríamos arriesgar que Telam parece tener un perfil volcado a noticias políticas y del ámbito internacional. Destacan términos como: consulado, integración, guatemala, exhorto, subsecretario, exteriores. Sin embargo, esto no resulta concluyente.

## 

## 

# Consigna 2: Modelado de tópicos - LDA

*¿Cuáles son los tópicos principales en el corpus? ¿Pueden evidenciar diferencias en cada uno de los medios? Explicar qué método se utilizó para responder la pregunta, cuáles son los supuestos del mismo. Generar las visualizaciones más adecuadas para responder a las preguntas*

Para hacer modelado de tópicos utilizamos como herramienta el modelo probabilístico utilizado en procesamiento de lenguaje natural llamado Latent Dirichlet Allocation (LDA).

Este modelo se basa en dos principios. Por un lado, entiende a cada *documento* que conforma el corpus como una *mezcla de temas*. Por otra parte, cada *tema* se representa como una *mixtura de palabras.*

El LDA estima ambos mediante el cálculo de la probabilidad de que cada palabra pertenezca a un determinado tópico y la probabilidad de que cada documento esté compuesto por ciertos tópicos.

Para su implementación, se parte de una representación del corpus en forma de Matriz Documento-Término, a partir de la cual se infieren distribuciones de tópicos sobre documentos y de palabras sobre tópicos.

A partir de la base de datos de artículos periodísticos preprocesada, se realizó el modelado de tópicos con LDA. Para eso, transformamos la matriz en TFM y luego modelamos utilizando diferentes valores de k (parámetro que indica el número de tópicos que se quieren extraer del corpus) de forma tal de poder realizar una comparación de resultados y optar por la mejor opción de acuerdo a la consistencia de tópicos encontrados. Se evaluaron 4 casos diferentes: 7, 10, 12 y 17 tópicos. En principio, se partió de los k 7 y 10 considerando que esas suelen ser, aproximadamente, la cantidad de secciones principales de los diarios. Luego se corrió el modelo para 12 y 17 tópicos de forma tal de obtener mayor especificidad.

A los fines de este trabajo, se optó por el modelado de 12 tópicos para caracterizar los bloques temáticos del corpus, dado que se obtiene mayor diferenciación y especificidad en las temáticas al mismo tiempo que se mantiene la consistencia y coherencia en cada uno de ellos. La opción por 17 tópicos presenta diferenciaciones que resultan excesivas para nuestros objetivos (por ejemplo, desarrolla dos tópicos sobre Fútbol, distinguiendo entre torneos y clubes locales y selección nacional).

**Gráfico 4. Modelo LDA - Palabras con mayor probabilidad de pertenecer a cada Tópico.**



Fuente: Elaboración propia

A partir de la observación de palabras con mayor probabilidad de generarse en cada tópico (matriz beta) y realizando una exploración “manual” de artículos periodísticos para cada caso (mediante la probabilidad de la matriz gamma), se elaboró una **clasificación de cada bloque temático** en vistas a poder caracterizar los 12 tópicos del corpus:

1- Relatos identitarios, de costumbres e interés general: este bloque conformado por palabras como “casa”, “gente” o “ciudad” contiene relatos acerca de costumbres propias de algunas zonas o historias particulares de distintos espacios y organizaciones. También describe historias de fechas típicas y personajes de la literatura, entre otros temas de interés general.

2- Desarrollo social, derechos, educación y salud: En este bloque podemos encontrar artículos periodísticos que hacen referencia a políticas sociales vinculadas con el desarrollo social (alimentación, comedores comunitarios, políticas de niñez y adolescencia, etc.), al mismo tiempo que noticias que hacen referencia a la educación y a la salud. Contiene también aquellas notas destinadas a tratar los reclamos por los derechos de las mujeres.

3- Política Internacional: en este bloque resaltan palabras como China, Brasil, Venezuela, EEUU y sus dirigentes políticos. Se nombran países con los que Argentina mantiene relaciones comerciales y políticas, además de ser naciones que han estado presentes en la agenda internacional de la época.

4- Elecciones nacionales: En este bloque podemos observar que las palabras con mayor probabilidad de pertenecer aluden a elementos propios de las elecciones nacionales (“paso”, “frente” o “campañas”) como así también se nombra a los candidatos de las elecciones presidenciales del 2019. Como el gráfico de nube indica mediante el tamaño de sus palabras, las que están contenidas en este tópico son las palabras con mayor peso en el corpus.

5- Tecnología y redes sociales: Algunas palabras presentes en este tópico son “empresa”, “marca”, “sistema”, “producto” así como también los nombres de algunas redes sociales. Algunos artículos periodísticos con mayor probabilidad de pertenecer a este código, refieren a presentaciones de celulares, artículos de tecnología o autos.

6- Chimentos y farándula: algunas palabras protagonistas de este bloque temàtico son vida, redes, casa, pareja. La indagación de artículos periodísticos con mayor probabilidad de pertenecer a este tópico mostró noticias alusivas a relaciones interpersonales y otros aspectos de la vida personal de famosos y personajes de la farándula nacional, así como relatos de hechos producidos en programas televisivos tales como “Bailando por un sueño”.

7- Arte, cultura y espectáculo: tal como lo indican las palabras con más probabilidad de aparecer en este bloque temático, los artículos periodísticos con mayor gamma de este tópico hacen referencia a eventos artísticos vinculados con el teatro, la música y el arte audiovisual.

8- Fútbol y deportes: como puede observarse en el Gráfico 7, las palabras con mayor presencia refieren a torneos y a los clubes con mayor cantidad de hinchas de Argentina, como así también referencias a la selección nacional de fútbol.

9- Siniestros viales e inclemencias climáticas: en este tópico, las palabras más relevantes son “policía”, “ciudad”, “persona”, pero al hacer la indagación de artículos periodísticos se notó que el bloque temático está vinculado con accidentes viales e incidentes ambientales.

10- Agricultura y ganadería: Las palabras predominantes en este bloque temático son aquellas alusivas a los productos derivados de la producción agrícola y ganadera local, como “soja” o “carne”, acompañadas de términos como “Argentina”, “mundo” y “producción”

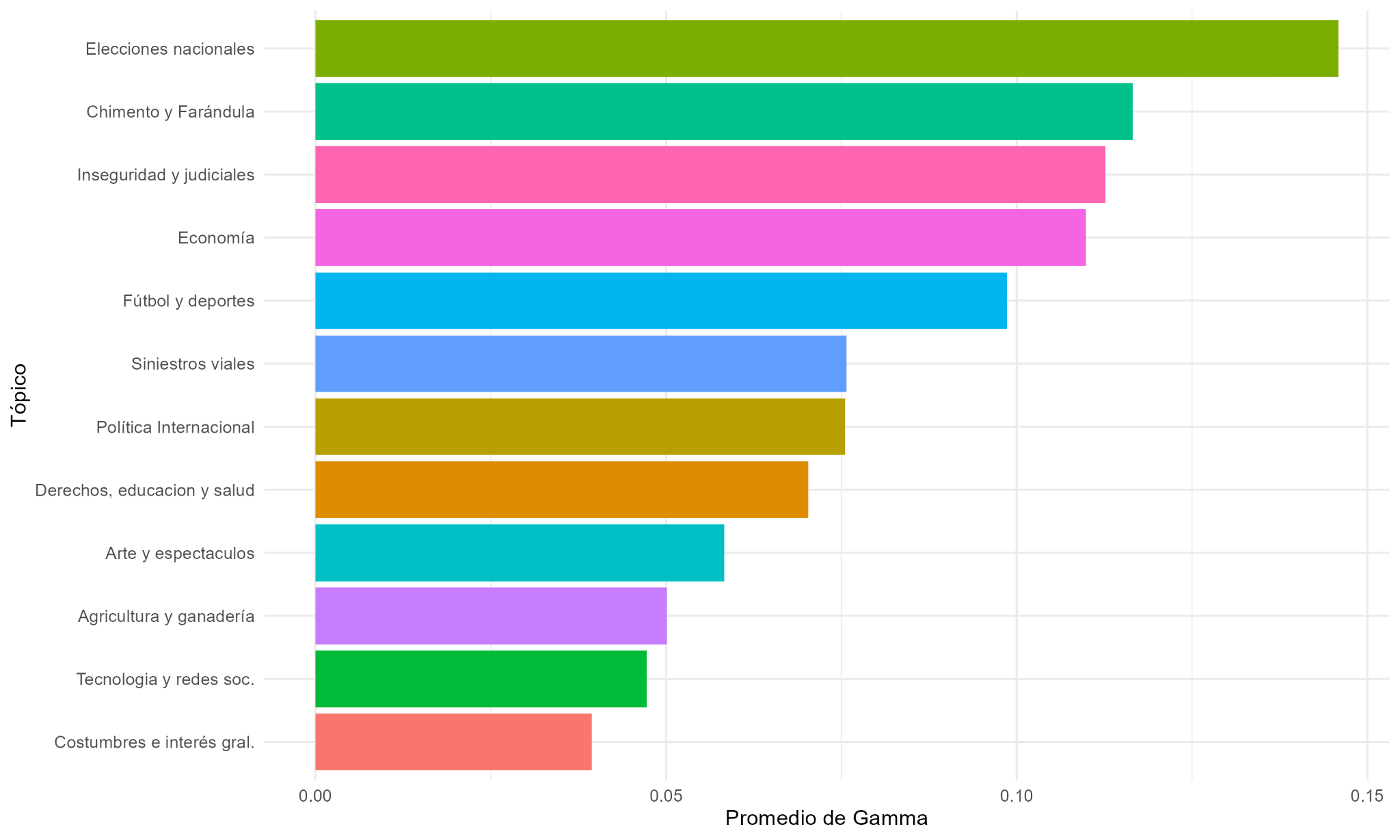
11- Economía y política económica: este tópico contiene palabras referidas a la economía local, tales como “mercado” y “Argentina”. Específicamente las noticias con mayor probabilidad de pertenecer a este tópico se relacionan con la política cambiaria, como lo indican las palabras “dólar”, “pesos”, “gobierno”.

12- Inseguridad, judiciales y policiales: los términos con mayor probabilidad de pertenecer a este bloque son aquellos relacionados con el sistema judicial (caso, fiscal, juez, causa) y con casos policiales y de inseguridad (víctima, policía). Este tópico contiene noticias tanto de casos judiciales vinculados con los gobiernos y la política, como casos de inseguridad en barrios y ciudades.

Ahora bien, si bien distinguimos 12 tópicos no todos tienen el mismo nivel de relevancia. Como se visualiza en el siguiente gráfico, para el conjunto de artículos periodísticos de esta muestra de ocho medios nacionales (Clarín, Crónica, Infobae, La Nación, MinutoUno, Página 12, Perfil y Télam) los **principales tópicos** desarrollados entre julio y septiembre del 2019 fueron, en primer lugar, las Elecciones Nacionales (con un 14.6% de documentos con probabilidad de pertenecer a este tópico), seguido por Chimentos y farándula (con 11.7% documentos con probabilidad de formar parte de este bloque), Inseguridad, casos judicializados y policiales en tercer lugar (con 11.3%), y Economía en cuarto puesto (con 11,0%).

Los tópicos con probabilidad de contener entre el 7 y el 9.9% de los artículos periodísticos son: Fútbol, Siniestros viales e inclemencias climáticas, Política Internacional y Derechos, educación y salud.

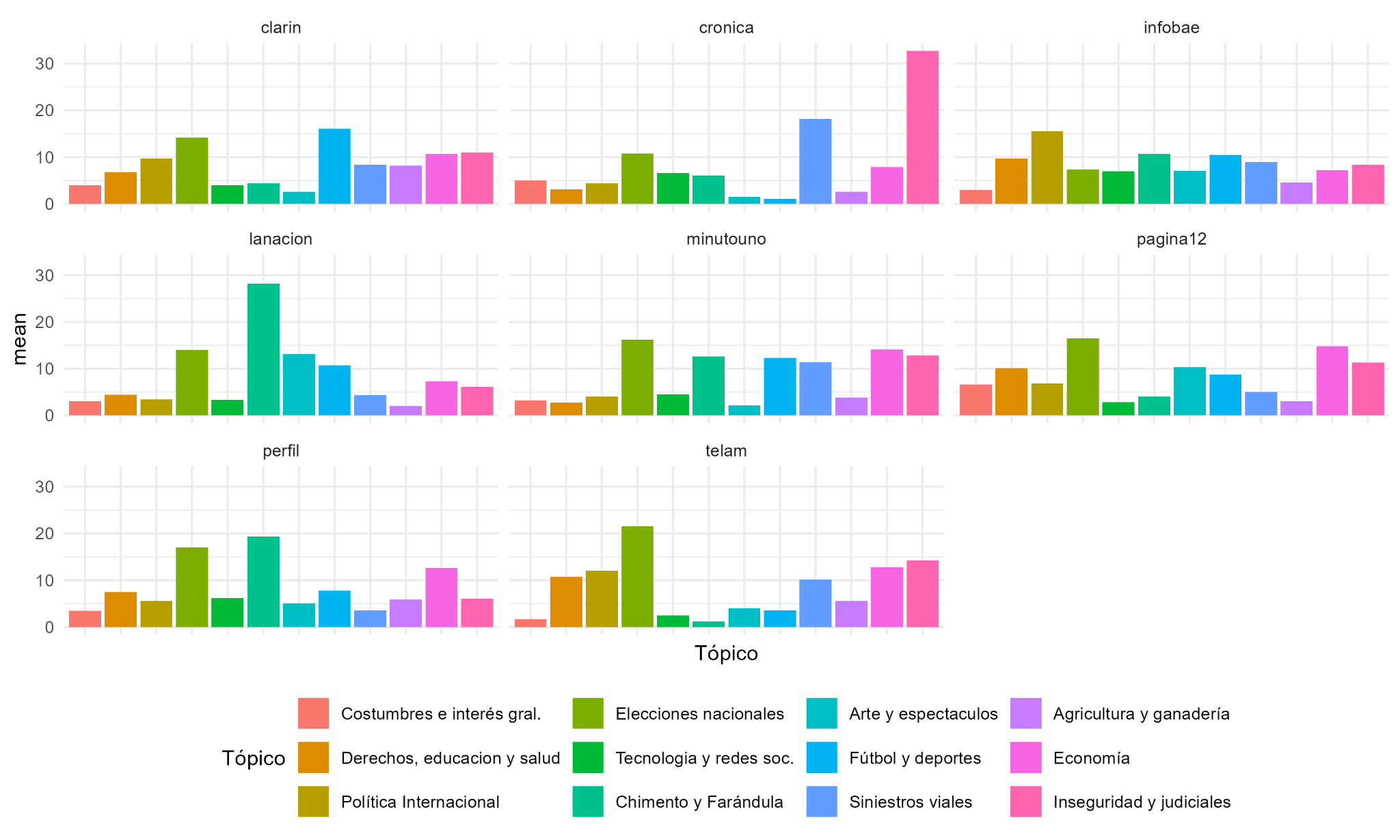
Mientras que aquellos tópicos con probabilidad de generarse en menos del 5.8% de las noticias son Arte y espectáculos, Agricultura y Ganadería, Tecnología y redes sociales y Costumbres e interés general.

**Gráfico 5. Tópicos por orden de relevancia según promedio gamma**

Fuente: elaboración propia

Si analizamos cada medio en particular, observamos diferencias a la hora de priorizar bloques temáticos en sus artículos periodísticos. Como puede observarse en el siguiente gráfico, en la muestra de noticias obtenida para los ocho medios nacionales mencionados algunos medios poseen una mayor concentración de notas con probabilidad de ser clasificadas bajo un bloque temático en particular, como es el caso de Crónica con el tópico de Inseguridad y casos judiciales (32,7%) o de Télam con las elecciones nacionales (21,52%), mientras otros poseen una distribución “más pareja”. Infobae es el medio que posee el menor rango entre su probabilidad mínima y máxima de artículos clasificados según tópicos, con una diferencia de 12,4 puntos porcentuales entre ambas.

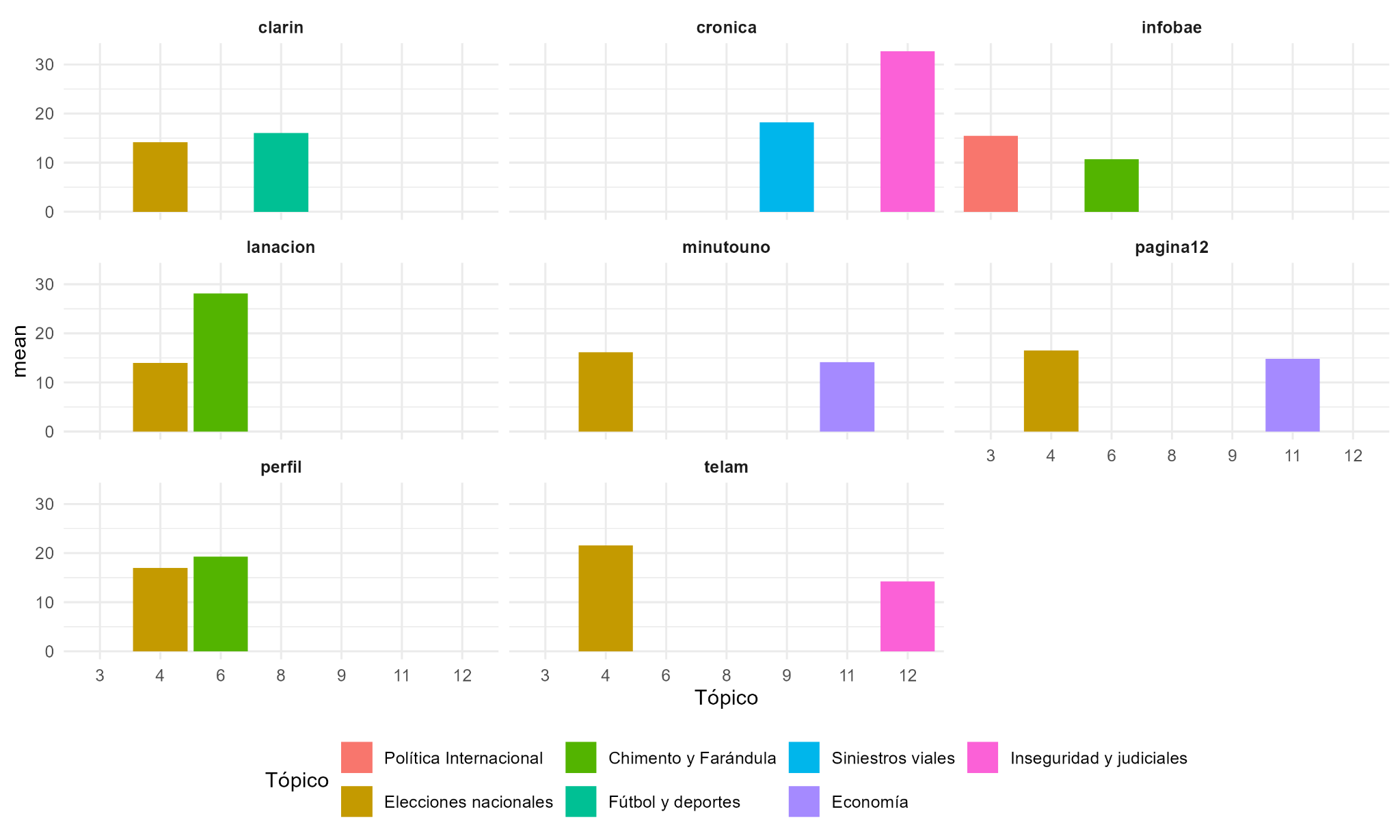
**Gráfico 6. Tópicos por medio.**



Fuente: Elaboración propia

Con el objetivo de poder visualizar cuales son los principales tópicos por medio, a continuación se presenta un gráfico donde pueden verse cuáles son los dos tópicos de los doce que tienen mayor probabilidad de ser tratados en cada medio.

**Gráfico 7. Principales tópicos por medio**



Fuente: Elaboración propia

Para Clarín, Fútbol y deportes y Elecciones nacionales son los principales, mientras que para Crónica son Inseguridad y judiciales y Siniestros viales e incidencias climáticas. Para Infobae son Política Internacional y Chimento y farándula. La Nación y Perfil tienen como principales bloques temáticos los de Chimentos y farándula y Elecciones Nacionales, mientras que MinutoUno y Página 12 tienen en su lugar a las Elecciones Nacionales y Economía. Por último Télam trata primordialmente a las Elecciones Nacionales y en segundo lugar los casos de Inseguridad y judiciales.

Los tópicos fundamentales de algunos medios coinciden con las palabras más relevantes detectadas en las métricas tf-idf : para Crónica, aquellas palabras vinculadas a hechos delictivos y noticias policiales eran las predominantes. De la misma manera, para La Nación, algunas palabras relacionadas con la farándula local aparecían como las más informativas.

# Consigna 3 y 4: Modelos de clasificación

*A continuación, seleccionar las noticias vinculadas a algún tópico relevante (por ejemplo, “Elecciones”) y construir un clasificador para predecir la orientación del diario. Utilizar alguno de los modelos de clasificación vistos a lo largo de la Diplomatura (regresión logística, random forest, etc.). Utilizar como features el* [*“Spanish Billion Word Corpus and Embeddings”*](https://crscardellino.net/SBWCE/)*, analizado en clase (pueden descargar el embedding en formato .bin del link). ¿Qué resultados arroja el modelo? ¿Es posible mediante el texto de las noticias conocer la línea editorial del diario? Generar las visualizaciones y tablas correspondientes para una correcta evaluación del modelo.*

*Diseñar un prompt para que Gemini (el LLM que usamos en clase) pueda realizar la tarea del punto anterior. Extraer una muestra de unos 800 artículos usados en el punto anterior y clasificarlos mediante Gemini. Comparar los resultados de ambos modelos. ¿Cuál funciona mejor? Generar las métricas y visualizaciones para comparar ambos modelos. ¿Cuáles podrían ser las causas de ambos comportamientos?*

Con el objetivo de predecir la orientación de un diario a partir de sus notas, entrenamos y testeamos distintas versiones de regresión logística, usando para *feature extraction* los embeddings provistos por la cátedra, y también experimentamos con Random forest. Finalmente hicimos pruebas tanto de “zero shot classification” como de “few shot classification” con Gemini, el LLM de Google. En todos los casos empleamos “conservador”, “neutro” y “progresista” como las categorías objetivo para la clasificación.

A raíz de las experimentaciones hechas, concluimos que los desempeños de los modelos testeados no son tan exitosos a la hora de predecir la orientación del medio a partir de los textos. Al menos no con la base de datos usada (y su respectivo etiquetado), las categorías usadas (“conservador”, “neutro” o “progresista”) y los tipos de modelo o enfoques usados.

Sin embargo, en los distintos ensayos llevados adelante se dilucidaron distintos factores que aumentan o disminuyen la probabilidad de correcta predicción. A continuación, se exponen algunos de estos factores.

Para probar nuestros modelos hemos elegido, en una primera instancia, los sets de notas pertenecientes a los dos tópicos más nutridos: tópico 4 - elecciones nacionales (1026 observaciones) y tópico 12 - inseguridad y judiciales. Posteriormente, nos hemos quedado sólo con el primero con el objetivo de poder ir realizando más pruebas.

## Regresión logística

### Ligera incidencia de los tópicos

En las primeras experimentaciones con regresión logística multinomial, se comprobó que el tópico elegido para el entrenamiento tiene una ligera incidencia en la capacidad del modelo de predecir la orientación de un medio a partir de sus textos. Esto puede tener que ver con que hay tópicos que son más tendientes a develar una orientación y también con que algunos tópicos, en este caso el número 4, tienen mayor cantidad de notas periodísticas.

| **Regresión logística m. - tópico 4** | **Regresión logística m. - tópico 12** |
| --- | --- |
|  |  |
|  |  |

El modelo entrenado con el tópico 4 (Elecciones nacionales), tuvo métricas ligeramente superiores al modelo entrenado con el tópico 12 (Inseguridad y judiciales).

La métrica *accuracy*, que mide la proporción de observaciones correctamente clasificadas (por lo que mientras más cercano a 1 sea el valor, indica una mejor clasificación) si bien es bajo para ambos, en el caso del tópico 4 es levemente mayor. La métrica *precission*, que mide la proporción de predicciones “positivas” que realmente pertenecen a esa clase, posee valores en la misma dirección: donde una alta precisión está dada por valores cercanos 1, tenemos métricas menores a 0,5, aunque un poco más altas para el tópico 4.

### Incidencia de categorías

En todos los intentos de regresión logística multinomial llevados adelante comprobamos que había menos probabilidad de categorizar correctamente notas etiquetadas como “neutro” o “progresista” que como “conservador”.

En uno de los intentos, se dejó la columna “medio” en la base que se usó como insumo para el entrenamiento del modelo. Siendo que en la base cada medio es invariablemente de una orientación, se esperaba que al testear el modelo, este pudiera fácilmente categorizar las notas. Aún en este escenario evidenciamos que el modelo tiene dificultades para categorizar lo progresista y neutro, como se deja ver el siguiente gráfico, donde la curva roc de categoría progresista (verde) y neutro (azul) indican mucho peor desempeño que la de categoría conservadora (roja).

|  |  |
| --- | --- |

### OTRAS PRUEBAS: Incidencia de la cantidad de categorías

**Apilando regresiones logísticas binarias**

Generamos un dataset a partir de las notas del tópico 4, “”Elecciones nacionales”, para poder conocer el porcentaje de posibilidad de que cada observación perteneciera a una determinada categoría. Una vez calculado cada porcentaje por separado, pusimos a competir estos para realizar nuestra predicción final, escogiendo aquella orientación que tuviera una probabilidad más alta.

Compartimos los resultados de cada regresión logística por separado.

| **Conservador (one vs. all)** | **Neutro (one vs. all)** | **Progresista (one vs. all)** |
| --- | --- | --- |
|  |  |  |

Como vemos, las mejores métricas las tenemos en el *recall*, lo que indicaría que nuestros modelos son buenos para predecir positivos.

Sin embargo, cuando ponemos a competir las probabilidades de pertenecer a una determinada orientación, los resultados no son buenos. Nos encontramos con que la mayoría de las métricas son mejores para el modelo de regresión logística multinomial, excepto para el *accuracy*.

| **Regresión logística binaria apilada** | **Regresión logística multinomial** |
| --- | --- |
|  |  |

Luego de hacer esta prueba, comprendimos que la regresión multinomial hace un proceso similar pero este experimento nos ayudó a evaluar cada categoría por separado y ensayar escenarios binarios.

**Eliminando observaciones neutras**

La otra prueba que realizamos, también a partir del tópico 4 (que es el que cuenta con mayor cantidad de artículos), fue eliminar aquellas observaciones que pertenecen a medios categorizados como neutros.

Los resultados obtenidos fueron mejores en todas las métricas que los obtenidos a través de la regresión logística multinomial.

| **Regresión logística binaria (sin neutros)** | **Regresión logística multinomial** |
| --- | --- |
|  |  |

Tanto la experimentación con la regresión logística binaria apilada como con la regresión logística eliminando la categoría neutro, nos permitieron evidenciar que usar dos en vez de tres categorías aumenta las posibilidades de predecir correctamente orientación a partir del texto. Siendo que, por un lado en las pruebas multinomiales el modelo tuvo dificultades con la predicción de las categorías “progresista” y “neutro” y que, por el otro, en estos experimentos binarios dio buenos resultados eliminar la categoría neutro así como predecir lo progresista aisladamente, podemos identificar que la categoría neutro representa una ambigüedad y posiblemente no tenga fronteras claras, particularmente con la categoría progresista.

## Random Forest

Realizamos un único experimento con Random Forest. Los resultados no fueron mejores en relación a la regresión logística multinomial

## 

## LLM-Gemini

Con Gemini, llevamos adelante dos pruebas con dos prompts, uno de zero shot y otro de few shot. Como se ve a continuación la performance mejora levemente con few shot.

| **Zero shot** | **Few shot** |
| --- | --- |
| *"A continuación vas a recibir un texto de una nota periodística.*  *Quisiera que la clasifiques como conservadora, progresista o neutra usando las siguientes categorías:*  *-conservador: la nota indica una postura conservadora*  *-progresista: la nota indica una postura progresista*  *-neutro: la nota no da señales de ninguna postura.*  *No justifiques tu respuesta, ni des información adicional, sólo contesta con una de estas tres categorias: conservador, progresista y neutro*  *"Quisiera que expliques paso a paso tu razonamiento.",*  *"",*  *"La salida debería tener el siguiente formato:",*  *"",*  *"clasif: seguido de la clasificación",*  *"expl: seguido de la explicación",*  *"",*  *"Este es el texto:"* | *"A continuación vas a recibir un texto de una nota periodística.",*  *"Quisiera que la clasifiques como conservadora, progresista o neutra usando las siguientes categorías:",*  *"",*  *"- conservador",*  *"- progresista",*  *"- neutro",*  *"",*  *"Este es un ejemplo conservador:", ej\_conservador,*  *"Este es un ejemplo progresista:", ej\_progresista,*  *"Este es un ejemplo neutro:", ej\_neutro,*  *"",*  *"Quisiera que expliques paso a paso tu razonamiento.",*  *"",*  *"La salida debería tener el siguiente formato:",*  *"",*  *"clasif: seguido de la clasificación",*  *"expl: seguido de la explicación",*  *"",*  *"Este es el texto:"* |
|  |  |

## 

## Conclusiones

En las experimentaciones realizadas se obtuvo mejores resultados con la regresión logística multinomial que con Gemini para la tarea de predecir categoría.

| **Mejor intento con regresión logística m.** | **Mejor intento con Gemini (few shot)** |
| --- | --- |
|  |  |

Ambos modelos, sin embargo, tuvieron dificultad para la correcta predicción. Las experimentaciones hechas con abordajes binarios nos permiten deducir que, en el caso de la regresión logística multinomial, uno de los principales obstáculos es la categorización de la base inicial. La base presenta una categorización de orientación de acuerdo a los medios, pero estos, por más que tengan una línea editorial definida, pueden tener una cierta diversidad de opiniones en sus notas. En consecuencia, la orientación de un medio no necesariamente se hereda o traspasa a sus notas, lo que significa que en nuestra base tendremos muchas notas etiquetadas de una cierta manera que quizás no las representa del todo. Esto lo pudimos comprobar con una exploración rápida de los falsos positivos de nuestros intentos. Esta problemática se acrecienta al no haber rasgos tan distintivos en notas que están etiquetadas como neutras o progresistas y la diferencia entre ambas es difusa. Siendo que los resultados con random forest no fueron mejores, entendemos que esta situación probablemente afectará el desempeño de cualquier modelo.

En el caso de Gemini, los resultados son aún peores porque, según las justificaciones que dio el modelo en su categorización, se pudo comprobar que tiene sus propias nociones de lo que es “neutro”, “conservador” y “progresista” y aún cuando se le dieron ejemplos de lo que constituiría presumiblemente un texto “conservador”, “neutro” o “progresista”, Gemini prioriza sus conceptos.

Con Gemini identificamos una oportunidad de mejora en próximos experimentos: se lo podría usar para hacer una nueva categorización individual de cada artículo. Sería una herramienta apropiada para esta tarea, pues parece ser más idónea para el aprendizaje no supervisado que para aprendizaje supervisado. Luego de esta categorización individual por artículo, se podrían evaluar las categorías prevalentes por medio o tendencias por medio (la categoría que aparezca con más frecuencia para un medio dado). Esta tendencia calculada sería lo que se compare con la orientación dada en la base. De esta forma se podrían sortear las problemáticas que surgen de la categorización de orientación según medio aprovechando las fortalezas de los LLM.

1. El corpus entregado por la Diplomatura para realizar el trabajo final, constituye una muestra aleatoria del corpus construido 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” [↑](#footnote-ref-0)
2. En el ejercicio nº2, luego de hacer la distinción de tópicos y la exploración de los tópicos más relevantes por medio, se verá que “Chimentos y farándula” es el principal bloque temático de este medio. Estas palabras están vinculadas a eso. [↑](#footnote-ref-1)