IMPLEMENTACIONES EN MINERÍA DE DATOS Y PROCESAMIENTO DEL LENGUAJE NATURAL SOBRE PUBLICACIONES EN PORTALES DE NOTICIAS: UN ENFOQUE EXPLORATORIO PARA LA GENERACIÓN DE HIPÓTESIS EN LA INTERPRETACIÓN SEMIÓTICA

Francisco Alomar
(Tecnicatura Universitaria en Inteligencia Artificial, Facultad de Ciencias Exactas, Ingeniería y Agrimensura de la Universidad Nacional de Rosario, Argentina)
franciscojalomar@gmail.com

Natalia Raimondo Anselmino
(Conicet y Centro de Investigaciones en Mediatizaciones de la Universidad Nacional de Rosario, Argentina)
nraimondo@conicet.gov.ar

Dolores Sollberger
(Facultad de Ciencias Económicas y Estadística de la Universidad Nacional de Rosario, Argentina)
dolores.sollberger@fcecon.unr.edu.ar

Irene Lis Gindín
(Conicet y Centro de Investigaciones en Mediatizaciones de la Universidad Nacional de Rosario)

Recibido: 11 de marzo de 2026
Aceptado: 15 de abril de 2026

RESUMEN

Este artículo presenta avances de una aproximación empírica (de alcance exploratorio) a un conjunto de notas publicadas en seis portales argentinos de información general y a los comentarios de usuarios-lectores publicados allí a través de Facebook. Lo compartido es parte de una investigación sobre la *plataformización* de los discursos acerca de lo público-común, circunscripta a un caso de "narcoterrorismo" en la ciudad de Rosario (Argentina) durante 2024, que lleva a cabo un equipo multidisciplinario.

Los hallazgos aquí expuestos derivan de la aplicación de algoritmos de minería de datos y procesamiento del lenguaje natural (PLN) que incluyen la vectorización de los textos mediante *embeddings*, su reducción de dimensionalidad y la obtención de agrupamientos algorítmicos, proyectados en gráficos. Se considera, además, que estos procedimientos (desempeñados en el marco de una estrategia de combinación metodológica denominada como *semiodata*) operan como detonantes de inferencias abductivas generadoras de hipótesis de trabajo.

El recorrido propuesto habilita a proponer una vía alternativa de ingreso al análisis sociosemiótico de los discursos mediatizados y la aplicación de un nivel de observación, que conjuntamente muestran cierta fecundidad para la distinción de delimitaciones que atañen al nivel de los géneros discursivos, mediante la visualización de distancias geométricas. En este sentido, las implementaciones algorítmicas serían potencialmente útiles para la identificación de determinadas disparidades invariantes ligadas a propiedades enunciativas de orden genérico, al menos en lo que respecta a la discriminación entre géneros primarios y secundarios, mientras que otras diferencias (relativas a la variedad de géneros periodísticos, por ejemplo) no habrían sido captadas computarizadamente.

PALABRAS CLAVE: semiodata**,** minería de datos, procesamiento del lenguaje natural, portales de noticias, géneros discursivos.

IMPLEMENTATIONS OF DATA MINING AND NATURAL LANGUAGE PROCESSING ON NEWS PORTALS PUBLICATIONS: AN EXPLORATORY APPROACH TO HYPOTHESIS GENERATION IN SEMIOTIC INTERPRETATION

ABSTRACT:

This article presents preliminary results from an empirical, exploratory approach to a set of news articles published on six Argentine general-interest news portals, together with the comments posted there by reader–users through Facebook. The material presented here forms part of a broader research project on the platformization of discourses about the public/common, focused on a case of "narcoterrorism" in the city of Rosario (Argentina) during 2024, conducted by a multidisciplinary research team.

The findings reported here derive from the application of data mining and natural language processing (NLP) algorithms, including the vectorization of texts through embeddings, dimensionality reduction, and the identification of algorithmic groupings projected into graphical representations. These procedures—implemented within a methodological strategy that combines approaches under the framework referred to as semiodata—are understood to operate as triggers for abductive inferences that generate working hypotheses.

The analytical path proposed here makes it possible to outline an alternative entry point for the sociosemiotic analysis of mediatized discourses and to apply a specific level of observation which, taken together, proves fruitful for distinguishing boundaries at the level of discursive genres through the visualization of geometric distances. In this sense, algorithmic implementations may be potentially useful for identifying certain invariant disparities linked to enunciative properties of a generic order—at least with regard to the distinction between primary and secondary genres—whereas other differences (such as those related to the variety of journalistic genres) do not appear to have been captured computationally.

KEYWORDS: semiodata, data mining, natural language processing, news portals, discursive genres.

1. Introducción

En marzo de 2024 aconteció en Rosario (Provincia de Santa Fe, Argentina) una secuencia de hechos violentos adjudicados a delincuentes asociados al narcotráfico que conmocionaron a la ciudadanía por varias semanas y dejaron secuelas de diverso orden. Es sabido que hace casi dos décadas la ciudad es foco del tránsito ilegal de estupefacientes, donde bandas criminales disputan territorios y poder (Yates, 2014; Bovino, 2016; Sampó, 2017; Bartolomé y Ventura Barreiro, 2019; Iazzetta y Gaiero, 2025). La novedad, en esta ocasión, emergió como un clivaje en la operatoria regular del comercio delictivo: se asesinó a trabajadores escogidos al azar y se atacaron instituciones y espacios de uso común de manera *random* como mecanismo de coerción frente a medidas de control carcelario adoptadas por el Gobierno provincial. Un clima de conmoción irrumpió y generó incertidumbre que rápidamente se expandió por el ecosistema mediático, alcanzando además una repercusión a nivel nacional, esto es, ganando espacio en medios porteños que habitualmente centran su contenido informativo en lo que sucede en Buenos Aires. Lo ominoso de los acontecimientos desbordó los marcos de significación habitual, y en el intento de semiotizar aquello que se vivía comenzaron a proliferar (tanto en el discurso de los medios como en las expresiones de los diversos funcionarios públicos vinculados al caso) denominaciones tales como "ola de sangre", "terror", o, incluso, "narcoterrorismo"[1]. En efecto, esta última calificación conllevó la intervención del Gobierno

[1] A propósito de esa designación, *Chequeado* (organización argentina dedicada al chequeo de la información o *fact-checking*), publicó el 14 de marzo de 2024 la siguiente nota donde se revisa el empleo de "narcoterrorismo", un término cuyo primer registro mediático data de la década de 1980: https://chequeado.com/el-explicador/violencia-narco-en-rosario-que-es-el-

nacional y significó, por tanto, que estos sucesos trascendieran el alcance de la acción delictiva puntal.

La magnitud mediática de tales acontecimientos y su definición como un problema público se tornó de interés para un abordaje interdisciplinar en el marco de un proyecto de investigación[2] que toma dichos sucesos para la construcción de un caso con carácter instrumental (Archenti, 2018), cuyo dominio empírico comprende un conjunto de discursividades vinculadas a esos hechos que han sido publicadas en diversas plataformas mediáticas (Raimondo Anselmino et al., 2025). Un proyecto que tiene como propósito comprender cómo el discurso sobre lo público-común se plataformiza (van Dijck, Poell y de Waal, 2018) al circular a través de distintos entornos como X, Facebook, Instagram, WhatsApp o los portales de noticias en la web. Como parte de ello, esto es, como hito intermedio de esa meta final de mayor alcance, lo que aquí se presenta es producto de la exploración inicial de la colección de notas publicadas en portales, en función de caracterizar las estrategias enunciativas empleadas por la prensa online de información general respecto del tópico en cuestión, para posteriormente poder compararlas con lo reconocido en las demás plataformas.

Estimulados por la composición multidisciplinaria del equipo de trabajo y orientados, a su vez, por la motivación de experimentar con técnicas y métodos que enriquezcan el análisis semiótico convencional (que, ciertamente, tiene límites cada vez más estrechos cuando se trata de producir conocimiento sobre las mediatizaciones contemporáneas), se adoptó una estrategia metodológica denominada *semiodata* (Raimondo Anselmino, 2022), que articula el estudio de discursos sociales con el empleo de herramientas y técnicas computacionales. Dicha estrategia de combinación metodológica es, por lo tanto, también un ejercicio que aspira a encontrar modos innovadores que refuercen la interpretación sociosemiótica de los fenómenos de producción de sentido en la actualidad.

El insumo base de este artículo está dado, entonces, por las noticias que trataron el tema y sus respectivos comentarios, cuyos mecanismos y criterios de recolección han sido descritos pormenorizadamente en Raimondo Anselmino et al. (2025). El período circunscrito comprende del 2 de marzo al 6 de mayo de 2024 y los portales relevados son seis en total: tres locales (*La Capital*, *Rosario3*, *El Ciudadano & y la región*) y tres de alcance nacional (*Clarín*, *Infobae*, *La Nación*), junto a los comentarios de lectores-usuarios de los medios regionales que intervinieron en las noticias desde sus perfiles de Facebook.[3] La recolección se realizó mediante técnicas de *web scraping*, sistematizando los textos y metadatos en un formato estructurado.

El corpus y dataset así constituido para este trabajo es *input* de implementaciones de minería de datos y PLN. Los vectores construidos a partir de los textos (*embeddings*), son representaciones numéricas que se proyectan en un espacio de menor dimensión que, a su vez, se agrupan mediante el algoritmo de aprendizaje no supervisado *k-means*, diseñado para delimitar agrupamientos (*clusters*).

Dada nuestra inclinación semiótica, este ejercicio exploratorio procura encontrar vías alternativas de observación que nutran la interpretación del corpus, habilitadas por algoritmos

---

narcoterrorismo-la-ley-rico-y-la-guerra-contra-las-drogas/ [02/03/2026). También puede leerse una problematización de esa expresión en un informe del Centro de Estudios Legales y Sociales (CELS, 2018).

[2] El proyecto (código 80020220700094UR) se encuentra acreditado por el Programa para la Investigación Universitaria Argentina (PRINUAR) y es ejecutado en la Universidad Nacional de Rosario (UNR), con sede en el Centro de Investigaciones en Mediatizaciones (CIM) de la Facultad de Ciencia Política y Relaciones Internacionales.

[3] El proyecto abarca un dominio empírico más amplio que incluye, en total, cinco corpus con sus respectivos *datasets* donde (además de las notas ya aludidas) se compendian publicaciones realizadas en las cuentas oficiales de Facebook, Instagram y X de los mismos portales informativos (Cfr. Gindin et al., 2026). También se colectaron, en esas mismas plataformas, publicaciones de funcionarios públicos pertenecientes a distintos niveles del Estado (municipal, provincial y nacional) que fueron protagonistas durante los sucesos, así como mensajes de audio, imágenes y videos compartidos por la ciudadanía a través de WhatsApp (Cfr. Musante et al., 2025).

de procesamiento y modelización utilizados (clustering, embeddings y reducción de dimensionalidad), con el afán de que los resultados así producidos colaboren con la construcción de hipótesis de trabajo, más que con su validación o refutación.
Consideramos, conjeturalmente, que los hallazgos derivados de la aplicación de los algoritmos de minería de datos y PLN cuando son proyectados en gráficos operan como detonantes de inferencias abductivas (en el sentido peirceano del término) generadoras de hipótesis. Desde este punto de vista, se entiende que la obtención de agrupamientos dentro de un espacio vectorial, claramente diferenciados mediante una representación visual, no sólo expresa la cuantificación de diferencias textuales, sino que además da acceso a un nivel de observación que (en la instancia de interpretación) ha mostrado utilidad para distinguir géneros discursivos. En otras palabras, entendemos que, en conjunto, los procedimientos dispuestos serían potencialmente útiles para detectar automáticamente diferencias que, en principio, atañen a (y pueden explicarse según) lugares de la organización textual (Steimberg, 2000) que implican clasificaciones sociales de lo discursivo mediatizado.
Para acometer lo expuesto hasta aquí, nuestra investigación ha afrontado los siguientes objetivos específicos: I) Iniciar una aproximación empírica al dataset construido, ensayando la articulación entre procedimientos automatizados y artesanales; II) Aplicar minería de datos para obtener agrupamientos de notas y comentarios, basados en distancias (algoritmo *k-means*) entre vectores (*embedding*) y reducción de dimensionalidad; III) Exponer los resultados de modo que puedan ser visualmente interpretados; IV) Construir una interfaz exploratoria de los gráficos que permita analizar en profundidad el corpus textual; V) Elaborar hipótesis de trabajo que den cuenta de las distancias y agrupamientos obtenidos y expliquen abductivamente la correlación hacia interpretaciones provistas desde un análisis de los discursos disponibles[4].
Para compartir lo expuesto, a continuación se desarrolla el marco teórico-metodológico que orientó el trabajo realizado, se pasa revista al proceso de construcción del corpus y a la metodología algorítmica empleada, así como a la implementación computacional correspondiente. Por último, algunos avances derivados de lo observable a raíz de dicho procesamiento.

2. Principales lineamientos del enfoque teórico-metodológico

Para cubrir los objetivos propuestos, nos ubicamos teórica y metodológicamente en la intersección entre el empleo de herramientas y técnicas computacionales aplicadas al análisis de datos no estructurados (lingüísticos) y el estudio empírico de la puesta en discurso desde un punto de vista sociosemiótico (Verón, 1998). Esta estrategia interdisciplinaria combinada se denomina *semiodata* (Raimondo Anselmino, 2022), porque integra el análisis sociosemiótico de los discursos mediáticos -que, como se advierte en Fernández (2023: 65), ya supone toda una metodología específica- con la aplicación de técnicas computacionales de minería de datos y procesamiento de lenguaje natural.
En esta línea, se retoma como antecedente directo el estudio desarrollado por Raimondo Anselmino, Rostagno y Cardoso (2021), en donde se aplicó un algoritmo de *clustering* sobre un corpus de posteos publicados en Facebook por páginas pertenecientes a los diarios argentinos *Clarín* y *La Nación*. En dicha aproximación, también experimental, se probó la capacidad de los métodos automatizados para acompañar el análisis semiótico, especialmente en escenarios que involucran altos volúmenes de datos y variabilidad dentro de la producción discursiva digital. El enfoque metodológico planteado en ese trabajo direcciona las tareas del análisis sociosemiótico convencional y la exploración automatizada de textos hacia un objetivo común, reconociendo que el procesamiento algorítmico no sustituye ni alterna con la interpretación teórica del corpus, sino que permite generar agrupamientos y visualizaciones que funcionan como estímulos o detonadores abductivos (Aliseda, 1998) para la formulación de hipótesis interpretativas. Más allá del antecedente mencionado, no se han identificado trabajos que empleen de manera

[4] Sobre este último objetivo, quedará para una publicación aparte lo que corresponde puntualmente a la interpretación semiótica de los tres agrupamientos reconocidos mediante el procesamiento con *k-means*. Esto es así porque su desarrollo desborda los intereses de este artículo y excedería además lo que puede exponerse en el espacio dispuesto por la revista.

comparable técnicas de embedding, reducción de dimensionalidad y clustering con el objetivo explícito de aproximar textos desde el marco teórico de la sociosemiótica. Esta ausencia relativa de antecedentes directos subraya la motivación y el aporte exploratorio del presente artículo.

El presente artículo se construye sobre esa base metodológica, extendiendo la aplicación de procedimientos a un corpus más amplio y en otro contexto de estudio, proponiéndose examinar hasta qué punto la sinergia producida entre las técnicas de *embeddings*, reducción de dimensionalidad y *clustering* puede ofrecer perspectivas interpretativas novedosas sobre las configuraciones discursivas que emergen desde la cobertura mediática de un acontecimiento.

En relación con el rol que desempeña la abducción en la construcción de hipótesis, se recupera la versión más madura de los planteos que al respecto propuso Charles S. Peirce (compendiados en Peirce, 1987), tal como han sido sistematizados por diversos autores (Genova, 1996; Santaella, 1998; Niño, 2001; Nubiola, 2001). En la definición que el lógico brinda en años posteriores al 1900, la abducción es, en tanto primer paso del razonamiento científico, el tipo de inferencia mediante la cual se proponen hipótesis (plausibles y probables) para explicar hechos observados.

Por otra parte, en lo que atañe a las técnicas de minería de datos involucradas, se comienza por recordar que el PLN se ocupa del análisis computacional del habla y la escritura en lenguas naturales (español, inglés, etc.), a diferencia de los lenguajes artificiales, creados deliberadamente para propósitos específicos, como el lógico-matemático, químico o musical, por mencionar algunos. Su objetivo es extraer información de los textos mediante algoritmos capaces de reconocer estructuras léxicas, gramaticales y sintácticas, así como realizar tareas más complejas tales como traducción automática o generación de texto (Jurafsky y Martin, 2024).

La minería de datos intenta, a su vez, descubrir patrones, regularidades o relaciones significativas dentro de grandes volúmenes de información, a menudo no detectables mediante la observación directa o contrarias a la intuición (Zaki y Meir, 2020). Siendo así, en el presente trabajo se aplicaron de modo combinado técnicas de PLN y minería de datos. Se utilizaron *embeddings* para transformar los textos en representaciones vectoriales de alta dimensionalidad a partir de los cuales, por ejemplo, es posible correlacionar su proximidad geométrica con significados semánticos similares. También se empleó *Uniform Manifold Approximation and Projection* (UMAP), que reduce la dimensionalidad de esos espacios vectoriales, facilitando su visualización y, de esta manera, dando acceso a una manera distinta de observar las unidades de análisis comprendidas en el corpus y sus relaciones.

Para identificar agrupamientos entre vectores representados en un espacio de baja dimensionalidad y permitir, en este caso, detectar regularidades o diferencias significativas en conjuntos textuales se optó por el algoritmo *k-means*, diseñado por MacQueen (1967). Mediante aprendizaje no supervisado el mismo opera sin etiquetar los datos y permite agruparlos según proximidad obtenidas a través de cálculos matemáticos.

Si bien *k-means* actúa minimizando la distancia hacia centroides de los vectores conformando *clusters*, cuando se aplica sobre *embeddings* estas distancias pueden adquirir una dimensión semántica. En otras palabras, los vectores que representan fragmentos textuales próximos en un espacio tienden a compartir significados, temas o posiciones enunciativas semejantes. De este modo, la "clusterización" puede comprenderse como una aproximación a los textos semántico-geométrica, donde las configuraciones espaciales resultantes, siguiendo la hipótesis planteada en el apartado anterior, no sólo cuantifican diferencias o similitudes textuales, sino que también habilitan la formulación de nuevas interpretaciones provisorias, al menos en el nivel que comprende a los géneros discursivos presentes en los textos, y por qué no, además, sobre los modos de enunciación. A partir de la comprensión del funcionamiento de los algoritmos y de la representación gráfica de sus resultados, entonces, se configura una lectura geométrica fundada en relaciones de distancia que se ponen en correlación con las categorías semióticas. De este modo, en la interpretación de visualizaciones de resultados se articula interpretativamente un nivel computacional con otro semiótico.

Siguiendo a Verón (2004), entendemos que nuestro objeto de análisis (es decir, los discursos mediatizados en cuestión) es un objeto multideterminado, porque sus "propiedades resultan del entrecruzamiento de una pluralidad de niveles de determinación diferentes" (p. 201). Precisamente, uno de esos niveles es, tal como ampliamos en el apartado 3.4, el correspondiente

a las delimitaciones genéricas, que atañen a "las clasificaciones sociales de lo discursivo" (Fernández, 2023, p. 34). Y la relevancia y centralidad que implica la detección de rasgos de genericidad ha sido ya argumentada suficientemente por referentes locales como Oscar Steimberg (1998) o Elvira Narvaja de Arnoux (2009). Como explica Bitonte (2025): "toda producción discursiva se enmarca en un género, y los medios de comunicación social construyen la realidad social a través de ellos. Los géneros son formas de categorización, y categorizar -como nos enseñó Peirce- es una capacidad semiótica que nos permite ordenar los modos de ser y de conocer las cosas" (p. 178).

3. Procedimientos e implementaciones

3.1 Construcción del corpus

Como se anticipó en la introducción, el corpus de análisis procesado se conforma con 596 publicaciones provenientes de seis portales de información general, tres con asiento en la ciudad de Rosario y otros tres de Buenos Aires y con alcance nacional (*La Capital*, *El Ciudadano & y la región* y *Rosario3*, por un lado; *La Nación*, *Clarín* e *Infobae*, por el otro), junto a los comentarios que lectores-usuarios hicieron en los sitios a través de sus perfiles de Facebook. La construcción del dataset se llevó a cabo mediante un proceso mixto de recolección automatizada y selección manual, tal como se describe pormenorizadamente en Raimondo Anselmino et al. (2025).

Básicamente, se implementaron técnicas de *web scraping* adaptadas a cada portal, empleando las bibliotecas *Selenium*, para automatizar la navegación y la interacción *web*, y *Newspaper3k*, que facilitó la extracción estructurada de textos (título, bajada y cuerpo de cada nota). Asimismo, se extrajeron los comentarios vinculados en las notas de portales locales, procesando las secciones incrustadas (*iframes*) de Facebook[5].

La selección de notas para el estudio del caso se guió por un conjunto de sintagmas nominales relacionados a los hechos de violencia ocurridos en el período de tiempo designado. El sistema de selección así diseñado rechazó de manera adecuada las notas, pero el *threshold* marcó un sesgo hacia aceptar muchas como falsos positivos, dado que una lectura más minuciosa las calificaba como no pertinentes. Por ello, se procedió a filtrar manualmente el corpus recolectado. Tras este trabajo (que además involucró limpieza y depuración de los datos) se construyó un corpus estructurado en un archivo *.csv, al que puede accederse de modo abierto a través de Raimondo Anselmino y Gindin (2025). Allí, cada registro corresponde a alguna nota periodística con sus comentarios correspondientes, y cuyos atributos son: *medio* (portal fuente), *url* (enlace a la nota), *fecha* (YYYY-MM-DD), *titulo*, *bajada*, *cuerpo_texto*, y *comentarios_fb* (lista de comentarios asociados a la nota; esta es la única columna que contiene valores NaN).

3.2 Procedimiento algorítmico

En esta sección se explica cómo fueron transformadas las notas y los comentarios en vectores (*embeddings*), sobre los que se aplicó una reducción de dimensionalidad mediante UMAP, previa a la ejecución del algoritmo no supervisado *k-means*[6]. A medida que se despliega la

[5] Los *inline frame* de un sitio son elementos html que posibilitan la inserción de otro documento html dentro de la misma página web. Esta interpolación es producto y síntoma, a la vez, de la creciente complejidad de las mediatizaciones contemporáneas. Como señala Fernández (2021), estudiamos una época en la cual "los medios masivos tradicionales diseminan sus emisiones en diversas plataformas interaccionales. A su vez, las plataformas interaccionales aumentan su presencia en los medios masivos tradicionales" (306-307).

[6] El procesamiento de los datos, la generación de gráficos y los distintos *outputs* fueron realizados mediante *scripts* escritos en *Python*, implementados en *Jupyter Notebook* y diseñados para su ejecución en *Google Colaboratory*. Asimismo, el código fue desarrollado en el marco de la pasantía que Francisco Alomar llevó adelante en el CIM como estudiante de la Tecnicatura Universitaria en Inteligencia Artificial (TUIA) de la UNR (2024-2025), dentro del proyecto de investigación en el que se inscribe el presente artículo. Se puede acceder al código mediante el

secuencia de pasos, se presentan gráficos para mostrar los resultados del procesamiento computacional, como así también para justificar el número de *clusters* mediante la visualización de métricas que permiten hacer una elección más precisa de la cantidad.

3.2.1 Vectorización y reducción de dimensionalidad

La vectorización se utilizó mediante la biblioteca *Sentence Transformer*. Los atributos *titulo*, *bajada*, *cuerpo_texto* y *comentarios_fb* de cada registro fueron convertidos a un vector de 512 dimensiones. Posteriormente, los *embedding* se normalizaron con *Min-Max Scaling* al rango [-1, 1], con el objetivo que todas las dimensiones incidan de manera equitativa durante el proceso de *clustering* de *k-means*, dado que el algoritmo se basa en distancias euclídeas y es sensible a la escala de los datos. Es decir, la normalización permite comparar vectores muy disímiles en sus magnitudes.

El siguiente paso fue la aplicación de UMAP, seleccionado por su capacidad para conservar tanto las relaciones locales como globales del espacio original y "desplegar" estructuras complejas en dimensiones menores (Géron, 2023).[7] Reducir la dimensionalidad se torna fundamental por dos razones. En primer lugar, para mitigar los efectos de la llamada *maldición de la dimensionalidad* (Bellman, 2003), fenómeno por el cual, en espacios de alta dimensión la distancia entre puntos tiende a no diferenciarse, dificultando la detección de agrupamientos significativos. En consecuencia, los centroides de *k-means* perdían representatividad y los *clusters* serían menos precisos. En segundo lugar, la reducción permite obtener representaciones que favorecen la comprensión visual, ya que un espacio de 512 dimensiones resulta inabordable gráficamente. Al proyectar los vectores en dos y tres dimensiones, es posible examinar rápidamente la estructura interna de los datos, sus relaciones, lo que se liga directamente con la hipótesis planteada.

Para las notas de portales, se calculó un único vector promedio combinando el título, la bajada y el cuerpo del texto, de modo que cada noticia represente un solo punto y pueda compararse directamente con los comentarios realizados a través de Facebook. También, todos los comentarios que aparecieron en las notas fueron vectorizados en un único vector de 512 dimensiones, y reducidos dimensionalmente mediante UMAP, siguiendo el criterio anterior de poder ser graficados, en su conjunto, mediante un único punto. A partir de este procesamiento, se obtuvieron los gráficos sobre el conjunto de notas dispuestos en las Fig. 1 y 2.

---

siguiente enlace: https://github.com/Fran251184/pasantia_TUIA_CIM_Francisco_J._Alomar/tree/main

[7] La calidad de la reducción que posibilita UMAP se evaluó mediante la métrica *trustworthiness* (*Scikit-Learn*, módulo *manifold*), que calcula en qué medida las vecindades originales se preservan tras la proyección. Valores cercanos a 1 indican una alta fidelidad estructural. Tras comparar otros métodos, como *Principal Component Analysis* (PCA), *t-Distributed Stochastic Neighbor Embedding* (t-SNE) e *Isometric Mapping* (Isomap), UMAP mostró un mejor desempeño en esta métrica, obteniendo en todos los gráficos un *score* mayor a 0.9.

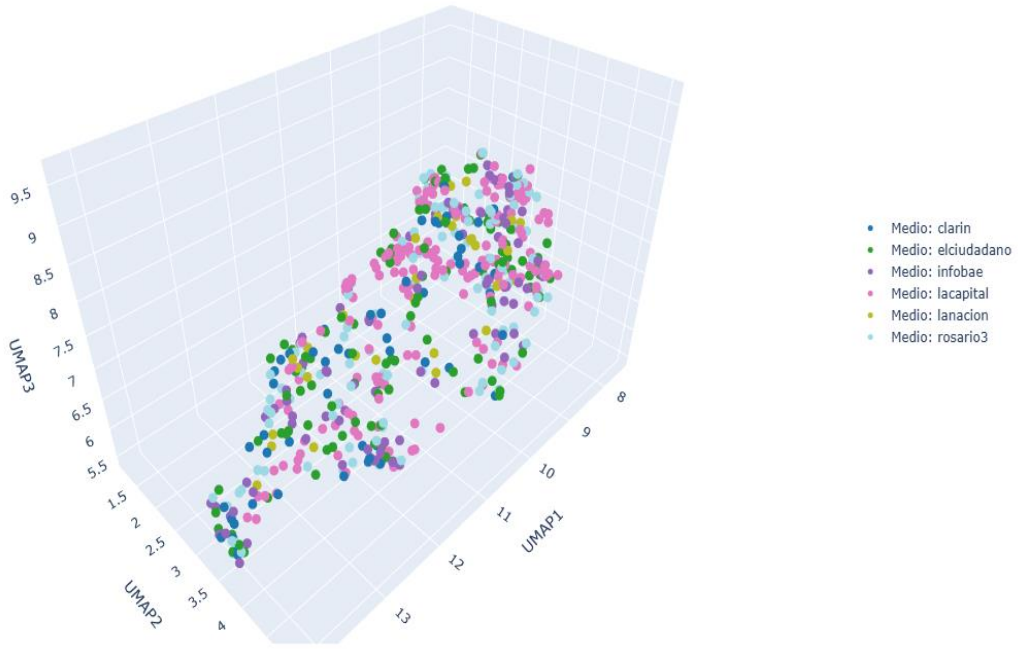


Fig. 1 - Reducción dimensional a 3 componentes mediante UMAP, con agrupamiento de vectores por medio.

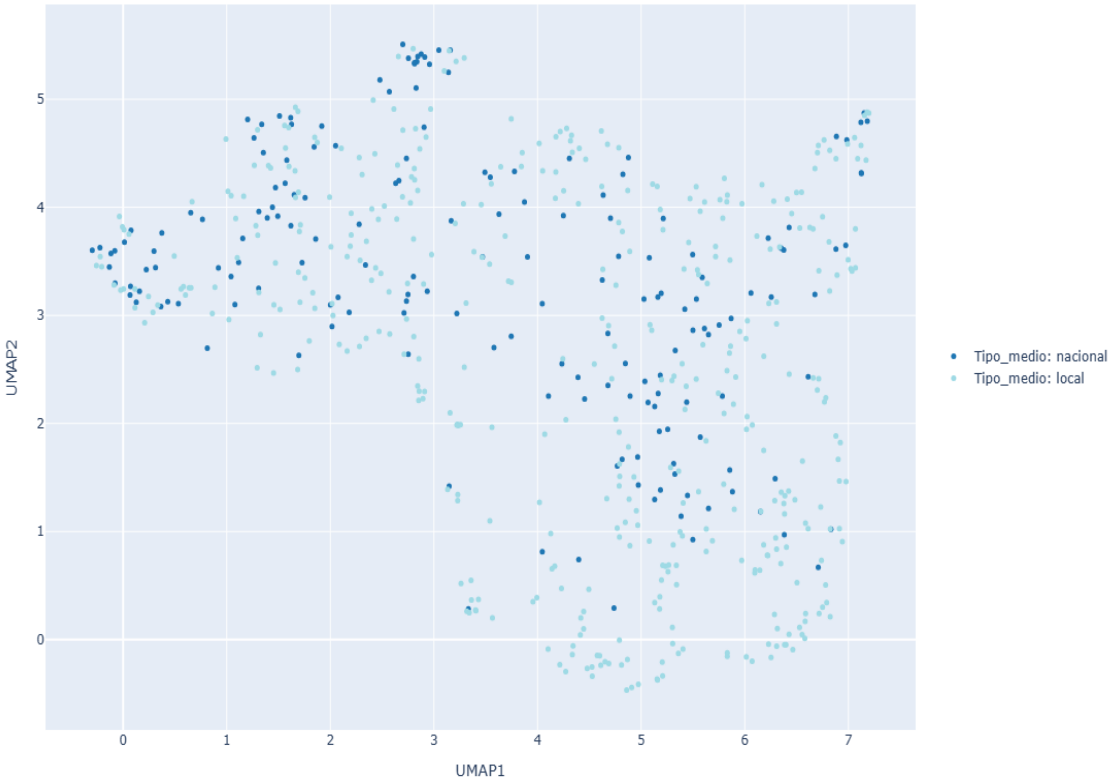


Fig. 2 - Reducción dimensional a 2 componentes mediante UMAP, con agrupamiento de vectores por tipo de medio (nacional/local).

Los gráficos resultantes, en tres dimensiones (3D) con un etiquetado por medio relevado (Fig. 1) y en dos dimensiones (2D) con un etiquetado por tipo de medio (Fig. 2), muestran las distribuciones obtenidas. El etiquetado inicial, por medio y tipo de medio, lo denominamos "trivial" dado que a los fines del procesamiento no permite interpretar cuestiones que distingan claramente conjuntos de datos. Sí podemos, por ejemplo, establecer a partir de la Fig. 2 que la cobertura nacional fue considerablemente menor que la local, lo cual si bien es un dato inferible no deja de ser una obviedad, porque de modo regular la producción mediática local es mayor a la nacional respecto de lo que concierne a la ciudad de Rosario, como así también, cuestiones internacionales tendrán menor tratamiento que las nacionales. Ahora bien, podríamos vislumbrar el sentido: "esto nos sucede a nosotros" (local, propio, consecuencia, mayor producción), versus "esto les ocurre a ellos" (nacional, ajeno, consecuencia, menor producción). Por su parte, el gráfico de la Fig. 1 permite hacer menos interpretaciones, aunque sí se evidencia que la variable *tipo de medio* no explica ninguna diferenciación en el conjunto, sino más bien una mezcla de las notas donde, tal vez, sólo se distingue que los portales *La Capital* y *Rosario3* han tenido una cobertura relevante de los acontecimientos en comparación con el resto.

3.2.2 Implementación de *k-means* y evaluación de métrica para la determinación del número de *clusters*

El algoritmo *k-means* se utilizó para dividir los textos en conjuntos vectoriales según su proximidad espacial. En este método, el parámetro *k* representa la cantidad de *clusters* que el modelo debe formar. A partir de *k* el algoritmo selecciona centroides iniciales y, mediante iteraciones sucesivas, asigna vectores a los más cercanos y actualiza las posiciones (de centroides) hasta alcanzar una estabilidad. El objetivo consiste en minimizar la suma de las distancias cuadráticas entre los puntos y centroides, optimizando la compactación interna de los *clusters*.

En la implementación se establecieron los siguientes hiperparámetros:

- *n_cluster:* cantidad de grupos a formar (*k*);
- *random_state:* garantiza reproductibilidad;
- *n_init=50:* ejecuta el algoritmo con diferentes inicializaciones para evitar mínimos locales;
- *init='k-means++':* mejora la convergencia, al seleccionar centroides iniciales de manera más eficiente.

La elección del número de *k* resulta crucial dado que define el grado de diferenciación y condiciona la interpretación posterior. Para estimar el valor más adecuado utilizamos diferentes métricas, guiados por el principio de parsimonia -usualmente conocido como *la navaja de Ockham*-, que orienta a preferir la explicación más simple y suficiente. Téngase en cuenta, además, que el principio de parsimonia o ley de simplicidad está presente también en la perspectiva semiótica peirceana como criterio que regula y acompaña la validación de las conjeturas interpretativas (Cfr. Eco, 1992, p. 99).

De esta manera, se evaluaron dos enfoques principales. En primer lugar, determinamos la inercia (que mide la compactación interna de los *clusters*) en función de *k*, graficamos la relación y aplicamos el método del codo. Mediante esta técnica se busca identificar el punto o zona de inflexión en la curva donde la reducción de la inercia deja de ser significativa (Géron, 2023) (Fig. 3).

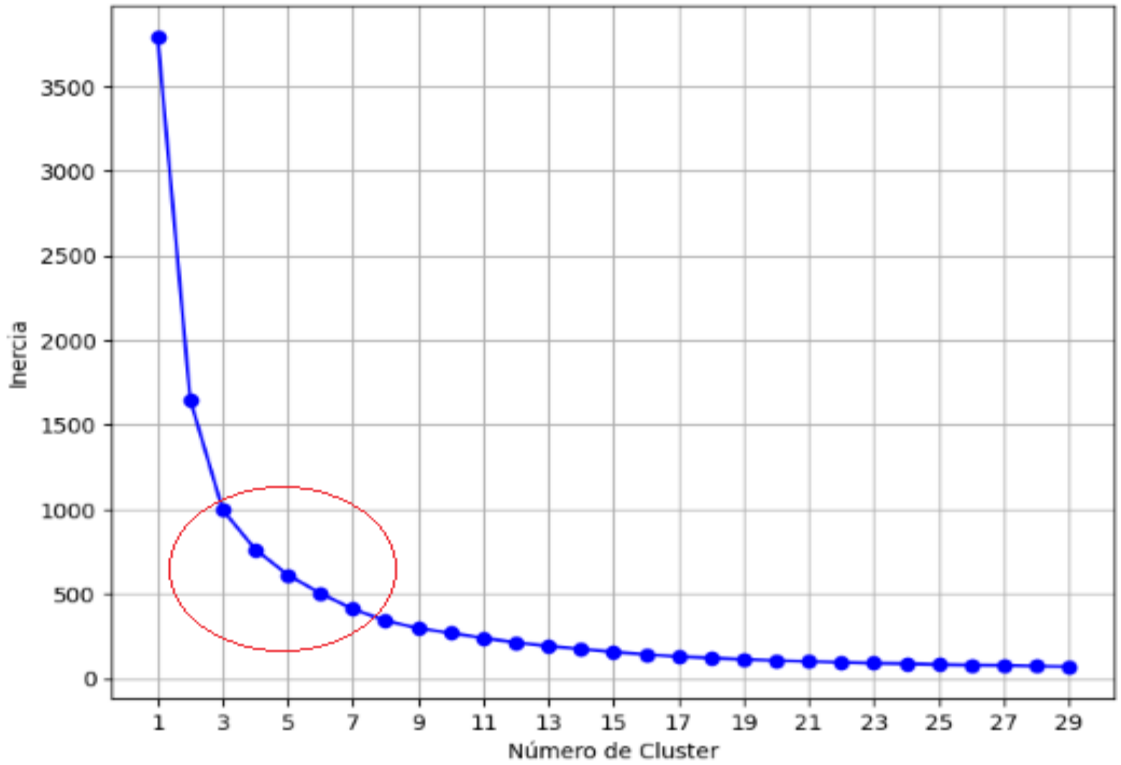


Fig. 3 - Método del codo para la determinación del número óptimo de *clusters*.

Como puede apreciarse en la figura Fig. 3, los puntos dentro de la elipse sugieren un rango razonable entre 3 y 7 clusters. Sin embargo, la inercia representa una limitación dado que siempre disminuye a medida que *k* aumenta, por lo que si bien ayuda a orientarnos hacia la elección de *k*, no resulta un criterio suficiente. Por esta razón, se complementa la evaluación con el segundo enfoque, la métrica *silhouette* (Fig. 4).

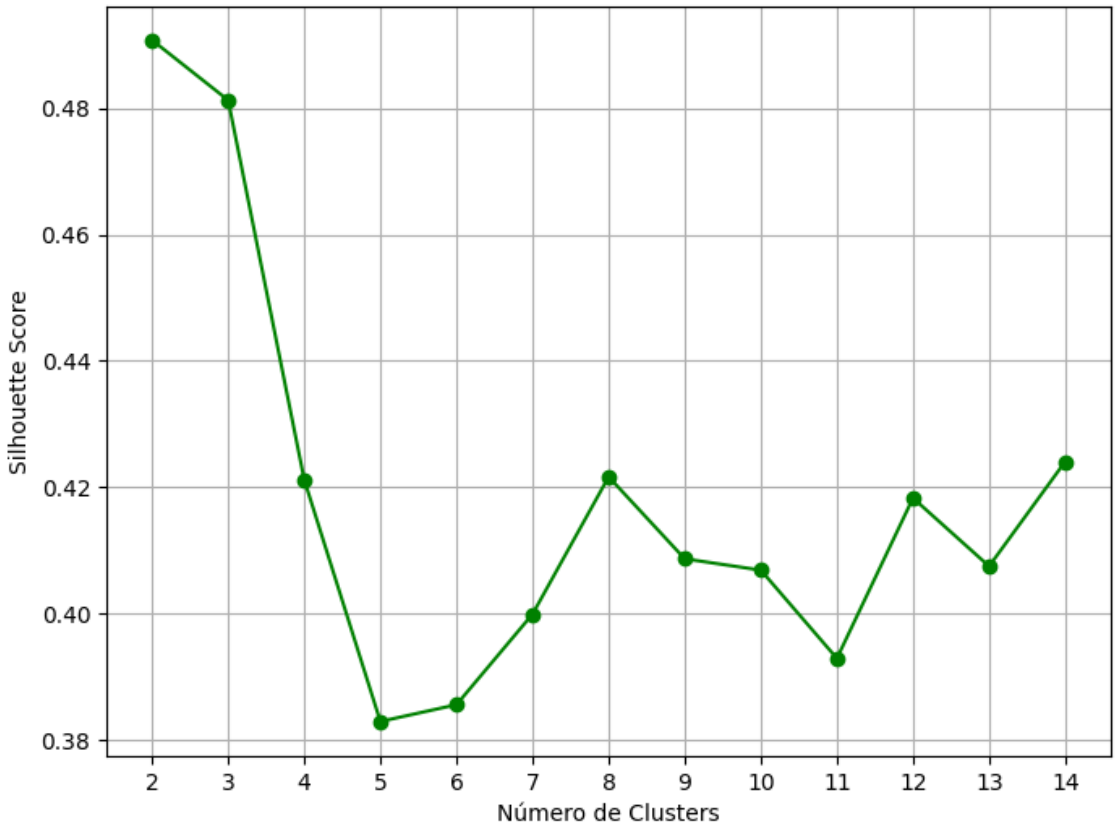


Fig. 4 - *Silhouette score* para la determinación del número óptimo de *clusters*.

A diferencia de la inercia, esta métrica considera simultáneamente la cohesión interna y la separación entre grupos. El rango de variación de *silhouette* es entre [-1,1], donde valores cercanos a 1 indican *clusters* bien definidos y distantes entre sí; alrededor de 0 sugieren solapamiento o ambigüedad entre los conjuntos. En cambio, valores negativos indican una separación deficiente o directamente errores en la asignación de vectores a centroides. Al observar el gráfico expresado en la Fig. 4, apreciamos que $k = 2$ presenta el valor máximo, aunque, comparado con $k = 3$, muestra una diferencia marginal (<0.02), por lo que nos inclinamos hacia la elección de 3 *clusters* según esta curva de *silhouette*.
En conjunto, los resultados de ambas métricas respaldan una elección de $k = 3$. Consideramos que este valor, a priori, denota un equilibrio adecuado entre simplicidad y diferenciación suficiente para hacer una interpretación de las agrupaciones de los textos del corpus (que, como señalamos antes, será desarrollada y discutida en otro artículo).

3.3 Interfaz exploratoria y visualización de datos

Habiendo elegido 3 *clusters* para el conjunto de notas de portales, al representarlos en dos dimensiones queda la distribución por conjuntos tal como es expresada en la Fig. 5.

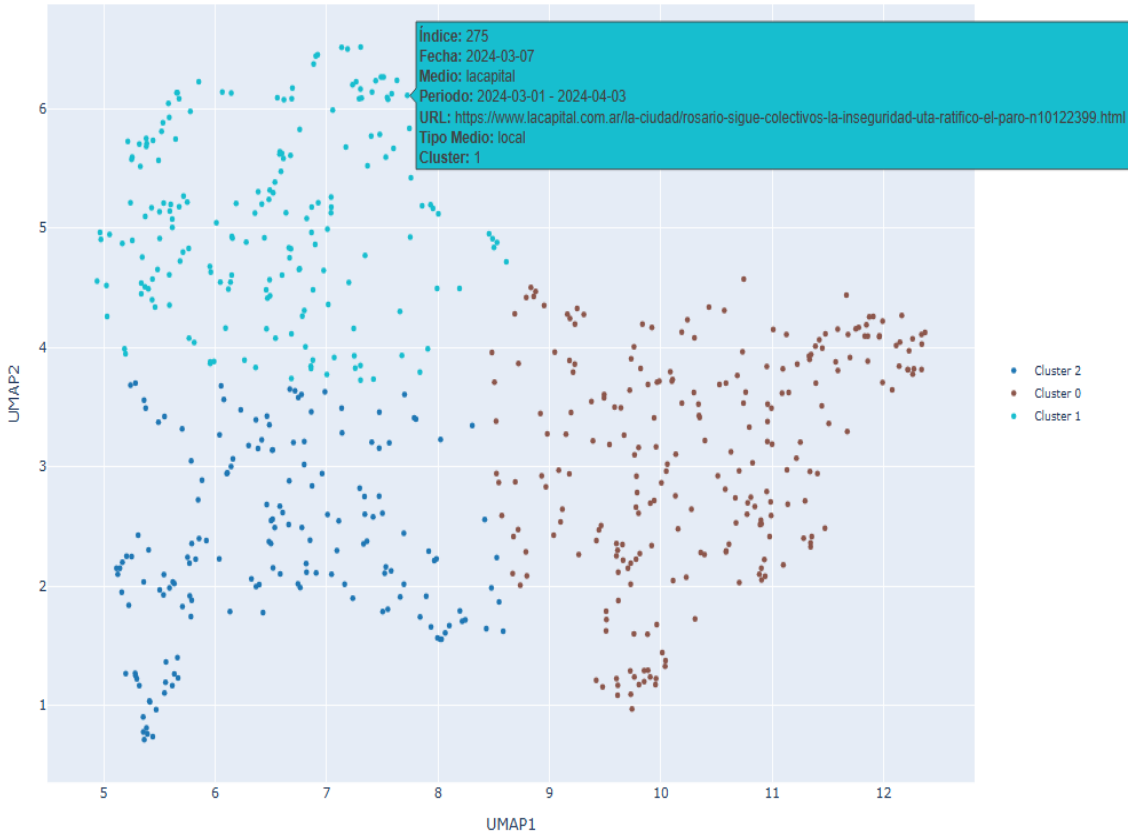


Fig. 5 - *Clustering* con k-means sobre *embeddings* promediados de notas de portales reducidos a 2 componentes mediante UMAP.

La partición obtenida en la Fig. 5 mediante k-means (k = 3), establece tres *clusters* en el espacio reducido por UMAP. Estas agrupaciones no deben comprenderse como categorías interpretativas establecidas, sino como regularidades producidas por el procesamiento algorítmico que habilitan una nueva instancia de observación del corpus. A diferencia de los etiquetados "triviales" (por medio, tipo de medio y fecha) que no permitieron identificar diferenciaciones significativas en la distribución de los textos, la formación de agrupamientos introduce una organización alternativa del conjunto que podría aportar indicios relevantes para el análisis interpretativo. En este sentido, los *clusters* funcionan como puntos de partida para la exploración abductiva de los textos, por lo que es necesario presentar la interfaz que se expresa a continuación.

Como se indicó en la introducción, uno de los objetivos propuestos fue construir una interfaz exploratoria de los datos para facilitar el acceso a (y la observación de) las textualidades e información que se relaciona a cada vector. La interfaz funciona del siguiente modo. Al posicionarse sobre cualquiera de los puntos graficados en la figura 5, emerge una ventana donde se despliegan metadatos de dicho vector. Para acceder al conjunto completo de datos asociados al vector, se ingresa su respectivo índice. Por ejemplo, al introducir la expresión "134", obtenemos el resultado dispuesto en la Fig. 6, donde se expone el cuerpo del texto correspondiente (que a los fines de la presentación de este artículo ha sido resumido con la expresión [...]).

"

```
Ingrese un índice entre 0 y 594: 134

========================================
```

```
    INFORMACIÓN DEL ÍNDICE SELECCIONADO
==========================================
Índice: 134
Fecha: 2024-03-10
Medio: elciudadano
Periodo: 2024-03-01 - 2024-04-03
URL:  https://www.elciudadanoweb.com/javier-milei-hablo-sobre-rosario-
estamos-enfrentando-a-un-grupo-de-narcoterroristas/
Tipo Medio: local
Número de Cluster: 0

==========================================
         INFORMACIÓN DEL ARTÍCULO
==========================================
Título:
Javier Milei habló sobre Rosario: “Estamos enfrentando a un grupo de
narcoterroristas”

Bajada:
El presidente de la Nación compartió una foto de Bruno Bussanich, el
playero asesinado a sangre fría el sábado, junto con un mensaje en el
que envió sus condolencias a las familias de las víctimas de esta
semana

Cuerpo del texto:
El presidente Javier Milei se pronunció sobre la escalada de violencia
vinculada al narcotráfico en Rosario. “No vamos a parar hasta que haya
Justicia”, escribió en un extenso posteo en su cuenta de X, en el que
habló específicamente del asesinato de Bruno, un playero de 25 años,
ejecutado de tres tiros. Además, culpó al “kirchnerismo y al
socialismo” de haber entregado la provincia de Santa Fe a la
delincuencia [...]

Comentarios de Facebook:
['Milei: Ud no esta capacitado ni tiene el critero para manejar esta
situacion. Ni su ministra ni el gobernador.Condolencias... uds son una
verguenza. Cinicos. Irresponsables. Pusilánimes', 'Más terrorista k
vos no hay□', 'Mister X , lo piensa arreglar por celular!!!', '20.000
pobres nuevos por dia entrega al "mercado" este enfermo psiquiatrico .
Incestuoso pervertido sin corazon y gran cipayo.', 'Míralo al loquito
psiquiátrico haciendo declaraciones ,y echando más nafta al fuego,
impresentable, che pullaro, javkin y cia que aliado que tienen y uds
le siguen chupando las bolas va a terminar mal también uds mentirosos
,']

==========================================
    □ SI QUIERE BUSCAR LA INFO DE OTRO ARTÍCULO, POR FAVOR EJECUTE
NUEVAMENTE EL CÓDIGO. □
==========================================
”
```

Fig. 6 - Ejemplo de *output* de inspección de nota de portal con metadatos, texto y comentarios de Facebook.

La interfaz arroja las textualidades relacionadas no sólo al índice de la nota ingresado, sino que además expone los comentarios realizados por lectores-usuarios a través de Facebook vinculados a ella, en el caso que los hubiere.

El procesamiento algorítmico generado también permite acceder visualmente a los comentarios de las notas, como así también a la relación y distancia espacial que mantienen con ellas, permitiendo en el gráfico interactivo hacer *zoom* y acercarse a cada punto representado (Fig. 7).

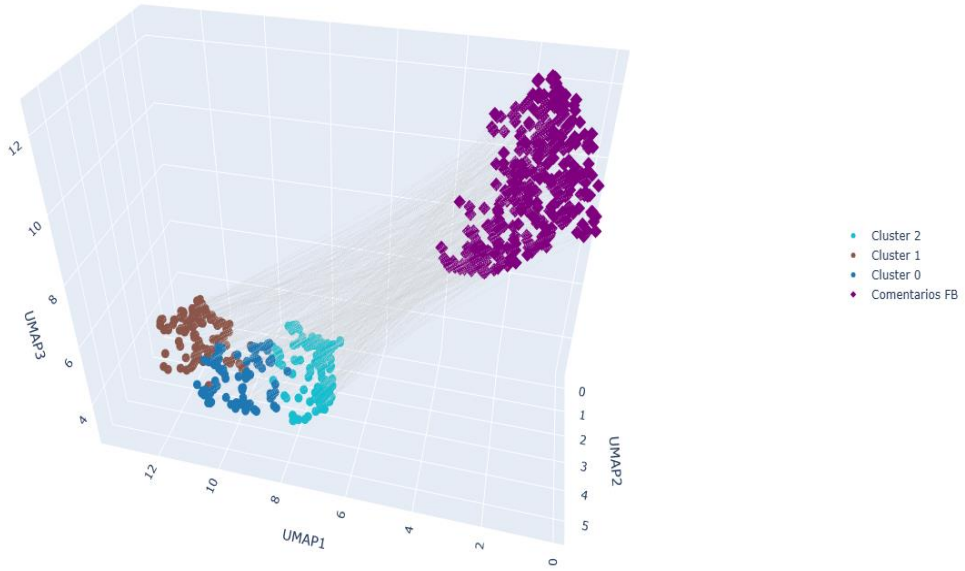


Fig. 7 - Proyección en tres componentes de *embeddings* reducidos con UMAP: *clusters* y comentarios de Facebook.

4. Primeras observaciones y elaboración de hipótesis para el análisis

En el punto 2 de este documento se ha recuperado, en conexión con el modo de conocer propuesto por los enfoques semióticos de base peirceana, el rol que desempeña la abducción en la construcción de hipótesis de trabajo, que se constituyen en impulsoras del quehacer científico. También se ha señalado, que dicho proceso inferencial se desencadena, creativamente, como respuesta a hechos observados que ameritan una explicación y se constituyen, por ello, en el estadio inicial de un proceso de investigación.

Retomando las implementaciones desplegadas en los apartados precedentes, en principio y tras los resultados de las interfaces gráficas y textuales es posible, a grandes rasgos, sostener al menos las siguientes cuestiones. En primer lugar, los comentarios de los lectores-usuarios se encuentran claramente distanciados del conjunto total comprendido por las notas publicadas en los portales. En segundo lugar y de modo contrario, los tres agrupamientos de notas periodísticas no tienen una separación tán nítida entre sí, denotando cercanía entre ellos y fronteras un tanto difusas.

Sería razonable pensar que la discriminación observada mediante dicha visualización (Fig. 7) puede deberse, en parte, a la extensión diferencial que existe entre uno y otro tipo de texto, en lo que hace a cantidad de caracteres. También podría adjudicarse esa distancia a propiedades inherentes al software mediante el cual se ha realizado la edición y publicación de cada agrupamiento de textos, puesto que los comentarios de los lectores-usuarios (a diferencia de las notas periodísticas) se insertan dentro de la página web de cada portal (tanto en el caso de *La Capital*, como *Rosario3* o *El Ciudadano & la región*) mediante un *iframe* de Facebook, y eso impacta inevitablemente en las condiciones de producción (Verón, 1998) que dejan *huellas* en las intervenciones del público de los medios online. Reparamos en esto, porque coincidimos con Zelcer (2021) cuando advierte (recuperando a Manovich, 2012) que el software intermediario (o

las aplicaciones) constituye un nivel de ultradeterminación, es decir, condiciona o regula (restringiendo y habilitando procedimientos) la producción discursiva que emplaza. Como explica el autor: cuando se trata de medios y soportes digitales, "las posibilidades y restricciones discursivas que se instauran provienen en buena medida del software o aplicación que interviene en cada caso, que regula los intercambios que pueden tener lugar, en relación con los distintos ejes (...): las materias de expresión, las presencias y ausencias del cuerpo, las construcciones temporales y espaciales y las estructuras del intercambio" (Zelcer, 2021: 344).

Ahora bien, es plausible (y probable mediante ensayos ulteriores) intuir, a la manera de un *insight* para Peirce (2024), que el algoritmo identificó automáticamente también como relevante ciertas disparidades invariantes ligadas a propiedades enunciativas de uno y otro conjunto. Podríamos agregar, incluso, que dichas propiedades se inscribirían en distintos tipos de géneros discursivos, si recuperamos la clásica distinción de Bajtín (1998) entre géneros discursivos primarios y secundarios, esto es, entre aquellos que son más sencillos, breves, espontáneos y con posible respuesta inmediata y otros detrás de los cuales hay un proceso de mayor elaboración.

Proponemos entonces, a partir de la configuración espacial obtenida en la Fig. 7, interpretar la distancia entre los dos conjuntos (notas y comentarios) como reflejo de su distancia genérica. Entendiendo al género discursivo no como una clasificación taxonómica cerrada, sino como un principio de organización textual socialmente reconocido (Steimberg, 1998). En este sentido, la distancia geométrica entre vectores podría leerse, en clave abductiva, como producto de las diferencias en los modos de configuración enunciativa dentro del corpus.

Por un lado, entonces, encontramos los comentarios, que para Sal Paz (2009) constituyen un género discursivo estructurante de los medios digitales, y ciertamente es un emergente de la mediatización contemporánea (Bitonte, 2025). En ellos se manifiesta una enunciación dialogal propia de este espacio de participación de los lectores en los periódicos digitales, donde además se hacen presentes atributos de la oralidad[8] e, incluso, es usual el empleo de emojis como recurso paralingüístico. Por otro lado, se visualizan las notas publicadas por los portales, que tienen un registro más impersonal propio del discurso de información de la prensa escrita y otra organización textual que es constante. Es decir, en el agrupamiento de unidades textuales correspondientes a las notas, se observa la presencia de la forma más típica de la enunciación periodística mediante la cual el medio se posiciona como un enunciador "objetivo" que da cuenta de un estado de cosas "sin resonancias orales de la escritura" (Steimberg, 2000, p. 240) y "produce informaciones sobre un registro impersonal" (Verón, 1985, s/p), dirigidas a un destinatario genérico.

En síntesis, sería razonable proponer que el conjunto de implementaciones algorítmicas efectuadas estaría discriminando paquetes textuales cuya diferencia puede atribuirse a uno de los niveles de determinación de los discursos sociales mediatizados que desde la sociosemiótica veroneana se ha designado como *géneros-L*. Un género-L, para Verón (2004), "se caracteriza necesariamente por cierta disposición de la materia lingüística" (196), pero a su vez se encuentra sobredeterminado, en otro nivel, por el *género-P* (que hace a la clasificación de productos mediáticos dentro de zonas de competencia directa, como los portales de noticia por ejemplo) que lo contiene y por el *tipo de discurso* en cuestión (asociado a un campo de la praxis social y a determinadas instituciones, como el político, el publicitario, el científico, el discurso de información, etc.).

Por otra parte, vale agregar que dentro del marco teórico adoptado, los géneros discursivos no se reducen a conjuntos de rasgos formales aislados, sino que articulan regularidades de composición, tematización y posicionamiento enunciativo que remiten a su vez a condiciones de producción y reconocimiento históricamente estabilizadas. Desde esta perspectiva, la relativa proximidad entre los tres subconjuntos de notas (visibles tanto en figuras 5 como 7) sugiere que, aún cuando el algoritmo distingue diferencias y regularidades internas, estas se inscriben dentro de una misma matriz aglutinada por el discurso informativo institucionalizado, que es propio de los portales periodísticos.

[8] Dice Gurevich (2018), en su estudio sobre Facebook: "el comentario es el lugar donde se aloja la voz" (p. 85).

Cabe reiterar, además, que el análisis pormenorizado acerca de qué pueden significar los agrupamientos obtenidos -y en qué medida pueden ser interpretados- excede los alcances del presente artículo y será desarrollado en una publicación posterior, cuando se aborde de manera específica la caracterización cualitativa de cada conjunto. Por ahora, es pertinente advertir que ciertas diferenciaciones internas propias del discurso periodístico -como aquellas tradicionalmente asociadas a modalidades informativas, interpretativas o de opinión que circunscriben géneros periodísticos- no se manifiestan inmediatamente con nitidez en la proyección obtenida. Esto, sospechamos, podría sugerir que no todas las clasificaciones genéricas estabilizadas resultan igualmente sensibles a (o han sido captadas por) esta implementación algorítmica.

Téngase en cuenta que los géneros periodísticos, particularmente, constituyen tipos relativamente estables de discurso de actualidad propios del periodismo de prensa desde fines del siglo XIX. Desde entonces, dichos géneros han asumido diferentes caracterizaciones y sus fronteras se fueron transformando progresivamente, situación que se ha hecho incluso más notoria desde la digitalización y puesta en línea de los medios tradicionales (Rončáková, 2017). En efecto, en los portales de prensa online es difícil encontrar publicaciones que continúen respetando la canónica estructura de pirámide invertida propia de una *noticia* (de Fontcuberta, 2011), siendo que antaño era una de las marcas distintivas del género. Aún así, sigue siendo posible distinguir una noticia de otros géneros periodísticos convencionales (como la *entrevista*, la *crónica* o la *columna de opinión*, entre otros).

En función de corroborar nuestra corazonada de que las implementaciones realizadas computacionalmente no registran diferencias relativas a los géneros periodísticos existentes al interior del paquete constituído por las 596 notas coleccionadas, como parte de este acercamiento exploratorio al corpus se efectuó la observación directa de algunas unidades pertenecientes a los 3 clusters visualizados en Fig. 5, accediendo a las mismas a través de la interfaz presentada en el punto 3.3 (que nos permite navegar en cada uno de los *clusters*). De esta manera, durante la observación directa de unas 30 piezas (alrededor del 5% del corpus) escogidas aleatoriamente se detectó que al menos 4 no se adaptan a las convenciones reconocidas para la clase noticia, sino que podrían, más bien, ser consideradas dentro de otro tipo de géneros periodísticos. En concreto, se trata de las unidades/vectores comprendidas por los índices 409, 444, 533 y 544[9]. Esa labor artesanal permitió verificar que, efectivamente, existe diversidad de géneros periodísticos al interior de la colección total de notas de portales. Todavía más, se pudo confirmar que esa variedad de géneros se encuentra dispersa entre los diferentes agrupamientos logrados tras la aplicación del algoritmo *k-means*. Por consiguiente, puede afirmarse que, en principio, las diferencias concernientes a las clases de textos habituales en el discurso informativo no explicarían las relaciones de distancia y proximidad computacionalmente reconocidas por las implementaciones.

Finalmente, y antes de concluir este apartado, es importante resaltar que la lectura que vincula proximidad o distancia espacial con diferenciaciones semánticas o genéricas no constituye una

[9] La primera pieza mencionada (https://www.lacapital.com.ar/edicion-impresa/organizaciones-criminales-que-mueven-los-hilos-un-grupo-chicos-n10127832.html) corresponde a *La Capital* y podría bien clasificarse como una *columna de opinión*. En segundo lugar, el vector con índice 444 es una publicación realizada por el mismo portal (https://www.lacapital.com.ar/suscriptores/milei-y-pullaro-exploran-alianzas-necesidad-y-urgencia-n10123938.html), que presenta cualidades similares al caso anterior. Los dos últimos vectores registrados, 533 (https://www.rosario3.com/informaciongeneral/Cuanto-tienen-que-saber-los-ninos-y-ninas-sobre-la-ola-de-terror-en-Rosario-20240311-0036.html) y 544 (https://www.rosario3.com/politica/Maria-Eugenia-Vidal-apunto-al-origen-de-los-fondos-y-al-lavado-del-dinero-narco-en-Rosario-20240323-0027.html) contienen notas de *Rosario3*, que corresponderían en ambos casos al género periodístico definido como *entrevista*. Por un lado, el diálogo mantenido por el portal con una psicopedagoga consultada en calidad de especialista, quien brinda su mirada sobre el impacto emocional que los acontecimientos provocan en las y los niños de la ciudad; por otro lado, la transcripción de una entrevista radiofónica realizada sobre el tema a la diputada nacional María Eugenia Vidal.

extrapolación metodológica automática. Se trata de un efecto específico de esta implementación técnica articulado con nuestra interpretación sociosemiótica de los resultados. En consecuencia, los *clusters* deben ser entendidos como configuraciones espaciales que operan como disparadores abductivos para la formulación de hipótesis sobre diferenciaciones discursivas existentes en el corpus.

5. Apostilla de cierre sobre la aproximación presentada

Nuestros objetos de análisis constituyen grandes paquetes textuales digitales intersecados y configurados por múltiples lógicas, entre ellas la derivada del hecho (no menor) de que se trata de discursos que circulan a través de plataformas con base en Internet y, por lo tanto, sería ingenuo desconocer, entre esas determinaciones, el nivel del software que opera como intermediario (Manovich, 2012). Siendo así, la estrategia metodológica definida como semiodata (Raimondo Anselmino, 2022) parte de reconocer la necesidad de incorporar (tal como se ha hecho en nuestro caso) procedimientos computarizados al proceso de investigación por dos razones fundamentales. En primer lugar, para tener acceso a atributos o detalles de los discursos-objeto que son imperceptibles a simple vista, ya sea porque se encuentran en el nivel de la infraestructura subyacente, o bien porque ligan detalles que (aunque visibles para el ojo humano) pasan desapercibidos por su dispersión en el gran volumen de discursos estudiados, siendo entonces regularidades de significado registrables "solo en una escala mayor" (Campagno y Treleani, 2019: 8).

La segunda razón esgrimida, es que la semiodata se orienta a vehiculizar y agilizar el trabajo empírico sobre paquetes textuales que tienen dimensiones o una magnitud de difícil maniobra para el hacer semiótico convencional. Por lo tanto, el tipo de implementaciones como las aquí presentadas pretenden constituirse en un aporte metodológico que no sólo sofistica los procedimientos de construcción y procesamiento de un corpus, sino que también aspira a contribuir con el reconocimiento de patrones (en el sentido de advertir regularidades significativas) sobre un conjunto voluminoso de discursos.

Este enfoque híbrido supone, además, integrar (calibradamente) diferentes niveles de observación de los discursos digitales estudiados entre los que, por el momento, es posible discriminar al menos tres. Por un lado, la ineludible observación "cercana" y directa de las unidades de análisis bajo estudio en su *puesta en discurso* (Verón, 1998), algo que sólo es posible efectuarse sobre una cantidad relativamente acotada de piezas. Por otro lado, un segundo nivel que implica observar los discursos datificados, es decir, estructurados en *datasets* (que a su vez pueden asumir diferentes formatos, sobre los que por el momento no nos explayaremos). Por último, lo desplegado en este artículo ha permitido mostrar que las implementaciones que integran minería de datos y PLN, particularmente la vectorización textual mediante *embedding* y reducción de dimensionalidad (UMAP) proyectadas en visualizaciones gráficas, resultan fecundas para la elaboración de hipótesis que distingue géneros discursivos, notas de portales y comentarios de Facebook, a partir de la correlación hallada entre la distancia de conjuntos de vectores y la interpretación de sus funcionamientos discursivos.

Paralelamente, no desconocemos que este enfoque híbrido tiene también sus debilidades y limitaciones, entre las que al menos podemos enumerar algunas evidentes. En primer lugar -y tal como notamos tempranamente en Raimondo Anselmino, Cardoso y Rostagno (2018)- lo engorroso de constituir y sostener (sobre todo en un contexto de desfinanciamiento del sistema científico argentino) un equipo multidisciplinario que cuente con investigadoras e investigadores expertos en diferentes disciplinas y capacidad para cooperar, quienes además puedan colaborar empleando un metalenguaje común. En segundo lugar, la necesidad de sortear las resistencias institucionales que suelen aparecer en distintos campos del saber (tantos aquellos pertenecientes a las ciencias sociales como los tradicionalmente ubicados dentro de las llamadas "ciencias exactas") cuando se propone revisar o discutir ciertas convenciones o hábitos imperantes, e incluso, las dificultades de romper con los propios automatismos incorporados en las trayectorias de quienes conforman los equipos de trabajo. En tercer lugar, la inestabilidad que supone la aplicación de las técnicas y los procedimientos computacionales, dadas las constantes transformaciones a las que son sometidas (a un ritmo cada vez más vertiginoso) las plataformas mediáticas cuyos discursos pretendemos estudiar (situación a la que no son ajenos

los portales de noticia). En cuarto lugar, los múltiples puntos ciegos implicados en el empleo de algoritmos cuyo funcionamiento (en algunos casos, como puede ser el de UMAP, por ejemplo) no resulta plenamente transparente o intuitivamente reconstruible, lo cual no quita valor a su empleo, pero repercute en la entidad (por ahora experimental) que se otorga a los resultados producidos, cuyas determinaciones no siempre podemos explicar.

Para concluir, y en relación con las tareas pendientes para la indagación en torno a las notas que sobre el "narcoterrorismo" en la ciudad de Rosario se recolectaron en los portales de noticia bajo estudio, adelantamos que el próximo paso procurará la interpretación sociosemiótica de los tres agrupamientos (*clusters*) reconocidos mediante el procesamiento con *k-means* que ha sido ilustrado en la Fig. 5, para posteriormente caracterizar las estrategias enunciativas empleadas por la prensa online de información general respecto del tópico en cuestión.

6. Referencias

Aliseda, A. (1998) La abducción como cambio epistémico: C. S. Peirce y las teorías epistémicas en inteligencia artificial. *Analogía*, 12, pp. 125-144.

Archenti, N. (2018) Estudio de caso/s. En Marradi, A., Archenti, N. y Piovani, J.I. Manual de metodología de las ciencias sociales. Buenos Aires: Siglo XXI.

Bajtín, M. (1998). *El problema de los géneros discursivos. Estética de la creación verbal*. Buenos Aires: Siglo XXI.

Bartolomé, M. y Ventura Barreiro, V. (2019) Narcotráfico en América del Sur más allá del bloque andino: los casos de Argentina y Brasil. Revista de Relaciones Internacionales, Estrategia y Seguridad, 14(1). Doi: https://doi.org/10.18359/ries.3760

Bellman, R. (2003). *Dynamic Programming*. Nueva York: Courier Dover Publications.

Bitonte, M.E. (2025) Medios, tipos y géneros discursivos en la sociedad posdigital. vigencia y proyecciones de la Teoría de los Discursos Sociales. En Cingolani, G. y Cossia, L. (eds.) *Eliseo Verón, el desvío del sentido*. Rosario: Editorial UNR.

Bovino, B. (2016) Aspectos socioculturales del narcotráfico: ayer Medellín, hoy Rosario. I+D Revista de Investigaciones, 8(2), pp. 46-53.

CELS (2018) *La guerra interna: cómo la lucha contra las drogas está militarizando América Latina*. Buenos Aires: Centro de Estudios Legales y Sociales, CELS.

Compagno, D. y Treleani, M. (2019) Introduction to Meaningful Data/Données signifiantes. *Semiotica*, Vol. 2019, nº 230, 1-17. https://doi.org/10.1515/sem-2018-0141

De Fontcuberta, M. (2011). *La noticia. Pistas para percibir el mundo*. Paidós.

Eco, U. (1992) *Los límites de la interpretación*. Buenos Aires: Lumen.

Fernández, J.L. (2021) *Vidas mediáticas. Entre lo masivo y lo individual*. Buenos Aires: La Crujía.

Fernández, J.L. (2023) *Una mecánica metodológica para el análisis de las mediatizaciones*. Buenos Aires: La Crujía.

Géron, A. (2023). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). O'Reilly Media. Recuperado de: https://www.clc.hcmus.edu.vn/wp-content/uploads/2017/11/Hands_On_Machine_Learning_with_Scikit_Learn_and_TensorFlow.pdf (27/02/2025)

Génova, G. (1996). Los tres modos de inferencia. A*nuario Filosófico XXIX/3 del Grupo de Estudios Peirceanos*, UNAV (pp. 1249-1263), Pamplona. Disponible en: http://www.unav.es/gep/AF/Genova.html

Gindin, I., Aráoz, V. y Raimondo Anselmino, N. (2026) Metodología de trabajo sobre un corpus. Recolección, procesamiento y descripción de posteos en Facebook e Instagram.

Iazzeta, M. y Gaiero, M. (2025) Narcotráfico y política. Gestión de la seguridad pública y violencia en Rosario, Argentina (2013-2015). URVIO. Revista Latinoamericana de Estudios de Seguridad, N.° 41, enero-abril. Doi. doi.org/10.17141/urvio.41.2025.6277

Jurafsky D. y Martin J.H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing,Computational Linguistics, and Speech Recognition*. University of Colorado at Boulder, EEUU. Recuperado de: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 5(1), 281-297. https://projecteuclid.org/proceedings/berkeley-symposium-on-mathematical-statistics-and-probability/proceedings-of-the-fifth-berkeley-symposium-on-mathematical-statistics-and/Chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992

Manovich, L. (2012) *El software toma el mando*. Traducción de *Software Takes Command* por Everardo Reyes-García.

Mohammed J. Zaki, Wagner Meira (2020). *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Jr. Cambridge University Press. 2do Edition. Recuperado de: https://pzs.dstu.dp.ua/DataMining/bibl/mohammed_j_zaki_wagner_meira_jr_data_mining_and_analysis_fun.pdf

Musante, M.C., Echecopar, C., Raimondo Anselmino, N. (2025) Matriz para análisis discursivo de intervenciones sobre asuntos públicos en Whatsapp. *Animus, Revista Interamericana de Comunicação Midiática*, 24(52) e024009, publicación continua. https://doi.org/10.5902/2175497793501

Narvaja de Arnoux, E. (2009) El análisis del discurso como campo interdisciplinario. En *Análisis del discurso. Modos de abordar materiales de archivo*. Buenos Aires: Santiago Arcos.

Niño, D. (2001). Peirce, abducción y práctica médica. A*nuario filosófico XXXIV/1 del Grupo de Estudios Peirceanos*, UNAV, Pamplona. Disponible en www.unav.es/gep/AF69/AF69Ninio.pdf

Nubiola, J. (2001) La abducción o lógica de la sorpresa. *Razón y Palabra*, Nº 21. Disponible en: http://www.razonypalabra.org.mx/anteriores/n21/21_jnubiola.html

Peirce, C.S. (1987) *Obra Lógico-Semiótica*. Madrid: Taurus.

Peirce, C.S. (2024 [1903]) Cómo teorizar (sobre la selección de hipótesis). En *Claves semióticas*. Compilación de Sara Barrena. Buenos Aires: Cactus.

Raimondo Anselmino, N., Cardoso, A. y Rostagno, J. (2018) Articulación artesanal-computacional para el estudio interdisciplinario de posteos en cuentas de Facebook. Relato de una experiencia asequible. *Simposio Argentino sobre Tecnología y Sociedad, 47º JAIIO (Jornadas Argentinas de Informática)*, Buenos Aires, pág. 128-140. Disponible en: http://47jaiio.sadio.org.ar/sites/default/files/STS-12.pdf (06/04/2026).

Raimondo, F., Rostagno, L., & Cardoso, A. (2021). Aplicación de técnicas de clustering para el estudio sociosemiótico sobre géneros periodísticos en fanpages de Clarín y La Nación. *Del prudente saber y el máximo posible de sabor*, 23(14), 77-103. Recuperado de: https://pcient.uner.edu.ar/index.php/dps/article/view/1137/1294 (27/02/2025).

Raimondo Anselmino, N. (25 de julio de 2022). Semiodata: una estrategia combinada para el estudio de la producción de sentido en el estadio actual de la mediatización. Conferencia realizada en el marco del Seminario "Arqueologías de la mediatización. Tiempos, espacios y tecnologías del mundo actual". https://www.youtube.com/watch?v=7rP8k6YpHvo

Raimondo Anselmino, N. y Gindin, I. (2025) *Noticias relacionadas con narcoterrorismo en Rosario, Argentina (Marzo a Mayo de 2024)* [conjunto de datos]. CONICET. https://ri.conicet.gov.ar/handle/11336/260282

Raimondo Anselmino, N., Gindin, I., Sartorio, A., Cellone, F., Aráoz, V. y Alomar, F. (2025) Herramientas y técnicas computarizadas al servicio de la construcción de corpus y datasets. Desafíos metodológicos del análisis de discursos mediatizados en plataformas con base en Internet. En Aguilera Ontiveros, A. (coord.) *Nuevas fronteras en las ciencias sociales computacionales: Análisis del discurso, la opinión y la dinámica social*. México: AMIDI.

Rončáková, T. (2017). Contemporary short-form genres in weekly print media. *Informatologia*, 50 (3-4), 151-161. https://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=283242

Sal Paz, J. (2009) Comentario de lector: género discursivo estructurante de los "nuevos medios". *Actas del IV Coloquio Argentino de la IADA. Diálogo y diálogos*. 1 al 3 de junio, La Plata.

Santaella, L. (1998) La evolución de los tres tipos de argumento: Abducción, inducción y deducción. *Analogía Filosófica XII/1 del Grupo de Estudios Peirceanos*, UNAV, Pamplona. Disponible en http://www.unav.es/gep/AN/Santaella.html

Sampó, C. (2017) Narcotráfico y trata de personas, una muestra de cómo el crimen organizado avanza en argentina. Revista de Relaciones Internacionales, Estrategia y Seguridad, 12(2). Doi: https://doi.org/10.18359/ries.2774

Steimberg, O. (1998) Proposiciones sobre el género. En *Semiótica de los medios masivos. El pasaje a los medios de los géneros populares*. Buenos Aires: Atuel.

Steimberg, O. (2000) Naturaleza y cultura en el ocaso (triunfal) del periodismo amarillo. *CIC Cuadernos de Información y Comunicación*, 5.

Verón, E. (1985). El análisis del "Contrato de Lectura". Un nuevo método para los estudios de posicionamiento de los soportes de los media [Traducción al español de: L'analyse du «contrat de lecture»: une nouvelle méthode pourles études de positionnement des supports presse]. En Touati, E. *Les Medias: Experiences, recherches actuelles, aplications*. París: IREP.

Verón, E. (1998) *La semiosis social. Fragmentos de una teoría de la discursividad*. Buenos Aires: Gedisa.

Verón, E. (2004) Prensa gráfica y teoría de los discursos sociales: producción, recepción, regulación. En *Fragmentos de un tejido*. Buenos Aires: Gedisa.

van Dijck, J., Poell, T., y de Waal, M. (2018). The Platform Society. Public values in a connective world. Oxford University Press.

Yates, Caitlyn, "El crecimiento de narcotráfico en Rosario: violencia disciplinada y la resistencia social frente a este sistema / The Growth of Narcotrafficking in Rosario: Disciplined Violence and the Social Resistance to Resist this System" (2014). Independent Study Project (ISP) Collection. 2005. https://digitalcollections.sit.edu/isp_collection/2005

Zelcer, M. (2021) *Devenires de lo fotográfico. Imágenes digitales en los dispositivos contemporáneos*. Buenos Aires: Teseo.