Aplicación de técnicas de *clustering* para el estudio sociosemiótico sobre géneros periodísticos en *fanpages* de *Clarín* y *La Nación*

NATALIA RAIMONDO ANSELMINO | UNIVERSIDAD NACIONAL DE ROSARIO Y UNIVERSIDAD ABIERTA INTERAMERICANA-CONICET
nraimondo@conicet.gov.ar
ORCID 0000-0001-7217-8754

JOSÉ ROSTAGNO | UNIVERSIDAD TECNOLÓGICA NACIONAL
jrostagno@frro.utn.edu.ar
ORCID 0000-0001-8346-0261

ANA LAURA CARDOSO | UNIVERSIDAD TECNOLÓGICA NACIONAL
acardoso@frro.utn.edu.ar
ORCID 0000-0002-2867-3317

https://doi.org/10.33255/26184141/1137

Resumen

Se presentan en este artículo los hallazgos obtenidos mediante la aplicación de técnicas de clustering provistas por la minería de texto sobre un corpus de posteos de las cuentas que los diarios argentinos *Clarín* y *La Nación* poseen en Facebook: @clarincom y @lanacion. Se trata de un acercamiento que, a modo de ensayo, pone a prueba la riqueza de los métodos computacionales para acompañar combinadamente el estudio sociosemiótico de las publicaciones que los periódicos realizaron en sus respectivas *fanpages* entre los años 2010 y 2017.

El análisis planteado se efectuó sobre un conjunto de posteos clasificados como «Otros» en la variable género periodístico, la cual discrimina si el contenido al que el posteo reenvía es una noticia, un reportaje, una nota de opinión, etc. Se pretende, así, identificar la existencia, al interior de dicho conjunto, de agrupamientos de publicaciones con características comunes —ya sean éstas derivadas de regularidades temáticas, retóricas, enunciativas o de otra índole— no detectadas, previamente, durante la observación convencional. Todo ello en el marco de una investigación mayor, de tipo interdisciplinaria, orientada a producir conocimiento sobre la manera en que los medios indagados enuncian en la plataforma de Facebook y el tipo peculiar de vínculo que le proponen, allí, a sus usuarios-lectores-seguidores.

Los resultados así obtenidos no sólo han permitido avanzar en la caracterización de posibles nuevas clases de textos, propias de la performance del discurso de los medios en la plataforma estudiada —que, de ahora en más, podrían sumarse a los *moldes de previsibilidad social* ya reconocidos—, sino además, identificar una falla en el diseño de los instrumentos creados, a priori, para la pesquisa en cuestión.

Palabras clave: discursos, géneros periodísticos, técnicas de *clustering*

Application of clustering techniques for the socio-semiotic study of journalistic genres on Clarín and La Nación's fanpages

Abstract

This paper presents the findings obtained through the application of clustering techniques provided by text mining on a corpus of posts from the accounts that the Argentine newspapers *Clarín* and *La Nación* have on Facebook: @clarincom and @lanacion. It is an approach that, by way of trial, test the wealth of computational methods to accompany in combination the socio-semiotic study of the publications that the newspapers made in their respective fanpages between 2010 and 2017.

The proposed analysis was carried out on a set of posts classified as "Others" in the journalistic genre variable, which discriminates if the content to which the post forwards is news, a report, an opinion note, etc. Thus, it is intended to identify the existence, within said set, of groupings of publications with common characteristics —whether these are derived from thematic, rhetorical, enunciative or other regularities— not previously detected during conventional observation. All this within the framework of a larger, interdisciplinary investigation, aimed at producing knowledge about the way in which the investigated media enunciate on the Facebook platform and the peculiar type of link that they propose, there, to their users-readers-followers.

The results thus obtained have not only made it possible to advance in the characterization of possible new classes of texts, typical of the performance of the media discourse in the studied platform —which, from now on, could be added to the already recognized social predictability molds— but also to identify a flaw in the design of the instruments created, a priori, for the research in question.

Keywords: discourses, journalistic genres, clustering techniques

INTRODUCCIÓN

Según Eliseo Verón, la novedad específica introducida por Internet consiste en haber producido una *revolución del acceso*, más específicamente, del acceso de los actores socio-individuales a los discursos que circulan en la red. Desde el inicio del proceso histórico de mediatización, «nunca antes el surgimiento de un dispositivo técnico de comunicación había provocado en tan poco tiempo movimientos que atraviesan a la vez los campos económico, tecnológico, político, social y cultural de nuestros viejos Estados-naciones» (Verón, 2013: 277).

Para las investigaciones semióticas sobre el discurso de los medios tradicionales de comunicación, antaño conocidos como «masivos», dicha transformación conlleva desafíos metodológicos específicos derivados de los constantes y profundos cambios que se fueron sucediendo a partir del surgimiento de Internet[1]. Desde entonces, como afirma Bunz (2017): «El ámbito del que debemos tener un panorama se ha vuelto inabarcable debido a su digitalización» (Bunz, 2017: 45). El grado de sofisticación y complejidad de los fenómenos a estudiar, así como la multiplicación del volumen de información acumulada, requieren abordajes que articulen saberes en pos de una comprensión más cabal y adecuada de los objetos de estudio.

Siendo la prensa diaria el medio cuyo discurso nos interesa analizar vale recordar que, desde que la misma arribó a la web a mediados de la década de 1990, se encontró ante la necesidad de seguir los desplazamientos de un público lector cada vez más inasible y disperso. En muy pocos años, lo que por entonces era un sistema de medios «masivos» de comunicación[2] se vio alterado por la creciente complejidad y todas las empresas mediáticas debieron reestructurarse para sobrevivir. Paralelamente, los «lectores» de diarios pasaron a ser, también, «usuarios» y «seguidores» de cuentas en cada una de las *plataformas mediáticas* (Fernández, 2018) que se incorporaron al mercado. De ahí en más, como lo precisó recientemente en entrevista el Director General del diario argentino *La Nación*, la estrategia comercial no perdió de vista «a dónde la gente elige consumir contenido y tratar de ver cómo armar el modelo de negocio para que sea algo rentable» (Seghezzo, 2021). A medida que los lectores-usuarios-seguidores fueron adquiriendo nuevas habilidades o practicando otros espacios —primero la ya anticuada blogósfera y, luego, las plataformas mediáticas como Facebook, Twitter o Instagram— los periódicos digitales se adaptaron a esos cambios, de una u otra manera.

Es precisamente en esta encrucijada, que nos propusimos conocer cómo enuncian los diarios argentinos *Clarín* y *La Nación* en Facebook y qué tipo de vínculos —en términos de *contrato de lectura* (Verón, 1985)— les proponen allí a sus usuarios-lectores-seguidores, procurando saber, además, cómo ha ido variando eso a lo largo del tiempo. Para ello, se analizaron los posteos publicados en las *fanpages* @clarincom y @lanacion, articulando la perspectiva

sociosemiótica (Verón, 1998) con el empleo de herramientas digitales y métodos computacionales de minería de datos, e indagando sobre:

a) los modos de composición de los posteos, identificando componentes elementales —texto del post (que, además del texto lingüístico-verbal, incluye elementos paratextuales y paralingüísticos), enlaces a sitios web, imagen, video—, y sus relaciones a lo largo del período estudiado;

b) el contenido presentado y la frecuencia de publicación según franjas horarias (madrugada, mañana, tarde y noche);

c) las modalidades discursivas prevalentes en el texto del post y en el título del enlace, según tipos de modalidades intersujetos propuestas por Antoine Culiolí (como se cita en Fisher y Verón, 1986)[3] y;

d) las interacciones obtenidas (reacciones, compartir y comentar) y su relación con los atributos antes listados.

Se trabajó sobre dos períodos de tiempo: en principio, 2010-2015 y, posteriormente, se agregó 2016-2017. Como parte del análisis se caracterizó un subgrupo de posteos a partir de seis variables cualitativas *ad hoc* que fueron relevadas manualmente por parte del equipo de trabajo: texto propio, localización geográfica de la información (local, nacional o internacional), temática de referencia (política, economía, deportes, espectáculos, etc.), temporalidad de los acontecimientos presentados en las notas (pasado, presente, instante o futuro), modalidad discursiva prevalente y, género periodístico (noticia, reportaje, opinión, etc.). La clasificación de esta última variable permitió advertir la presencia significativa de publicaciones que no podían ser atribuidas a ninguno de los géneros tradicionales del discurso periodístico, ya sea impreso u *online* y que, por lo tanto, fueron clasificadas en la categoría «Otros». Lo cual nos llevó a conjeturar que esos géneros tradicionales no son suficientes para caracterizar cabalmente el discurso de los periódicos en Facebook. Siguiendo a Steimberg, consideramos que los géneros son, en un sentido discursivo amplio, «reguladores de la circulación de los textos» (1998: 5), siendo posible observar, «cuando se expande el dispositivo social —técnico y espectatorial— de un nuevo medio» (Steimberg, 1998: 35), dos posibles tendencias: por un lado, procedimientos de adscripción a «moldes de género» previamente consolidados —definidos por Steimberg como *moldes de la previsibilidad social*— y, por otro lado, procedimientos de emergencia e instalación de nuevos géneros. Cuando sucede esto último, por lo general, los «nuevos géneros» recaen en (o revelan) la diferencia técnica constitutiva del medio (o, en este caso, de la plataforma) en cuestión y las «posibilidades de contacto que lo definen» (1998: 36). Esos moldes, agrega Larrondo Ureta, «dan forma a los textos periodísticos para que sean identificables tanto por parte del periodista, como del público que los recibe (...) [y] son el resultado de una lenta elaboración histórica vinculada al desarrollo de cada

uno de los medios de comunicación que han surgido con el paso de los años» (2008: 168).

Se presentan en este artículo, entonces, los resultados de la aplicación de las técnicas de *clustering* provistas por la minería de texto, cuyo empleo tuvo como finalidad colaborar con la posibilidad de identificar si existen, al interior del subconjunto de posteos clasificados como «Otros», agrupamientos de publicaciones con características comunes —ya sean éstas derivadas de regularidades temáticas, retóricas, enunciativas o de otra índole— no detectadas durante la observación convencional. A eso se llega, por ejemplo y como también proponen Touileb y Salway (2014: 642), cuando los algoritmos permiten reconocer fenómenos interesantes —que no serían evidentes para un investigador que observa convencionalmente el material— para un análisis posterior más detallado.

Este acercamiento tiene carácter experimental y procura dilucidar, en el marco de una investigación interdisciplinaria, la riqueza de los métodos computacionales y de la minería de textos para acompañar el estudio semiótico de los discursos sociales mediatizados. De este modo, y en concordancia con el proceso emergente que Berry (2011: 12) denomina como giro computacional o que Lazer et al. (2009, 721: 723) distinguen como ciencias sociales computacionales, esta indagación se asume como el ensayo de una manera combinada de abordar el estudio de la configuración discursiva de las publicaciones que los diarios realizan en sus *fanpages*. Si bien la incorporación de esos métodos computacionales al análisis de corpus textuales es algo que se viene practicando desde hace ya algunos años en distintos contextos académicos —sobre todo, en idioma inglés—, vale señalar que hay escasos antecedentes de ello en lo que concierne al lenguaje natural en español y, menos aún, motorizados desde una perspectiva semiótica.

CONFIGURACIÓN DEL CORPUS BAJO ANÁLISIS

Tal como se explicó en Raimondo Anselmino et al. (2018: 132), un aspecto preliminar a todo análisis de los discursos consiste en discernir cómo se conformará el corpus sobre el cual trabajar, es decir, en palabras de Barthes esa «colección finita de materiales, determinada previamente por el analista, con cierta (inevitable) arbitrariedad» (1993: 80). Esa construcción, como advierten Gindin y Busso, «se erige como resultado y condición de una serie de interrogantes que guían nuestros trabajos, como ese conjunto significante sobre el que ponemos a trabajar nuestras hipótesis y marcos teóricos» (2018: 31). Por ello, un corpus se determina, como también lo señalan las autoras, luego del acceso —por lo general, exploratorio y aproximado— a un conjunto mayor del que los materiales recopilados son parte. Pero es evidente que, en la actualidad, esa corre-

lación entre la parte y el todo supone una relación de proporción, una escala, diferente.

A la hora de validar como satisfactoria y suficiente una determinada colección es usual emplear el *principio de saturación*, que Barthes explica de la siguiente manera:

> el corpus tiene que ser suficientemente amplio como para que se pueda suponer razonablemente que sus elementos saturan un sistema completo de semejanzas y de diferencias; es seguro que si se entresaca un conjunto de materiales se llega, al cabo de un cierto tiempo, a encontrar nuevamente hechos y relaciones ya aislados anteriormente (...); estas 'vueltas atrás' se hacen cada vez más frecuentes, hasta que se llega a un punto en que no se descubre ya ningún material nuevo: el corpus está entonces saturado (1993: 80-81).

Sin embargo, esa labor adquiere otras dimensiones —o, incluso, se imposibilita— en determinados contextos; esto sucede cuando, en vez de trabajar sobre universos más acotados compuesto, por ejemplo, por diarios impresos que tienen una limitada cantidad de páginas, nos enfrentamos a decenas o centenas de miles de posteos *online*. En otras palabras, a la hora de trabajar sobre poblaciones de discursos tan voluminosas como la que aquí nos atañe, la tarea de elaboración del corpus se complejiza aún más.

Es por todo lo antes dicho que se decidió operar con dos tipos de corpus: uno denominado *corpus de base* y, otro, *corpus total*. Entre estos, podría decirse, hay un cambio de escala que evita, como propone Manovich (2012), tener que escoger entre tamaño y profundidad.

En función de poder observar la producción rutinizada de la información que los diarios comparten en la plataforma pero, al mismo tiempo, propiciando cierta cuota de aleatoriedad, para el período 2010-2015 el corpus de base se confeccionó teniendo en cuenta los siguientes criterios: se relevó, por cada diario, todos los posteos publicados durante una semana completa por año, eligiendo distintos meses en forma alternada (esto es, un mes sí y un mes no) y variando también la semana escogida (por ejemplo, la primera semana completa de diciembre de 2011, la segunda de febrero de 2012 y así sucesivamente). Vale aclarar, que aunque *La Nación* abrió su *fanpage* en 2009, el recorte temporal comienza en la última semana de octubre de 2010 —porque es recién allí cuando la cuenta oficial de *Clarín*, dada de alta más tarde, presenta una actividad sistemática de publicación pasible de comparación— y termina el 31 de diciembre de 2015. Empleando, entonces, la herramienta digital *Netvizz*[4] se constituyó una colección (corpus de base 1) que comprende un total de 1.129 posteos —534 de *Clarín* y 595 de *La Nación*. Además, para sistematizar toda la información necesaria y clasificar los distintos atributos de interés, se enri-

queció la planilla de datos obtenida automáticamente con la inclusión de las seis variables cualitativas *ad hoc* que se completaron manualmente, luego de la observación directa y pormenorizada de cada uno de los posteos y el cotejo de los mismos con las notas publicadas en las versiones *online* de los diarios a las que dichos post reenvían, cuando hubiera enlace/link. A dicho conjunto de posteos, se sumó otro corpus (corpus total 1) —también recuperado mediante *Netvizz*— que comprende la población total de los 54 742 posteos publicados entre 2010 y 2015, 29 341 de *Clarín* y 25 401 de *La Nación*. Este segundo corpus tuvo la función de permitir colegir conjeturas y evaluar el grado de generalidad de los hallazgos realizados a partir del análisis pormenorizado del *corpus de base* con aquello que se deriva del estudio automatizado de la población completa de posteos (corpus total); todo esto, gracias al empleo de métodos computacionales sobre una vista minable armada con la ayuda de *MySQL* y el módulo para extracción, transformación y carga de datos de *Pentaho.*

Por su parte, la recolección de los discursos publicados durante 2016 y 2017 se hizo en 2018 a través de *BuscarPosteosFacebook* (Leale et al., 2020), una herramienta desarrollada *ad hoc* para sortear los inconvenientes suscitados ese año con *Netvizz*[5]. La misma permitió recolectar, nuevamente, dos conjuntos de materiales a analizar: un corpus total 2, que incluye los 15 298 posteos publicados por las cuentas entre el 1 de enero de 2016 y el 31 de diciembre de 2017 (9.726 de *Clarín* y 5.572 de *La Nación*), así como un corpus de base 2, con 670 posteos (407 de *Clarín* y 263 de *La Nación*). A diferencia de lo efectuado durante el período anterior, este último subconjunto comprende dos semanas por año por cada cuenta; se recogió una semana completa siguiendo los mismos criterios tenidos en cuenta para el corpus de base 1, más una semana construida eligiendo distintos meses en forma alternada (esto es, un mes sí y un mes no) y variando el día por cada año (empezando por el primer lunes de enero, el segundo martes de marzo, el tercer miércoles de mayo, el cuarto jueves de julio, etc.).

SOBRE LOS GÉNEROS PERIODÍSTICOS

Se parte de entender a los *géneros* como «clases de textos u objetos culturales, discriminables en todo lenguaje o soporte mediático, que presentan diferencias sistemáticas entre sí y que en su recurrencia histórica instituyen condiciones de previsibilidad en distintas áreas de desempeño semiótico e intercambio social» (Steimberg, 1998: 41).

Tal como se detalla en Raimondo Anselmino, Sambrana y Cardoso (2017), los géneros periodísticos en particular, en tanto tipos relativamente estables de discurso de actualidad propios del periodismo de prensa moderno, han asumido diferentes caracterizaciones en el transcurso de la mediatización y sus fronteras se difuminan cada vez más desde su digitalización y puesta en línea

(Rončáková, 2017 y 2019). No obstante ello, es posible apreciar la existencia de ciertas clases de textos que, en su interior, presentan «características comunes de forma y contenidos, es decir, unas normas y convenciones que incluyen leyes discursivas propias y ciertos rasgos lingüísticos obligatorios» (Parrat, 2008:11). Esto es así porque, como sostiene Steimberg, «los géneros existen e insisten en los medios, y también insisten esas clasificaciones que constituyen, de por sí, un objeto de investigación con interés propio, en tanto *interpretante* estabilizado en una región cultural» (1998:15) [el resaltado es nuestro]. Eso es lo que puede observarse, por ejemplo, en el estudio presentado en Rončáková (2017), donde se elabora una clasificación de nuevos géneros periodísticos —en ese caso en particular, operantes en revistas semanales que cubren asuntos sociales y políticos— sobre la base de cinco criterios constitutivos: tema, función, forma, composición y lenguaje.

En función de las ideas planteadas, se clasificaron los posteos presentes en ambos corpus de base, en su vinculación con los contenidos del diario al que cada uno remite, aplicando las siguientes opciones/valores: noticia, crónica, opinión, entrevista, reportaje, crítica, anuncio o posteo de saludo a usuarios, u «otros». Vale reiterar que dicha clasificación se llevó a cabo no sólo observando cada posteo en particular sino, además, considerando como otra unidad de observación al discurso informativo al cual reenvía el hipervínculo que cada publicación en Facebook suele contener y que, por lo general, linkea a algún contenido localizado en el sitio web del medio en cuestión. Se procedió de esa manera por considerar que, si bien hay posteos que no contienen link alguno esto es algo bastante infrecuente en ambas *fanpages*: se acerca, aproximadamente, al 1 % del corpus total 1, y al 0,32 % del corpus total 2. Y estos últimos casos suelen corresponderse, precisamente, con posteos identificados dentro del género «Anuncio o posteo de saludo a usuarios», propio de la plataforma Facebook, o con *posteos* ubicados como «Otros» y que son los que en este artículo analizaremos. Volveremos sobre esto más adelante.

Respecto de la clasificación de géneros de la que partimos, siguiendo a Peralta y Urtasun (2007:48) se consideró *noticia* a aquella unidad textual en la que se relata un «hecho nuevo de la realidad —entre todos los que acontecen— que los medios periodísticos consideran que es socialmente relevante y que por lo tanto merece ser comunicado». Se trata de un género en el que predomina (aunque cada vez menos) una estructura de pirámide invertida y «un estilo claro, directo» (de Fontcuberta, 2011:102). Se diferencia de la *crónica* en que esta última presenta «una estructura textual en la que predomina el tipo narrativo cronológico» (Peralta y Urtasun, 2007:37) y puede contener ciertos elementos valorativos pero secundarios al hecho a informar en sí.

Por su parte, la *opinión* se caracteriza por tener una dimensión argumentativa explícita, en tanto «proceso discursivo por el cual se llega a cierta conclusión y se la defiende o sostiene» (Peralta y Urtasun, 2007:18), mientras que la

crítica se corresponde con aquello que de Fontcuberta (2011: 133) nombra como artículo o comentario y consiste en una «exposición de ideas y juicios valorativos suscitados a propósito de hechos que han sido noticias más o menos recientemente» pero que, no obstante, «presenta un estilo literario muy libre».

La *entrevista* da cuenta de un diálogo entre un entrevistador y un entrevistado; esta no debe ser confundida con el *reportaje* —también llamado reportaje en profundidad—, entendido como la «explicitación de hechos actuales que ya no son estrictamente noticia (aunque a veces pueden serlo), que intenta explicar lo esencial de los hechos y sus circunstancias explicativas» (de Fontcuberta, 2011: 132) con un estilo casi literario.

Además de todos esos géneros tradicionales mencionados anteriormente, un acercamiento preliminar exploratorio al corpus, realizado a priori del proceso clasificatorio propiamente dicho, permitió reconocer e incorporar una nueva opción/valor de la variable género cuya singularidad y recurrencia significativa pudo comprobarse luego; dicha clase fue denominada como *anuncio o posteo de saludo a usuarios*. La misma es muy propia de las plataformas mediáticas como Facebook y se diferencia de los géneros que convencionalmente han sido considerados como periodísticos (de Fontcuberta, 2001) por las siguientes características:

a) no suele referir a un hecho noticiable, ni a ideas expresadas o defendidas argumentativamente;

b) puede no presentar enlace alguno (Imagen 1) —lo cual sucede con una frecuencia mucho mayor al resto de los posteos clasificados en otro tipo de género o al porcentaje de posteos sin link de ambos corpus totales— o, si tiene algún link, suele: I) si es enlace al sitio web del medio, compartir contenido lateral como, por ejemplo, el pronóstico del tiempo (Imagen 2) o; II) compartir enlace a contenido externo al periódico en cuestión como, por ejemplo, video en YouTube[6] (Imagen 3);

c) manifiesta una hibridación entre géneros discursivos primario y secundario, recuperando la clásica distinción de Bajtín (1998), es decir, entre aquellos géneros que son más sencillos, breves, espontáneos y con posible respuesta inmediata y otros detrás de los cuales hay un proceso de mayor elaboración;

d)articula, por lo tanto, peculiaridades de la comunicación interpersonal con otras propias de esos procesos de comunicación más complejos y;

e) se asocia a estrategias discursivas de homología —en tanto dimensión retórica que figura la coincidencia entre temporalidades textuales y experimentadas por la audiencia, tal cual ha sido definida en Morley (1996)—, de identificación y de establecimiento de lazos de tipo comunitario (Cfr. Raimondo Anselmino et al., 2018 y Raimondo Anselmino, 2019).

Imagen 1: Ejemplo de *Anuncio o posteo de saludo a usuarios*, sin enlace

Imagen 2: Ejemplo de *Anuncio o posteo de saludo a usuarios*, con enlace a sitio web del medio

Imagen 3: Ejemplo de *Anuncio o posteo de saludo a usuarios*, con enlace a YouTube

Es así que en esta dimensión de nuestro estudio se evidenció que en las *fanpages* analizadas predominan los denominados *géneros informativos* —que, como señala de Fontcuberta (2011), dan a conocer hechos—, en detrimento de los *géneros de opinión*, que dan a conocer ideas. Como era previsible, se detectó que la mayor parte los *posteos* de ambos diarios enlazan con el género más tradicional de la prensa de masas, la noticia, siendo esta presencia más destacada y con tendencia creciente en *Clarín* (54,12 % para el primer período y 67,57 % para el segundo) que en *La Nación* (41,68 % y 30,96 %, respectivamente), donde se observa una tendencia decreciente (ver Tabla 1). No obstante, un aspecto

de singular interés para el análisis que acá desarrollamos se desprende de los siguientes puestos en el *ranking* por géneros, que permiten reconocer la emergencia de los que podrían ser considerados nuevos moldes o clases de textos; en ellos se advierte la presencia de procedimientos de identificación ligados a estrategias de establecimiento de lazos —en cierto modo— comunitarios, como veremos luego en el apartado de cierre. En este sentido, se destacan tanto la clase denominada como *anuncio* o *posteo de saludo a usuarios* —con el 17,23 % de las publicaciones de *Clarín* y el 9,58 % de *La Nación*, durante el primer período indagado; dato que desciende significativamente luego, en 2016 y 2017, al 0,25 % en ambos casos— como la heteróclita clase de los «Otros». De hecho, esos *posteos* que en ambos corpus de base no pudieron ser identificados ni como noticia, crónica, crítica, entrevista, etc. representan, en el primer período, el 21,18 % en *La Nación* y el 9,18 % de *Clarín*, así como el 23,34 % de *La Nación* y el 11,55 % de *Clarín* para el segundo período. Es, precisamente, sobre un total de 317 posteos de esta última clasificación (175 correspondientes al corpus de base 1 y 142 al corpus de base 2) que se realizó el análisis aplicando técnicas de *clustering* de texto con el fin de identificar si existen grupos con relaciones comunes no triviales, es decir, no detectables a simple vista por la clasificación tradicional-artesanal propia del enfoque sociosemiótico, ya que su naturaleza proviene de las características intrínsecas de los textos analizados. Como advirtió Larrondo Ureta en un momento, incluso, más inaugural del periodismo *online*, «los géneros representan un área viva del conocimiento que demanda recapitular los logros del pasado y atender, al mismo tiempo, a los cambios que introducen en este ámbito nuevos modos como el ciberperiodismo» (2008:165). Es, precisamente, en función de darle sentido a esos cambios que este artículo propone también su aporte.

Género	Clarín - @clarincom				La Nación - @lanacion			
	2010-2015		2016-2017		2010-2015		2016-2017	
	Cant.	%	Cant.	%	Cant.	%	Cant.	%
Noticia	289	54,12 %	275	67,57 %	248	41,68 %	126	47,91 %
Otros	49	9,18 %	47	11,55 %	126	21,18 %	95	36,12 %
Crónica	30	5,62 %	36	8,85 %	32	5,38 %	11	4,18 %
Opinión	14	2,62 %	27	6,63 %	41	6,89 %	11	4,18 %
Reportaje	43	8,05 %	10	2,46 %	66	11,09 %	9	3,42 %
Entrevista	6	1,12 %	7	1,72 %	13	2,18 %	9	3,42 %
No corresponde	9	1,69 %	3	0,74 %	10	1,68 %	1	0,38 %
Crítica	2	0,37 %	1	0,25 %	2	0,34 %	0	0,00 %
Anuncio o posteo de saludo a usuarios	92	17,23 %	1	0,25 %	57	9,58 %	1	0,38 %
Totales	**534**	100,00 %	**407**	100,00 %	**595**	100,00 %	**263**	100,00 %

Tabla 1: Distribución de posteos según género periodístico

METODOLOGÍA

Como se ha anticipado, la estrategia metodológica combina, de modo interdisciplinario, las labores artesanales provenientes del análisis sociosemiótico (Verón, 1998) con el empleo de herramientas digitales y métodos computacionales que permiten la recopilación, el pre-procesamiento, el procesamiento y la visualización de cantidades masivas de datos y de metadatos. Utilizamos el término «combinación» como una tercera vía de articulación metodológica ante la diferencia entre convergencia y complementación, en el sentido propuesto por Bericat:

> Existe *convergencia* cuando se utilizan ambos métodos para el estudio de un mismo aspecto de la realidad social. Por su parte, la *complementación* se produce cuando la investigación genera dos imágenes del objeto analizado —una derivada de métodos cualitativos y otra de métodos cuantitativos— que iluminan diferentes dimensiones o aspectos. Finalmente, existe *combinación* cuando se trata de integrar de manera subsidiaria un método en el otro con el objeto de fortalecer su validez y compensar sus debilidades (como se cita en Piovani, 2018: 442-443; el resaltado es nuestro).

Es así que, como parte de un diseño metodológico que excede ampliamente a lo presentado en este escrito, se aplicaron técnicas de *clustering* propias de la minería de texto sobre el paquete de 317 posteos clasificados como «Otros» en la variable *género periodístico*. Se entiende por minería de texto a toda «aplicación de la lingüística computacional y del procesamiento de textos que pretende facilitar la identificación y extracción de nuevo conocimiento a partir de colecciones de documentos o corpus textuales» (Brun y Senso, 2004: 11). La misma es una variante o extensión de la minería de datos (Tan, 1999) que «implica la extracción de conocimiento a partir de datos masivos y las relaciones subyacentes que pueden existir entre ellos» (Arcila, Barbosa y Cabezuelo, 2016: 627). Por su parte, la técnica de *clustering* permite crear «agrupaciones entre documentos de forma desatendida —sin intervención del usuario—. Es decir, un programa informático decidirá qué grupos va a generar a partir de la similitud que calcule entre los documentos de la colección» (Brun y Senso, 2004: 16). Dichos agrupamientos se someten, luego, a un proceso inferencial e interpretativo, orientado por las categorías propias del análisis de los discursos sociales, que busca dar sentido a los hallazgos producidos automáticamente.

Sobre el proceso de trabajo con algoritmos, a continuación se detallan las tareas más relevantes realizadas conforme a las fases propuestas por Fayyad et al. (1996) en la *metodología Knowledge Discovery in Databases* (en adelante, KDD). Dichos autores definen a la misma como el proceso no trivial de identificación de patrones significativos en los datos que sean válidos,

novedosos, potencialmente útiles y comprensibles para un usuario; las distintas fases sugeridas son: 1) comprensión del dominio de aplicación; 2) extracción de la base de datos objetivo; 3) preparación de los datos; 4) minería de datos; 5) interpretación y; 6) empleo del conocimiento descubierto.

Como primer paso para preparar los datos[7], se determinó cuáles elementos, entre los que componen un posteo (Imagen 4), era conveniente conservar para la posterior aplicación de los algoritmos. Fueron así seleccionados tanto el *texto del post* como el *título del enlace*, siendo los únicos correspondientes a elementos textuales en sentido estricto.

Imagen 4: Componentes elementales de un posteo

Posteriormente, se removieron los links, emojis, signos de puntuación y caracteres especiales (numeral, arroba, etc.) y se llevó a cabo un proceso de *tokenización* del texto, que consistió en tomar el texto de cada posteo para transformarlo en una lista de palabras (Imagen 5). Esta forma de representar los datos permitió eliminar las denominadas palabras vacías (*stopwords)* que —como suele suceder con los artículos, las preposiciones y algunos pronombres como los relativos, por ejemplo— se repiten con mucha frecuencia y sin aportar valor, y normalizar el texto mediante un proceso de lematización-reducción de las formas gramaticales de las palabras a su base común con el fin de poder compararlas.

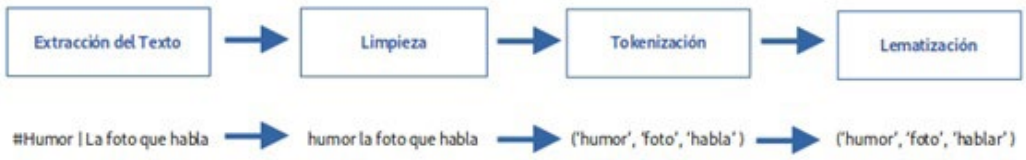

Imagen 5: Preprocesamiento de documentos

Luego, se aplicó la función TF-IDF —que, según Jones (1972), calcula la relevancia que tiene cada término en un texto, con respecto a una colección general— sobre las listas de palabras generadas para poder construir una matriz

de similaridad. Dicha matriz indica, para cada par de textos, qué tan similares son y es requerida como entrada por los algoritmos de *clustering*.

Respecto a los algoritmos empleados, en primera instancia se aplicó *K-Medias*, diseñado por MacQueen (1967), para el agrupamiento de los *posteos*. La idea principal del mismo es definir k centroides y luego tomar cada elemento y situarlo en el grupo más cercano. El próximo paso es recalcular el centroide (es decir, el centro geométrico) de cada grupo y volver a distribuir todos los *posteos* según el *centroide* más cercano; proceso que se repite hasta que ya no se produzcan más cambios.

Dado que este algoritmo no proporciona una cantidad de grupos óptima, es necesario definir la misma de antemano. De las varias formas de calcular la cantidad de grupos (valor K) adecuada se utilizaron dos diferentes:

-el método del codo (*elbow method*), que permite identificar cuánto mejora el agrupamiento a partir de la representación gráfica del vector de la suma de cuadrados de la varianza intra *clusters* para diferentes números de k; si el gráfico tiene forma de brazo, el valor correspondiente al codo es el valor k ideal.

-el método de la silueta (*silhouette*), donde el coeficiente de silueta contrasta la distancia media a elementos en el mismo grupo con la distancia media a elementos en otros grupos. Los valores de este coeficiente se encuentran en el intervalo [-1;1], donde los objetos con un valor de silueta alto están considerados bien agrupados, mientras que los objetos con un valor bajo pueden ser atribuidos a ruido o anomalías (Rousseeuw, 1987: 53).

En segunda instancia, se aplicó un método de ensamble de *clusters*, a través del cual se intenta combinar múltiples modelos de *clustering* para producir un mejor resultado en términos de consistencia y estabilidad que el obtenido por los algoritmos en forma individual (Alqurashi y Wang, 2019:1227). Así fue posible enriquecer los resultados agregando las variables *ad hoc* por fuera del género (esto es, texto propio, localización geográfica de la información, temática de referencia, temporalidad de los acontecimientos presentados), así como los elementos paratextuales (corchetes, hashtags, menciones y signos de interrogación) y paralingüísticos (emoticones y emojis).

Vale aclarar, que en el contexto de *machine learning*, un ensamble es generalmente definido como un sistema de aprendizaje automatizado que es construido con un set de modelos individuales trabajando en paralelo, cuyos resultados son combinados con una función de consenso para producir una única respuesta a un determinado problema (Alqurashi y Wang, 2019:1227). En el estudio aquí presentado se aplicó el algoritmo k-medias de forma individual a cada variable obteniéndose una partición de los datos —para cada elemento el grupo al que pertenece— y, luego, se construyó una nueva matriz de distancia en donde el valor Nij indica las veces que dos elementos comparten el mismo grupo. Esta

nueva matriz de distancia es la entrada necesaria para el algoritmo de *clustering* jerárquico que actúa como función de consenso y devuelve un nuevo agrupamiento. El *clustering* jerárquico funciona agregando aquellos pares de elementos más cercanos entre sí, por niveles, formando una jerarquía que puede visualizarse en un dendrograma —esto es, un diagrama de datos en forma de árbol—. Para determinar la cantidad de grupos factibles se trazan líneas horizontales y la cantidad de líneas verticales de la intercepción es el valor buscado.

Como se mencionó antes, una vez concluida la aplicación de la técnica de *clustering*, procuramos inferencialmente reconocer cuáles de los agrupamientos identificados de modo automático podían ser considerados fértiles para la determinación de especies textuales, tanto por su recurrencia significativa como por su disparidad invariante, desde el punto de vista de la comparación con las otras clases de géneros periodísticos ya incorporados a la clasificación. Esto es así porque, como lo señalan Günther y Quandt, al ser el trabajo con *clustering* un procedimiento de análisis textual totalmente automatizado y sin categorización previa, «el investigador tiene que interpretar los grupos para encontrar el vínculo entre los documentos agrupados y dar sentido a los resultados» [traducción propia] (2016: 77). En sintonía con ello, también Touileb y Salway acuerdan en que los análisis automatizados no son suficientes por sí solos y necesitan de una labor intelectual —que ellos denominan, poco felizmente, «manual inspections of the texts» (2014: 635)— que los complete.

HALLAZGOS DERIVADOS DE LA APLICACIÓN

Una vez completada la primera experiencia de aplicación (algoritmo *K-Medias*) se determinó, tanto por el método del codo (Gráfico 1) como por el método de la silueta (Gráfico 2), que el número potencial de grupos en el caso del texto del post era de cinco, mientras que en el caso del texto del enlace fue de tres grupos (Gráficos 1 y 3). Es decir, que el trabajo con ambos métodos sobre los dos componentes textuales escogidos permitía, efectivamente, detectar la tendencia de ciertos *posteos* clasificados como «Otros» a agruparse por presentar ciertas características intrínsecas comunes entre sí y establecer, a su vez, diferencias con los otros *clusters*.

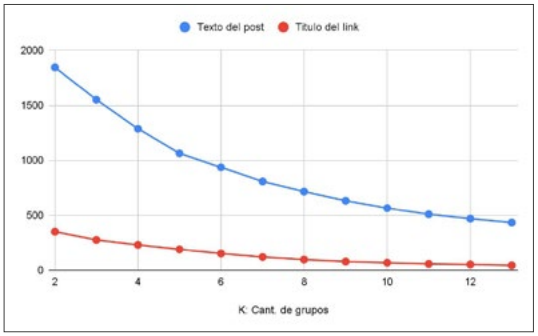

Gráfico 1: Varianza por cantidad de grupos

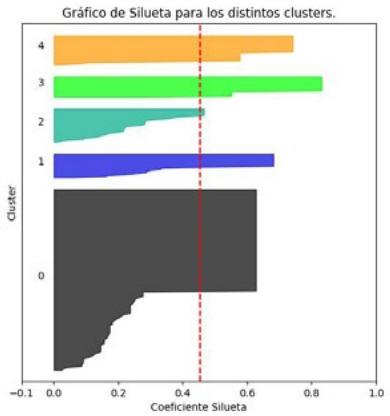

Gráfico 2: Silueta para texto del post

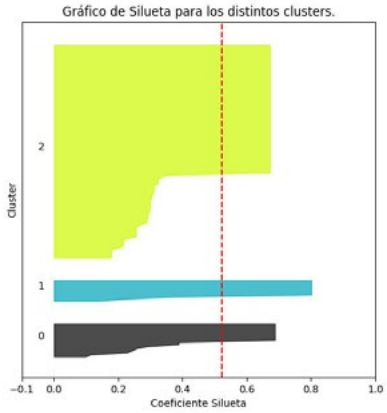

Gráfico 3: Silueta para título del enlace

Posteriormente, se procedió al análisis inferencial sobre los *clusters* generados. En el caso del texto del post, de los cinco grupos potenciales solo dos eran fértiles para el objetivo de la presente investigación, esto es, para reconocer posibles nuevas clases de textos discriminables como géneros. El primer grupo está compuesto por 20 posteos en los que la *fanpage* @lanacion comparte contenido de humor producido originalmente para el diario por el humorista gráfico Juan Matías Loiseau, conocido por el nombre artístico *Tute*. Además de contener de modo protagónico la imagen de la viñeta cómica, en esas publicaciones se observa siempre la presencia de la etiqueta #HUMOR y la mención con enlace a la cuenta que el humorista tiene en la misma plataforma (Imagen 6). Dicho hallazgo, en vez de reconocer un nuevo género, permitió advertir un descuido previo en la clasificación artesanal realizada sobre los corpus de base, en la cual ignoramos lo que Steimberg define como un lugar «de lo previsible en los espacios del género» (2001: 116) vinculado a la «*crónica crítica* de nuestra sociedad» (2001: 117). No habíamos considerado el género «humor» o «humor gráfico» a la hora de la codificación inicial e, incluso, no nos percatamos de este error sino hasta ver el agrupamiento devuelto por la labor automatizada. Ciertamente, en este punto el trabajo interdisciplinario nos permitió identificar y superar una falla en el diseño de los instrumentos de nuestra investigación. Al considerar, luego, de forma aunada la población de los *posteos* recopilados en el corpus total 1 y el corpus total 2, pudo comprobarse que un 1,20 % de las publicaciones que los diarios realizaron entre 2010 y 2017 (n=847) se corresponden con este género periodístico.

Imagen 6: Ejemplo de posteo del cluster de *humor*

Por otra parte, el segundo grupo está compuesto por 29 posteos (16 de @clarincom y 13 de @lanacion) en los que se presenta un video, una imagen o ambos elementos, constituyéndose en una unidad propia de eso que, en el campo del ciberperiodismo, suele definirse como contenido *multimedia*. Se trata de publicaciones donde *lo visual* es el centro y, a su vez, contienen un texto del post que invita expresamente a un usuario individual a interactuar con ese contenido mediante frases imperativo-interpelativas como «¡Mirá el video!» o «¡Mirá la foto!» (Imagen 7). Siguiendo la clasificación de modalidades enunciativas propuesta por Culiolí (como se cita en Fisher y Verón, 1986), puede observarse en el componente texto del post la predominancia de ese tipo enunciativo que se caracteriza como Modalidades-4. Son enunciados que se dirigen a un co-enunciador individual y anónimo (no se trata ya de un *lectorado* en sentido colectivo) que se propone como co-presente, co-temporáneo de la enunciación, y cuyo ejemplo más expresivo es, precisamente, el de la interpelación. De hecho, el 100 % de los posteos de este agrupamiento reconocido por la técnica de *clustering* presenta en algún lugar del texto del post la expresión conativa «mirá» así como, tras una búsqueda automatizada en ambos corpus totales, se pudo encontrar que son 3.868 *posteos* (esto es, el 5,52 % de la población completa) los que asumen estas cualidades.

Asimismo, en este tipo de publicaciones se hace evidente el fenómeno de hibridación entre esas dos posiciones de intercambio mediático que Fernández define como *espectatorial*, por un lado, e *interactiva*, por el otro. Según la explicación del autor, la primera de esas dos prácticas de intercambio dentro de la comunicación masiva es aquella «en la que los receptores tienen un lugar relativamente fijo frente al cual les llega la emisión de su mediatización elegida y/o aceptada» (2021: 223); mientras que en la posición interaccional «los emisores y receptores deben realizar actividades registrables y combinables con diversas movilidades o posiciones estacionarias para que el intercambio discursivo se produzca» (2021: 223).

Respecto del contenido visual o audiovisual que se comparte en este segundo grupo de posteos, este generalmente remite a las temáticas «espectáculo» o «entretenimiento», o versa sobre aquello que podría denominarse como «historias de interés humano» (Hughes, 1981), que buscan suscitar la sorpresa o la curiosidad de la audiencia en torno a un contenido trivial[8] que suele ser viralizable. Y, como lo señaló Hugues, en muchas ocasiones el interés humano en un contenido «radica en el comportamiento casi humano de un animal» (1981: 52), tal como sucede en el posteo ilustrado en la imagen 7.

Imagen 7: Ejemplo de posteo del cluster de *contenido multimedia*

En el caso del título del componente enlace (gráficos 1 y 3), el acercamiento interpretativo permitió observar que solo un grupo comparte características relevantes a simple vista (por ejemplo, los *posteos* que lo componen incluyen la palabra «video»), las cuales coinciden, parcialmente, con el segundo grupo obtenido del análisis de la variable texto del post.

Finalmente, al momento de la segunda experiencia de aplicación —método de ensamble de *clusters*—se obtuvo como salida el dendrograma (Gráfico 4) en el cual, trazando líneas horizontales, se puede determinar la cantidad de grupos. Para cantidades de grupos de dos a cuatro solo se pudo distinguir el conglomerado de *posteos* con contenido *multimedia viral* acompañado de frases imperativo-interpelativas, mientras que cuando se formaron cinco grupos se encontró nuevamente el conjunto de *posteos* de *humor*. Es decir que, en líneas generales, se obtuvieron los mismos resultados que los logrados en la primera experiencia de aplicación.

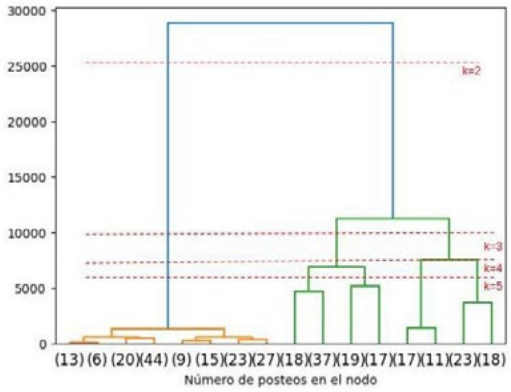

Gráfico 4: Dendrograma

Esta segunda experiencia nos permitió, además, reconocer relaciones significativas en otros dos agrupamientos. Un primer hallazgo novedoso respecto a la experiencia anterior fue el surgimiento de un *cluster* compuesto por *posteos* relacionados a la temática «Política», cuando la cantidad de grupos es igual o mayor a cinco. Más precisamente, los mismos versan sobre los incidentes ocurridos en 2017 en el Congreso de la Nación, durante los días en los que se trató en Argentina la denominada Reforma previsional. Son piezas que, al estilo de las *breaking news* televisivas, comparten un video grabado —o la captura, a posteriori, de una transmisión que fue antes compartida «en vivo»— que es subido directamente desde la plataforma[9]. Se muestran, de esta manera, los hechos ocurridos en un pasado inmediatamente reciente al momento de la publicación, cuya inscripción actual también es señalada por el tiempo verbal *presente* en un texto del post que reza expresiones como: «Así es el enfrentamiento entre los manifestantes», «Así son los incidentes afuera del Congreso». Sería deseable, en un futuro, continuar observando qué recurrencia y estabilidad tiene este tipo de procedimiento en el marco del discurso de los periódicos analizados en Facebook.

El segundo hallazgo como resultado del proceso de análisis obtenido en estas dos experiencias comprende un conjunto de 14 posteos que contaban con ciertas similitudes en su composición pero que no fue detectado por los algoritmos de agrupamiento, es decir, que se encontraban distribuidos en diferentes *clusters* y fueron reconocidos, luego, por la observación convencional. Todos estos *posteos* presentan una lista numerada de objetos, lugares, eventos, comidas (Imagen 8), con enlace que direcciona al sitio web del diario. Dicha modalidad, a la que en un artículo anterior llamamos *formato lista*, fue ya analizada en Raimondo Anselmino et al. (2018) en tanto una operación de encuadre presente dentro de los componentes texto del post y título de enlace, así como un procedimiento de estructuración del contenido informativo, cuya regularidad advertimos sobre todo a partir del año 2014. Una búsqueda realizada sobre los corpus totales de los períodos estudiados permitió reconocer, aproximadamente, que existen por lo menos 607 publicaciones (el 0,86 % de la población)

que asumen esta forma. Ciertamente, pese a que el trabajo con algoritmos no lo haya detectado, puede observarse aquí otra clase de textos a partir de la cual se moldea el contenido informativo, otorgándole cierta previsibilidad al desempeño semiótico y al intercambio social.

Imagen 8. Ejemplo de posteo con formato lista

CONCLUSIONES

La publicación de este artículo tiene un doble objetivo. En primer lugar, presentar una pequeña porción de los hallazgos producidos en el marco de una investigación de carácter interdisciplinar sobre la configuración discursiva de los posteos publicados en las *fanpages* de los periódicos argentinos *Clarín* y *La Nación*, entre 2010 y 2017; particularmente, aquello que atañe a la detección de agrupamientos dentro del conjunto de los 317 posteos considerados «Otros» en relación con la variable género periodístico. En segundo lugar, se propone mostrar cómo el análisis automatizado, a través de herramientas y procedimientos propios de la minería de textos —como es el caso de las técnicas de *clustering*—, puede colaborar con las labores de análisis sociosemiótico cuando se trata de estudiar paquetes voluminosos de discursos digitalizados y puestos en circulación a través de plataformas mediáticas.

Respecto del primer objetivo, la mencionada integración permitió efectivamente reconocer, en principio, la existencia de dos grupos bastante definidos vinculados al entretenimiento —uno ligado al *humor* y otro a la publicación *de contenidos multimedia* (frecuentemente, viral o viralizable) que procuran atraer la atención del usuario-lector-seguidor y fomentar su interacción; ambos conjuntos claramente diferenciados habían pasado inadvertidos al momento de la clasificación artesanal. En cuanto a los *posteos* que comparten las viñetas cómicas de Tute, estos pertenecen a un género periodístico que ya es clásico en la prensa gráfica (Cfr. Steimberg, 2001), pero que, erróneamente, había sido olvidado en la preparación de los instrumentos que guiaron la clasificación ini-

cial de ambos corpus de base. En cambio, puede afirmarse que las publicaciones del segundo agrupamiento reconocido sí desbordan los moldes establecidos tradicionalmente y constituyen una novedosa clase de textos.

Si, junto a Steimberg, consideramos que «es condición de la existencia del género su inclusión en un campo social de desempeños o juegos del lenguaje» (1998: 61), no debe asombrarnos que la performatividad (en el sentido de van Dijck, 2016) que la plataforma de Facebook le otorga al discurso de los periódicos bajo estudio derive en el establecimiento de nuevos tipos textuales —aunque para decidir si son o no nuevos géneros haya que esperar a determinar cuán estabilizados están y su grado de institucionalización. Esto es lo que sucede con el tipo *anuncio o posteo de saludo a usuarios*, ya identificado antes del proceso de clasificación propiamente dicho, pero también con la incorporación de ese otro tipo relativamente estable de discurso de actualidad que es novedoso por la fisonomía que asume —esto es, por ciertas características comunes de forma y contenido— en el marco del discurso mediático: el del *contenido multimedia viral*. Se advierte, así, que la praxis discursiva de los medios emplazada en Facebook genera, al menos hasta donde se ha podido estudiar, otros *moldes de previsibilidad social* que —aunque no son prevalentes en términos de cantidad en nuestros corpus de análisis— se suman al repertorio al que antaño la prensa nos tenía acostumbrados; y su presencia es significativa en relación con el tipo de vínculo enunciativo que, en términos de contrato de lectura, los diarios establecen allí con sus públicos lectores. En resumen, las publicaciones propias de aquello que definimos aquí como *contenido multimedia viral* tienen las siguientes características:

a) refieren a un contenido visual o audiovisual que se convierte, en sí mismo, en el acontecimiento noticiable, y más que brindar información sobre un hecho proponen compartir una experiencia espectatorial;

b) salvo excepciones, versan sobre temáticas vinculadas al espectáculo, el entretenimiento o el interés humano, procurando atraer la atención de los usuarios en torno a lo curioso, lo insólito o lo extravagante. Así proliferan imágenes y/o videos sobre: «un espectacular aterrizaje», «Minions furiosos», «australiano que "surfea" con su moto», «la aurora boreal en Islandia», «Cristiano Ronaldo se disfrazó de mendigo», «Carlos Tévez ayudando a una tortuga» o «cómo atacan a Chano Charpentier sus vecinos», entre otros;

c) suelen presentar enlace/link que, por lo general, genera tráfico hacia el sitio del medio en cuestión;

d) poseen un texto del post que individualiza e interpela a un co-enunciatario anónimo, ubicándose, así, más cerca de los géneros discursivos primarios —propios del intercambio cotidiano— que de los secundarios;

e) recuperando la clásica tipología de funciones del lenguaje según Jakobson (1975: 353), puede decirse que el texto del post vehiculiza rasgos conati-

vos del discurso de información —esto es, orientados hacia el destinatario— que, como ya se advirtió en Raimondo Anselmino et al. (2019), constituyen una propiedad peculiar del modo en que *Clarín* y *La Nación* enuncian en Facebook. La función referencial, por su parte, es cumplida generalmente por el texto del componente enlace;

f) como sucede con la clase *anuncio o posteo de saludo a usuarios*, puede asociarse al establecimiento de lazos de tipo comunitarios. Este tipo de vínculo es el que se advierte en publicaciones que le proponen a sus destinatarios experiencias que exceden el tradicional consumo de noticias e información y, en cierto punto, al ubicarse por fuera de una apelación raciocinante, están ligadas a un tipo de socialización asentada en «lazos prerracionales, como aquellos que nacen a partir del afecto, de los usos y de las interdependencias» (Honneth, 1999: 8).

Por otra parte, además de los dos agrupamientos reconocidos mediante el trabajo con algoritmos, durante el acercamiento interpretativo que se hizo tras los resultados arrojados por el método de ensamble de *clusters* fue posible distinguir otra clase de textos que no había sido detectada por la técnica de minería, aunque había llamado nuestra atención previamente (Raimondo Anselmino, et al., 2018). Denominamos a la misma como *listado de recomendaciones o recomendación enumerada.* Consiste en una especie derivada y amplificada del proto-género discursivo lista —cuyo rol en nuestras sociedades escriturales ha sido analizado por Goody (como se cita en Verón, 2015)—, que tiene también como antecedente mediático directo a un tipo de notas frecuentes en las revistas periódicas y semanarios. Las publicaciones de las cuentas de @clarincom y @lanacion que asumen este molde se destacan por los siguientes atributos:

a) promueven un tipo de organización textual con formato lista que se manifiesta tanto en el posteo como en la organización textual de la nota a la que este enlaza;

b) proponen al usuario una lectura ligera y secuenciada de elementos enumerados, en ocasiones al estilo de *tips* o sugerencias y recomendaciones útiles sobre un amplísimo espectro de temas, tales como: «Los 5 mejores vinos argentinos», «los 10 chistes prohibidos en los aeropuertos», «los 50 mejores restaurantes de Latinoamérica», «los 20 mejores murales del año», «las diez esquinas más peligrosas para cruzar» o «7 frutos rojos que no deben faltar en un plan alimentario saludable»;

c) remiten, generalmente, a un contenido publicado por fuera de las secciones tradicionales del periódico e, incluso, muchas veces vinculan con alguna de las otras unidades de negocio del medio como, por ejemplo, las revistas *Ohlalá*! o *Rolling Stone* en *La Nación* y *Vía Restó* o *Todo Viajes* en *Clarín*.

Para cerrar, y en relación con el segundo objetivo de este artículo, consideramos que el acercamiento cuasi-experimental mediante el cual nos propusimos poner a prueba la integración de métodos computacionales con las labores artesanales de análisis sociosemiótico ha mostrado resultados que son satisfactorios. Si bien el trabajo con algoritmos de *clustering* ha sido, podría decirse, un pequeño ejercicio de aplicación, consideramos que alcanza para seguir avanzando en el camino por consolidar una estrategia metodológica de combinación en el campo de estudios sobre la configuración que asumen los discursos publicados en plataformas mediáticas como Facebook.

Estamos, así, en la senda por delimitar un nuevo subcampo o área de especialización semiótica que se propone el estudio empírico de la puesta en discurso pero que —dado el tipo de discursos mediatizados que indaga— lo hace desde una cooperación de saberes con otras disciplinas, como la ingeniería en sistemas de información, la lingüística computacional, o la estadística. Lo hacemos poniendo énfasis en que se trata de una cooperación interdisciplinaria y no de una mera apropiación de esos saberes específicos. Como también lo sugieren Karlsson y Sjøvaag (2017: 6) para el caso de estudios en comunicación en general o sobre periodismo digital en particular:

> como es poco probable que la automatización de procedimientos metodológicos completos, sin la participación humana, sea suficientemente válida para el estudio científico social de las noticias (Mahrt & Scharkow, 2013; Zamith & Lewis, 2015), la cooperación interdisciplinaria es esencial en este desarrollo. Los investigadores también se enfrentan cada vez más a desafíos para capturar objetos digitales fluidos, 'en vivo' y 'en movimiento' (Karlsson, 2012), como el desplazamiento dinámico, las noticias en movimiento, los sistemas de gestión de contenido (Rodgers, 2015), los agentes inteligentes y los 'robots' (Anderson, 2011; Clerwall, 2014; van Dalen, 2012), cuyos comportamientos no se pueden congelar tan fácilmente (Karlsson y Strömbäck, 2010) [traducción propia].

Una semiótica que podríamos denominar, al menos provisoriamente, como *semiodata*. No sería una semiótica aplicada a la minería de datos sino, más precisamente, una semiótica combinada que integra (y se ve fortalecida por) ese otro saber.

Notas

1. En Raimondo Anselmino (2012) se describe una *ruptura de escala* (Verón, 2001) en el proceso histórico de mediatización, propiciada por las perturbaciones que introduce el conjunto Internet/dispositivos-móviles/redes-sociales. Dicha ruptura altera las relaciones establecidas entre las distintas instituciones de la sociedad posindustrial y el sistema de medios, e impacta sobre todo en las condiciones de credibilidad de los medios tradicionales.

2. Se explican las razones por las cuales es conveniente estudiar a los medios como parte de un sistema mayor siguiendo «la hipótesis de que las transformaciones de los diferentes soportes mediales no son autónomas, sino que se derivan fundamentalmente de los cambios dominantes en el sistema entendido como totalidad» (Valdettaro, 2008: 40).

3. Los principales resultados sobre las dimensiones a, b y c han sido publicados en los siguientes artículos: Raimondo Anselmino, Sambrana y Cardoso (2017), Raimondo Anselmino, Cardoso, Rostagno y Sambrana (2018), Raimondo Anselmino, Cardoso, Rostagno y Sambrana (2019) y Raimondo Anselmino (2019).

4. Aplicación que, al momento de la primera captura, nos permitió extraer datos de diferentes secciones de la plataforma de Facebook, a través de su Application Programming Interface (API), disponible en: https://apps.facebook.com/netvizz/, accedida por última vez el 16/04/2018. En la siguiente extracción, como veremos, se empleó una herramienta informática creada *ad hoc*.

5. A raíz del conflicto con Cambridge Analytica, a comienzos de 2018 Facebook realizó severos cambios en su política de privacidad, tras lo cual ninguna APP (como Netvizz) pudo utilizar la API de la plataforma para recolectar los posteos de una página pública. Por ello, decidimos desarrollar, entre otras, la herramienta informática *BuscaPosteosFacebook*, *un script* que accede automáticamente a la búsqueda avanzada de Facebook y recolecta de esa forma datos y metadatos de cada uno de los posteos resultados de la búsqueda.

6. Entre 2010 y 2017, de 122 posteos de este tipo solo 9 llevan al diario; los demás, o son contenido dentro del mismo Facebook o envían a Youtube.

7. El código fuente utilizado es de acceso libre y se encuentra en https://github.com/pepeleproso/clustertexto

8. Trivial en el sentido de que remite a aquello que se define como soft news, si se recupera la clásica distinción entre noticias duras —generalmente referidas a temas como política, economía o internacionales— y noticias blandas —circunscritas a secciones como espectáculos o sociedad.

9. Según se observa en ambos corpus totales, desde comienzos de 2015 tanto @clarincom como @lanacion suben directamente desde la plataforma de Facebook todos los videos que componen sus posteos; antes de esa fecha, era habitual que el contenido audiovisual de las publicaciones en las dos *fanpages* fuera subido desde YouTube.

Referencias bibliográficas

Alqurashi, T. y Wang, W. (2019). Clustering ensemble method. *Int. J. Mach. Learn. & Cyber*. 1227–1246. https://link.springer.com/article/10.1007/s13042-017-0756-7

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

Barberá, J. (2017). *Análisis de los factores asociados a la elección de estudios universitarios utilizando técnicas de agrupamiento*. Valencia: Escola Tècnica Superior d'Enginyeria Informàtica Universitat Politècnica de València.

Barthes, R. (1993). *La aventura semiológica*. Paidós.

Berry, D. (2011). The computational turn: thinking about the digital humanities. *Culture Machine*, Vol. 12. https://sro.sussex.ac.uk/id/eprint/49813/1/BERRY_2011-THE_COMPUTATIONAL_TURN-_THINKING_ABOUT_THE_DIGITAL_HUMANITIES.pdf

Brun, R. y Senso, J. (2004). Minería Textual. *El profesional de la información*, Vol. 13, N.°1. http://profesionaldelainformacion.com/contenidos/2004/enero/2.pdf

Bunz, M. (2017). *La revolución silenciosa. Cómo los algoritmos transforman el conocimiento, el trabajo, la opinión pública y la política sin hacer mucho ruido*. Cruce.

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

Fayyad, U., Piatetsky-Shapiro G. y Smyth P. (1996). *The KDD Process for Extracting Useful Knowledge from Volumes of Data*. New York: Communications Of The ACM.

Fernández, J.L. (2018). *Plataformas mediáticas. Elementos de análisis y diseño de nuevas experiencias*. La Crujía.

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

Fisher, S. y Verón, E. (1986). Théorie de l'énonciation et discours sociaux. *Etudes de lettres*, 211, 71-92. (Teoría de la enunciación y discursos sociales. Traducción al español de Mollinedo, S. para la cátedra Teorías y Medios de Comunicación, UBA).

Gindin, I. y Busso, M. (2018). Investigaciones en comunicación en tiempos de big data: sobre metodologías y temporalidades en el abordaje de redes sociales. Castellón. *adComunica. Revista Científica de Estrategias, Tendencias e Innovación en Comunicación*, N.° 15. https://www.e-revistes.uji.es/index.php/adcomunica/article/view/4965

Günther, E. y Quandt, T. (2016) Word Counts and Topic Models. Automated text analysis methods for digital journalism research. *Digital Journalism*, Vol. 4, 75-88. https://www.tandfonline.com/doi/10.1080/21670811.2015.1093270

Honneth, A. (1999). Comunidad. Esbozo de una historia conceptual. *Isegoría, Revista de Filosofía Moral y Política*, N.° 20. http://isegoria.revistas.csic.es/index.php/isegoria/article/view/89/89

Hughes, H. (1981). *News and the human interest story*. New Brunswick: Transaction.

Jakobson, R. (1975). Lingüística y Poética. En *Ensayos de lingüística general*. Seix Barral.

Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, Vol. 28 N.°1. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.115.8343&rep=rep1&type=pdf

Karlsson, M. y Sjøvaag, H. (2017) Rethinking Research Methods for Digital Journalism Studies. En Franklin, B. y Eldridge II, S. (Eds.) *The Routledge Companion to Digital Journalism Studies*. Routledge.

Larrondo Ureta, A. (2008) *Los géneros en la Redacción Ciberperiodística. Contexto, teoría y práctica actual*. Universidad del País Vasco.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutman, M., Jebara, T., King, G., Macy, M., Roy, D. y Van Alstyne, M. (2009). Computational Social Science. *Science*, Vol. 323, N.° 5915. https://pubmed.ncbi.nlm.nih.gov/19197046/

Leale, G., Raimondo Anselmino, N., Cardoso, A. L. y Rostagno, J. (2020). *BuscarPosteosFacebook* (Copyright RL-2021-01144224-APN-DNDA#MJ). Rosario: Consejo Nacional de Investigaciones Científicas y Técnicas, Universidad Nacional de Rosario, Universidad Tecnológica Nacional. https://github.com/Departamento-Sistemas-UTNFRRO/buscarPosteosFacebook

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). Trending: The promises and the challenges of big social data. En Gold, M. (Ed.). *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.

Morley, D. (1996). Comunicación doméstica: tecnologías y sentidos. *Televisión, audiencias y estudios culturales*. Amorrortu.

Parrat, S. (2008). *Géneros periodísticos en prensa*. Ciespal.

Peralta, D. y Urtasun, M. (2007). *La crónica periodística. Lectura crítica y redacción*. Crujía.

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

Raimondo Anselmino, N. (2012). *La prensa online y su público. Un estudio de los espacios de intervención y participación del lector en Clarín y La Nación*. Teseo.

——— (2019). Trayectorias y metamorfosis en la configuración discursiva de las publicaciones de los diarios argentinos Clarín y La Nación en Facebook (2010-2017). Exposición realizada en *el 14° Congreso Mundial de Semiótica IASS-IAS, Trayectorias*. Buenos Aires, 9 al 13 de septiembre.

Raimondo Anselmino, N., Cardoso, A., Rostagno, J. y Sambrana, A. (2018). El discurso de la prensa argentina en tiempos de algoritmos: una mirada diacrónica sobre la composición de posteos en las *fanpages* de Clarín y La Nación. *Âncora, Revista Latino-americana de Jornalismo*, Vol. 5(1) enero-junio. https://periodicos.ufpb.br/index.php/ancora/article/view/42043

——— (2019). Recursos paratextuales y paralingüísticos en las fanpages de los periódicos argentinos Clarín y La Nación. Atributos del discurso de la prensa en las redes. *Revista Perspectivas de la Comunicación*, Vol. 2, N.° 2, 245-280. http://revistas.ufro.cl/ojs/index.php/perspectivas/article/view/2028

Raimondo Anselmino, N., Cardoso, A., Rostagno, J. (2018). Articulación artesanal-computacional para el estudio interdisciplinario de posteos en cuentas de Facebook. Relato de una experiencia asequible. *Anales de las 47 JAIIO*. Buenos Aires. https://47jaiio.sadio.org.ar/sites/default/files/STS-12.pdf

Raimondo Anselmino, N., Sambrana, A. y Cardoso, A. (2017). Medios tradicionales y redes sociales en Internet: un análisis de los posteos compartidos por los diarios argentinos Clarín y La Nación en Facebook (2010-2015). *Revista Astrolabio*. Nueva Época, n.° 19, Monográfico «La experiencia de los públicos en América Latina», 32-68. https://revistas.unc.edu.ar/index.php/astrolabio/article/view/17787

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

Rončáková, T. (2019). *Žurnalistické žánre. Ružomberok*: Verbum.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, Vol. 20. https://www.sciencedirect.com/science/article/pii/0377042787901257

Steimberg, O. (1998). *Semiótica de los medios masivos. El pasaje a los medios de los géneros populares.* Atuel.

Steimberg, O. (2001). Sobre algunos temas y problemas del análisis del humor gráfico. *Signo y seña. Revista del Instituto de Lingüística*, N.° 12, 99-117. http://revistascientificas.filo.uba.ar/index.php/sys/article/view/5606

Seghezzo, F. (2021). *#35. La Nación y los desafíos para los medios de comunicación. Entrevista realizada por Leónidas Rojas* para Comscore Talks en español. https://www.comscore.com/lat/Insights/Podcast/Comscore-Talks-en-Espanol?utm_campaign=LATAM_REG_JAN2021_MT_COMSCORE_TALKS&utm_medium=email&utm_source=comscore_elq_LATAM_REG_JUN2021_MT_COMSCORE_TALKS_T35

Tan, A.-H. (1999). Text Mining: The state of the art and the challenges. *Proceedings, PAKDD'99 Workshop on Knowledge discovery from Advanced Databases* (KDAD'99), Beijing, 65-70. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.132.6973&rep=rep1&type=pdf

Touileb, S. y Salway, A. (2014). *Constructions: A New Unit of Analysis for Corpus-Based Discourse Analysis. PACLIC*, 28: 634-43. https://aclanthology.org/Y14-1072.pdf

Valdettaro, S. (2008). Algunas consideraciones acerca de las estrategias del contacto: del papel a la in-mediación de las interfaces. *Revista LIS, Letra, Imagen, Sonido.* Ciudad Mediatizada, N.° 1. https://publicaciones.sociales.uba.ar/index.php/lis/article/view/3616

Van Dijck, J. (2016). *La cultura de la conectividad. Una historia crítica de las redes sociale*s. Siglo XXI.

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

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

——— (2001). *Espacios mentales. Efectos de agenda* 2. Gedisa.

——— (2013). *La semiosis social, 2. Ideas, momentos, interpretantes*. Paidós.

——— (2015). Teoría de la mediatización: una perspectiva semio-antropológica. *CIC Cuadernos de Información y Comunicación*, Vol. 20, 173-182.

Fecha de recepción: 31/8/2021 | Fecha de aceptación: 5/10/2021