BIG DATA Y DATA MINING. UN ANÁLISIS CRÍTICO ACERCA DE SU SIGNIFICACIÓN PARA LAS CIENCIAS PSICOSOCIALES A PARTIR DE UN ESTUDIO DE CASO

GASTÓN BECERRA*, JUAN PABLO LÓPEZ ALURRALDE **

* Universidad de Buenos Aires(Argentina) , ** Universidad Nacional de Quilmes (Argentina)

gastonbecerra@sociales.uba.ar

{PSOCIAL} Revista de Investigación en Psicología Social

ISSN 2422-619X

Resumen. Big data y Data mining son presentados desde una perspectiva crítica que los caracteriza como una configuración social marcada por la disponibilidad de grandes volúmenes de datos –especialmente digitales– y de técnicas tendientes a su exploración. Luego, se presenta un caso de exploración de datos sobre 40.000 comentarios de Facebook acerca de noticias que tratan con la desaparición de Santiago Maldonado, donde además de informar los resultados, se explicitan los pasos seguidos tanto en la extracción de datos, en el preprocesamiento y el análisis. Finalmente, se evalúa su relevancia y significado para las ciencias sociales y psicosociales, a través de una discusión de sus mitos, y se argumenta en torno a su potencial metodológico, en la medida en que puedan ser integrados a programas de investigación, como el de las representaciones sociales.

Palabras Claves. Big data – Minería de datos–Estudios críticos de datos – Representaciones sociales – Santiago Maldonado

Abstract. We describe Big data and Data mining from a critical perspective that characterizes them as a social configuration marked by the availability of large volumes of data -especially digital- and techniques tending to their exploration. Then, we present a case of data mining on 40,000 Facebook comments for news articles regarding the disappearance of Santiago Maldonado, where in addition to reporting the results, we describe the steps followed in the extraction of data, preprocessing and analysis. Finally, We evaluate its relevance and meaning for the social and psychosocial sciences, through a discussion of its myths. There, we discuss its methodological potential, insofar as they can be integrated into particular research programs, such as that of social representations.

Keywords. Big data – Data mining –Critical data studies – Social representation theory – Santiago Maldonado

Enviado.01-09-2017 | **Aceptado.** 15-12-2017

El crecimiento de internet no sólo resulta en la constitución del mayor sistema de comunicación del que la historia humana tenga registro. Como consecuencia del uso y las interacciones en la red, se producen cuantiosos datos que patentan, entre otras, las conductas y elecciones de los usuarios, y de todo proceso conectado a la red. De esta manera, tanto los millones de posts que se publican a diario en redes sociales como Facebook, como los datos que producen sensores climáticos, o los registros de compras en un supermercado son plausibles de ser analizados por quienes los detentan. Así, mientras las redes sociales pueden servirse de la explotación de los datos generados por sus usuarios para delinear esquemas publicitarios segmentados, instituciones civiles y climatológicas pueden, en el segundo caso, estudiar la proyección del clima a futuro. La clave, en cada caso, estará en poder diseñar y desplegar las herramientas pertinentes para hacerse de nueva información. Estas prácticas forman parte de lo que se conoce como Big data y Data mining.

En las siguientes secciones se tratará lo siguiente: (1) una introducción conceptual a Big

data y Data mining; (2) la presentación del proceso y de los resultados de un caso de estudio propio sobre 40.000 comentarios de Facebook en torno a la desaparición de Santiago Maldonado utilizando técnicas de Data mining y; (3) reflexiones finales en torno a la significación de Big Data y Data mining para la investigación en psicología social.

Big data y Data mining

De acuerdo con *Internet Live Stats,* al día de hoy contamos con 3,8 mil millones de usuarios conectados a la red (casi un 50% de la población mundial) y con una cifra mucho más alta de dispositivos interconectados. En este contexto, vivimos en un mundo de una producción inconmensurable de datos, donde la noción de Big data cobra una presencia creciente en los titulares y comunicaciones de diarios, revistas y publicaciones científicas.

Big data es la posibilidad de trabajar con grandes bases de datos, construidas a partir de la extracción de datos y su procesamiento. Como señala la literatura, puede ser fácilmente caracterizado a partir de tres "V": volumen (haciendo referencia a grandes cantidades de datos), variedad (pues puede implicar varios tipos de datos), y velocidad (en la medida en la que los datos se acumulan progresivamente) (Berman, 2013). Estudios más recientes han propuesto incorporar una cuarta V: Veracidad (Jagadish, 2015). Las razones que motivaron esta nueva reflexión serán expuestas a lo largo de este trabajo.

Sin embargo, si nos quedamos con esta primera definición, algo se pierde. Y es que una mirada más amplia a Big data lo comprende como un escenario social caracterizado por la disponibilidad de vastos registros de datos y de nuevas capacidades técnicas de análisis en un contexto particular. Y esto es un fenómeno social de primera importancia. Por ello, junto con el avance del Big data –en gran parte gracias a su paulatino reconocimiento por parte de las ciencias sociales–, en los últimos años se ha ido constituyendo un sub-campo denominado *critical data studies* en la que se han popularizado los siguientes 7 mandatos (Kitchin & Laurialt, 2014; Dalton & Thatcher, 2014; Illiadis & Russo, 2016):

1. Sitúa el big data en el tiempo y espacio;
2. Expone que los datos son inherentemente políticos (y a qué intereses sirven);
3. Problematiza la compleja e indeterminada relación entre datos y sociedad;
4. Ilustra de qué formas el dato nunca es neutral;
5. Denuncia la falacia de que los datos "hablan por sí mismos" y que big data reemplazará a la investigación actual (con pocos datos);
6. Explora de qué manera este nuevo régimen de datos puede utilizarse al servicio de proyectos socialmente progresivos;
7. Examina de qué manera la academia se acerca a este nuevo régimen de datos y explora las oportunidades de dicho acercamiento.

Por su parte, el término Data mining remite al conjunto de prácticas y técnicas utilizadas para el procesamiento de estos datos. Más específicamente, es el proceso de indagación en torno a patrones que surgen del análisis de grandes volúmenes de datos. (Han, Kamber, & Pei, 2012). Como es obvio, los datos a ser analizados por estas herramientas son de carácter digital, y generalmente provienen del entramado que supone internet.

El método comprende, esquemáticamente, un proceso que puede dividirse en dos partes: pre-procesamiento y "minado" propiamente dicho. Durante el pre-procesamiento se realizan las tareas de limpieza, integración y selección de datos. Ejecutadas estas tareas, se procede con las operaciones consistentes, la mayoría de las veces, en la extracción de patrones y correlaciones. Por último, se exportan estos resultados a representaciones gráficas que terminan por sintetizar la información obtenida.

Los datos a ser analizados pueden clasificarse en dos grandes grupos: estructurados y no estructurados. Los primeros son la minoría y se generan –la mayoría de las veces– automáticamente. Los segundos –datos no estructurados– representan la gran mayoría de datos disponibles en el ciberespacio y se presentan generalmente en forma de texto e imagen (Gandomi & Haider, 2015). Tienen un gran valor potencial para las empresas del sector privado como para agencias de gobierno pues allí se patentan muchas de las preferencias y posiciones de los consumidores/ciudadanos. Mientras los primeros resultan de mecánicas básicas inherentes al funcionamiento de internet y no necesitan de la pro-actividad (ni conciencia) de los usuarios que los producen, los segundos son proyecciones patentes de la subjetividad de los usuarios, de sus opiniones, calificaciones, expresiones. Por ello, los datos no estructurados requieren técnicas de explotación más sofisticadas en el campo del procesamiento del lenguaje natural.

El desafío es estructurar estos datos no estructurados para poder procesarlos a través de rutinas de análisis automatizadas. En otras palabras, se trata de hacer accesible el lenguaje humano al lenguaje de las máquinas para su exploración.

Un caso de estudio de Data mining: los comentarios de noticias en torno a la desaparición de Santiago Maldonado

Para poder evaluar mejor las posibilidades y limitaciones del Data mining en el marco de las prácticas pertenecientes al dominio del Big data, propondremos en esta sección el recorrido paso a paso de la puesta en operación de una rutina de procesamiento del lenguaje natural, en general, y de análisis de sentimientos y opiniones, en particular, sobre una base de 40.000 comentarios adjuntados a noticias publicadas en Facebook relativas al “caso Maldonado”. Esta indagación tiene fines ilustrativos. En la tercera sección discutiremos algunos mitos en torno a Big data y Data mining, con especial énfasis en aclarar su significación para las ciencias sociales y psicosociales.

Para los siguientes análisis se utilizaron técnicas de procesamiento cuantitativo de lenguaje natural, por medio de rutinas programadas en lenguaje R, y basadas mayormente en el universo de librerías Tidy (Silge & Robinson, 2017).

Primer paso: extracción de datos

El primer objetivo que se buscó realizar en la rutina fue buscar "santiago+maldonado" en un buscador de internet, restringiendo la búsqueda a un conjunto de sitios de noticias de Argentina. Esto devolvió casi 1.300 noticias.

Luego, la rutina buscó esas noticias en Facebook como post publicados por los mismos diarios y recogió más de 40.000 comentarios dejados por los visitantes entre los meses de agosto y diciembre de 2017, que conformarán nuestro corpus de análisis

.

Gráfico 1 – Noticias colectadas con el criterio "santiago+maldonado"

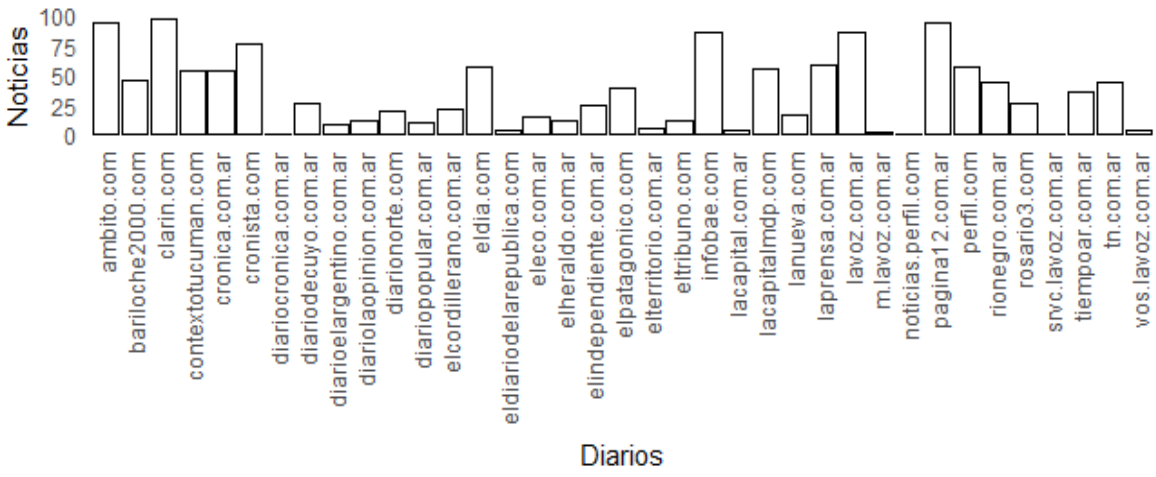
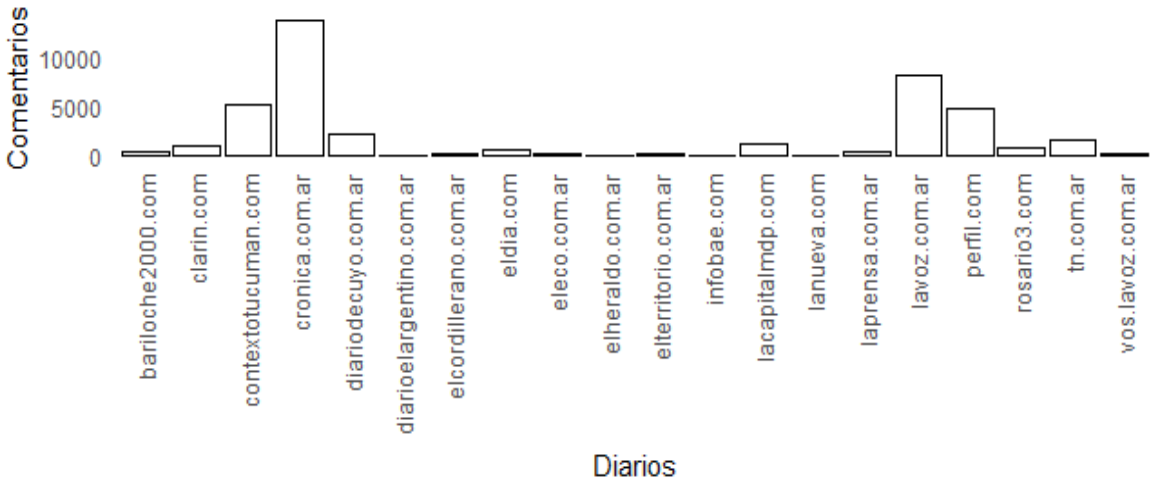


Gráfico 2 – Comentarios colectados donde se discuten las noticias



Una de las principales limitaciones de esta estrategia es que las distintas políticas de cada medio acerca de la difusión en redes sociales resultó en una distribución de comentarios muy desigual incluyendo la falta de representación de varios medios registrados en el paso anterior, y la sobre-representación de algunos otros (como el diario Crónica, Contexto de Tucumán, La Voz del Interior, o Perfil). Esto es algo que, para los fines demostrativos del caso en nuestro estudio, no será rectificado con mejoras en la rutina de adquisición de datos.

Segundo paso: preprocesamiento

El corpus de comentarios construido es un clásico ejemplo de una fuente de datos no estructurada: los elementos que lo integran no se ciñen a consignas o preguntas disparadoras, ni a limitaciones o filtros más allá de los impuestos por las reglas de uso de cada medio y de Facebook; algunos comentarios son intervenciones espontáneas mientras que otros son respuestas a comentarios anteriores; además de todas las licencias del lenguaje natural, no es infrecuente observar el uso de emoticones, elementos de hipertexto, links, y citas pegadas de otras fuentes, algunas en otros idiomas; su extensión varía entre 1 y 5.200 palabras, con una media de 8.

Previo a su procesamiento se corrió una rutina muy elemental de limpieza, consistente en el reemplazo de caracteres con tildes y en la eliminación de números, símbolos, y palabras comunes y poco significativas como artículos o conectores. Esto redujo las más de 10.000.000 palabras de todos los comentarios a aproximadamente 440.000.

Una mayor normalización de las palabras se lograría con rutinas de reemplazos de términos similares, sinónimos, acepciones locales, y términos de raíces similares -incluida la singularización de plurales-, o el reemplazo de términos genéricos, tareas que aquí no hemos emprendido.

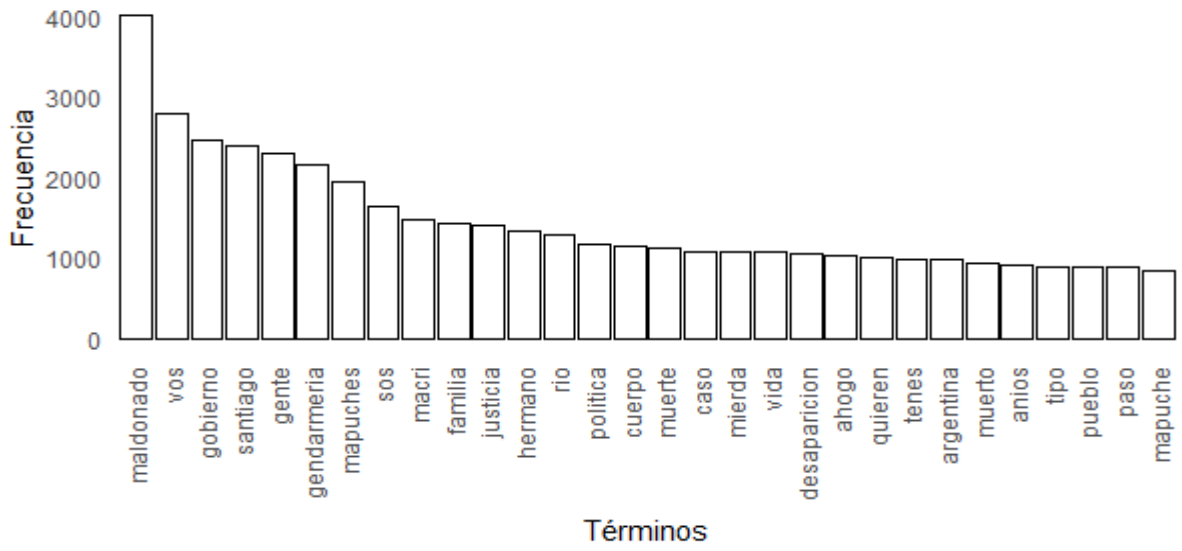
Luego, se eliminaron todas las puntuaciones. Esta es una decisión metodológica de un alto costo, ya que convierte a cada comentario en una "bolsa de palabras" sin una estructura semántica, al punto que no se distinguen diferentes oraciones, cambios de sujetos y objetos referidos, y demases recursos argumentativos. Los análisis elementales que presentamos en este estudio no dan cuenta de estos elementos. Sin embargo, existe una cuantiosa bibliografía que avanzan sobre cada una de estas limitaciones (Pang & Lee, 2008).

Un preprocesamiento importante que aquí hemos omitido es la detección y eliminación de comentarios de perfiles falsos u operadores que buscan influenciar la opinión pública, generalmente conocidos como *trolls* en los foros y redes sociales (DiPaola, 2017). Al respecto, hay diversos procedimientos sencillos que se pueden realizar en el nivel del análisis de patrones de posteos de usuarios, como considerar sus horarios, comentarios similares, perfiles con posteos por arriba de la media, entre otros.

El resultado de estos primeros procesamientos nos devuelve una distribución de palabras que

sigue la ley de Zapf, donde muy pocos términos tienen una frecuencia significativa –la moda es de poco más de 4.000 y corresponde a "Maldonado"–, mientras que la media es de 1.

Gráfico 3 – Términos más frecuentes

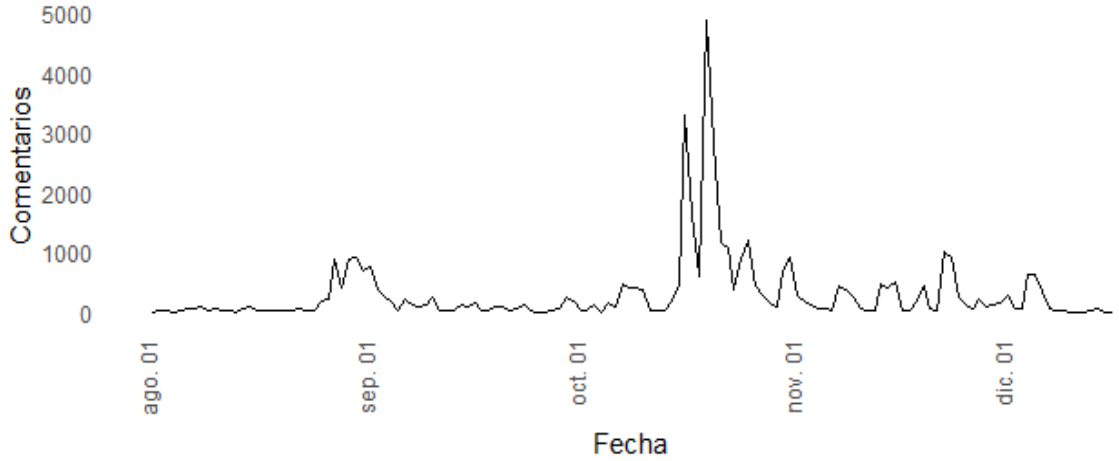


Tercer paso: análisis (procesamiento)

Existen diversas estrategias para analizar los datos colectados. Puede hacerse foco en los usuarios, para detectar cómo los distintos grupos de usuarios comentan e interactúan entre sí, como en los comentarios, para observar cómo diversos temas o tópicos –expresados como conjuntos de palabras– se desenvuelven en el tiempo.

En este estudio hemos optado por la segunda estrategia. Para ello, es importante considerar cómo se distribuye la base de comentarios en los distintos días desde el 1 de agosto de 2017.

Gráfico 4 – Distribución de comentarios por fechas



A pesar de la evidente disparidad en la muestra, creemos conveniente segmentar los comentarios en 2 grupos con un corte temporal en Octubre de 2017. Esto deja por un lado a los comentarios de agosto y septiembre de 2017, meses en las que toma estado público la desaparición de Santiago Maldonado, y durante los cuales se realizan dos marchas multitudinarias; y, por otro lado, los comentarios registrados desde octubre de 2017, meses en los que se encuentra el cuerpo de Maldonado en el Río Chubut comenzando así las investigaciones forenses.

Una primera forma muy elemental de ver estas transformaciones diacrónicas es analizando la diferencia de posición de las palabras más repetidas en ambos grupos del corte temporal. Así, se observa que hacia el final de la muestra ganan prominencia palabras relativas al peritaje sobre las causas de la muerte como mayores referencias relativas a los familiares de Santiago Maldonado y sus reclamos.

Tabla 1 – Diferencias en orden (posición por frecuencia) de palabras antes/después de Octubre/2017

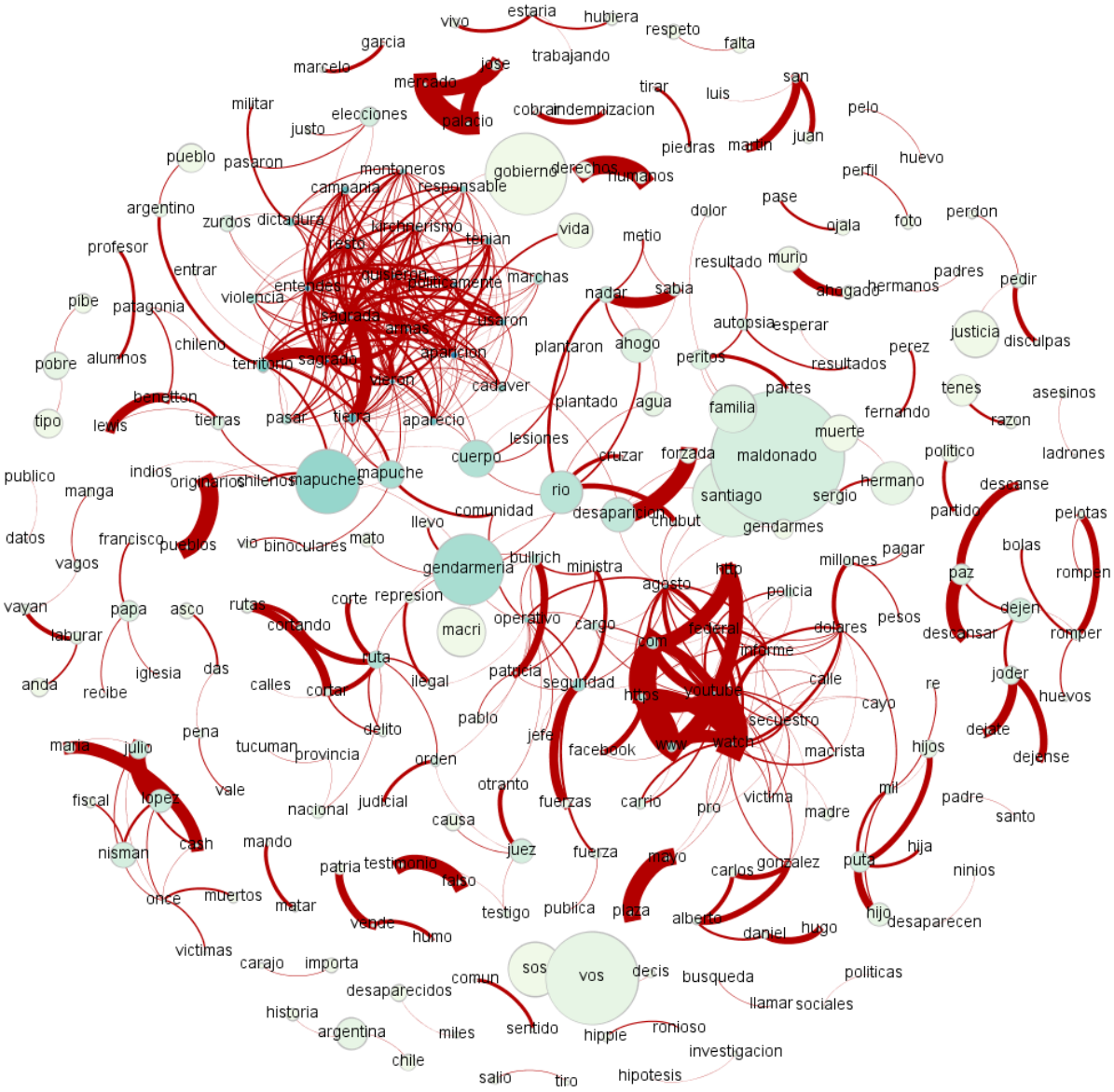
Palabra	Posición T1	Posición T2	Diferencia
Peritos	3690	36	-3654
Ahogo	1233	16	-1217
Papa	598	36	-563
Murio	467	24	-443
Paz	395	40	-355
Cuerpo	268	14	-254
Hermano	198	9	-189

Luego, para comprender cómo se estructura la discusión podemos comenzar por analizar las correlaciones de palabras, es decir, rastrear la presencia conjunta de las palabras en los distintos comentarios y cuantificar su peso. Algunas de estas correlaciones se deben a términos compuestos (e.g., “Sergio Maldonado”, “Buenos Aires”, etc.). Otras, no obstante, permiten empezar a observar campos semánticos, lo que se considerará como el principal criterio para el análisis de temas.

Estas relaciones pueden ser exportadas a una plataforma de análisis de redes -aquí utilizamos Gephi- para su visualización. Para construir este gráfico, primero filtramos las palabras de más de 100 repeticiones y calculamos sus correlaciones (coeficiente de Pearson); luego, se seleccionaron las 500 relaciones (edges) más fuertes sobre el

total de las más de 500.000 relaciones calculadas, las cuales correspondían a 261 palabras únicas (nodes). En el gráfico, el tamaño del círculo corresponde a la frecuencia de la palabra, el tono de verde a su centralidad (cantidad de relaciones con otras palabras), mientras que el grosor y tono de rojo refiere al peso de la relación entre las dos palabras conectadas.

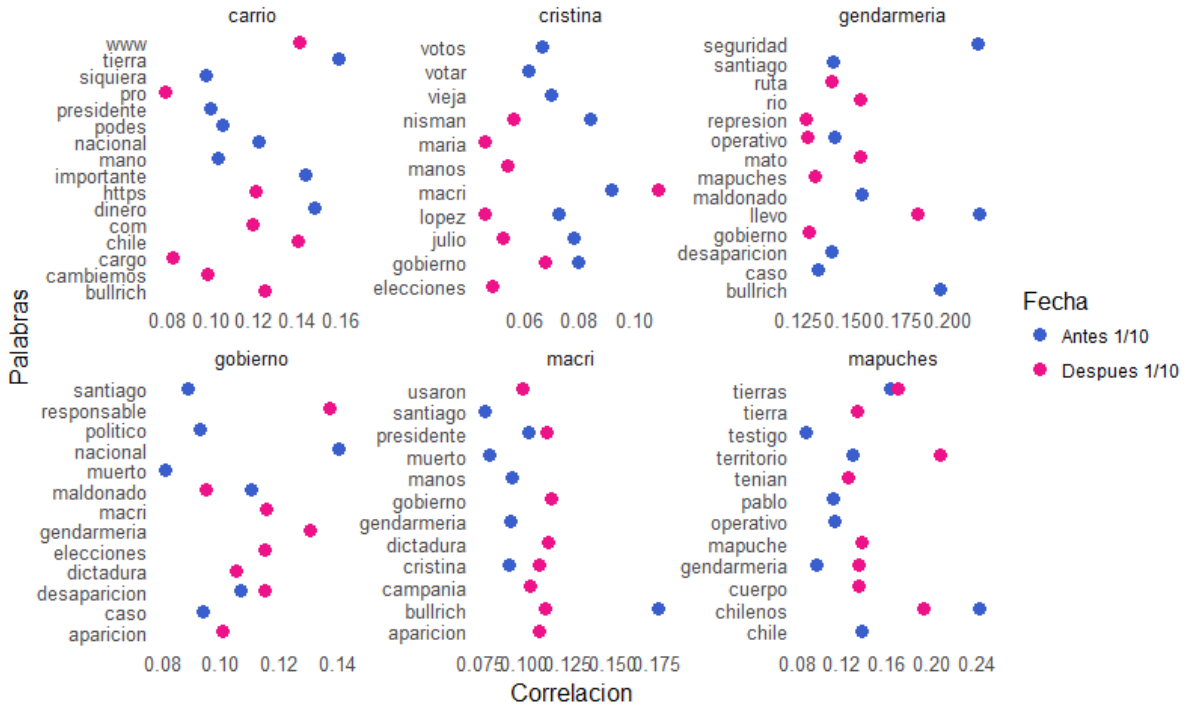
Gráfico 5 – Correlaciones de palabras



Si observamos las principales relaciones se notan secuencias de palabras que se pueden entender como campos semánticos que referencian distintos tópicos. Por ejemplo, en torno a Maldonado se observan Santiago, Sergio, familia, y gendarmes pero también las palabras relativas a las circunstancias de su muerte, como peritos, partes, desaparición, forzada, río, Chubut; otro cluster lo forman Macri, gendarmería, Bullrich, represión, ilegal, corte, ruta, delito; uno cercano lo forman Mapuches, Mapuche, territorio, tierra, sagrada, y cercano a ellos kirchnerismo, zurdos, dictadura, militar, montoneros, campora, gobierno, y derechos y humanos; o fiscal, Nisman, Jorge, López, María, Cash, Once, muertos y víctimas. Otros clusters más pequeños pero más fuertes son: Lewis, Benetton, tierras, patagonia, chileno; Francisco, papa, iglesia, recibe; así como también otras palabras que refieren a frases recurrentes en los intercambios entre foristas como manga, vagos, vayan, laburar; o dejen, romper, pelotas, bolas, huevos, descansar, paz.

Estas correlaciones también pueden ser observadas diacrónicamente, de forma tal de comparar los distintos contextos en los que se inserta un término en distintos momentos. Si retomamos aquí el corte temporal de nuestra muestra, se puede observar que, por ejemplo, “Cristina” correlaciona fuertemente con (Alberto) “Nisman” y con (Jorge) “Julio” “López” hacia el inicio de la muestra y que esta relación fue perdiendo fuerza hacia el final; o que entre “Gobierno” y “Responsable” y “Gendarmería” se genera una relación fuerte hacia el final de la muestra..

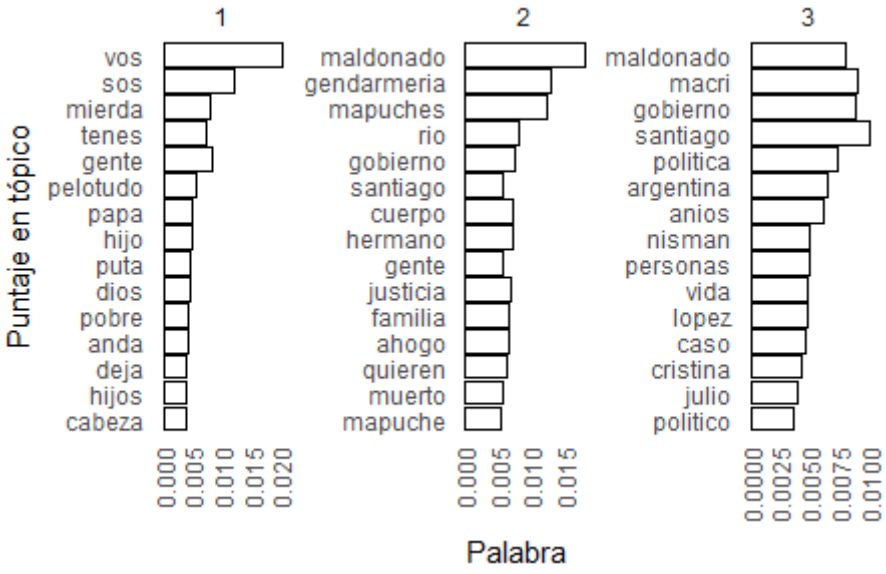
Gráfico 6 – Correlaciones de palabras antes/después de Octubre/2017



Luego, sobre la base de co-ocurrencias de palabras se puede hacer una clasificación de comentarios en diferentes temas (topic modelling), implementando técnicas elementales de inteligencia artificial. Estas técnicas se suelen dividir en dos grandes enfoques: con entrenamiento o supervisión (humana), y sin ella, ya sea porque la rutina clasifica los textos a partir de criterios y correcciones provistas por el investigador o analista, o porque lo hace exclusivamente sobre correlaciones positivas y negativas entre palabras.

En este estudio optamos por la segunda opción. Particularmente, se trabajó con el modelo Latent Dirichlet Allocation que genera un número determinado de tópicos a partir de asignar valores a las distintas palabras, y subsecuentemente, un score a cada comentario en cada tópico (con suma 1). Aquí predeterminamos el número de tópicos en 3, cuyas palabras de mayor puntaje se pueden observar en el gráfico 7.

Gráfico 7 – Principales palabras de cada tópico

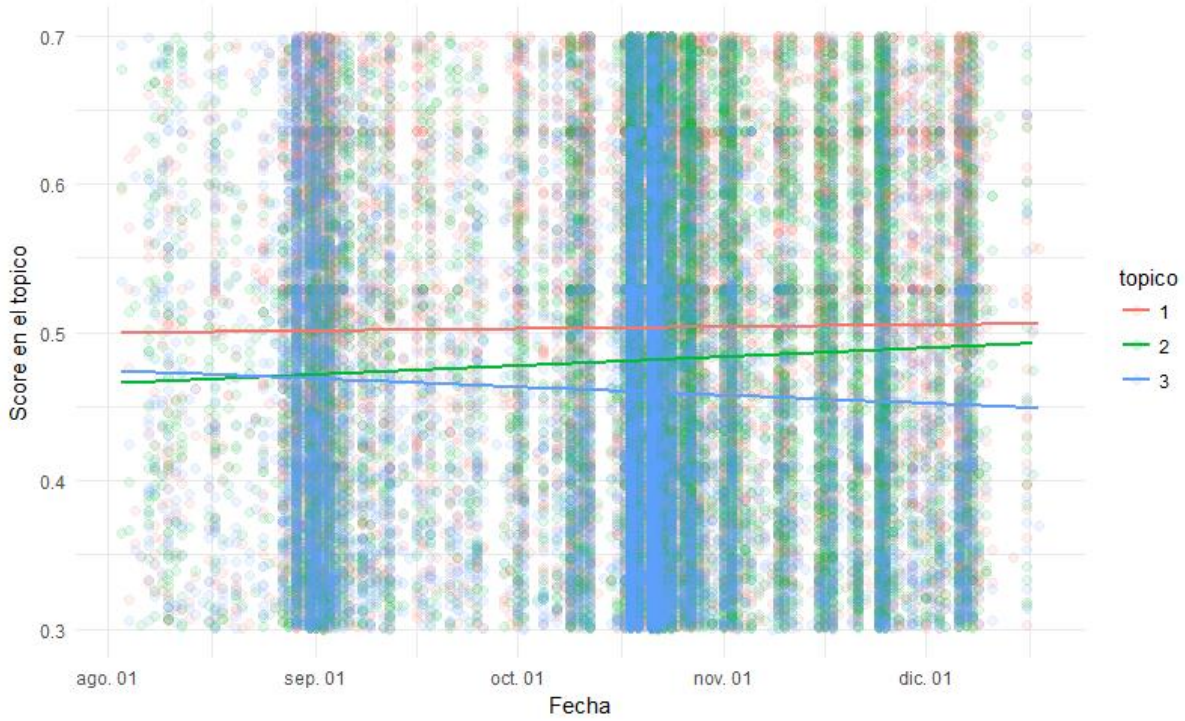


Se debe advertir que, como en este estudio no se realizaron tareas de limpieza ni de exclusión de comentarios "copiados y pegados" o con métricas fuera de la media, la constitución de los tópicos se encuentra fuertemente condicionado por mensajes tendenciosos y de trolls que no reflejan al grueso de los comentarios (los cuales tienen una distribución más pareja entre al menos 2 de los tópicos). Con estas salvedades, podemos buscar los comentarios que más puntaje obtuvieron en cada tópico para tener una idea de su sentido. Así, se puede observar que en el tópico #1 predominan los mensajes dirigidos a otros foristas, generalmente con insultos y descalificaciones; en el tópico #2 los comentarios tratan mayormente acerca de lo sucedido a Santiago Maldonado, generalmente con referencias al conflicto de las tierras de los

pueblos originarios y su interpretación política; en cuanto al tópico #3, el eje de la discusión se corre hace el trasfondo político y partidario.

Si se grafican los mensajes en los 3 tópicos (excluyendo los puntajes extremos de más de 0.7, y su contraparte de menos de 0.3 en algún tópico), se observa que la discusión del tópico #1 se mantiene constante en el transcurso del tiempo, mientras que la discusión del tópico #2 gana preeminencia, en la medida en que decrece la del tópico #3.

Gráfico 8 – Tópicos por tiempo y score de los mensajes



Finalmente, un último análisis que nos interesa presentar trata con la orientación semántica positiva/negativa de los comentarios. Este tipo de análisis tiene distintos nombres en la literatura, como "análisis de sentimientos", "de emociones", "de estados subjetivos", o "de polaridades" (Pang & Lee, 2008). La idea que subyace a todos ellos es que los mensajes se pueden clasificar por su orientación en diferentes escalas –bueno/ malo; positivo/ negativo; agradable/ desagradable–, o bajo alguna categoría vinculada a un estado emocional –como la Rueda de Plutchik que distingue entre enojo, anticipación, alegría, confianza, sorpresa, tristeza y disgusto–. Este tipo de análisis suele ser muy utilizado en el análisis de reseñas y críticas de productos, o en el análisis de expresiones que contienen valoraciones de índole política.

Aquí también priman dos abordajes: los que hacen uso de inteligencia artificial y entrenamiento, y los que se basan en léxicos con valoraciones. En cualquier caso la idea es que distintas palabras tienen diferentes valoraciones,

las cuales pueden ser calculadas de diversas formas para dar cuenta de la orientación global del comentario.

En este trabajo optamos por el segundo enfoque (uso de un léxico con valoraciones). Actualmente se cuenta con varias herramientas informáticas comerciales de análisis estadístico y procesamiento de fuentes cualitativas que incluyen este tipo de léxicos, incluso en castellano. La especificidad del lenguaje utilizado en el corpus de este estudio, y nuestro interés por explorar los distintos aspectos de esta indagación, nos motivó a construir nuestro propio léxico. Para ello, realizamos una breve encuesta donde se ofrecían 15 palabras (seleccionadas al azar y extraídas del corpus) a cada encuestado para que las valorase en el continuo que va desde negativo (-1) hasta positivo (+1), pasando por neutral (0). La encuesta fue compartida en Facebook y en sólo 2 días obtuvo más de 450 respuestas que arrojaron valoraciones para las primeras 1.000 palabras más significativas del corpus, de las que aquí hemos tomado la mediana. En esta toma se consultó además por datos sociodemográficos, lo que permitiría segmentar el léxico por sexo, edad o niveles de estudios, aunque esto no se consideró en los análisis.

Es importante destacar que las valoraciones del léxico reflejaron mayor presencia de palabras negativas. El gráfico 9 muestra la distribución de palabras por valoración (con base 0.1), mientras que el gráfico 10 muestra las principales 10 palabras de valoración más extrema en el polo positivo y el negativo.

Grafico 9 – Distribución de palabras por valoraciones en léxico

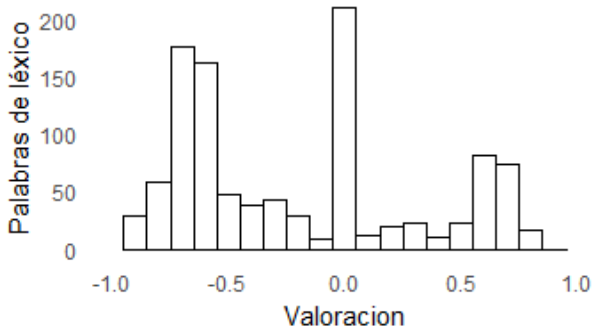
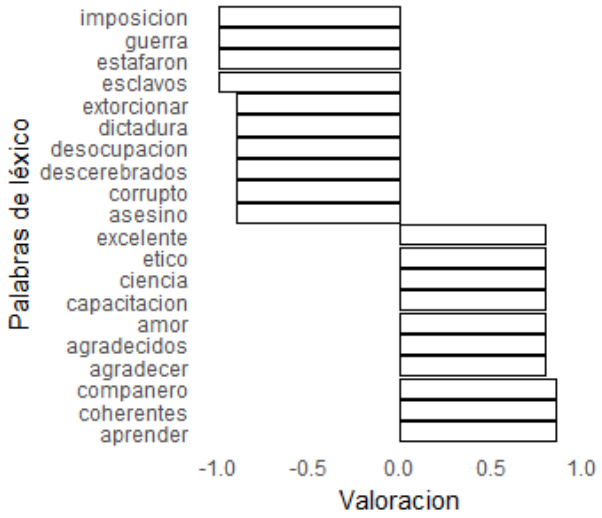
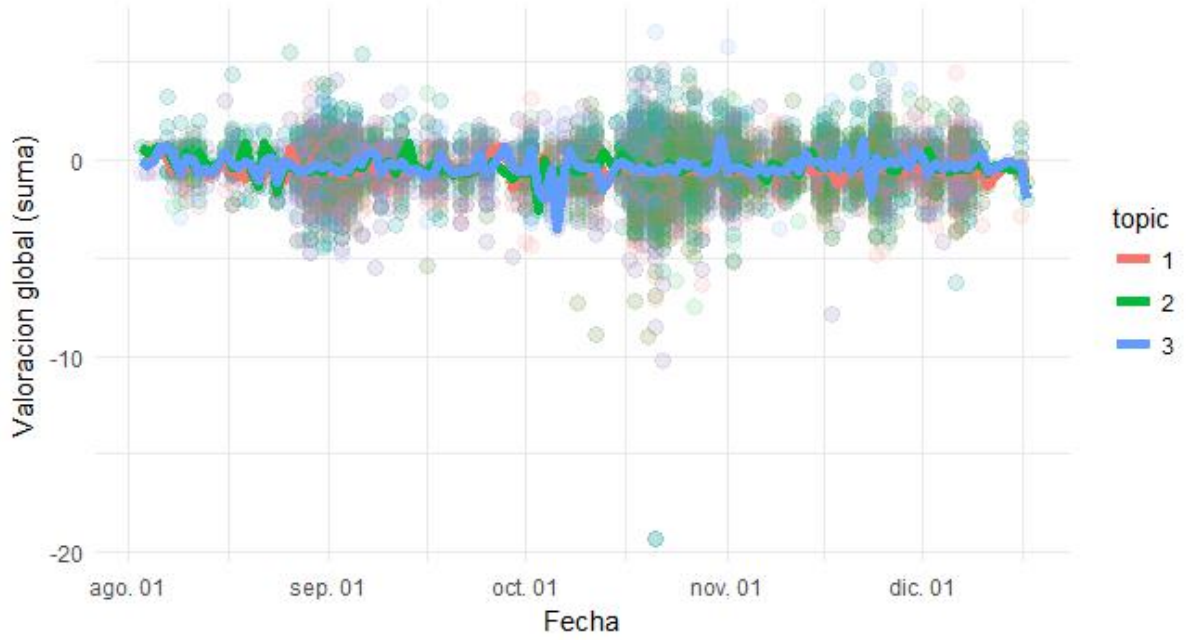


Gráfico 10 – Palabras de valoraciones extremas en léxico



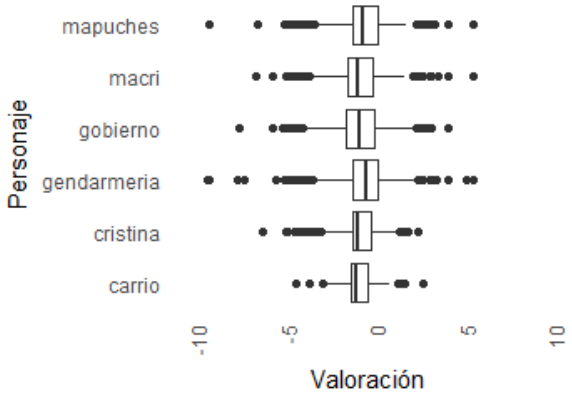
Para calcular la valoración positiva/negativa global de cada comentario hay diversas estrategias y algoritmos. Aquí hemos optado por el más sencillo de todos: sumar los puntajes de las palabras (con valoración en el léxico) que aparecen en cada comentario. El siguiente gráfico muestra las valoraciones de cada comentario (la línea indica el valor medio) por tópico y por fecha.

Gráfico 11 – Valoración global de comentarios, por tópico y fecha



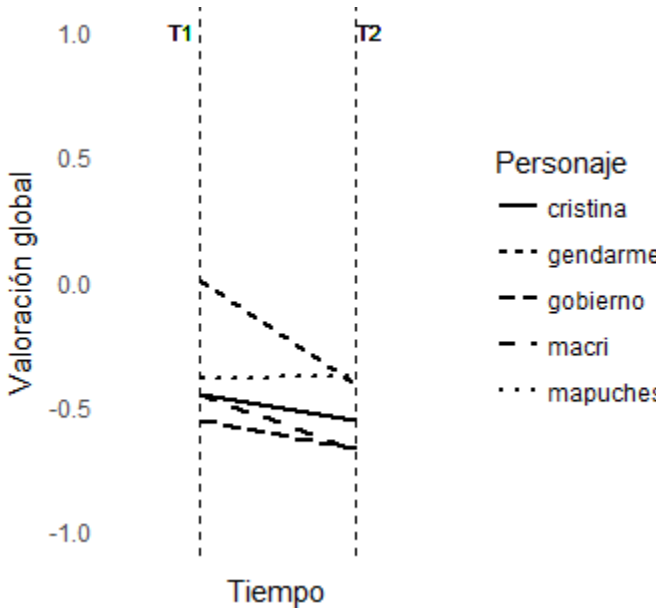
De todas formas, más interesante que un cálculo global puede resultar preguntarse por las valoraciones de los comentarios en los que se mencionan a ciertos personajes. Sin embargo, aquí se debe tener en cuenta la limitación introducida por la rutina de limpieza que eliminó las puntuaciones que permitirían distinguir oraciones: si en un mismo comentario se califica positivamente a un sujeto en una oración, y luego se descalifica a otro, en la valoración global -sin puntuaciones- ambos sujetos quedan asociados a los mismos calificativos. Todas estas limitaciones y otras -como la consideración de negaciones y frases que invierten la orientación de la oración ("shifters"), el uso de términos que amplifican/reducen las valoraciones ("amplifiers") se pueden subsanar con rutinas más complejas (Cambria, Das, Bandyopadhyay, & Feraco, 2017).

Grafico 12 – Comentarios con valoración por personajes



Además, este análisis se puede combinar con las consideraciones anteriores para observar si la valoración global en la que se inserta un personaje varía en el tiempo (gráfico 13), o entre distintos tópicos.

Gráfico 13 – Valoración media de comentario por personaje mencionado, antes/después de octubre/2017



Relevancia del Big data y Data mining para las ciencias sociales y psicosociales

En esta sección nos interesa retomar el debate en torno a la significación del big data y de las técnicas de data mining para las ciencias sociales y psicosociales. Como suele ocurrir con innovaciones que se originan en otros campos -máxime cuando estos se divulgan y masifican rápidamente por fuera de la academia (Gandomi & Haider, 2015)-, en la significación del big data abundan mitos que deben ser cuestionados. En un sugerente trabajo que abre el número inicial de *Big data & Society*, Kitchin (2014) ha señalado varios de estos mitos –de clara raíz empirista–, entre los cuales destacamos:

1. en big data se trabaja con un volumen de datos sin precedentes en las ciencias, lo que obligaría a repensar la relación entre casuística e investigación;
2. el uso de algoritmos y rutinas estandarizadas para el análisis de los datos evita los riesgos de interpretaciones tendenciosas y cargadas de la subjetividad del intérprete;
3. las técnicas de data mining permiten descubrir patrones y relaciones directamente en los datos, de modo que no se necesitan teorías, modelos e hipótesis;
4. los análisis trascienden los dominios de conocimientos específicos, de modo que cualquier persona (con un cierto conocimiento de estadística, visualización de datos y programación) puede llevar adelante los análisis e interpretar los resultados.

En lo que queda de esta sección, discutimos estos 4 mitos, ilustrando nuestros argumentos con nuestra experiencia de indagación en los comentarios acerca de la desaparición de Santiago Maldonado.

Big data no es necesariamente "big"

El primer mito asegura que big data trata con un volumen de datos que, de por sí, obliga a replantearse la relación entre la casuística y la investigación científica. Este mito, a su vez, está asentado en una especie de "fetichismo del tamaño" que es tan central en big data que llega al punto de ser, para algunos, su criterio definitorio.

Esta fijación por el tamaño es nociva por diferentes razones, como por ejemplo, por

pretender imponer una consideración que sólo es relevante a algunos agentes -como Google o Facebook-, y porque termina por reducir el fenómeno a un problema de capacidad tecnológica, bloqueando consideraciones más importantes, como su veracidad (Jagadish, 2015; Gandomi & Haider, 2015).

En cuanto a los volúmenes de datos, más sensato que imponer medidas a priori es limitarse a evaluar las casuísticas en relación a los estándares de cada campo (Burrows & Savage, 2014). Por caso, nuestro corpus de 40.000 comentarios es muy inferior en volumen a las bases de registros de sensores climáticos o de biotecnología, que suelen estar en el orden de los millones pero se diferencia de los cientos que suelen ser utilizados en estudios de opinión.

De todas formas, el escenario límite en el que el volumen de datos involucrados podría implicar un replanteo de la metodología tradicional de investigación empírica en ciencias sociales y psicosociales, sería en torno a aquellos casos donde se cuenta con información para todo un universo, evitando así los problemas derivados del muestreo. Pero es difícil imaginarse dichos casos. E incluso si así fuere, es esperable que junto con el volumen de la información, también se incremente su variedad, de modo que se da una especie de compensación entre la reducción de los posibles errores muestrales y el incremento de posibles errores de medición (Mayer-Schonberger & Cukier, 2013).

Volviendo a nuestro caso, los 40.000 comentarios se colectaron en 1 sola jornada de trabajo -y la rutina es reutilizable de modo tal que sólo es necesario cambiar el input (un criterio de búsqueda en internet) para generar una nueva base para otro estudio- pero si quisiéramos conseguir información más significativa que la expuesta, necesitaríamos rutinas de control, limpieza y normalización de datos cuyo esfuerzo de creación sería mucho más costoso.

El uso de algoritmos no asegura objetividad

El segundo mito que nos interesa discutir afirma que las herramientas de data mining pueden producir interpretaciones exentas de contenido subjetivo. Este mito se sustenta en la idea de que existe una objetividad "procesual", cuyos defensores antes observaban en las técnicas cuantitativas y matemáticas, y que hoy presumen transversales a los algoritmos de exploración de datos (Gaukroger, 2012). Sin embargo, este mito choca con la evidencia de que las rutinas introducen transformaciones en los datos, y que las mismas dependen de decisiones que se toman en cada paso, desde la adquisición hasta los análisis. En nuestro caso, quisimos mostrar un ejemplo de exploración de datos sin introducir mayores transformaciones, de modo que optamos siempre por la forma más sencilla de construir la rutina. Y sin embargo, tomamos innumerables decisiones centrales para los resultados expuestos, tales como: construir la muestra de comentarios en base a resultados de un motor de búsqueda de internet cuyos métodos de *rankeo* no conocemos ni controlamos; recolectar la información a través Facebook cuya política de uso limita considerablemente la adquisición; no incluir tareas de normalización de términos, lo que

hubiera permitido dar mayor visibilidad a ciertos agentes por la vía de incluir también otras designaciones; realizar los análisis en el nivel de las palabras y no en el de las oraciones; el uso de algoritmos específicos en diversos momentos, los cuales son productos de largas discusiones en el campo del procesamiento del lenguaje natural; el uso de un léxico de elaboración propia, y no de uno de los varios disponibles comercialmente (Henriquez-Miranda & Guzman, 2016); y, asimismo, varias especificaciones que tomaron la forma de parámetros, como indicar a la rutina que queremos 3 tópicos, o limitar los análisis de correlación a palabras de más de 250 repeticiones.

Hay, además, una particularidad de los estudios de big data que debe ser señalada. Gran parte de los datasets que se utilizan para los análisis han sido generados como subproductos de procesos de registro más grandes y para diversos fines. Es decir, que a diferencia de una base de datos en la que se colectan las respuesta a una encuesta diseñada por un investigador, abudan bases creadas sin un propósito claro y sin una pregunta especificada por el investigador. Al respecto, desde las ciencias sociales se pueden señalar dos grandes problemas que merecerían un mayor debate. En primer lugar, se podría señalar que si bien los dataset no responden a una pregunta de investigación, lo que resulta en su forma desestructurada y eventualmente motiva el data mining, sería un error creer que no hay detrás de ellos una "intencionalidad de registro". Por caso, aquí hemos analizado comentarios pero también podríamos haber trabajado con audios, videos, u opiniones en blogs y foros, los cuales no sólo responden a una intención de expresión de su "autor" -lo que generalmente querremos indagar-, sino que también se rigen por los formatos y las reglas de las plataformas que alojan dicho contenido (Gandomi & Haider, 2015). En este sentido, si se cree que se está "dejando hablar" a los datos, cabe la pregunta de si acaso esa voz no sea la de los medios que los distribuyen. En segundo lugar, y haciendo caso omiso de lo anterior, se podría pensar que dichos datos adquieren un estatus similar al de los registros de observaciones no-intrusivas y que, por esta vía, son mas "naturales" ya que no se encuentran condicionados por los instrumentos de registro (del investigador). Este es un punto importante a considerar, especialmente para las metodologías que dependen de tematizaciones y expresiones que dan cuenta de la subjetividad del respondente.

El análisis de Big data tiene más sentido dentro de las ciencias

El tercer mito que queremos discutir es un corolario del anterior pero sus ramificaciones son profundas, al punto que es hoy una de las principales posiciones en las disputas en torno al big data. Chris Anderson, ex-editor de la influyente revista de divulgación *Wired*, se convirtió en vocero de esta posición al sostener:

> El método científico se basa en hipótesis comprobables. Estos modelos, en su mayor parte, son sistemas visualizados en la mente de los científicos. Luego se prueban los modelos y los experimentos confirman o falsifican modelos teóricos de cómo funciona el mundo.

> Esta es la forma en que la ciencia ha funcionado durante cientos de años. … Pero frente a los datos masivos, este enfoque de la ciencia – hipótesis – modelo – prueba se está volviendo obsoleto. … Ahora hay una mejor manera. Los petabytes nos permiten decir: "La correlación es suficiente". Podemos dejar de buscar modelos. Podemos analizar los datos sin hipótesis sobre lo que podría mostrar. Podemos arrojar los números a los clusters de computación más grandes que el mundo haya visto y permitir que los algoritmos estadísticos encuentren patrones donde la ciencia no puede. (Anderson, 2008)

Anderson no está solo en esta posición. Viktor Mayer-Schönberg y Ken Cukier, autores de *Big data. A revolution that will transform how we live, work and think*, sostienen en su introducción:

> Como humanos hemos sido condicionados para buscar causas, a pesar de que la búsqueda de la causalidad a menudo es difícil y puede llevarnos por el camino equivocado. En un mundo de grandes datos, por el contrario, no tendremos que fijarnos en la causalidad; en su lugar, podemos descubrir patrones y correlaciones en los datos que nos ofrecen información nueva e invaluable. Las correlaciones pueden no decirnos con precisión por qué algo está sucediendo, pero nos alertan de que está sucediendo. Y en muchas situaciones esto es suficiente (2013, p.13)

No podemos negar el valor de las correlaciones, ni mucho menos su aporte en áreas donde hay un monitoreo constante y donde lo que se requiere es una acción rápida, ya sea para una recomendación inteligente de productos, o ya sea para detectar un brote epidémico (e.g., Ginbsger, et. al, 2009). Pero aquí nos interesa discutir la significación del Big data y de las técnicas de Data mining para la ciencia, un ámbito donde la afirmación "correlaciones son suficientes" es una verdadera confesión de valores epistémicos, una que supone una renuncia a las pretensiones de comprensión, explicación y análisis causales. La disputa es si queremos generar conocimiento, o simplemente quedarnos con información. Si no se está dispuesto a renunciar al conocimiento, se debe entender las técnicas del Data mining como herramientas metodológicas insertas en un diseño y una tradición de investigación más amplia, guiado por una teoría general, e inserta en las discusiones propias de una disciplina científica.

Para una exploración como la que aquí hemos expuesto, un candidato ideal es la teoría de las representaciones sociales, uno de los programas más extendidos de investigación en psicología social. Esta teoría tiene su origen hace más de 50 años, y su objeto es de dar cuenta de una modalidad de conocimiento que se elabora en la interacción social cotidiana, y que hace inteligible la realidad social y física (Moscovici, 2001; De Rosa, 2013). Entre sus principales características se encuentra una cierta insistencia por la apertura conceptual –su autor, Sergei Moscovici, ha llegando al extremo de negarse a dar una definición canónica para las mismas–, con varias líneas de diálogo con otras teorías y disciplinas –especialmente la psicología, la sociología y la antropología–, y una pluralidad metodológica que tal que le permite indagaciones con enfoques tanto cualitativos como cuantitativos, y por medio de diversas

técnicas. Incluso en un nivel muy programático y preliminar, se pueden señalar algunas convergencias entre este programa -especialmente en su versión estructural y en la formulación de Abric (2001)- y nuestro caso:

- Las representaciones sociales tienen un contenido –que se define por cierta información que se recorta o privilegia y por las actitudes que le otorgan una orientación valorativa– y una cierta estructura u organización en base a la jerarquía de los elementos –hay un componente central prominente en torno al cual se asocian otros componentes periféricos–. En nuestro caso, el contenido y la estructura se puede observar en los términos y los campos semánticos que fueron detectados, mientras que el componente actitudinal –si bien muy elementalmente registrado aquí– se puede rastrear en lo que hemos llamado "orientación semántica" (para una propuesta metodológica sobre técnicas asociativas que comparte este entendimiento "polar" de las actitudes, véase De Rosa (2002)). Por su parte, la estructura queda en evidencia a través de los análisis de correlaciones que aquí hemos visualizado como una red, aunque en futuros trabajos se podrían ensayar otras visualizaciones sobre la lógica de centros/periferias.
- Una dimensión de análisis importante en las representaciones sociales corresponde a las transformaciones de las mismas en el tiempo. Particularmente, los elementos del núcleo central no tienen mayores variaciones, mientras que los del sistema periférico son más permeable a los contextos y las variaciones. Las comparaciones que aquí hemos hecho en relación a las tematizaciones y los tópicos, y también en torno a correlaciones antes/después entre ciertos elementos puede ser una forma (elemental) de tratar con esta dimensión.
- Otra dimensión de análisis es la que vincula la representación a los grupos sociales que las discuten. En nuestro caso, esta dimensión de análisis fue desestimada, en gran parte por las limitaciones que Facebook impone sobre los datos de los comentaristas. Pero esta no es una limitación inherente a la técnica sino a nuestras decisiones particulares. Hay casos de sobra de buenos análisis, especialmente en redes sociales, que llegan al punto incluso de señalar "voceros" para los distintos grupos.
- Uno de los métodos más extendidos en esta perspectiva es el uso de técnicas asociativas (Abric, 2001, pp. 59-64) en las que se le pide al entrevistado que indique todos los términos, expresiones o adjetivos con los que vincula un cierto objeto de representación, para luego analizar su contenido y su organización, en términos de frecuencias, rangos y órdenes de aparición, y la importancia dada a cada elemento (algo que suele pedirse al entrevistado que aclare en un segundo momento). En nuestro caso ya hemos trabajado con la frecuencia de las términos de los comentarios; el rango de aparición podría ser fácilmente calculado dentro del contexto semántico de los comentarios (siendo éste, además, un

elemento que suele perderse en las técnicas asociativas y que aquí puede reinsertarse en el análisis); la importancia, no obstante, puede ser entendida como una función de las correlaciones. Esto es algo que el mismo Abric sugiere (2001, p. 62-63), aunque también podrían pensarse otras formas de tratar la importancia, ya no sólo representativa sino también comunicacional, a partir de evaluar qué elementos son re-tematizados en sucesivas respuestas.

Si esta propuesta es plausible, su contexto sería el de la digitalización de las representaciones sociales, y la indagación trataría con cómo las nuevas formas de comunicación digital se están convirtiendo en marco y soporte a la construcción y transformación de representaciones (Wahlström, 2012), lo que a su vez supone a internet como un agente o espacio de socialización (Simkin & Becerra, 2013)

Big data requiere de especialistas (científicos) disciplinares

El último mito que nos interesa discutir sostiene que para hacer data mining se requiere sólo de un entrenamiento en estadística y programación sin necesidad de contar con formación en algún dominio específico.

La respuesta a este mito debe haber quedado clara en el tratamiento de la ciencia en los párrafos anteriores: se trata de querer reducir al "investigador científico" por un "analista" -o un data scientist, o "científico de datos"-. Por otro lado, si aceptamos como objetivo epistémico deseable la búsqueda de interpretaciones y explicaciones, las mismas necesitan de teoría general y específica, la cual se inserta en dominios de indagación científica. En nuestro caso hemos intentado presentar pocas interpretaciones pero las mismas se hacen evidentes en las decisiones: así, por ejemplo, preferimos observar "personajes" políticos porque entendemos que una forma particular en la que se estructura la discusión política en nuestro contexto.

Conclusiones

Con todo, las limitaciones y malos entendidos asociadas a las nociones de Big data y Data mining no derrumban las potencialidades metodológicas que comportan como prácticas autónomas o para la labor investigativa de las disciplinas científicas. Al respecto de estas últimas, aunque el auge de la acumulación y análisis de datos no resulte en una "revolución paradigmática", sí resulta pertinente y necesario pensar los nuevos horizontes que un mundo hiperconectado plantea como posibles para la construcción de conocimientos.

Por lo tanto, antes de sucumbir frente a los cantos de sirena que decretan el fin y la inutilidad de las metodologías vigentes, disciplinas como la psicología social, en particular, y las ciencias sociales, en general, encuentran en el Big data y en Data mining un repertorio de datos y técnicas de análisis con los que enriquecer su quehacer investigativo. Si tales elementos son introducidos en diseños enmarcados en teorías generales provenientes de tradiciones de investigación más amplias, semejante fórmula puede potenciar la

producción de conocimientos en el ámbito de disciplinas científicas. Pero para ello, es necesario adoptar una posición crítica en torno a los mitos y malos entendidos asociados a estas ideas.

Referencias

Abric, J.-C. (2001). *Prácticas sociales y representaciones*.. México D.F.: Presses Universitaires.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7), 16-07.

Berman, J. (2013). *Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information*. Waltham: Elsevier.

Burrows, R., & Savage, M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society*, *1*(1), https://doi.org/10.1177/2053951714540280

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). *A Practical Guide to Sentiment Analysis*. New York: Springer.

Dalton, J., & Thatcher, J. (2014). What does a critical data studies look like , and why do we care? Seven points for a critical approach to "big data." *Society and Space*, 1–12

De Rosa, A. S. (2002). The associative network: a technique for detecting structure, contents, polarity and stereotyping indexes of the semantic fields. *European Review of Applied Psychology*, *52*(3–4), 181–200.

De Rosa, A. S. (2013). *Social Representations in the "Social Arena"* Sussex: Routledge.

DiPaola, E. (2017). Decir la verdad. El troll y la producción de lo público. *Sociales en debate*. 12. 49–58

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gaukroger, S. (2012). *Objectivity. A very short introduction*. New York: Harvard University Press.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining. Concepts and techniques*. Waltham: Elsevier.

Henriquez Miranda, C., & Guzman, J. (2016). A review of Sentiment Analysis in Spanish. *Tecciencia*, *12*(22), 35–48. https://doi.org/10.18180/tecciencia.2017.22.5

Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society*, *3*(2), https://doi.org/10.1177/2053951716674238

Jagadish, H. V. (2015). Big Data and Science: Myths and Reality. *Big Data Research*, *2*(2), 49–52. https://doi.org/10.1016/j.bdr.2015.01.005

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

Kitchin, R., & Lauriault, T. P. (2014). Towards critical data studies : Charting and unpacking data assemblages and their work. *Geoweb and Big Data*, 1–19.

Liu, B. (2015). *Sentiment Analysis*. *Computational Linguistics*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139084789

Mayer-Schonberger, V., & Cukier, K. (2013). *Big data. A revolution that whill transform how we live, work, and think*. Ontario: Eamon Dolan/Houghton Mifflin Harcourt.

Moscovici, S. (2001). *Why a theory of social representations? Representations of the social: Bridging theoretical traditions*.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000001

Silge, J. & Robinson, D. (2017). *Text mining with R.* New York: O'Reilly

Simkin, H., & Becerra, G. (2013). El proceso de socialización . Apuntes para su exploración en el campo psicosocial. *Ciencia, Docencia Y Tecnología*, *XXIV*(47), 119–142. Retrieved from http://www.redalyc.org/pdf/145/145298840005.pdf

Wahlström, M. (2012). *The Digitalisation of Social Representations: The Influence of the Evolution of Communication Technology on the Development of Shared Ideas*. *... Department of Social Research 2012:*