Procesos de (auto) aprendizaje de herramientas digitales en la investigación social: nubes de palabras en Twitter

Ferreyra, Silvana [1], Juares Wanda [2], Borthiry Emanuel [3], Emiliano Calomarde [4], Milagros Dolabani [5]

emmanuelborthiry@gmail.com; emiliano.mdq19@gmail.com; milagros.dola@gmail.com; silvanaferreyra82@gmail.com; wandajuares@gmail.com

[1, 2, 4, 5] INHUS-CeHIS, CONICET

[3] CeHIS- UNMdP

Línea temática 4. Metadatos. Datos sobre datos

(Buscar y que nos busquen a través de nuestras palabras)

Palabras clave

Twitter, Minería de textos, Ciencias sociales computacionales, Nube de palabras, Stopwords

Resumen

El trabajo con documentación nacida digital, cada vez más usual en nuestra labor cotidiana, configura nuevos interrogantes y demanda una mirada crítica que desnaturalice el proceso de construcción de datos. A partir de un informe sobre los avances del proyecto de investigación bianual “Comunidades virtuales: historia, prácticas e imaginarios sociales” pertenecientes a la Universidad Nacional de Mar del Plata (UNMdP), proponemos discutir aspectos fundamentales para la reproducción y modularidad del proceso de aprendizaje y aplicación de herramientas digitales en el análisis de corpus textuales voluminosos. Centrándonos en la red social Twitter se pretende deslindar los caminos, bifurcaciones y virajes, en su mayoría resultado de la autodidaxia, haciendo hincapié en algunas

estaciones: a) especificidades e interdisciplinariedad en la conceptualización de metodología y técnicas para la investigación sobre redes sociales; b) aprendizaje para el uso de programas y codificación, c) “cacharreo” y dinámica del escritorio, d) crítica de los algoritmos utilizados, y e) interacción con la producción computacional (visualizaciones). Los retos de nuestras prácticas investigativas sobre Twitter y política conforman los hilos que unen a esta ponencia en la que presentamos los objetivos de la investigación; el informe sobre el proceso formativo; y exponemos avances hacia un posible modelo de trabajo sistematizado y reproducible.

Introducción

El trabajo con documentación nacida digital, cada vez más usual en nuestra labor cotidiana, configura nuevos interrogantes y demanda una mirada crítica que desnaturalice el proceso de construcción de datos. En esta ponencia buscamos reconstruir los procesos de auto-aprendizaje ligados al empleo de herramientas de análisis digitales aplicadas a la investigación social, en el marco de nuestras trayectorias en el proyecto de investigación “Comunidades virtuales. Prácticas, historia e imaginarios sociales”, de la Universidad Nacional de Mar del Plata (UNMdP).

Como experiencia concreta, dentro de un campo de investigación relativamente joven, el aporte que intentamos realizar responde a la necesidad de socializar y discutir errores y aciertos que, a partir de un caso concreto, ejemplifiquen el uso de Twitter en investigaciones sociológicas. Puntualmente, el propósito de la presente ponencia versa en la realización de un ejercicio de reflexividad de nuestras prácticas de investigación, parte de procesos que raramente resultan homogéneos y lineales, y que aquí intentaremos reponer.

Asimismo, pretendemos reflexionar sobre la formación del/a cientista social respecto a temáticas digitales, un campo de investigación que no siempre resulta familiar para personas formadas en áreas lejanas a la programación o las ciencias exactas, y que suele producir extrañamiento. Centrándonos en los cruces entablados entre la red social Twitter y la política a escala local, el trabajo con grandes corpus textuales nos condujo a familiarizarnos con técnicas de minería de textos como la frecuencia de palabras y la obtención de nube de etiquetas, acercándonos al contenido de grandes volúmenes de texto desde una lectura distante (Moretti, 2015). Tomamos como caso testigo la investigación realizada sobre narrativas políticas presentes en biografías de

seguidores en Twitter de candidatxs a Intendente en General Pueyrredón para la campaña de 2019 (Juares, 2021).

Como investigadorxs en ciencias sociales nos guían preguntas-problemas de investigación, y suponemos a las técnicas y herramientas computacionales como medios para responderlas. No obstante, en la aproximación al objeto digital es un paso central pensar sobre la estructura de datos y los algoritmos que procesan o arrojan lecturas posibles. A la hora de pensar los estudios digitales Fussey y Roth (2020) sostienen que para la sociología digital la relación teoría - metodología se muestra mucho más intrincada en el proceso de investigación.

En este sentido realizamos una inmersión en el campo de la Social Big Data y las Humanidades digitales por un lado, y algunas aproximaciones a trabajos del campo de la Ingeniería y ciencia de datos, los que suelen tener mayor reflexión en torno a procesos algorítmicos. Como sostiene Gimena del Rio:

> Nada de lo que nos muestran las Humanidades digitales nos es completamente desconocido, sin embargo, el elemento diferencial que traen consigo es permitirnos ver más allá de la recepción humana de textos o imágenes. Aquí están los datos –masivos, ocultos o intuidos– que emergen y dialogan con los textos y con nosotros a través de la tecnología, pero que a través de nuestra mirada humana construyen sentido (2019:55)

Como ha señalado Schwandt (2019) en su reflexión de métodos digitales para el análisis de campos semánticos en Historia, uno de los mayores desafíos estriba en poder distanciarnos del conocimiento semántico existente, encontrando en los métodos del área de las humanidades digitales nuevas maneras de afrontar ese desafío. ¿Cómo podemos, a pesar de esto, llegar de manera exitosa a nuevas interpretaciones? El cálculo de frecuencias de las palabras, de la distancia entre palabras y de las coocurrencias ofrece un panorama del texto que está libre de interpretaciones, puesto que la computadora es “semánticamente ciega” y nos ayuda a poner en cuestión nuestras presuposiciones (Schwandt, 2019:162).

Considerando los antecedentes descritos, nos propusimos indagar las descripciones de biografías en tanto estrategias de autoidentificación para el conjunto de internautas seguidores de cada unx de lxs candidatxs consideradxs, con el objetivo de caracterizar y comparar sus prfiles en base a conceptos, palabras e ideas claves sobre cómo se definen. En base a dicho objetivo, nos preguntamos: ¿Qué jergas aparecen? ¿Qué distancias y similitudes ideológicas se presentan entre los adeptos de los distintos canditatxs en relación a sus narrativas? ¿Qué elementos identitarios son afines a cada comunidad de seguidores? En última instancia, nos preguntamos sobre las marcas de la polarización política en las biografías de seguidores de

candidatos con perfil nacional en contraste con aquellos que muestran un perfil vecinal.

La realización de ese trabajo a partir de nubes de palabras vio emerger un compendio de problemas que se fueron resolviendo a medida que avanzábamos. En cuanto a la apropiación de esta técnica para explorar las narrativas en Twitter, nos preguntamos ¿cuál es la diferencia entre la realización de nubes de las biografías de lxs seguidorxs o de sus tweets? En relación a la limpieza de los textos ¿Qué sucede con los *emojis*?, ¿y con las palabras compuestas? ¿Cómo debe tratarse el género? ¿Qué modos de supervisión se nos ocurren para atender a las ironías y sentidos no figurados?

En el primer apartado recopilamos algunas reflexiones sobre el proceso de trabajo en las humanidades digitales, haciendo énfasis en la pérdida de control en el proceso de investigación como riesgo del trabajo con grandes volúmenes de datos. En la segunda sección describimos algunas características de la plataforma Twitter y sus potencialidades como objeto de la investigación académica. Nos detenemos en las características generales de la recolección de tweets y damos cuenta de nuestros primeros pasos en este trabajo a partir del lenguaje R. En el último apartado mostramos las decisiones en torno al proceso de preprocesamiento de datos para el armado de nubes de palabras a partir de las biografías de seguidores a políticxs locales en Twitter. Nos interesa reflexionar sobre las tensiones entre destrezas técnicas y control del proceso de investigación a partir del “cacharreo” con scripts para minería de textos.

Metodologías digitales- humanidades digitales

Yuk Hui (2014) ha caracterizado al proceso de creciente proliferación de datos como constitutivo de un sistema digital que acelera la capacidad de producir, almacenar y organizar datos. Esa aceleración es acompañada por la producción de metadata, permitiendo el procesamiento de grandes cantidades de datos con computadoras, estableciendo nuevas conexiones y redes que se extienden de plataforma en plataforma y de una base de datos a otra (Hui, 2014).

Si bien el uso de computadoras en las humanidades tiene una larga historia, recién ahora las herramientas computacionales para el procesamiento y análisis de información comienzan a tener una mayor legitimidad. Este hecho debe asociarse a la abundancia de materiales para la investigación social (Rosenzweig, 2003), que ha consolidado una inclinación hacia los métodos digitales y la proliferación de trabajos que, desde múltiples enfoques, se suscriben al campo de las humanidades digitales.[1] La discusión sobre aquello

[1] Los métodos digitales pueden ser definidos por el uso de tecnologías informáticas en de procesamiento y visualización, son técnicas para el análisis de las condiciones culturales usando datos en línea (Rogers, 2013b).

que define a este espacio de trabajo no está cerrada: hay quienes sostienen que su objeto son las tecnologías de la información y la comunicación en sí mismas, mientras que otrxs explican que su rasgo particular es la utilización de las herramientas informáticas. En términos generales podemos señalar que las humanidades digitales poseen métodos y perspectivas relacionadas con el proceso de digitalización, donde lo impreso no es el único medio (Pons, 2013).

Por su parte, el uso de las computadoras como instrumentos o herramientas heurísticas dio lugar a una serie de efectos epistemológicos e institucionales que demandan, al menos, una breve mención. Las oportunidades que ofrece la informática en términos investigativos, como la capacidad de construcción y procesamiento de corpus inimaginables décadas atrás, no debe invisibilizar la posible pérdida de control del investigador sobre una de las etapas centrales de su trabajo. Tampoco la necesidad de un perfeccionamiento técnico extra que pocas veces es incluido en los currículos de las carreras humanísticas.

En el plano institucional, la experimentación con la informática por parte de las humanidades inicia en la década de 1980, con una serie de proyectos en el Norte Global que estuvieron centrados en pensar la aplicabilidad de herramientas y análisis estadísticos en el quehacer académico. En los últimos años, producto del avance de la digitalización, ha adquirido prioridad el aprendizaje sobre técnicas que permitan abordar grandes cantidades de datos. En este escenario, se produjo un avance de cursos cuyos programas de estudio adoptaron una perspectiva enfocada en la programación y que, en ocasiones, relegó la reflexión sobre las implicancias que tienen los algoritmos en los procesos investigativos.

Asimismo, se reconoce cierto rechazo hacia las matemáticas, dada la dificultad que supone la programación para lxs humanistas, y un retorno a la reflexión sobre las oportunidades informáticas en el campo disciplinar, puesto que ha sido difícil alcanzar un equilibrio entre el campo disciplinar y la informática (Crymble, 2021). En este sentido, mantenerse comprometidxs con la adquisición de las destrezas tecnológicas, descuidando los propios procesos de investigación, puede convertirse en una práctica recurrente. En estos casos, dominar la tecnología se convierte en un fin en sí mismo, en lugar de un medio para la consecución de propósitos mayores como la producción de conocimiento (Cohen et. al, 2008). Además, es menester señalar que la creciente oferta de espacios para aprender a codificar no resuelve de manera completa las dificultades del proceso didáctico. La curva de (auto) aprendizaje sigue siendo muy empinada y las posibilidades de construir inusitadas preguntas de investigación siguen siendo socavadas por el nuevo conjunto de habilidades que demandan las humanidades digitales. En este marco, se torna necesario propiciar espacios de sociabilidad y aprendizaje de nuevas técnicas de recolección, procesamiento e interpretación de los materiales.

Finalmente, en una dimensión epistemológica, la digitalización introdujo cambios que tienen consecuencias importantes sobre los métodos y la producción de conocimiento. Si bien sus resultados pueden ser visualmente impresionantes e intuitivamente convincentes, el estado metodológico y epistemológico de su producción aún no está claro (Rieder y Röhle, 2012). Aunque la utilización de algoritmos y la presentación de los datos mediante nuevas técnicas de visualización no implica necesariamente un método más transparente, y aunque la escritura o explicitación de los códigos no garantiza la inteligibilidad de las maneras en que los algoritmos producen los datos, la motivación por incursionar en los métodos digitales no debe verse socavada, reivindicando la importancia de un trabajo técnico que no pierda de vista la reflexividad metodológica.

Twitter como plataforma para la investigación social a través de R

Desde sus orígenes en 2006, Twitter se ha convertido en uno de los ejemplos más exitosos de red de microblogging. Este canal de comunicación dinámico y efectivo para la transmisión de mensajes ha generado un modelo multiplataforma caracterizado por la sencillez, velocidad, movilidad y comunidad (Orihuela, 2011). La sencillez se manifiesta en el conjunto acotado de funciones que presenta el sitio: escribir mensajes de hasta 280 caracteres, responder a tuits, retuitear contenidos de otros y seguir perfiles.

Los usos políticos de Twitter lo vuelven un medio legítimo de producción y difusión de los discursos, y por ende, un nicho de construcción de sentidos culturales-políticos (Ferreyra, Reclusa, Juares y Perez Rubini, 2019). Esta caracterización llevó a que los estudios sociales encontraran inicialmente en el análisis del contenido del tuit un terreno fértil para indagar acerca de la opinión pública. Particularmente se centran en las estrategias de comunicación política, el estilo y narrativa que se articula en el mensaje, usos y estéticas, estrategias de segmentación de usuarios y manipulación informativa con *fakes news* y operativos de *trolls*, entre otros elementos discursivos propios de la red (Galup, 2019; O´Neil, 2017; Woolley y Howard, 2018). En Argentina, las investigaciones sobre Twitter y política constituyen un campo reciente donde convergen estudios sobre discursos políticos en perfiles de candidatxs, como aquellos que analizan los efectos de la polarización política (Slimovich, 2012; Castelo, 2014; Calvo, 2015; Annunziata, Ariza y March, 2017; Ventura, 2018; Calvo y Aruguete, 2020).

La impresión que la cultura política realiza sobre los modos de intercambio online, remarca en ese sentido la necesidad de problematizar las esperanzas democráticas depositadas inicialmente en la web, donde Twitter no se convirtió en un espacio para el estudio del debate o la deliberación, sino más bien en un ámbito que propicia un tipo de “socialidad en red” (Rogers, 2013a). De esta manera, es menester considerar que la expansión de su acceso no se tradujo

en una participación igualitaria, puesto que la discusión en línea sigue siendo dominada por grupos hegemónicos, un hecho que insta a romper con la presunción sobre la universalidad de la experiencia digital y a reconocer su importancia en términos culturales.

En efecto, la comprensión de la sociabilidad en Twitter no puede escindirse de una historia socio-técnica más amplia que, en los últimos años, se ajusta a los parámetros de una nueva etapa del desarrollo tecnológico definido por un conjunto de aplicaciones basadas en comunidades y servicios de redes comúnmente denominada web 2.0 (O'Reilly, 2009). Concretamente, Twitter forma parte de la convergencia de un conjunto de compañías, plataformas, tecnologías y culturas que co-evolucionan en el tiempo, un entramado entre medios conectivos y nuevas sociabilidades que se ha denominado cultura de la conectividad (Van Dijck, 2016). Desde esta perspectiva, las plataformas son entendidas como infraestructuras que performan actividades sociales mediadas por la arquitectura computacional, las capacidades de acción/creación de lxs usuarixs y las condiciones de cada servicio. Tal como relatan Burgess y Baym (2020) en la biografía sobre esa red social, hasta el año 2009 sus propietarios no definieron estrategias respecto a cómo monetizar el servicio, dedicándose a conseguir inversiones para desarrollar un modelo de negocios adaptado a la plataforma, construyendo un amplio público de usuarixs para luego aumentar sus flujos de ganancias. A partir del año 2010 se produce un distanciamiento de Twitter respecto al paradigma de innovación abierta característico de la web 2.0: su modelo se centraliza, se introducen mecanismos publicitarios, así como mayores controles a la experiencia y contenidos producidos por lxs usuarixs. En ese derrotero, esta red social fue adoptando una lógica cada vez más utilitarista, caracterizada por la manipulación y venta de datos, en desmedro de ciertos imaginarios que visualizaban al sitio como un lugar de discusión desde una lógica colaborativa.

Dichos cambios también implicaron un control más férreo sobre sobre la interfaz de programación de aplicaciones (API), un punto estrechamente vinculado con la investigación científica, puesto que restringe el acceso a los datos por parte de lxs investigadorxs, algo que ha comenzado a flexibilizarse a partir de la aparición de la API v2. Como contrapunto, es necesario aclarar que el acceso a la información en Twitter es significativamente más abierto que el de otras redes sociales como Facebook (Murthy, 2012). En efecto, el sitio dispone de cantidades masivas de tweets que pueden ser raspados y analizados, hecho que convierte a la plataforma en objeto de creciente producción académica, propiciando el surgimiento de nuevos campos de investigación que desde las ciencias sociales y las humanidades digitales lo observan como un "sensor social" de eventos en tiempo real, asociados a las tendencias en la plataforma (Van Dijck, 2016).

Actualmente, para recolectar tweets es necesario poseer una cuenta, luego registrar una aplicación personal (API) en *Twitter Developer* que habilita a hacer minería de datos. La recolección devuelve una enorme cantidad de datos en formato JSON (*Javascript Object Notation*), obtenibles tanto en flujo directo (*streaming*), como del archivo histórico (API SEARCH). En este último caso, resulta fundamental recolectar los tweets del evento que nos interesa en un plazo de aproximadamente una semana. En concreto, al trabajar de forma gratuita la cantidad de descargas y el plazo de extracción se encuentran regulados por la empresa, lo que implica una primera opacidad en el proceso. Los métodos por los cuales Twitter toma la muestra no son totalmente transparentes, aunque estudios puntuales la han considerado representativa al compararla con aquellos que permiten obtener todo el caudal de tweets (statuses/firehose) (Morstatter, Pfeffer et al, 2013)

La extracción y procesamiento de los datos pueden realizarse mediante aplicaciones dedicadas a su ejecución[2], o bien utilizando softwares de lenguajes de programación como *R* o *Python*, entre otros. Para este trabajo nos servimos del software *R studio*, punto donde aparece uno de los núcleos problemáticos iniciales del autoaprendizaje, pues requiere una base de entrenamiento y familiarización con lenguajes de programación resumible en la frase “qué es esa cosa llamada R”. El software *R* es mencionado por la bibliografía recorrida como una herramienta de gran eficacia no solo por su gratuidad, sino también por ser un programa de código abierto que permite la colaboración de desarrolladores de todo el mundo. Distintas disciplinas científicas utilizan este software, específicamente para análisis de datos desde las Ciencias Sociales y Humanidades digitales (Uridinez y Cruz, 2020).

Este camino no nos libera de la opacidad de los algoritmos, pero sí nos permite otra escala de entendimiento sobre el proceso de tratamiento de los datos. Para quienes provenimos de las ciencias sociales, con casi nulo conocimiento informático a partir de nuestra formación de grado, la inmersión en estas técnicas puede resultar tediosa y frustrante en sus inicios, viéndonos frente a un artefacto que no entendemos en absoluto y que pareciera estar sumamente alejado de la experiencia. Este extrañamiento conlleva una forma diferente de experimentar el método de investigación, en la que quizás estemos muchas horas sin resultados publicables. Para familiarizarnos inicialmente con el lenguaje y su entorno realizamos algunos cursos virtuales introductorios en plataformas como Coursera y Udemy.

Pero a medida que avanzamos, una práctica que facilitó de manera notable nuestro “autoaprendizaje” fue la existencia de diversas comunidades online del lenguaje R que comparten sus códigos, discusiones sobre determinados paquetes, asesoría, entre otros. Los tutoriales más detallados fueron nuestras

[2] Para realizar wordclouds: La herramienta digital Wordle (http://www.wordle.net/), Voyant tools https://voyant-tools.org/

primeras alternativas, en especial aquellos publicados en *The Programming Historian*[3] *o Rpubs*[4]*.* Tras la lectura y comparación entre diversos scripts que nos guiaron para la realización de distintos procedimientos, también recurrimos a GitHub[5] o Stackoverflow[6], por mencionar algunas de las comunidades donde encontramos respuestas a nuestras inquietudes.

En nuestra investigación, dimos los primeros pasos acompañadxs del anexo del libro de Ernesto Calvo (2015) sobre la anatomía política de Twitter en Argentina, donde el investigador visibiliza lo que habitualmente conocemos como "cocina" de la investigación con TwitteR, un paquete que provee una interfaz para trabajar con la API de Twitter. Aunque aquí no brindaremos mayores detalles sobre ese proceso, uno de los primeros desafíos fue la actualización constante que implica la descarga de datos, con un seguimiento cotidiano de las modificaciones en las condiciones de acceso y las actualizaciones de los distintos paquetes desarrollados en R para el trabajo con Twitter. Si bien en los trabajos iniciales utilizamos TwitterR, posteriormente adoptamos rtweet, un paquete que brinda más posibilidades de análisis y cuyo uso se ha extendido de modo creciente en la comunidad.

En el apartado siguiente procuraremos avanzar en una descripción detallada de nuestro trabajo con nubes de palabras a partir de las biografías de 90000 seguidores de políticos en Twitter. Apostamos a construir un género que, aunque se acerca al tutorial por algunos detalles técnicos, se centra en la reflexión sobre el proceso de construcción de datos, tanto a partir de la intervención humana como la maquínica.

"Cacharreo" con nubes de palabras

"Cacharrear", "trastear", eran las palabras que escuchábamos de nuestrxs colegas iberoamericanxs cuando buscaban describir el trabajo que los cientistas sociales realizaban frente a sus computadoras a la hora de trabajar con los datos. Nos pareció un verbo atractivo para iniciar la reflexividad sobre nuestras propias prácticas, pues daba cuenta del aprendizaje a partir de aciertos y errores que veníamos atravesando, y que ahora sabíamos parte de una experiencia más colectiva.

Los primeros ejercicios de nubes los realizamos con tweets de los timelines de los candidatxs a intendente para General Pueyrredón en 2019, con el propósito de trazar la agenda política, su plataforma electoral, y las principales ideas que presentaban sus espacios políticos.[7] En estos ejercicios fueron surgiendo los

[3] https://programminghistorian.org/
[4] https://rpubs.com/
[5] https://github.com/
[6] https://stackoverflow.com/
[7] En esta línea el trabajo de González Bengoechea, Fernández Muñoz, y García Guardia (2019) para las elecciones españolas ofició como guía.

primeros inconvenientes. Por un lado, el trabajo no aportaba información distinta a la que se podía obtener sobre agenda electoral de otras fuentes, tales como la prensa o las plataformas de campaña. Por otro, dado que se trataba de candidatxs locales con escasa producción de tweets[8], los corpus resultantes eran demasiado "pequeños" para este modo de visualización.

Tras este primer ensayo decidimos trabajar con un dataset construido con el conjunto de seguidores de lxs candidatxs políticos. Ese giro se dio de modo paralelo a un viraje en nuestras inquietudes de investigación, ya no orientada hacia la agenda, sino volcada a una inquietud que cruzaba preguntas del campo de la sociología política con el del estudio de las plataformas digitales. ¿Qué significa seguir en Twitter? Según Barberá, Jost, Nagler, Tucker y Bonneau (2015) lxs usuarixs de Twitter prefieren seguir a los políticos −o a otras figuras de peso− cuya posición en la dimensión ideológica latente es similar a la suya. ¿Qué podíamos aprender sobre la polarización política en Twitter a partir de un panorama de las biografías de los seguidores? ¿Qué diferencias aparecían entre candidatos de partidos nacionales y candidatos vecinalistas? Para construir el corpus identificamos los seguidores de cada político y guardamos para cada id la descripción de su perfil. No todxs los usuarios poseen descripción, solo una parte ellos la escribía y esto ya era un indicador de niveles de participación y expresión presentes en la red. Las personas se ubican en el espacio social y digital mediante palabras, imágenes, gestos, imaginaciones, deseos, miedos, odios. El perfil de Twitter aparece como una superficie de inscripción en la que nos jugamos el "yo" como un lugar de enunciación. El "yo" que se expresa en las redes digitales representa siempre un yo situado en la estructura, y desde cada posición habla una condición política, social, geográfica, de género, religiosa, etc. (Reguillo, 2017).

En la última década, la proliferación de una nueva forma de metadatos (datos creados a partir de datos) incluye a las personas en su producción, puntualmente mediante la creación y aplicación de etiquetas: palabras clave asignadas a una información por su creador, la persona que la comparte o incluso los usuarios finales. Para representar las etiquetas con el fin de facilitar la recuperación efectiva de los contenidos, necesitamos una interfaz atractiva y fácil de usar, y la nube de etiquetas deviene una opción plausible para ese propósito. Kaptein (2012) demuestra en sus investigaciones como la técnica nube de palabras permite aproximarnos a Twitter en dos sentidos: por un lado, tener una búsqueda general o mirada a modo "resumen" de los principales tópicos; a su vez agrupar diferentes tópicos y compararlos para aplicar análisis de contenido, sentimientos, entre otros. (Khusro, 2018). Una nube de etiquetas transforma el campo semántico en una visualización mediante la ponderación de las etiquetas teniendo en cuenta su frecuencia de uso. Por lo general, la

[8] El rango de tweets publicados de los 10 precandidatxs con usuario de Twitter iba desde 548 hasta 38467 en julio 2019. En todos los casos se trataba de números bajos si lo comparamos con políticos nacionales como @CFKArgentina (15.300) o @mauriciomacri(10.200). La API permite bajar 90.000 tweets por usuario.

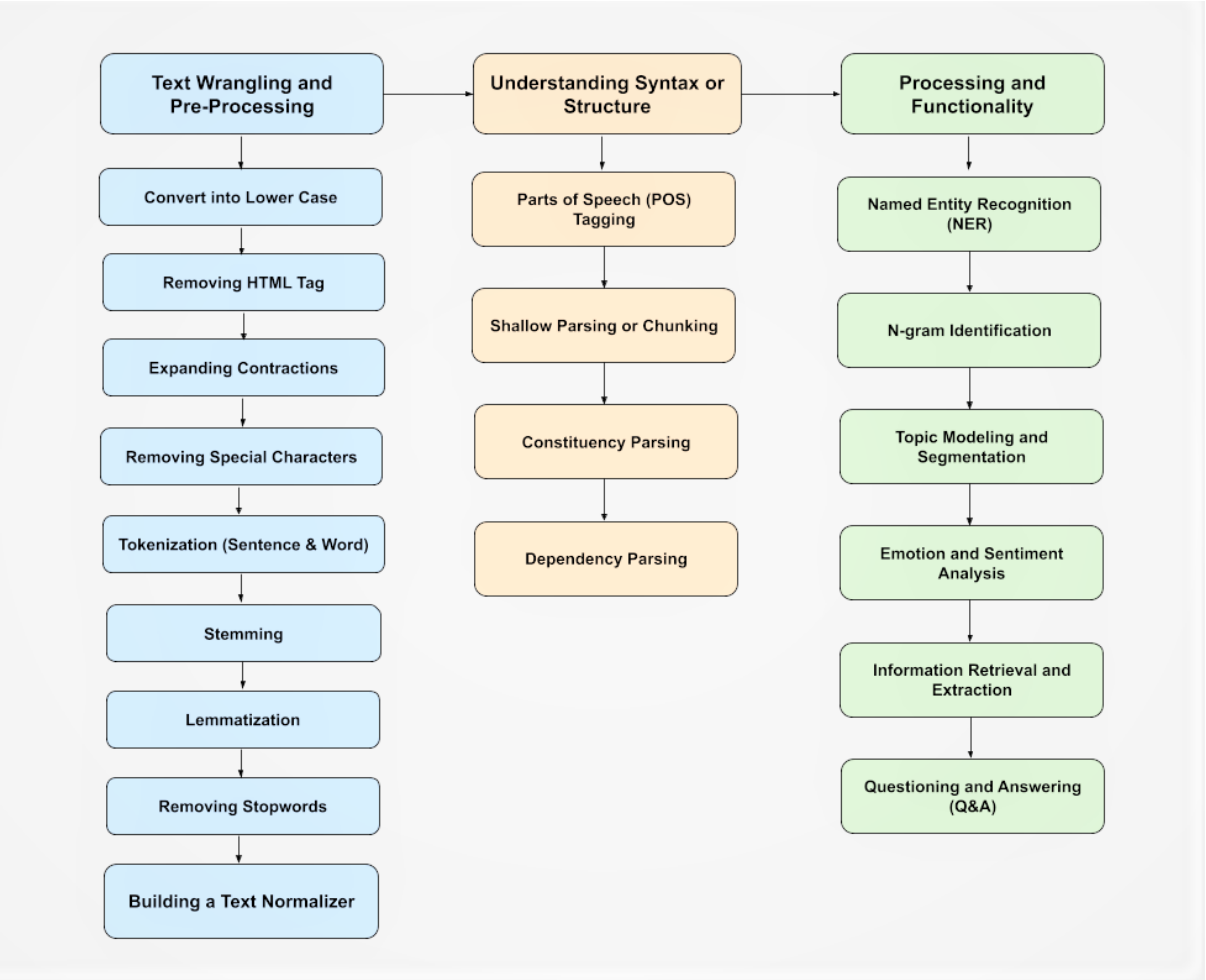
nube de etiquetas es sólo una lista alfabética de las etiquetas más populares cuya reiteración se muestra mediante el tamaño de la letra. (Hernández Fernández, 2015).

Las nubes que realizamos fueron efectuadas a partir del paquete w*ordcloud2,* que permite crear nubes interactivas y con mayor flexibilidad y estética que su antecesor wordcloud.[9] En las nubes que construimos las palabras se acomodan de manera aleatoria, aunque existe una fuerza gravitatoria que focaliza en el centro del círculo las palabras con mayor frecuencia y aleja los términos menos utilizados. Aunque existen opciones para modificar la forma de la nube, no hemos alterado estas opciones de visualización, así como tampoco las opciones de color, cuyas variaciones han sido indagas a partir de la teoría psicológica sobre la percepción visual.

El aspecto de esta técnica en el que aquí nos concentraremos se vincula con el preprocesamiento o curaduría de los datos. Como dijimos la nube se realiza sobre un corpus textual, pero para que la máquina pueda leerlo y arrojar resultados requiere un trabajo “demasiado humano” de preparación del texto. Es decir, debemos construir un corpus de datos que pueda conversar con las máquinas, depurar y ajustar la información. Como muestra el siguiente diagrama, el preprocesamiento es tan solo una parte de los múltiples canales que debemos recorrer para realizar Procesamiento de Lenguaje Natural (en inglés NLP).

[9] Alonso, Julio. (2020). Una introducción a la construcción de Word Clouds (para economistas) en R. Disponible en: https://www.researchgate.net/publication/341829699_Una_introduccion_a_la_construccion_de_Word_Clouds_para_economistas_en_R

Figura 1. Procesamiento e Lenguaje Natural, paso a paso.



Fuente: https://suneelpatel-in.medium.com/nlp-pipeline-building-an-nlp-pipeline-step-by-step-7f0576e11d08

Para realizar las nubes nos concentramos únicamente en el canal celeste. Para llevar adelante esta operación en R Studio construimos un script. Estos códigos o instrucciones son recetas que permiten ejecutar la orden que necesitamos realizar con respecto a la recolección y análisis de datos. No somos programadorxs sino cientistas sociales y recurrimos al trabajo colaborativo y comunitario con otros colegas para co-crear los códigos que requerimos. La reproducción y adaptación de otros *script* mediante “copiar/pegar/crear” conforman parte de las estrategias desplegadas para llegar a nuestro objetivo, una parte fundamental de la experiencia en esta etapa inicial. En esta instancia, la escritura del código consistió en recuperar “modelos” que se aproximaron a

nuestras búsquedas y reescribir las partes que no eran coincidentes con nuestros objetivos.

Veamos el script con el que empezamos a realizar una serie de pruebas.

```
#Nube de palabras con tuits

require(rtweet)        # extracción de datos de twitter
require(tm)            # limpieza de texto
require(stringr)       # remueve caracteres
require (qdapRegex)    # remueve urls
require(wordcloud2)    # nubes

## Almacenar API KEYS (completar con sus propias claves)

api_key <- "xxxxxxxxxxxxxxxxxxx"
api_secret_key <- "xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"
access_token <- "xxxxxxxxxxxxxxxxxxx"
access_token_secret <- "xxxxxxxxxxxxxxxxxxxx"

## Autenticación con token propio

token <- create_token(
  app = "xxxxxxxxx",
  consumer_key = api_key,
  consumer_secret = api_secret_key,
  access_token = access_token,
  access_secret = access_token_secret)

get_token()


#Descarga de seguidores de un candidatx

#FerRaverta
Ravertaseguidores_flw <- get_followers("FerRaverta", n = "all")
Ravertaseguidores_flw_data <- lookup_users(Ravertaseguidores_flw$user)


# Extraer datos de la biografía de seguidores de Raverta

text <- Ravertaseguidores_flw_data$description

#Código para limpiar la variable "text".

text_limpio <-
  text %>%
  str_remove("\\n") %>%                   # remueve saltos de línea
  rm_twitter_url() %>%                    # remueve URLS
  rm_url () %>%
  str_remove_all("#\\S+") %>%             # remueve hashtags
  str_remove_all("@\\S+") %>%             # remueve @ mentions
  removeWords(stopwords("spanish")) %>%   # Remueve stopwords
  removeNumbers() %>%
  stripWhitespace() %>%
  removeWords(c("amp"))                   # Otras palabras para eliminar

# Hace un corpus de palabras con el texto que pasamos en la variable "text" y arma el data.frame con las
frecuencias de palabras para el wordcloud.

textCorpus <-
  Corpus(VectorSource(text_limpio)) %>%
  TermDocumentMatrix() %>%
  as.matrix()
textCorpus <- sort(rowSums(textCorpus), decreasing=TRUE)
textCorpus <- data.frame(word = names(textCorpus), freq=textCorpus, row.names = NULL)

# hace el wordcloud y muestra
wordcloud <- wordcloud2(data = textCorpus, minRotation = 0, maxRotation = 0, ellipticity = 0.6)
wordcloud
```

```
#Adaptación de este código: https://mikeyharper.uk/creating-twitter-wordclouds/
```

En primer lugar, descargamos los datos de twitter con el paquete rtweet y trabajamos con esos archivos en formato de tablas o *dataframes*. En este caso, cada fila contiene los datos de un *follower* o seguidor del candidatx, mientras que las columnas organizan las distintas variables (nombre de usuario, fecha de creación de la cuenta, descripción, entre otras). Para este trabajo nos interesa trabajar con la columna *description*, que es la biografía de cada seguidor.

A continuación, tenemos que avanzar con la limpieza del texto. Pero veamos antes el resultado con el texto en crudo de una nube de palabras de los seguidores del Gustavo Pulti, candidatx del partido vecinalista Acción Marplatense. En lo sucesivo, vamos a utilizar como ejemplo este dataset.

Figura 1. Nube de palabras a partir de biografías de seguidores de Gustavo Pulti. Texto sin procesar.

Los textos de tweets son textos muy “ruidosos”, razón por la cual partimos de un *script* que había sido desarrollado para este fin específico, aunque necesitamos realizar varios “agregados”. Nos familiarizamos con las instrucciones que ofrecen los paquetes ™ y stringr para este fin, ya que son los elegidos en el script que tomamos como punto de partida.[10]

Lo primero será quitar números, espacios en blanco y signos de puntuación. Este último proceso no estaba en el código original pero decidimos incorporarlo, aunque cuidando que esa eliminación no pegue palabras que

[10] https://cran.r-project.org/web/packages/stringr/index.html y https://cran.r-project.org/web/packages/tm/index.html

estaban separadas, lo que por supuesto nos ocurrió en la primera prueba. También, por lo que observamos en otros códigos con los que hacemos dialogar al que elegimos como parámetro, suele recomendarse pasar todo el texto a minúsculas para evitar que el programa reconozca como distintas palabras iguales. Ese problema, evitar que palabras que consideramos iguales sean percibidas por la lectura maquínica como diferentes, es el principal inconveniente con el que nos cruzamos una y otra vez.

Un aprendizaje particular tiene que ver con el uso de las expresiones regulares, patrones que se utilizan para hacer coincidir combinaciones de caracteres en cadenas. Así, por ejemplo, mientras tm tenía una instrucción específica para remover números [tm::removenumbers], en otras librerías que también se utilizan para limpieza de texto debemos conocer la expresión "[[:digit:]]". Familiarizarse con este tipo de expresiones es un paso importante en el auto-aprendizaje, ya que permite ir desprendiéndose lentamente de las funciones preconfiguradas hacia la programación de aquellas que resuelvan las necesidades específicas de nuestra investigación.

Otro problema específico de la limpieza de textos en castellano es la cuestión de los acentos. Dado que el script que utilizamos de modelo está pensado para el idioma inglés, tampoco contemplaba esta situación. Si bien Twitter permite utilizar tilde en los hashtags, los usuarios suelen no prestar tanta atención a estas reglas ortográficas.[11] Nos decidimos por eliminar los acentos, ya que parecía la opción menos costosa para evitar la identificación de palabras iguales como diferentes. Para este fin utilizamos una función preconstruida, que reemplaza cada vocal con acento (á,é,í,ó,ú) por una igual sin tilde.

Otra cuestión a destacar es que el script que utilizamos como base estaba pensado para limpiar texto de tweets, mientras que nuestra intención era trabajar con las biografías de los seguidores. En ese sentido, si bien eliminar URL, hashtags y menciones podía ser una decisión plausible, no implicaba el mismo volumen de información. Por otro lado, si eliminamos aquí el contenido del hashtag no era posible, como en el caso de los tweets, recuperar esa información con otra columna del dataframe general, razón por la cual nos decidimos por remover únicamente el signo #.

Como parte de la limpieza de los datos también es necesario eliminar las palabras que no tienen significado propio, para ello se requiere construir un diccionario que incluya dichas palabras. Los *stopwords*, son diccionarios negativos, es decir nos permiten quitar términos no deseados para el análisis. La ausencia de investigaciones sobre stopwords resulta en el uso intensivo de diccionarios preexistentes, los cuales no son apropiados en diferentes

[11] Un ejemplo de esta tendencia puede apreciarse con la existencia de campañas para promover el buen uso del lenguaje en redes sociales. Véase https://www.lanacion.com.ar/cultura/sin-excusas-hay-que-usar-tilde-en-twitter-nid1828486/

contextos textuales (Murphy Choy, 2012). En efecto, muchas librerías de R ofrecen diccionarios pre armados, aunque la oferta en español siempre es menor. En el script que empezamos a trabajar se utilizaba la del paquete ™, el cual contiene 306 palabras. Si lo comparamos, por ejemplo, con otro diccionario revisado[12], que contiene 608 palabras, este parece bastante escueto. El problema de la extensión de las listas de stopwords es un debate cuya resolución parece ubicarse en el terreno de la práctica. Una alternativa a las listas de stopwords genéricas es la confección de un listado específico para cada texto, tanto a partir de la frecuencia de palabras simple como de otras metodologías. Algunas de las propuestas más complejas apuntan a la frecuencia inversa del documento, la probabilidad media, la probabilidad de varianza, medidas de entropía y métodos como los términos de adyacencia o el muestreo aleatorio basado en términos (Burns, 2018:7). Se trata de alternativas diversas, que buscan formalizar el proceso arbitrario de decisiones con el que conformamos la lista.

En este trabajo combinamos la lista genérica, que copiamos a continuación, con un ajuste en la última línea del código de limpieza, la que permite incorporar algunos términos, identificados a partir de la frecuencia.

[12] https://github.com/Alir3z4/stop-words/blob/master/spanish.txt

Figura 2. Lista de stopwords en el paquete tm.

```
Loading required package: NLP
  [1] "de"         "la"         "que"        "el"         "en"
  [6] "y"          "a"          "los"        "del"        "se"
 [11] "las"        "por"        "un"         "para"       "con"
 [16] "no"         "una"        "su"         "al"         "lo"
 [21] "como"       "más"        "pero"       "sus"        "le"
 [26] "ya"         "o"          "este"       "sí"         "porque"
 [31] "esta"       "entre"      "cuando"     "muy"        "sin"
 [36] "sobre"      "también"    "me"         "hasta"      "hay"
 [41] "donde"      "quien"      "desde"      "todo"       "nos"
 [46] "durante"    "todos"      "uno"        "les"        "ni"
 [51] "contra"     "otros"      "ese"        "eso"        "ante"
 [56] "ellos"      "e"          "esto"       "mi"         "antes"
 [61] "algunos"    "qué"        "unos"       "yo"         "otro"
 [66] "otras"      "otra"       "él"         "tanto"      "esa"
 [71] "estos"      "mucho"      "quienes"    "nada"       "muchos"
 [76] "cual"       "poco"       "ella"       "estar"      "estas"
 [81] "algunas"    "algo"       "nosotros"   "mi"         "mis"
 [86] "tú"         "te"         "ti"         "tu"         "tus"
 [91] "ellas"      "nosotras"   "vosotros"   "vosotras"   "os"
 [96] "mío"        "mía"        "míos"       "mías"       "tuyo"
[101] "tuya"       "tuyos"      "tuyas"      "suyo"       "suya"
[106] "suyos"      "suyas"      "nuestro"    "nuestra"    "nuestros"
[111] "nuestras"   "vuestro"    "vuestra"    "vuestros"   "vuestras"
[116] "esos"       "esas"       "estoy"      "estás"      "está"
[121] "estamos"    "estáis"     "están"      "esté"       "estés"
[126] "estemos"    "estéis"     "estén"      "estaré"     "estarás"
[131] "estará"     "estaremos"  "estaréis"   "estarán"    "estaría"
[136] "estarías"   "estaríamos" "estaríais"  "estarían"   "estaba"
[141] "estabas"    "estábamos"  "estabais"   "estaban"    "estuve"
[146] "estuviste"  "estuvo"     "estuvimos"  "estuvisteis" "estuvieron"
[151] "estuviera"  "estuvieras" "estuviéramos" "estuvierais" "estuvieran"
[156] "estuviese"  "estuvieses" "estuviésemos" "estuvieseis" "estuviesen"
[161] "estando"    "estado"     "estada"     "estados"    "estadas"
[166] "estad"      "he"         "has"        "ha"         "hemos"
[171] "habéis"     "han"        "haya"       "hayas"      "hayamos"
[176] "hayáis"     "hayan"      "habré"      "habrás"     "habrá"
[181] "habremos"   "habréis"    "habrán"     "habría"     "habrías"
[186] "habríamos"  "habríais"   "habrían"    "había"      "habías"
[191] "habíamos"   "habíais"    "habían"     "hube"       "hubiste"
[196] "hubo"       "hubimos"    "hubisteis"  "hubieron"   "hubiera"
[201] "hubieras"   "hubiéramos" "hubierais"  "hubieran"   "hubiese"
[206] "hubieses"   "hubiésemos" "hubieseis"  "hubiesen"   "habiendo"
[211] "habido"     "habida"     "habidos"    "habidas"    "soy"
[216] "eres"       "es"         "somos"      "sois"       "son"
[221] "sea"        "seas"       "seamos"     "seáis"      "sean"
[226] "seré"       "serás"      "será"       "seremos"    "seréis"
[231] "serán"      "sería"      "serías"     "seríamos"   "seríais"
[236] "serían"     "era"        "eras"       "éramos"     "erais"
[241] "eran"       "fui"        "fuiste"     "fue"        "fuimos"
[246] "fuisteis"   "fueron"     "fuera"      "fueras"     "fuéramos"
[251] "fuerais"    "fueran"     "fuese"      "fueses"     "fuésemos"
[256] "fueseis"    "fuesen"     "siendo"     "sido"       "tengo"
[261] "tienes"     "tiene"      "tenemos"    "tenéis"     "tienen"
[266] "tenga"      "tengas"     "tengamos"   "tengáis"    "tengan"
[271] "tendré"     "tendrás"    "tendrá"     "tendremos"  "tendréis"
[276] "tendrán"    "tendría"    "tendrías"   "tendríamos" "tendríais"
[281] "tendrían"   "tenía"      "tenías"     "teníamos"   "teníais"
[286] "tenían"     "tuve"       "tuviste"    "tuvo"       "tuvimos"
[291] "tuvisteis"  "tuvieron"   "tuviera"    "tuvieras"   "tuviéramos"
[296] "tuvierais"  "tuvieran"   "tuviese"    "tuvieses"   "tuviésemos"
[301] "tuvieseis"  "tuviesen"   "teniendo"   "tenido"     "tenida"
[306] "tenidos"    "tenidas"    "tened"
```

El último paso de la limpieza es la eliminación de los emojis. Pero ¿es buena idea eliminarlos? Los emojis brindan información tanto o más relevante sobre el contenido que la que nos proporcionan las palabras. Avanzamos en la tokenización para ver el modo en que puede afectar a nuestro estudio sobre frecuencia de palabras. El cuadro nos muestra que necesitamos eliminar o separar los emojis si queremos que el lenguaje no tome como diferentes palabras que son iguales.

Figura 3. Tabla de frecuencias.

	word	freq
7683	😍peronista	1
7684	😍pesca😍	1
7685	😍😘💋	1
7686	😒buscadora	1
7687	😘fulbol	1
7688	😘😘	1
7689	😚	1
7690	😜lic	1
7691	😜😛altamente	1
7692	😳😃😘arquitectainformarmeprogramas	1
7693	😻handball😻	1

Cuando optamos por eliminar los emojis (o tal vez mejor, separarlos para futuros análisis), un problema técnico que suele presentarse es que se eliminen también otros caracteres que no queríamos borrar, como por ejemplo, las letras con acentos y la ñ. Por esta razón, incluimos en nuestra función para remover acentos el reemplazo de ñ por n, admitiendo como costo que algunas palabras no se lean correctamente.

Finalizado el proceso de limpieza, seguimos con la tokenización, es decir, dividimos el texto en las unidades que lo conforman, entendiendo por unidad el elemento más sencillo con significado propio para el análisis en cuestión, en este caso, las palabras. En tm el producto de la tokenización es una matriz que nos permite agrupar los términos iguales y así determinar su frecuencia. Como ya dijimos, la linealidad del proceso es bastante mentirosa, ya que el desafío suele ser ir y volver varias veces sobre la cadena de caracteres inicial, para resolver distintas situaciones que se presentan.

La visualización de la matriz en la nube nos ayuda a identificar problemas en este trabajo. El desafío es encontrar un equilibrio entre la perfección que resuelva todos los inconvenientes, incluso dedicando mucho tiempo a correcciones que tienen poca repercusión en nuestro análisis, y la tentación de no introducir modificaciones, obnubilados por las potencialidades de esta visualización.

Figura 4. Nube de palabras a partir de biografías de seguidores de Gustavo Pulti. Texto procesado versión 1.

En esta nube, un primer problema aparecía con las denominaciones que contienen varias palabras. En ese caso, la ciudad de Mar del Plata. La contracción “del” fue eliminada con las stopwords y mar y plata fueron identificadas como palabras distintas. En efecto, es posible que en algunas biografías “mar” o “plata” no refieran al nombre de la ciudad, pero creemos que sí en la mayoría de los casos. También vemos que en otros casos los usuarios han decidido colocar mdp en la bio para designar a la ciudad (y seguramente ahorrar caracteres).

Las decisiones que podemos tomar dependen aquí de los objetivos de nuestra investigación. Si nuestro propósito fuese determinar el uso de abreviaturas en el lenguaje de los tuiteros, mdp debería considerarse por su peso específico. Pero en este proyecto nos proponíamos verificar la existencia de un espacio público virtual local y un internauta vecinx, cuya vinculación con las problemáticas locales lo despegaba, al menos parcialmente, de las dinámicas más polarizantes de la red. El peso de los diferentes modos de registrar lo local es algo que nos interesa recuperar, más allá de las contracciones y las separaciones de palabras. En ese sentido, una opción era intervenir el corpus antes de iniciar la limpieza y transformar “Mar del Plata” en mdp. La otra opción era convertir todo en mardelplata, aunque parece mejor evitar las palabras extensas, para obviar las distorsiones que el largo de los términos genera sobre la atención de las palabras en la nube.

Figura 5. be de palabras a partir de biografías de seguidores de Gustavo Pulti. Texto procesado 2.

Otro problema se relacionaba con las inflexiones de género y número de sustantivos y adjetivos, algo que podría resolverse con una lematización. En lingüística, este proceso se define como agrupación de las diferentes formas flexionadas de una palabra para que puedan analizarse como un solo elemento. Analicemos la situación a partir de una cuestión concreta. En la visualización de la nube nos llamó la atención la cantidad de datos sobre la ocupación de los usuarixs. Por ejemplo, en la base de seguidores de Alejandro Martinez (candidatx a intendente por el Frente de Izquierda), podemos observar que hay múltiples variaciones en torno a la palabra abogado, lo que se resolvería lematizando.

520	abogacia	2
84	abogada	9
956	abogadas	1
957	abogades	1
46	abogado	14

Sin embargo, perderíamos marcas sobre el uso del lenguaje inclusivo, un aspecto muy relevante para una investigación como la nuestra, preocupada por marcas identitarias. En cualquier caso, lematizar parece una opción apropiada siempre que se preserve la base de palabras para recuperar otros datos. Un próximo paso de trabajo debería profundizar en los algoritmos de lematización en español y el desigual desarrollo de los distintos paquetes con este fin. Algunas pruebas nos muestran que el stemming del paquete tm genera demasiados errores, unificando las palabras anteriores como abog. Por otra parte, la lemmatización a partir de librerías como udpipe, que tokeniza y diferencia las palabras semánticamente (verbos, sustantivos, adjetivos, adverbios) antes de lematizar, parece una alternativa más prometedora.

Un último problema del análisis textual en twitter, que ha sido señalado de manera recurrente en la bibliografía, es que estas técnicas no permiten captar las ironías. (Larrosa, 2013; Sanchez, 2017) Volviendo a la nube de Gustavo Pulti, es relevante el lugar que ocupa el vocablo peronista. El significado del vocable en la nube nos es transparente si consideramos que se trata de los seguidores de un candidato de un partido vecinalista que en 2015 fue elegido como intendente formando parte de la boleta del Frente para la Victoria, pero en 2019 se presentó como alternativa al Frente de Todxs. Justamente, el problema de este tipo de aproximaciones es que perdemos contexto, ya que los adjetivos están separados de otros conceptos, lo que nos lleva a tener dificultades para derivar significado y advertir connotaciones positivas o negativas en los términos. (Graham, Millingan, 2016: 75) ¿Qué significa aquí ser peronista? ¿Tiene el mismo significado que en la nube de Fernanda Raverta, candidata por la coalición que integra el Partido Justicialista? Una alternativa que se nos ocurre para aclarar las posibles ambigüedades es analizar la frecuencia del prefijo anti respecto a la de peronista. La relación entre 12 anti y 89 peronista en la nube de seguidores de Pulti sugiere que no hay una asociación directa.

Una alternativa interesante para complejizar el análisis puede ser evaluar los contextos de cada palabra. Uno de los modelos alternativos que mayor impacto ha tenido en la literatura sobre nubes de palabras es el de los tag-clouds semánticamente agrupados, propuestos por Hassan Montero y Herrero-Solana (2010). En este modelo los tags no se encuentran ordenados alfabéticamente o aleatoriamente, como es usual, sino agrupados en clusters alineados verticalmente. Dentro de cada cluster los tags se ordenan por alineación horizontal en base a su similitud semántica - similitud calculada sobre la co-ocurrencia relativa entre tags, o coeficiente de Jaccard-. De este modo, se pretende que el usuario pueda detectar e inferir relaciones semánticas entre tags (y clústers de tags), con el objetivo de facilitar tareas de búsqueda exploratoria. (Hassan Montero y Herrero Solana, 2010,19).

Si bien esta alternativa parece interesante, nos encontramos frente a la falta de destreza técnica para llevarla adelante con R Studio. El cacharreo nos familiariza con el intento aprehender cual es el producto de cada función o instrucción. Cuando damos un paso más e intentamos entender los procedimientos que se ejecutan para ese resultado nos vemos frente al compromiso, en ocasiones, de modificar esas instrucciones para lograr mejores resultados. Algunas veces, como en el caso de las nubes agrupadas semánticamente, este proceso puede ser estar más allá de nuestro umbral de conocimientos. En esas ocasiones la combinación con otras técnicas, cómo el análisis de los contextos de las palabras, pueden funcionar como atajos.

Reflexiones finales

Las nubes de palabras son una herramienta de visualización de textos que permite al investigador realizar análisis descriptivos de un texto. Si bien es usada para diferentes corpus textuales existen puntos de tensión sobre para que y como incorporarla. Para quien conoce el corpus de datos parecería no decir mucho. Para quien no conoce nada, es un descriptor de lo aún no explorado. En nuestra experiencia, las nubes fueron una interesante herramienta para comparar comunidades de internautas que seguían a distintxs políticos en una campaña local. Las nubes nos contaron algo sobre las identidades políticas en la red, pero también sobre la acción de seguir.

A partir de esta herramienta, en esta ponencia nos propusimos desnaturalizar el proceso de construcción de datos para confeccionar nubes de palabras, realizadas a partir de las biografías de los seguidores de politicxs en twitter. Las reflexiones siguieron varios caminos paralelos. Por un lado, recuperamos un recorrido general de lecturas sobre plataformas, tecnologías y culturas que guiaron nuestra aproximación a Twitter. Por otro, describimos paso a paso el preprocesamiento de textos que efectuamos para armar las nubes de palabras. Nos propusimos conceptualizar y “desencriptar” las modalidades de codificación de tipo “caja negra” que suelen acompañar el uso acrítico de las herramientas que utilizamos. En efecto, las dificultades en torno al auto-aprendizaje de las técnicas ocuparon de tal modo nuestra agenda de trabajo que pospuso la reflexión en torno a los procesos que se desencadenan en la escritura de cada línea de código. La imposibilidad de avanzar en las tareas a partir de fallas en el código suele concentrar la atención en el objetivo “hacerlo correr”, perdiendo el control sobre el proceso de trabajo.

Aquí nos propusimos dar cuenta de esas tensiones, entre el aprendizaje de las técnicas y el control del experimento de trabajo; admitiendo que en muchos casos la relación con la técnica no fue transparente. Esto tuvo una ventaja y un problema. La ventaja es que conocimos mucho sobre cada método a partir de retocar los parámetros por medio de prueba y error. La desventaja es que se introducen condiciones al corpus que -aunque puedan ser enumeradas-

introducen ediciones en los discursos analizados. Se trata de un ejercicio que no resulta violento en la interpretación tradicional, pero en la interpretación con herramientas computacionales amenaza con descontrolar el experimento. Así, cuando descarto los emojis descubro que se borran los acentos y vuelvo a eliminarlos o cuando elimino los puntos tengo que atender a no pegar palabras. La reproducción del código línea por línea y la revisión constante de los resultados es indispensable, aunque no garantiza un control completo.

En la limpieza y tokenización buscamos (de mínima) que las palabras iguales no sean tomadas como diferentes y (de máxima) que las palabras similares sean consideradas como iguales. Es posible que la omisión de algunos pasos no modifique sustancialmente nuestras conclusiones finales sobre los seguidores de políticos, aunque la visualización de la nube a partir del texto en crudo no deja dudas sobre la relevancia del preprocesamiento de los datos. Si no procesamos los datos la técnica no sirve, si los procesamos con muchas idas y vueltas corremos el riesgo de alterar los resultados. Algunos ejercicios deben ser colocados en stand by frente a la imposibilidad técnica de concretarlos.

La opción de los arreglos, antes que la producción desde cero de los encodings que necesitamos, nos ubica frente a otro problema: el cacharreo naturaliza el abusivo anglocentrismo de las herramientas. En esta línea pueden mencionarse varios aspectos: las dificultades para eliminar emojis y mantener acentos, el escaso desarrollo de listas de stopwords, los problemas específicos de la lematización en español.

En resumen, grandes volúmenes de datos, destreza técnica, problemas teóricos y algoritmos se cruzan de manera constante en las reflexiones sobre el proceso de trabajo en humanidades digitales. Evitar que el aprendizaje técnico del manejo de los datos se transforme en el único foco de atención parece ser un desafío tan grande para los cientistas sociales como el de transformarse en programadores.

Pero antes que introducir esa última posibilidad como interrogante –algo frecuente en los debates sobre humanidades digitales–, nos preguntamos sobre los caminos actualmente posibles en la capacitación técnica para poder analizar documentos nacidos digitales y grandes corpus textuales provenientes de algunas redes sociales. Esos caminos son escasos y, como esperamos haber expuesto en este trabajo, consisten en tips, anotaciones, código compartido, tutoriales, manuales y cursos breves de aprendizaje individual y en los que la autoevaluación o la evaluación por resultados son la norma. Además de esas características, existe otra constante: a mayor complejidad, mayor distancia con las disciplinas de partida; y por el contrario, a mayor proximidad, menor progresión en el aprendizaje. El tipo de aprendizaje que venimos desarrollando, más intrumental que conceptual, nos aleja de procedimientos sofisticados y poderosos para producir análisis textual (redes neuronales, por

ejemplo), y eso definitivamente es un problema para la investigación social, en la medida en que, no en la técnica, pero sí en los objetivos, el comportamiento social es un horizonte compartido entre los enfoques tradicionales y los computacionales.

Bibliografía

Annunziata, R., Ariza, A., March, V. (2017). “Gobernar es estar cerca”. Las estrategias de proximidad en el uso de las redes sociales de Mauricio Macri y María Eugenia Vidal (pp.71-93). Revista Mexicana de Opinión Pública. Año 12, Num.24

Barberá, P., J. T. Jost, J. Nagler, J. A.Tucker y R. Bonneau (2015) “Tweeting From Left to Right”. *Psychological Science*, Vol. 26, N °10.

Burgess, J., & Baym, N. K. (2020). Twitter: A biography. NYU Press.

Calvo E., Aruguete N. (2020). *Fake news, trolls y otros encantos. Cómo funcionan (para bien y para mal) las redes sociales*. Buenos Aires: Editorial Siglo XXI.

Calvo, E. (2015). *Anatomía política de Twitter en Argentina. Tuiteando #Nisman*. Buenos Aires: Capital Intelectual.

Castelo, S. (2014). #PolíticosViolentos. Un análisis de la agresión en el discurso político en Twitter. Revista SAAP. Publicación de Ciencia Política de la Sociedad Argentina de Análisis Político, 8(2),609-629.[fecha de Consulta 22 de Junio de 2021]. ISSN: 1666-7883. Recuperado en: https://www.redalyc.org/articulo.oa?id=387139302009

Cavallin, C. (2009). Del Twitter como plaza o cómo se configuran los nuevos espacios para el periodismo cultural. Anuario Electrónico de Estudios en Comunicación Social "Disertaciones", 2 (2), Artículo 4. Disponible en la siguiente dirección electrónica: http://erevistas.saber.ula.ve/index.php/Disertaciones/

Cohen, D. J., Frisch, M., Gallagher, P., Mintz, S., Sword, K., Taylor, A. M., ... & Turkel, W. J. (2008). Interchange: The promise of digital history. *The Journal of American History*, 95(2): 452-491.

Criado, J. I.; Martínez, G.; Silvan, A. (2012): “Twitter en campaña: las elecciones municipales españolas de 2011”, *RIPS*, vol. 12, nº 1, pp. 93-113.

Crymble, A. (2021). *Technology and the Historian: Transformations in the Digital Age* (Vol. 1). University of Illinois Press.

Del Rio Riande, M. G. (2019). La mirada humana. La mirada crítica. Telos, 112. Recuperable en : https://telos.fundaciontelefonica.com/telos-112-cuaderno-central-humanidades-en-un-mundo-stem-gimena-del-rio-la-mirada-humana-la-mirada-critica/

Ferreyra, S.; Reclusa, A.; Juares, W. y Perez Rubini, B. (2019). Twitter y #Corrupción: un balance sobre problemas y técnicas para pensar la política en hashtags. Actas II Jornadas de sociología Universidad Nacional de Mar del Plata, Mar del Plata.

Fussey, P., y Roth, S. (2020). Digitizing Sociology: Continuity and Change in the Internet Era. *Sociology, 54*(4), 659–674. doi: https://doi.org/10.1177/0038038520918562

Galup, L. (2019). *Big Data & Política. De los relatos a los datos, persuadir en la era de las redes sociales*. Buenos Aires: Ediciones B.

González Bengoechea, A, Fernández Muñoz, C. y García Guardia, M. L. (2019). Uso institucional o partidista de Twitter: análisis comparativo de los perfiles de Moncloa, Ayuntamiento de Madrid y sus partidos gobernantes. *Communication & Society.* 32(1): 19-38.

Hassan Montero y Herrero Solana (2010), Usabilidad de los tag-clouds: Estudio mediante eye-tracking, *Scire*, 16:1, pp. 15-33

Hernández Fernández, C. (2015). Nuevos recursos para la investigación cualitativa: Software gratuito y herramientas colaborativas. *Opción*. 31(5): 453-471

Hui, Yuk (2012). What is a Digital Object? Metaphilosophy, 43 (4):380-395.

Humberstone Morales, J. (2018). Pescando información en el océano de datos de Twitter. *Realidad Y Reflexión*, (47): 47-57. https://doi.org/10.5377/ryr.v0i47.6273

Juares, W. (2020) Redes globales para la política local. Candidatxs y seguidores en Twitter durante la campaña a intendente en General Pueyrredon 2019 (tesis de licenciatura) Universidad Nacional de Mar del Plata.

Kaptein, R.. (2012). Using wordclouds to navigate and summarize twitter search results. CEUR Workshop Proceedings. 909. 67-70.

Khusro, S., Jabeen, F. & Khan, A. (2018) Tag Clouds: Past, Present and Future. *Proc. Natl. Acad. Sci., India, Sect. A Phys. Sci., 91*,369–381. doi: https://doi.org/10.1007/s40010-018-0571-x

Larrosa, J. M. (2013). Un ejercicio de estudio de una acción colectiva: El 8N en Twitter. https://ri.conicet.gov.ar/handle/11336/2040

Lovink, G. (2019). *Tristes por diseño. Las redes sociales como ideología*. Bilbao: Consoni.

Molina, J. L. (2001). *El análisis de redes sociales. Una introducción*. Barcelona: Ediciones Bellaterra.

Moretti, F. (2015). El matadero de la literatura. *Lectura distante*. Buenos Aires: CFE.

Murphy Choy (2012). Effective Listings of Function Stop words for Twitter. *International Journal of Advanced Computer Science and Applications.* 3 (6): 8-11.

Murthy, D. (2012). Towards a Sociological Understanding of Social Media: Theorizing Twitter. *Sociology*, *46*(6), 1059–1073. doi: https://doi.org/10.1177/0038038511422553

Morstatter, F.; Pfeffer, J.; y otros (2013) Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. Disponible en: https://arxiv.org/abs/1306.5204

O´Neil, Cathy (2017). *Armas de Destrucción Matemática. Cómo el Big Data aumenta la desigualdad y amenaza la democracia*. Madrid: Capitan Swing.

O'Reilly, T. (2009). What is web 2.0. " O'Reilly Media, Inc.".

Orihuela, J. L. (2011). Mundo Twitter. Barcelona:Alienta.

Pons, A. P. (2013). *El desorden digital: guía para historiadores y humanistas*. Madrid: Siglo XXI Editores.

Sanchez, E. (2017) Observatorio de conversaciones en Twitter. (Trabajo de grado en Ingeniería Informática) Universidad politécnica de Valencia.

Reguillo, R. (2017) P*aisajes insurrectos. Jóvenes, redes y revueltas en el otoño civilizatorio.*Barcelona, España: Nuevos emprendimientos editoriales.

Rogers, R. (2013a). Debanalizing Twitter: The transformation of an object of study. In Proceedings of the 5th annual ACM web science conference (pp. 356-365).

Rogers, R. (2013b). *Digital methods*. Boston: MIT press.

Röhle, B. R. T. (2012). *Digital methods: Five challenges. In Understanding digital humanities*. London: Palgrave Macmillan.

Rosenzweig, R. (2003). Scarcity or abundance? Preserving the past in a digital era. *The American historical review.* 108 (3): 735-762.

Silke Schwandt (2020). Métodos digitales para la semántica histórica. Tras el rastro de los conceptos en corpus digitales. *Conceptos Históricos*, 5 (8): 160-196.

Slimovich, A. (2012). El Facebook de los gobernantes. El caso de Cristina Fernández de Kirchner y de Mauricio Macri. En M. Carlón y A. Fausto Neto (Comps.), Las políticas de los internautas. Nuevas formas de participación (pp. 137-154). Buenos Aires, Argentina: La Crujía.

Urdinez, F. and Cruz, A. (eds.) (2020). *Political Data Science Using R: A Practical Guide*. CRC Press.

Van Dijck, J. (2016). *La cultura de la conectividad. Una historia crítica de las redes sociales*. Buenos Aires: Siglo XXI.

Ventura, A. (2018). ¿Cómo analizar discursos de 140 caracteres? Propuesta metodológica para el estudio del discurso político de campaña en Twitter a

partir del análisis estratégico del discurso con una perspectiva multimodal y crítica. *CHIMERA: Romance Corpora And Linguistic Studies, 5*(2), 275-287. doi:http://dx.doi.org/10.15366/chimera2018.5.2.006

Woolley S. C. and Howard P. (2018) Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media. New York, NY : Oxford University Press.