ISSN 2013-9004 (digital); ISSN 0210-2862 (paper)
https://doi.org/10.5565/rev/papers.2967

El 14F a Instagram. Una proposta d'articulació de tècniques de raspat web i anàlisi de xarxes

Jordi Morales-i-Gras
Oriol Sánchez-i-Vallès
Network Outsight
morales.jordi@gmail.com; sanchezv.oriol@gmail.com

Recepció: 04-03-2021
Acceptació: 15-07-2021
Publicació: 13-01-2022

Resum

En aquest article analitzem la campanya electoral del 14 de febrer de 2021 al Parlament de Catalunya a través de la conversa dels partits a Instagram, una de les plataformes digitals amb més usuaris registrats i també una de les menys ateses en la recerca sociològica sobre els mitjans socials. Hem aplicat una sèrie de tècniques de raspat web ètic i legal per adquirir les dades, que han estat recuperades, processades i emmagatzemades en una base de dades relacional. Posteriorment, hem aplicat tècniques de mineria de dades i algoritmes d'aprenentatge no supervisat orientats, per una banda, a l'anàlisi descriptiva i exploratòria de la conversa i, per l'altra, a l'elaboració de xarxes de coocurrències lèxiques que ens permeten fer una anàlisi sobre el discurs que articulen dels partits. Partint d'una metodologia inductiva i estructural, hem caracteritzat diversos aspectes del relat que han construït els partits catalans en la campanya electoral: els relatius a les seves pràctiques de publicació de continguts, a l'acollida de les seves audiències i als diferents usos de *hashtags* i paraules que han dut a terme. Més enllà del cas d'anàlisi concret i de la caracterització dels relats dels partits durant la campanya política del 14F i les seves diferències internes, amb aquest article també pretenem posar sobre la taula un model d'estudi basat en tècniques d'anàlisi de dades massives aplicables i replicables en qualsevol escenari de dades adquirit mitjançant tècniques de raspat web ètic i legal que garanteix l'autonomia investigadora dels científics/ques socials.

Paraules clau: anàlisi de xarxes socials; Instagram; raspat web; dades massives

Abstract. *The 14F on Instagram. A proposal for articulation of web scraping and network analysis techniques.*

In this article we analyse the election campaign of 14 February 2021 in the Catalan Parliament through the parties' conversation on Instagram: one of the digital platforms with the most registered users and one of the least attended in sociological research on social media. We have applied a few ethical and legal web scraping techniques to acquire the data, which have been retrieved, processed and stored in a relational database. Subsequently, we have applied data mining techniques and unsupervised learning algorithms oriented, on the one hand, towards the descriptive and exploratory analysis of the conversation, and on the other, towards the elaboration of networks of lexical co-occurrences that allow us to apply an analysis on the discourse articulated by the parties. Using an inductive and structural methodology, we have characterised various aspects of the narrative constructed by the Catalan parties in the electoral campaign: aspects relating to their content publication practices, the reception of their audiences and the different uses of hashtags and words they have made. Beyond the specific case of analysis and the characterisation of the parties' narratives during the 14F political campaign and their internal differences, with this article we also aim to put on the table a model of analysis based on big data analysis techniques applicable and replicable in any data scenario acquired through ethical and legal web scraping techniques that guarantee the research autonomy of social scientists.

Keywords: social network analysis; Instagram; web scraping; big data

Sumari

1. El *big data* en campanyes polítiques i la invisibilització d'Instagram

Els processos electorals i els referèndums que han tingut lloc durant la segona dècada del segle XXI s'han desenvolupat també en un l'escenari digital, per complementar i ampliar els espais que eren habituals en les campanyes polítiques d'aquesta mena. La campanya de Barack Obama per a la presidència dels EUA de 2008 va marcar un punt d'inflexió a partir del qual ja mai més cap partit ni moviment deixaria de banda els anomenats «mitjans socials» (en anglès, *social media*) a l'hora de planificar, executar i avaluar una campanya política (Metzgar i Maruggi, 2009).

En un primer moment, des del món acadèmic i des de l'àmbit de la investigació social aplicada, es van mobilitzar recursos intel·lectuals, tècnics i econòmics amb la intenció d'utilitzar les «dades massives» (en anglès, *big data*) que proporcionen les plataformes digitals per generar coneixement social. Molt significativament, van sorgir diverses propostes que tractaven de substituir o, com a mínim, complementar les anàlisis demoscòpiques i els sondeigs electorals

mitjançant tècniques com l'anàlisi de xarxes socials (d'ara endavant, AXS) o aproximacions amb models estadístics i d'intel·ligència artificial, amb la intenció d'identificar patrons en grans volums de dades que permetessin predir la conducta electoral o l'opinió de manera general i generalitzada (Gayo-Avello, 2013). El sociòleg Vincent Mosco (2014) va anomenar «positivisme digital» aquest grup de propostes que pressuposen, implícitament o explícitament, que els mitjans socials constitueixen quelcom equiparable a una rèplica o una mostra d'un suposat «món real» o, fins i tot, un laboratori esterilitzat en el qual es pot observar netament i sense biaixos tota mena de fenòmens mitjançant tècniques eminentment quantitatives. El físic i periodista Chris Anderson, editor de la revista de divulgació científica i tecnològica *Wired*, representa una de les versions més extremes d'aquest positivisme digital, que menysprea el paper de les teories científiques i de les explicacions causals:

> The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all. (Anderson, 2008)

El positivisme digital continua sent avui la ideologia més finançada en el món de les dades massives, un espai dominat per perfils tecnòlegs, físics, matemàtics, actuaris i executius amb MBA obtinguts a escoles de negocis (Fuchs, 2017). Però, tal com ho veiem nosaltres, és molt probable que aquesta ja no sigui la tendència majoritària en el món de la ciència social ni en l'opinió pública. Va ser a partir d'escàndols com el de l'empresa Cambridge Analytica i de les victòries polítiques de Trump, Bolsonaro o del Brexit —difícils d'entendre sense el paper dels mitjans socials— que la comunitat científica, la premsa i la societat en general vam centrar cada cop més l'atenció en qüestions com la creació i la disseminació de notícies falses a través dels mitjans socials i en la manipulació massiva (Grinberg et al., 2019), o fins i tot en el fenomen que es va donar a conèixer com la postveritat, que va ser la paraula de l'any 2016, segons el diccionari d'Oxford (2016). És fàcil argumentar que, vistes des de 2021, moltes d'aquelles primeres propostes de predicció electoral o de mesura demoscòpica semblen enormement ingènues en les seves assumpcions i absurdes en les seves ambicions. A aquestes alçades, ja no tenim cap dubte que els mitjans socials no constitueixen una rèplica ni encara menys una mostra d'un suposat «món real». Avui sabem positivament que l'esfera pública digital és una arena política altament disputada i un espai per a la influència i la manipulació social (Masip et al., 2019).

Més enllà de les opinions personals i professionals que pugui tenir cadascú, resulta evident que les relacions de l'acadèmia i del món de la investigació aplicada amb els mitjans socials i les dades massives no són avui les mateixes que fa deu anys. La segona dècada del segle XXI ha suposat un viatge des d'aquella mena de positivisme ingenu —tot i que segur que no per a tothom— cap a un

tipus d'utilitarisme escèptic (Morales-i-Gras, 2020). És a dir, avui sabem que els mitjans socials són espais enormement complexos —i a voltes confusos— en els quals és massa arriscat treure segons quines conclusions. Alhora, però, també sabem que són espais privilegiats per a la generació de coneixement social i, sobretot, per dur-hi a terme operacions d'enginyeria social (per exemple, per conscienciar sobre problemes concrets, per difondre iniciatives, per generar debat públic, per crear i mantenir climes d'opinió, etcètera).

Tots els canvis en la mentalitat acadèmica i en l'opinió pública han vingut acompanyats d'una sèrie de transformacions legals, com és el cas de l'aplicació del Reglament General de Protecció de Dades (RGPD) a la Unió Europea a partir del maig de 2018. D'altra banda, la majoria empreses d'internet han implementat una sèrie de modificacions molt importants en les condicions d'accés a les dades massives. D'aquesta manera, un clar efecte pervers del suposat increment de la privacitat i la seguretat de les dades a internet ha estat que les interfícies de programació d'aplicacions (API, segons la sigla en anglès) de les plataformes digitals proporcionen avui moltes menys dades que les que proporcionaven abans de 2015. Això és així fins i tot per als usuaris de tipus acadèmic, que, al cap i a la fi, són els més indicats per estudiar les dinàmiques de desinformació i manipulació social i per proposar mecanismes correctius. Tal com exposa el sociòleg Axel Bruns (2019), l'*APIcalipsi* ha estat especialment dura en plataformes com Facebook o Instagram, que han tancat l'aixeta de les dades gairebé totalment, la qual cosa ha provocat que aplicacions com Netvizz o Instagram Network —totes dues creades i mantingudes pel grup de recerca Digital Methods Initiative de la Universitat d'Amsterdam, que proporcionaven dades de Facebook i d'Instagram— hagin deixat de funcionar. Per totes aquestes raons, cada cop més investigadors/es socials, tant des de l'acadèmia com des de la recerca aplicada, opten per l'aplicació de tècniques de raspat web (*webscrapping*) per a l'obtenció de dades socials per dur a terme els seus projectes (Veltri, 2019). Aquest article n'és un exemple.

Més endavant entrarem en qüestions ètiques, legals i tècniques al voltant del raspat web. Abans, però, és important considerar quins han estat els efectes que les restriccions en les API de les plataformes han tingut en la recerca acadèmica i en la generació de coneixement social. En aquest punt, cal assenyalar com un problema greu la sobrerepresentació de Twitter per damunt de les altres plataformes en la recerca acadèmica i, alhora, una molt profunda subrepresentació de plataformes com Instagram (Matamoros-Fernández i Farkas, 2021). Tot just ara, la comunitat cientificosocial estaria començant a corregir aquest fet, malgrat el seu gran nombre d'usuaris, uns 16 milions a l'Estat espanyol, segons diversos estudis,[1] en contrast amb els 7,5 milions d'usuaris de

1. Entre aquests estudis, consten el «VI Informe de los perfiles en redes sociales de España», dut a terme per la consultora The Social Media Family (2020), o el «Digital 2020 España», de la consultora We Are Social (2020). Els dos informes afirmen treballar amb extrapolacions de dades proporcionades per les mateixes plataformes i, per tant, no podem garantir que les xifres que ofereixen siguin exactes.

Twitter al mateix territori. La política relativament oberta de Twitter d'accés a dades a través de la seva API oficial contrasta amb l'hermetisme i l'opacitat de Facebook o Instagram. Això té com a resultat una situació semblant a la del famós acudit de l'home ebri que perd les claus de casa a la porta del bar i les busca sota el fanal. Tot sovint, els/les investigadors/es socials ens veiem obligats a cercar totes les respostes a les nostres preguntes a la zona que està il·luminada (i. e., Twitter), en lloc de provar d'il·luminar els territoris més foscos on intuïm que podria haver-hi respostes molt interessants a preguntes que no gosem fernos (per exemple, Facebook, Instagram, TikTok, etcètera).

La invisibilització d'Instagram en la recerca acadèmica s'explica en gran part per la major dificultat en l'adquisició de dades de manera sistemàtica, però també és molt probable que qüestions com la seva orientació audiovisual o l'estil més aviat frívol dels seus continguts hi tinguin alguna cosa a veure (Highfield i Leaver, 2016). Dit d'una altra manera, és molt possible que els/les científics/ques socials, periodistes, divulgadores i altres perfils amb un capital cultural elevat es trobin més còmodes a Twitter que a Instagram i que, per tant, apliquin el seu biaix a l'hora d'escollir la plataforma sobre la qual fer recerca. Explorar la hipòtesi anterior va més enllà del que ens proposem en aquest article, però creiem que és important considerar-la com a argument a favor d'incloure més mitjans socials en el radar de la ciència social.

L'objectiu d'aquest article és doble. D'una banda, volem caracteritzar el relat que els diferents partits polítics catalans han articulat a Instagram per a la campanya del 14F. Per fer-ho, utilitzarem tècniques quantitatives descriptives sobre les dades i establirem estratègies d'anàlisi més aviat qualitatives sobre el discurs dels partits. Ho portarem a terme mitjançant tècniques d'AXS aplicades a xarxes de coocurrències d'unitats lèxiques: una xarxa de *hashtags* i una xarxa de paraules. D'altra banda, la nostra intenció també és contribuir a l'elaboració d'estratègies generals —que vagin, per tant, més enllà de l'anàlisi de xarxes aplicada al discurs electoral— que permetin als investigadors/es socials guanyar autonomia en la recerca sobre continguts publicats en mitjans socials com Instagram, per deixar de dependre tant de les decisions —sovint molt arbitràries— que prenen les diferents empreses d'internet sobre l'accés a les dades, ja que quasi sempre deixen de banda aspectes com el bé comú i el progrés científic.

2. Aspectes ètics, legals i tècnics del raspat web de dades socials massives

Les tècniques de raspat web han estat i són crucials per al disseny d'internet tal com el coneixem avui. De fet, molts dels serveis en xarxa més populars estan totalment o parcialment basats en aquestes tècniques, en què s'inclouen tot tipus d'agregadors de continguts, comparadors de productes i, fins i tot, indexadors i cercadors com Google o Bing. En essència, el raspat web consisteix a extreure informació d'una pàgina web de manera automatitzada i simulant la navegació d'un humà. D'aquesta manera, mitjançant tècniques de raspat web, és possible visitar una web recurrentment (per exemple, cada setmana, cada

dia, cada hora, cada minut, etcètera) i extreure'n la informació que conté per injectar-la a una base de dades sobre la qual es té un accés complet.

Per tant, el raspat web no és una tècnica intrínsecament il·legítima o il·legal —de fet, és fonamental per al bon funcionament de la majoria de serveis d'internet—, però s'ha d'entendre que no totes les pràctiques que es poden dur a terme mitjançant raspat web són legítimes i legals. Malgrat que hi ha diferències substancials entre les lleis aplicables a diferents territoris, com a mínim hi ha dues qüestions que cal tenir en compte èticament i legalment: 1) la propietat intel·lectual dels textos, imatges o vídeos que es puguin extreure d'una web i 2) el tractament de la informació de caràcter personal que es pot extreure d'una web (Landers et al., 2016). Si centrem les qüestions anteriors en el marc d'Instagram, pel que fa a la propietat intel·lectual dels continguts als quals es pot accedir, la investigadora o l'investigador han de considerar la normativa pròpia de la plataforma, que prohibeix explícitament la recuperació no autoritzada d'informació en les seves condicions d'ús. En conseqüència, cal assumir que no és possible crear comptes destinats a recopilar informació de manera automàtica si no es disposa d'un permís explícit de l'empresa. Així doncs, no podem accedir a dades d'Instagram mitjançant raspat web a través d'usuaris registrats. L'única informació que es pot extreure d'Instagram amb raspat web és aquella que la plataforma fa visible per als robots indexadors, és a dir, la disponible per als navegants que no han fet *log-in* a l'aplicació i que no requereix de l'acceptació de les condicions d'ús de la plataforma.

Pel que fa el tractament de la informació de caràcter personal, el reglament de referència no és el condicionat d'ús d'Instagram, sinó el RGPD, el qual estipula clarament que els i les responsables i encarregats del tractament de dades hem de tenir una actitud proactiva en la protecció de les dades de caràcter personal dels individus, com poden ser noms d'usuari, telèfons o correus electrònics, entre d'altres (Demetzou, 2019). En aquest sentit, i sempre que no hi hagi alguna bona raó basada en l'interès legítim que ens permeti articular una estratègia alternativa (una excepció típica podrien ser les recerques que es proposen la identificació d'*influencers* o de líders d'opinió, que ens pot oferir el marc legal i ètic oportú per identificar aquest tipus d'usuaris), és necessari dur a terme procediments d'anonimització irreversible d'aquestes dades personals, per exemple, agregant casos en les bases de dades o aplicant-hi multiplicacions matricials, en lloc de disposar de dades i metadades individuals. En aquest punt, és molt important entendre que l'analista ha de tenir una actitud proactiva en l'anonimització de les dades que s'obtenen mitjançant raspat web —com també passa amb les dades que provenen d'API oficials— i que en cap cas es poden difondre en el mateix format en el qual s'han obtingut. Respecte a aquest últim punt, també s'ha d'entendre que el reglament només es pot aplicar sobre persones físiques, i que, per tant, sí que es poden emmagatzemar, processar i difondre dades d'empreses, institucions o partits polítics extretes de plataformes com Instagram:

> Los principios de la protección de datos deben aplicarse a toda la información relativa a una persona física identificada o identificable (…). Por lo tanto los

> principios de protección de datos no deben aplicarse a la información anónima, es decir información que no guarda relación con una persona física identificada o identificable, ni a los datos convertidos en anónimos de forma que el interesado no sea identificable, o deje de serlo. En consecuencia, el presente Reglamento no afecta al tratamiento de dicha información anónima, inclusive con fines estadísticos o de investigación. (Diario Oficial de la Unión Europea, 2016: 119/5)

Un cop contrastada la viabilitat ètica i legal d'un projecte de recerca, és important tenir en compte una sèrie de qüestions tècniques que afecten la implementació del raspat web. Extreure informació d'una pàgina web és una tasca molt senzilla quan està escrita en llenguatge HTML, però és una operació força més complicada quan s'utilitza llenguatge JavaScript, com és el cas d'Instagram. Una dificultat addicional és que la plataforma podria bloquejar les adreces IP des de les quals provem d'accedir a les dades. Per superar totes aquestes limitacions, és altament recomanable utilitzar serveis intermediaris com ApiFy, Octoparse, Luminati, ProxyCrawl o ScrapeHero, que permeten implementar operacions de raspat web anònimament, sense *log-in* i a través d'una xarxa de *proxys* distribuïts per tot el món. Típicament, aquests serveis facturen en funció del nombre de pàgines consultades, a raó de fraccions de cèntims de dòlar per pàgina. En aquesta investigació hem utilitzat el proveïdor Luminati per consultar diàriament nou pàgines públiques d'Instagram, amb un cost de $0,005 per pàgina durant 20 dies. Per tant, el cost de les dades d'aquesta investigació ha estat de $0,9 o 0,75 €.

3. Metodologia

En aquest article volem mostrar i fer valer una sèrie de tècniques d'extracció de dades socials massives de la plataforma Instagram, i oferir també propostes d'anàlisi descriptives i exploratòries basades en tècniques d'anàlisi estadística i d'AXS. Les dades de l'estudi han estat obtingudes a través del proveïdor Luminati amb tècniques de raspat web sobre una sèrie de perfils d'Instagram. Seguidament, expliquem el procés segons el qual les dades han estat capturades, processades i posteriorment analitzades.

— Les dades es van generar, recuperar i emmagatzemar durant la campanya electoral de les eleccions al Parlament de Catalunya del 14 de febrer de 2021, entre el 29 de gener i el 17 de febrer de 2021, amb l'objectiu de capturar l'increment de *likes* o comentaris en els posts fins a cinc dies després del final de la campanya, el 12 de febrer. Per tant, l'anàlisi va del 29 de gener al 12 de febrer de 2021, període que correspon a la campanya electoral, deixant de banda la jornada de reflexió i el dia de l'elecció.
— Cada dia a les 00:00 es recuperaven els dotze[2] posts més recents de cada un dels nou partits polítics que es presentaven a les eleccions, i que van ser

2. Per defecte, el proveïdor de dades retorna dotze posts per consulta a Instagram. El màxim nombre de posts publicats per un sol partit en 24 hores va ser de nou i, per tant, no vam haver de revisar aquest paràmetre.

considerats per la Corporació Catalana de Mitjans Audiovisuals i la majoria de mitjans de comunicació, a través dels seus usuaris oficials: ciutadanscs, cupnacional, encomupodem, esquerrarepublicana, juntspercat, pdemocratacat, ppcatala, socialistes_cat i voxbarcelona.[3]

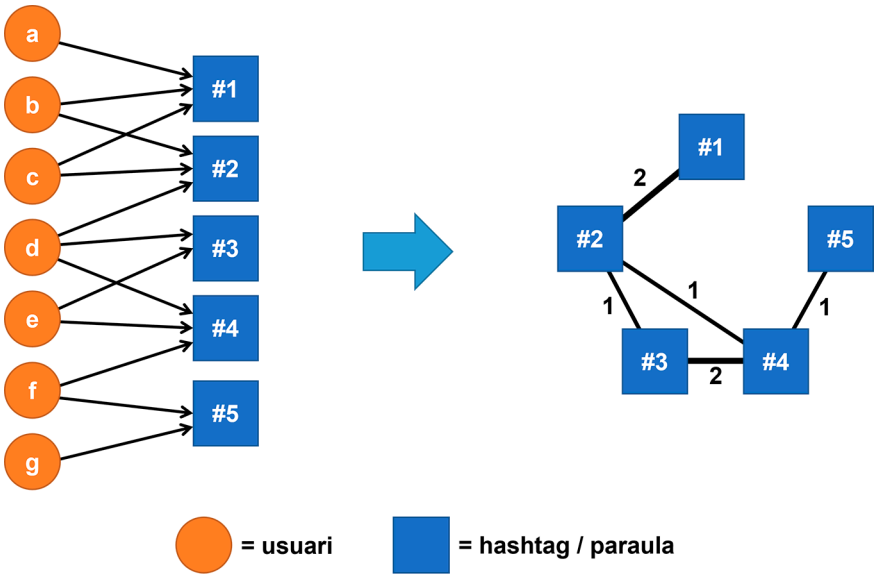
— Els camps de dades que es recuperaven diàriament eren els següents: nom curt de l'usuari, text del post, nom llarg de l'usuari, enllaç a la imatge, descripció de l'usuari, URL externa de l'usuari, nombre de seguits per l'usuari, nombre de seguidors de l'usuari, nombre de posts de l'usuari, nombre de *likes* en el post, nombre de visualitzacions del vídeo del post, nombre de comentaris en el post, condició de verificació de l'usuari, dia i hora del post, URL de la imatge del post, URL del vídeo del post, URL del post i URL de les miniatures (*thumbnails*).
— Les dades es generaven en format semiestructurat JSON i, posteriorment, eren ingerides per Apache NiFi (2018), estructurades i inserides a una base de dades relacional PostgreSQL (2021), en la qual es duien a terme les operacions d'eliminació de duplicats o de segmentació de textos requerides per a les anàlisis posteriors (i. e., aïllament de *hashtags* i aïllament de paraules clau). Els camps de dades considerats en les anàlisis són l'URL del post com a identificador únic, el nom curt de l'usuari, el text del post, el nombre de *likes* i el nombre de comentaris.

El conjunt de les dades de l'estudi són 325 posts publicats per nou partits polítics durant quinze dies que han sumat un total de 221.581 *likes* i 8.364 comentaris. Per tal d'obtenir les diverses xarxes sobre les quals aplicar les tècniques d'AXS, s'han practicat una sèrie de multiplicacions matricials amb el *software* Pajek (1998) que permeten passar d'una xarxa de dos modes a una xarxa d'un sol mode. Així, una xarxa d'usuaris i *hashtags* pot ser transformada en una xarxa de *hashtags* vinculats en funció del nombre de vegades que diversos usuaris els han utilitzat (figura 1). En total, s'han aplicat dues operacions d'aquest tipus: 1) per transformar una xarxa d'usuaris i *hashtags* en una xarxa de *hashtags* i 2) per transformar una xarxa d'usuaris i paraules en una xarxa de paraules.

Les xarxes d'un mode sintetitzades (i. e., xarxa de *hashtags* i xarxa de paraules) són xarxes de coocurrències lèxiques no dirigides i ponderades, de manera que els vincles entre els nodes no tenen direcció perquè venen determinats pel nombre d'usuaris que els diferents *hashtags* o paraules comparteixen, però sí que tenen pesos perquè poden haver aparegut conjuntament en més d'un post. A l'hora d'analitzar el rol dels diversos nodes, hem tingut en compte la centralitat d'intermediació creada per Linton Freeman (1977) i optimitzada per Ulrik Brandes (2001), que es troba disponible en el *software* Gephi (Bastian et al., 2009), en el qual hem analitzat les xarxes. Els nodes amb més centralitat

3. Vox no té un perfil d'Instagram d'àmbit català, ja que treballen només amb usuaris estatals i provincials. En aquesta investigació, hem considerat més adequat incloure l'usuari provincial de Barcelona que l'usuari de Vox estatal, per no veure'ns obligats a considerar també els usuaris estatals de partits com el PP, el PSOE, Podem o Ciutadans.

Figura 1. Transformació d'una xarxa de dos modes en una xarxa d'un mode



Font: elaboració pròpia.

d'intermediació són aquells que estan més ben connectats en una xarxa, des de la perspectiva que es troben enmig d'un gran nombre de les rutes geodèsiques (i. e., les més curtes possibles) entre tots els parells de nodes connectats de la xarxa. D'altra banda, també hem utilitzat l'algoritme d'identificació comunitària Louvain (Blondel et al., 2008), basat en la mètrica de modularitat de Mark Newman (2006) i optimitzat mitjançant el paràmetre de resolució disponible a Gephi (Lambiotte et al., 2009). L'algoritme genera una partició de xarxa agrupant els nodes, maximitzant els vincles intragrup i minimitzant els intergrup, cosa que proporciona un índex de modularitat entre 0 i ±1 que serveix per avaluar la qualitat de la partició (les xifres superiors a 0,3 són interpretades com a matemàticament significatives).[4] Les xarxes sintetitzades s'han visualitzat amb l'algoritme Force Atlas 2 de Gephi (2014), que apropa o allunya els nodes d'un graf en funció de la intensitat dels seus vincles.

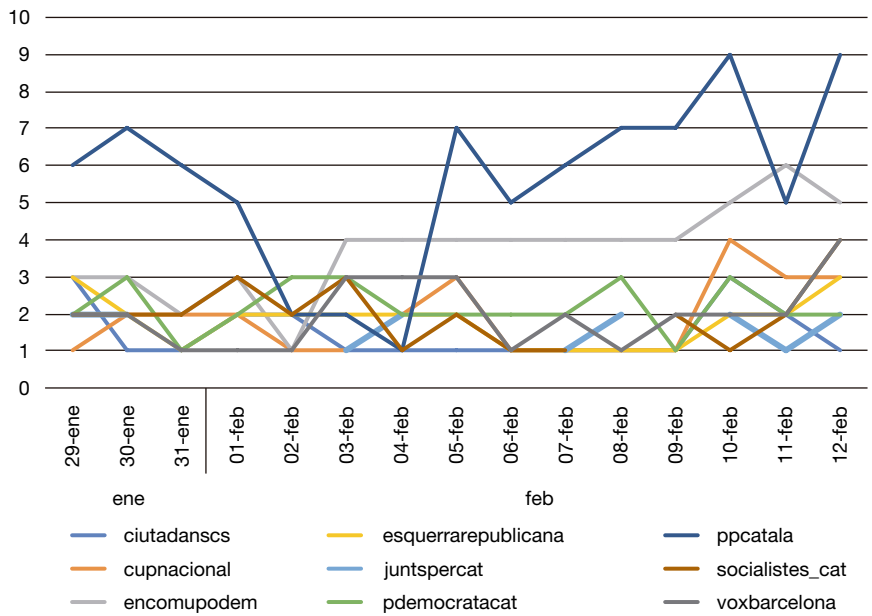
4. La xifra de 0,3 va ser proposada per Newman (2006) com a convenció general, de la mateixa manera que en tasques de validació hipotètica es considera el valor p de 0,05 o 0,001. Però la significació matemàtica que proporciona la mètrica de modularitat en un algoritme de detecció comunitària no pot ser interpretada sota els mateixos paràmetres que la significació del valor p en estadística inferencial. Per la seva naturalesa no supervisada i inductiva —diferent, per tant, dels models supervisats i hipoteticodeductius—, l'algoritme Louvain optimitza els seus resultats mitjançant l'avaluació permanent de la modularitat implementant només aquelles operacions que milloren l'indicador. Això vol dir que l'algoritme està dissenyat per obtenir una solució òptima amb una modularitat positiva i elevada. Per tant, no és correcte interpretar la modularitat com un mètode de validació de resultats extern a la lògica de l'algoritme.

L'estratègia analítica articulada és, per tant, de tall exploratori i inductiu, tal com sol ser habitual en molts estudis que se serveixen de l'AXS (Lozares, 1996). Per començar, s'efectuaran una sèrie d'anàlisis exploratòries descriptives (i. e., encreuament de variables, anàlisi de correlació i anàlisi k-Means) per obtenir una panoràmica general de les dades i de les grans tendències que dibuixen. Per fer-ho, s'utilitzaran els programes MS Excel i Orange Datamining per Python3 (Demsar et al., 2013). Posteriorment, se sintetitzaran diferents xarxes a partir de diferents blocs d'informació provinents de la base de dades capturada, s'aplicaran les mètriques i els algoritmes de xarxes corresponents, i es procedirà a l'exploració de les xarxes i a la seva interpretació mitjançant diferents estratègies d'encreuament de dades. Al cap i a la fi, la metodologia aplicada és un exemple més de la hibridació entre les mirades quantitatives i qualitatives que demanden els entorns de dades massives.

4. Anàlisi de dades

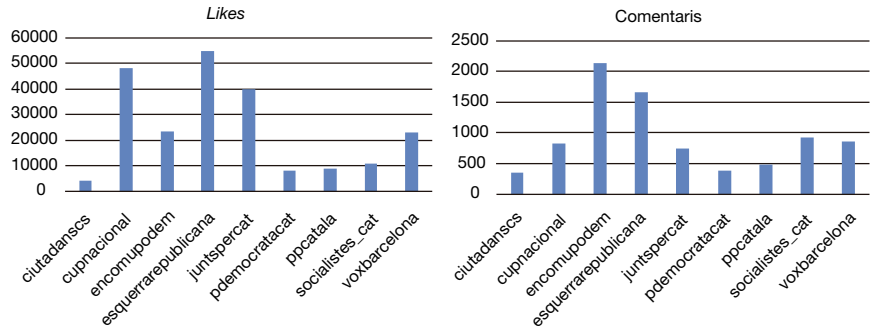
Les dades d'aquest estudi reflecteixen l'activitat de les nou candidatures principals a les eleccions del 14F al Parlament de Catalunya a Instagram. Des d'un punt de vista merament descriptiu, cal destacar que els partits que més han publicat a Instagram durant la campanya electoral han estat el PP (84 publicacions, 5,6 al dia) i ECP (56 publicacions, 3,7 al dia). Aquests dos partits també són els que han publicat de manera més irregular cronològicament (figura 2),

Figura 2. Cronologia del nombre de publicacions per partit



Font: elaboració pròpia amb MS Excel.

Figura 3. Nombre de *likes* i comentaris per partit



Font: elaboració pròpia amb MS Excel.

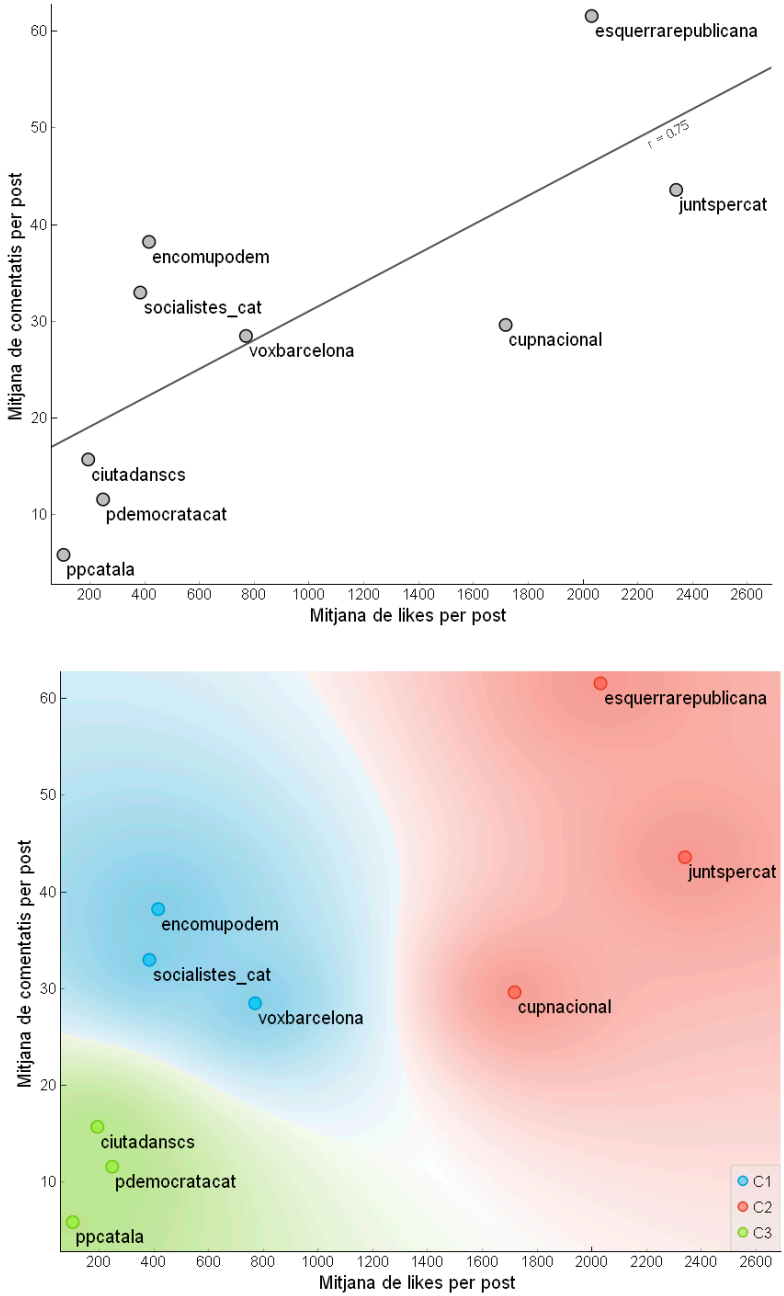
entenem que per adaptar-se als diferents esdeveniments que succeïen durant la campanya. La resta de partits han oscil·lat entre les 17 i les 33 publicacions totals (entre 1,1 i 2,2 al dia) i han optat per la regularitat cronològica.

El nombre de publicacions és un indicador que ens apropa al concepte d'intensitat en l'activitat digital dels partits, però que no ens diu gairebé res sobre l'acollida d'aquests continguts per part de l'audiència d'Instagram. Al contrari, indicadors com el nombre de *likes* o el nombre de comentaris sí que ens aproximen a aspectes relatius a les reaccions de l'audiència als continguts publicats pel partit (figura 3). Els tres principals partits independentistes (ERC, CUP i Junts) són clarament els usuaris que més *likes* han recollit. Entre tots tres reuneixen el 64,47 % dels *likes* de la campanya, fet que deixa entendre que l'audiència independentista és la més activa a Instagram. En un segon terme, ens trobem amb opcions tan diferents com ECP o Vox, que sumen el 21 % dels *likes* de la campanya, amb xifres generals molt semblants, properes als 23.000 *likes*. Finalment, tenim la resta de partits, que han recollit menys del 15 % dels *likes* i disposen de xifres totals per sota dels 11.000 *likes*. Pel que fa a les xifres dels comentaris, cal destacar que els dos partits que més n'han recollit són ECP i ERC, ja que entre tots dos tenen el 45,47 % del total. La resta de partits es reparteixen el 54,53 % dels comentaris restants, amb xifres totals que oscil·len entre els 300 i els 1.000 comentaris en total.

Una manera de posar en context totes les dades anteriors és comparar les xifres mitjanes de *likes* i comentaris que cada partit ha aconseguit en els seus posts. Si ens fixem en la relació entre la mitjana d'aquestes dues variables pels posts de cada partit (figura 4), podem observar amb nitidesa que totes dues estan fortament correlacionades[5] (r = 0,75): com més *likes* per post reculli un partit polític català en campanya electoral, més comentaris per post aconseguirà

5. El coeficient de correlació de Pearson (r) mesura la dependència lineal entre dues variables quantitatives i contínues. Pren valors entre 0 i ±1 en funció de la intensitat i la direcció de la relació.

Figura 4. Relació entre la mitjana de *likes* i la mitjana de comentaris per post



Font: elaboració pròpia amb Orange Datamining per Python3.

també. També hem aplicat l'algoritme no supervisat k-Means[6] sobre les dades amb l'objectiu d'identificar grups de partits similars, i s'ha seleccionat la clusterització amb el valor Silhouette[7] més elevat (0,53). La clusterització resultant classifica els partits en tres segments diferents: ECP, PSC i Vox conformen un grup amb les audiències moderadament actives, amb força comentaris i no tants *likes* per post (C1, blau); ERC, Junts i la CUP es perfilen com els usuaris amb les audiències més actives a Instagram, amb molts *likes* i molts comentaris per post (C2, vermell), i finalment Cs, el PDeCat i el PP es posicionen com el clúster amb menys *likes* i menys comentaris als seus posts (C3, verd).

4.1. Anàlisi d'una xarxa de hashtags

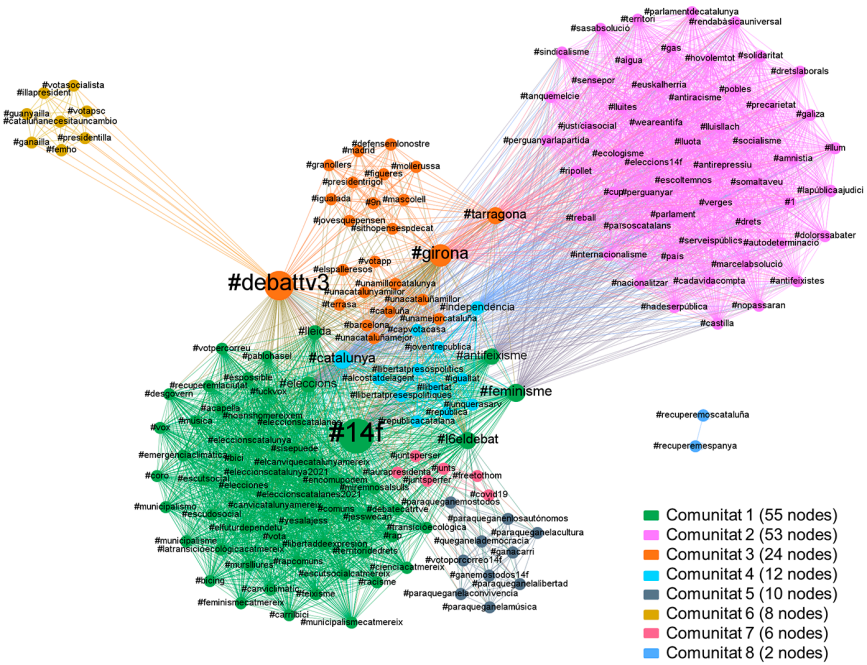
La primera proposta d'anàlisi de xarxes consisteix a transformar la conversa capturada en una xarxa de *hashtags*. Per dur-la a terme, s'han aplicat una sèrie de transformacions sobre les dades emmagatzemades a la base de dades relacional PostgreSQL i, posteriorment, amb Pajek.

1. S'han transformat els textos dels posts en bosses de paraules a través de l'expressió regular *regexp_split_to_table*, que permet segmentar cadenes donant lloc a múltiples files amb codis de longitud variables (*varchar*).
2. Posteriorment, s'han filtrat només les paraules que comencen amb el caràcter # (els *hashtags*) i s'ha fet una nova consulta a la base de dades per recuperar les columnes *autor* i *hashtag* en minúscules per evitar duplicitats.
3. S'ha transformat la consulta anterior en una xarxa de dos modes mitjançant el *software* Txt2Pajek (Pfeffer et al., 2013). S'ha introduït la xarxa de dos modes a Pajek i s'ha transformat en una xarxa d'un mode: una xarxa de *hashtags* vinculats pels autors que comparteixen amb els vincles ponderats segons el nombre de cops que han aparegut conjuntament.
4. La xarxa de *hashtags* d'un sol mode s'ha exportat de Pajek i importat a Gephi, on s'ha interpretat com a xarxa no dirigida. Amb Gephi, s'han identificat les comunitats amb l'algoritme Louvain i s'ha calculat la centralitat d'intermediació de cada node.

La xarxa de *hashtags* sintetitzada consta de 170 nodes i 3.966 arestes ponderades (figura 5). Els *hashtags* amb una centralitat d'intermediació superior a zero són aquells que han estat utilitzats per més d'un partit. Des d'aquesta perspectiva, podem identificar que els *hashtags* comuns i compartits de la

6. L'algoritme k-Means agrupa els casos en diferents clústers en funció de la seva proximitat amb els valors mitjans o centroides. L'algoritme genera diversos resultats possibles que poden ser comparats amb una mètrica com el valor Silhouette.
7. El valor Silhouette contrasta la distància mitjana entre els elements dins d'un clúster i la distància mitjana amb els elements dels altres clústers, i pren valors entre 0 i ±1 en funció de la intensitat de la similitud o la diferència entre els elements d'un clúster. És un indicador que es pot fer servir per validar la consistència interna dels clústers generats amb algoritmes no supervisats com k-Means.

Figura 5. Xarxa de *hashtags*

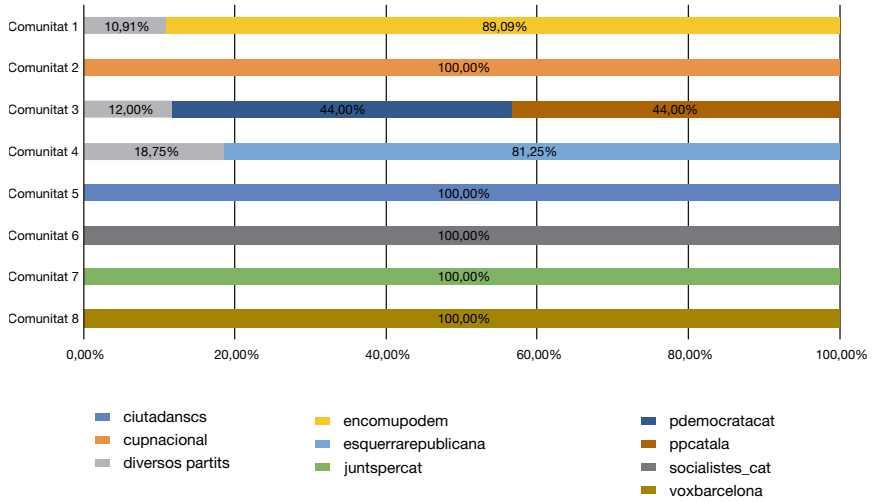


Font: elaboració pròpia a partir de Gephi.

campanya han estat la mateixa cita electoral i els seus debats televisius (#14f, #debattv3, #l6eldebat, #eleccions), alguns elements territorials (#girona, #catalunya, #tarragona i #lleida) i també *hashtags* que fan referència a reivindicacions i lluites polítiques transversals i que ocupen més d'un partit (#feminisme, #antifeixisme i #independència).

S'han detectat en la xarxa de *hashtags* un total de vuit comunitats diferents amb una xifra de modularitat de 0,494 i, per tant, matemàticament significativa. Tal com es pot veure en la gràfica de més avall (figura 6), les comunitats de la xarxa tendeixen a solapar-se amb els perfils dels partits que han elaborat els posts, la qual cosa implica que les comunitats en el graf tendeixen a reflectir els discursos d'un sol partit cada una, i que aquells partits que han fet servir molts *hashtags* són els que tenen la comunitat més gran en la xarxa sintetitzada. Hi ha dos tipus d'excepcions a la tendència anterior. Per una banda, s'observa que les comunitats 1 i 4, protagonitzades per *hashtags* d'ECP i d'ERC, comptenentre un 10 % i un 12 % de *hashtags* que han estat utilitzats per diversos partits (#feminisme ha estat utilitzat per ERC, ECP i la CUP, i #antifeixisme per ECP i la CUP), però que formen part de la comunitat dels comuns o dels republicans, perquè n'han fet un ús més intensiu. Per una altra banda, també hi ha la notable excepció de la comunitat 3, en la qual podem veure *hashtags*

Figura 6. Percentatge de *hashtags* de cada comunitat segons el partit



Font: elaboració pròpia amb MS Excel.

utilitzats pel PP i pel PDeCAT. Això és degut al fet que aquests dos partits van desplegar una estratègia molt semblant a Instagram utilitzant *hashtags* amb un component territorial important (per exemple, #girona, #tarragona) i fent servir molt intensivament el *hashtag* del #debattv3.

El que més crida l'atenció de la xarxa en un primer moment és la mida dels diversos clústers. És molt destacable la diferència en el nombre de *hashtags* utilitzats per partits com la CUP o ECP, per una banda, i Vox, per l'altra. Mentre que els cupaires i els comuns aposten clarament per la generació abundant de discurs i per la pluralitat i la diversitat de *hashtags* —n'arriben a utilitzar fins a 61 o 58, respectivament—, Vox opta per una estratègia que passa exclusivament per dos *hashtags*, que en realitat són complementaris en el seu discurs: #recuperemoscataluña i #recuperemespanya. Així, des del punt de vista de l'anàlisi de *hashtags* que estem duent a terme en aquesta secció, el simplisme del discurs de l'extrema dreta contrasta amb la riquesa del discurs dels partits situats més a l'esquerra de l'espectre polític català. La resta de partits utilitzen xifres de *hashtags* que oscil·len entre els tretze i els setze (PP, PDeCAT i ERC) o entre els set i els nou (Junts i PSC), la qual cosa es pot interpretar com un terme mitjà en la complexitat discursiva palesada en els *hashtags*, i que suggereix també un esforç més gran a acotar el discurs electoral respecte als partits més d'esquerres, sense arribar a l'extrem de simplificació de l'extrema dreta.

4.2. Anàlisi d'una xarxa de paraules

La segona aproximació que proposem per analitzar mitjançant eines d'AXS les dades capturades a Instagram durant la campanya electoral del 14F és

transformar la conversa en una xarxa de paraules. Com en el cas anterior, les transformacions sobre les dades s'han aplicat mitjançant consultes amb llenguatge SQL i amb Pajek.

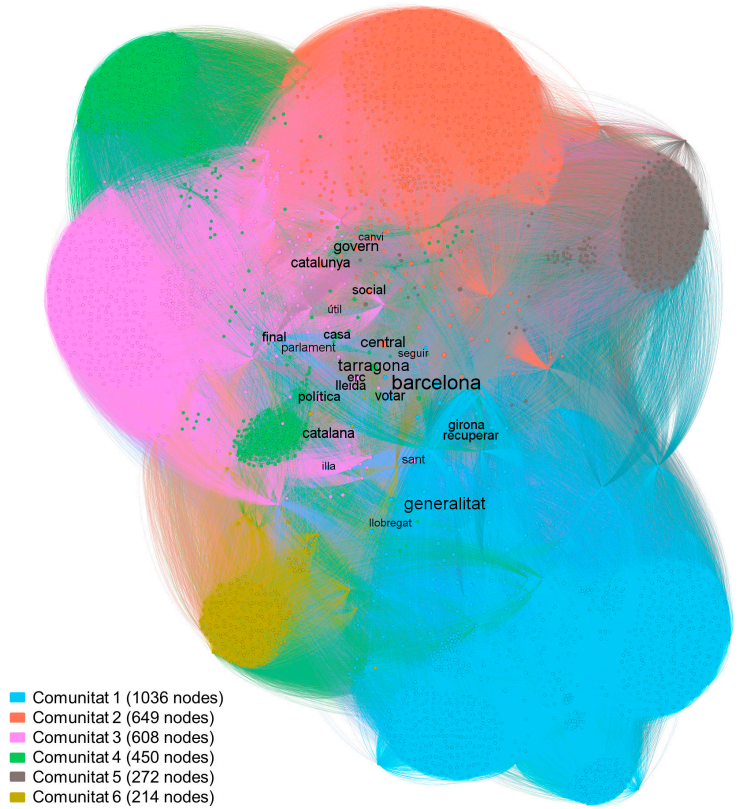
1. S'ha utilitzat la funció *regexp_split_to_table* de PostgreSQL per segmentar les cadenes de text i transformar-les en codis de longitud variables (*varchar*).
2. S'ha utilitzat el comandament SQL *where* per deixar fora de la consulta tots els mots que eren *hashtags*, URL o paraules buides.[8] També s'ha decidit mantenir i no excloure les paraules que comencen amb @ al corpus de text (i. e., les mencions).
3. S'ha transformat la consulta de PostgreSQL en una xarxa de dos modes amb l'ajuda de Txt2Pajek, i després s'ha transformat la mateixa matriu en una xarxa d'un sol mode amb Pajek: una xarxa de paraules vinculades pels autors dels posts amb els vincles ponderats segons el nombre de cops que han aparegut conjuntament.
4. La xarxa s'ha exportat de Pajek i s'ha importat a Gephi. Com en l'exercici anterior, s'ha interpretat com a xarxa no dirigida i ponderada, i s'han aplicat els algoritmes d'identificació comunitària i pel càlcul de la centralitat d'intermediació de cada node.

La xarxa de paraules sintetitzada consta de 3.229 nodes i ni més ni menys que 1.105.098 arestes ponderades. Es tracta, per tant, d'una xarxa força densa en relacions (i. e., el 21,2 % dels vincles possibles es donen en la xarxa). Les paraules amb més centralitat d'intermediació són les que apareixen al centre de la xarxa (figura 7). Paraules com *Barcelona*, *Generalitat* o *social* les han fet servir entre set i nou partits, mentre que altres com *recuperar*, *seguir* o *canvi* les han utilitzat entre cinc i sis partits. Entre les paraules amb més centralitat d'intermediació, no s'observa cap tendència política destacable, tan sols elements territorials, de govern i llocs comuns del discurs polític i electoral, ja que, al cap i a la fi, són les paraules que comparteixen cinc o més dels nou partits. Si aïllem les paraules amb més centralitat d'intermediació que han estat utilitzades per només dos o tres partits, aleshores sí que trobem elements compartits més suggeridors. És el cas de la paraula *economia*, que ha estat utilitzada per ECP i pel PDeCAT, o de la paraula *educació*, feta servir pels dos anteriors i també pel PP. Un altre cas particularment interessant és l'ús molt intensiu de la paraula *YouTube* per part de la CUP i d'ECP, i ocasionalment també pel PP.

Com que els nodes són paraules, les diferents comunitats de la xarxa anterior poden ser interpretades com a unitats de discurs. Es tracta de clústers de paraules utilitzades de manera conjunta pels partits i que bastei-

8. L'exclusió de paraules buides és una tècnica de processament del llenguatge natural que serveix per excloure aquells mots que no aporten informació rellevant en un text. Les paraules buides excloses en aquest treball són les proposades en català, castellà i anglès en el projecte d'Alireza Savand (2014), publicat a GitHub sota llicència Creative Commons 4.0.

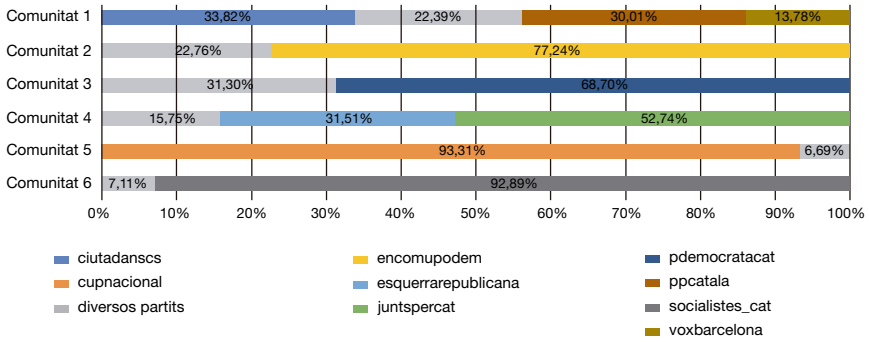
Figura 7. Xarxa de paraules



Font: elaboració pròpia a partir de Gephi.

xen el seu relat en campanya. La partició comunitària en sis clústers que ha estat generada amb l'algoritme Louvain té una xifra de modularitat de 0,475, i és per tant matemàticament significativa. Com en l'exercici anterior, en aquest també resulta molt interessant observar quins partits han publicat les paraules que conformen cada comunitat. Per fer-ho, hem tingut en compte aquelles paraules utilitzades per més d'un partit i aquelles que només han estat usades per un dels nou. Tal com es pot observar a la gràfica de barres agrupades (figura 8), la primera comunitat i la quarta són les úniques conformades per paraules que provenen de més d'un partit: la comunitat 1 la constitueixen les paraules publicades en els posts de Ciutadans, el PP i Vox, i la comunitat 4, les paraules publicades en els posts d'ERC i Junts. Resulta evident que aquests dos grups de partits han articulat discursos relativament semblants o, com a mínim, que pivoten sobre els mateixos eixos. Molt probablement, això és degut al fet que es tracta de partits que competeixen per nínxols molt semblants. La resta de comuni-

Figura 8. Percentatge de paraules de cada comunitat segons el partit



Font: elaboració pròpia amb MS Excel.

tats de la xarxa es nodreixen fonamentalment de les paraules publicades en els posts d'un sol partit: la comunitat 2 amb les paraules d'ECP, la comunitat 3 amb les del PDeCAT, la comunitat 5 amb les de la CUP i la comunitat 6 amb les del PSC. Les dades suggereixen que aquests últims quatre partits han articulat discursos més diferenciats que la resta de partits. Finalment, cal destacar que la presència de paraules compartides en cada clúster oscil·la entre el 6,69 % i el 7,11 % (i. e., els casos de la comunitat 5 i la 6), i el 31,3 % (en el cas de la comunitat 3).

Un últim exercici que proposem per dur a terme en la xarxa de paraules és la generació de taules només amb les 20 paraules i emoticones que han estat utilitzades per un sol partit en cada un dels clústers. Per poder-ne mesurar l'acollida de l'audiència, hem quantificat el seu èxit en funció de la suma de *likes* acumulats en els posts en els quals han aparegut les paraules (taules 1-3). D'aquesta manera, podem veure quines són les estratègies dels diferents partits per marcar un perfil propi. De la taula de la comunitat 1 (Ciutadans, PP i Vox) i de la 4 (ERC i Junts), cal destacar-ne que són comunitats compostes i, per tant, hi trobem la presència prominent dels noms propis de cada candidatura. També la taula de la comunitat 5 (CUP) és plena de noms propis, probablement per emfatitzar el caràcter coral i el paper dels lideratges compartits en l'esquerra independentista. Tornant a les comunitats compostes, com a elements diferenciadors en el seu discurs, destaquen paraules com *barrios*, *carpas* + *informativas* o *simpatizantes* en la comunitat 1, i paraules com *preses*, *exiliats*, *emocionant* o *climàtic* en la comunitat 5. De la resta de comunitats, totes amb paraules d'un sol partit, destaquen mots com *desgovern*, *cures*, *protegir* o *feminisme* (comunitat 2: ECP), *impost*, *lleialtat*, *estabilitat* o *fiscalitat* (comunitat 3: PDeCAT), *solidaritat*, *tendresa*, *lluitant* o *feixista* (comunitat 5: CUP), i finalment paraules com *rescatar*, *tornar*, *reencuentro*, *independentistas* o *mujeres* (comunitat 6: PSC).

Taula 1. Les 20 paraules amb més *likes* acumulats en les comunitats 1 i 2

	Comunitat 1		Comunitat 2	
Rànquing	Paraula	*Likes* acumulats	Paraula	*Likes* acumulats
1	@garriga_ignacio	6.446	@jessicaalbiach	13.029
2	@carrizosacarlos	4.736	bio	7.928
3	@alejandrotgn	4.388	enllaç	6.160
4	sabadell	4.011	@adacolau	4.745
5	@ivanedlm	3.986	comú	4.037
6	@inesarrimadas	3.770	TOP	3.631
7	👏	3.646	web	3.374
8	@santi_abascal	3.401	directe	3.067
9	carpas	3.218	@rosalluchbramon	2.811
10	informativas	3.218	segur	2.805
11	compañeros	3.134	desgovern	2.727
12	garriga	2.708	tenir	2.667
13	@rociomonasteriovox	2.687	@yolanda_carmela	2.664
14	pie	2.540	💙	2.253
15	@vox_es	2.373	@joancgallego	2.132
16	@juangarrigadomenech	2.321	ecp	2.130
17	@monicalora_	2.321	SOON	1.942
18	recuperemos	2.138	protegir	1.929
19	barrios	2.106	🎗	1.773
20	jorge	2.028	@iglesiasturrionpablo	1.730

Font: elaboració pròpia.

Taula 2. Les 20 paraules amb més *likes* acumulats en les comunitats 3 i 4

	Comunitat 3		Comunitat 4	
Rànquing	Paraula	*Likes* acumulats	Paraula	*Likes* acumulats
1	@angelschaconfeixas	12.880	@junqueras	30.083
2	pdecat	7.886	oriol	18.619
3	@artur_mas	4.348	republicana	13.777
4	@joanaortega_a	2.960	@carlespuigdemont	13.681
5	@marcsolsonaaixala	1.705	celebrat	12.480
6	@jaumedulsat	1.670	polítics	11.748
7	@bonvehidavid	1.654	preses	11.511
8	@marcarza	1.504	@forcadellcarme	10.774
9	impost	1.312	@joseprulliandreu	9.367
10	oportunitats	983	l'acte	9.343
11	turisme	974	@jordialapreso	9.343
12	@jordimasquef	873	república	8.581
13	pels	866	laura	8.292
14	intel·ligent	831	borràs	8.292
15	estabilitat	798	💙	7.342
16	gironina	790	vulnerabilitat	7.113
17	votaré	787	fiscalia	6.996
18	@marc_castellsb	774	secretari	6.705
19	pensen	754	comuns	6.477
20	lleialtat	738	tribunal	6.314

Font: elaboració pròpia.

Taula 3. Les 20 paraules amb més *likes* acumulats en les comunitats 5 i 6

	Comunitat 5		Comunitat 6	
Rànquing	Paraula	*Likes* acumulats	Paraula	*Likes* acumulats
1	sabater	12.435	@salvador_illa	2.708
2	fernàndez	11.253	tornar	2.286
3	⏭	10.940	fem-ho	1.514
4	david	9.931	👇	1.470
5	dolors	8.548	mujeres	1.050
6	estrada	8.501	rescatar	993
7	laia	7.179	independentistas	846
8	dani	6.980	reencuentro	778
9	cup-uncpg	6.961	l'hora	747
10	juvillà	5.479	farta	744
11	natàlia	5.479	socialista	733
12	sànchez	5.479	imprescindible	729
13	vehí	5.479	míting	725
14	gabriela	5.479	hagámoslo	692
15	actuacions	5.479	socialista	680
16	ginestà	5.479	progressista	680
17	band	5.479	40000	662
18	llach	5.322	actual	640
19	tendresa	5.161	tiempo	577
20	solidaritat	4.777	protecció	570

Font: elaboració pròpia.

5. Discussió

L'anàlisi implementada en aquest article ha estat feta des d'una lògica inductiva. Les tècniques aplicades són descriptives i exploratòries (per exemple, les visualitzacions de cronologies, de taules de freqüències mitjançant columnes, de proporcions mitjançant barres apilades, les estadístiques descriptives bàsiques o l'anàlisi de correlacions visualitzada amb un núvol de punts), o es tracta d'algoritmes no supervisats que permeten descobrir patrons inicialment desconeguts en les dades (per exemple, l'algoritme k-Means, l'algoritme Louvain o la visualització de xarxes amb l'algoritme Force Atlas 2). En conjunt, es pot argumentar que la nostra perspectiva d'anàlisi és matemàtica —i, per tant, numèrica i quantitativa—, però, alhora, hem utilitzat lògiques i recursos que de manera general s'han aplicat sobretot en la recerca qualitativa, com pot ser l'estratègia inductiva que parteix de l'observació i no d'un seguit de preguntes explícites o d'hipòtesis (tal com passa amb tècniques com l'enquesta o l'experimentació). A més a més, gran part de la informació analitzada i del coneixement generat en aquesta investigació correspon amb coneixements molt més habituals en la recerca purament qualitativa, com és el cas de l'anàlisi de discurs. Per tant, un primer punt a tenir en compte, i que considerem que pot tenir un abast paradigmàtic, és que les perspectives metodològiques híbrides poden ser enormement fecundes en entorns de dades massives. En conseqüència, el tipus de

competències analítiques que requereixen les perspectives empíriques sobre les dades massives són simultàniament matemàtiques i fenomenològiques.

Hem pogut observar que les pràctiques dutes a terme pels diferents partits a Instagram durant la campanya per a les eleccions al Parlament de Catalunya del 14 de febrer són d'allò més diverses. Els dos partits que més han publicat són també els que ho han fet de manera més irregular: el PP i ECP. En canvi, els partits que han rebut més *likes* han estat els principals partits independentistes: ERC, CUP i Junts. Un primer element a constatar és, per tant, la hipermobilització independentista que hem pogut observar a Instagram. Però, si ens fixem en el volum de comentaris, obtindrem un matís diferent: els dos partits que n'han recollit més han estat ECP i ERC. A diferència dels *likes*, que poden interpretar-se com un indicador bastant clar d'adhesió particular o de suport al contingut d'un post, les dades suggereixen que els comentaris podrien contenir matisos importants, probablement derivats del fet que poden ser positius o negatius i poden albergar contingut dialògic o, fins i tot, dialèctic. En aquest sentit, és molt interessant observar que dos dels partits que passada la campanya s'han situat en la centralitat mediàtica i informativa (i. e., ERC, per tenir la clau de volta en la formació del govern, i ECP, per ser una de les dues forces que podrien aliar-se amb ERC i posar fi als pactes de govern en clau independentista)[9] són també els que van acumular més comentaris durant la campanya. Aquest fet suggereix que el volum de comentaris, més enllà de llegir-se com un element d'adhesió o suport, podria estar relacionat amb elements de centralitat política i mediàtica.

Quan mirem les xifres mitjanes de *likes* i comentaris que han rebut els partits catalans durant la campanya del 14F, observem una forta correlació ($r = 0{,}75$) entre la mitjana de *likes* que reben els partits en els seus posts i la mitjana de comentaris, és a dir, aquells partits que acostumen a rebre més *likes* són també els que solen rebre més posts. Mitjançant l'anàlisi k-Means, i amb un valor Silhouette força elevat (0,53), hem pogut diferenciar tres patrons que agrupen els partits en funció de les seves xifres mitjanes de *likes* i comentaris: els partits ECP, PSC i Vox comparteixen un patró de molts comentaris i no tants *likes*; ERC, Junts i CUP es perfilen com a partits amb molts *likes* i molts comentaris, i Cs, el PDeCAT i el PP són les formacions del clúster caracteritzat per pocs *likes* i pocs comentaris. Val a dir que la caracterització anterior només té sentit quan comparem les xifres mitjanes de *likes* i comentaris dels partits entre ells mateixos, i sense tenir en compte factors externs a Instagram com l'èxit electoral.

En aquest article hem volgut anar més enllà d'una caracterització general de les dades i hem articulat dues propostes analítiques fonamentades en l'AXS. Hem sintetitzat dues xarxes diferents —una de *hashtags* i una de paraules—, la qual cosa ens ha permès dur a terme una anàlisi de discurs basada en la coocurrència d'elements: la utilització conjunta de les diferents unitats lèxiques per part dels partits polítics catalans durant la campanya del 14F a les eleccions del Parlament de Catalunya.

9. Vegeu, per exemple, CCMA (2021).

La primera xarxa —de *hashtags*— ens ha permès observar l'escassedat de *hashtags* comuns ideològicament connotats durant tota la campanya. Això vol dir que, més enllà de *hashtags* genèrics o televisius (per exemple, #14f, #debattv3, #l6eldebat) o de *hashtags* que fan referència al territori (per exemple, #tarragona, #girona, #lleida), ens hem trobat amb molts pocs *hashtags* de tipus ideològic (per exemple, #feminisme, #antifeixisme, #independència) que hagin estat utilitzats per més d'un partit polític. Dins de l'escassedat, ECP i ERC són els partits que més capitalitzen aquests *hashtags* comuns, la qual cosa també permet fer-ne una lectura en clau de centralitat política: ERC i ECP són els partits que més parlen d'allò del que també parlen la resta, i així aconsegueixen arrossegar aquests *hashtags* comuns cap als seus clústers. El mateix passa amb el PP i el PDeCAT, però en aquest cas els *hashtags* que aconsegueixen capitalitzar són els de tipus genèric o territorial, la qual cosa suggereix més aviat similitud en l'estratègia —o fins i tot falta d'originalitat— més que no pas centralitat política.

Un darrer element que hem pogut observar en la xarxa de *hashtags* és la complexitat discursiva dels partits més d'esquerres (CUP i ECP han utilitzat prop de 60 *hashtags* diferents cada un), i el simplisme i reduccionisme de l'extrema dreta (Vox n'ha utilitzat només dos de diferents). La majoria de partits se situen en xifres que oscil·len entre els set i els setze *hashtags* diferents, els quals no permeten una lectura nítida en clau esquerra-dreta i més aviat suggereixen un esforç de simplificació del missatge en la comunicació política.

La segona xarxa sintetitzada —la de paraules— resulta complementària respecte a l'anterior pel que fa l'anàlisi de discurs dels partits en campanya. Les característiques formals de la xarxa són força diferents de les anteriors, ja que es mostren molts més nodes i connexions. Un efecte important d'aquesta diferència de mida respecte a la xarxa de *hashtags* és que aquí comptem amb un bon nombre de paraules compartides, mots que han estat utilitzats per més d'un partit o fins i tot per tots. Entre les paraules compartides per cinc o més partits durant la campanya, no s'observen elements gaire interessants, sinó que més aviat són molt previsibles: territorials, de govern o llocs comuns del discurs polític en campanya electoral (per exemple, *Generalitat*, *Barcelona*, *recuperar*, *seguir*, *canvi*). En canvi, si ens fixem en aquelles paraules compartides per només dos o tres partits, sí que ens trobem amb elements d'interès que testimonien els diferents relats dels partits. Es tracta, per exemple, de l'ús de la paraula *economia* per part d'ECP i el PDeCAT, o de l'ús de la paraula *educació* per part dels dos anteriors i pel PP. En aquest sentit, també podem observar que la CUP i ECP mencionen amb certa insistència la paraula *YouTube*, per introduir així la importància del relat transmèdia.

L'anàlisi de clústers com a unitats de discurs permet identificar dos pols multipartit conformats per Cs, PP i Vox, per una banda, i ERC i Junts, per l'altra. Aquests dos grups de partits han mantingut discursos molt semblants entre si o que, com a mínim, han pivotat sobre els mateixos elements lèxics. Els primers destaquen per l'ús de mots típics de les campanyes i en castellà (per

exemple, *barrios*, *plaza*, *carpas informativas*, *compañeros*, *escuchando*), i els segons ho han fet sobre continguts altament emocionals per a l'independentisme (per exemple, *preses i presos polítics*, *surt* [de la presó], *tribunal*, *fiscalia*). La resta de partits, ECP, el PDeCAT, la CUP i el PSC, han mantingut relats més propis, encara que tots han aconseguit arrossegar cap al seu clúster una sèrie de paraules compartides amb els altres. D'ECP es pot destacar l'ús de paraules crítiques amb el govern de la Generalitat i de suport a la sanitat pública o al feminisme (per exemple, *desgovern*, *cures*, *protegir*, *feminisme*); el PDeCAT es distingeix per un discurs altament econòmic i molt poc emocional (per exemple, *impost*, *lleialtat*, *estabilitat*, *fiscalitat*); la CUP capitalitza paraules que tenen a veure amb la tradició d'esquerres més combativa i assenyala l'extrema dreta com a problema greu (per exemple, *solidaritat*, *tendresa*, *lluitant*, *feixista*), i el PSC articula el seu relat amb paraules que apel·len a la concòrdia i al feminisme (per exemple, *rescatar*, *tornar*, *reencuentro*, *mujeres*), però també és destacable que ho fa sense deixar d'assenyalar l'adversari i utilitzant molt la llengua castellana (per exemple, *independentistas*).

6. Conclusions

En aquest article perseguíem dos objectius diferents. El primer consistia a caracteritzar diversos aspectes de la conversa a Instagram relativa al relat electoral dels partits polítics en les eleccions del 14 de febrer de 2021 al Parlament de Catalunya, com la intensitat en l'activitat d'elaboració del discurs o com l'acollida de l'audiència d'aquest discurs, i molt especialment elements relatius al discurs articulat pels partits. L'anàlisi de dades s'ha implementat a través de tècniques d'AXS i d'encreuament de dades. Tal com hem explicat, després de recuperar i emmagatzemar les dades mitjançant tècniques de raspat web aplicades seguint criteris ètics (i. e., considerant qüestions relatives a la propietat intel·lectual, a la sobirania de les dades i al dret a l'anonimat de les persones físiques) i legals (i. e., adaptant-nos a la normativa pròpia d'Instagram i al RGPD), l'estratègia empírica desplegada ha estat de tall exploratori i inductiu. Això vol dir que no hem partit d'un conjunt d'hipòtesis particulars, sinó que hem aplicat una sèrie de proves i algoritmes sobre les dades per identificar-hi patrons ocults a simple vista que ens permeten descobrir determinats aspectes de la conversa digital. Bona part d'aquests algoritmes han estat de tipus no supervisat, com és el cas tant dels algoritmes k-Means com Louvain, que s'han aplicat sobre conjunts de dades sintetitzades a partir de la base de dades original: sobre la mitjana de *likes* i comentaris rebuts per cada partit i sobre dues xarxes de coocurrències lèxiques elaborades a partir de les publicacions dels partits.

El segon objectiu que ens plantejàvem era força més genèric: contribuir a l'elaboració d'estratègies que permetin als investigadors/es socials guanyar autonomia en la generació de coneixement en mitjans socials sovint invisibilitzats per una sèrie de raons que ja hem apuntat (per exemple, l'accés complicat a les seves dades o els propis biaixos dels científics/ques socials). En

aquest sentit, aquest article constitueix un exemple de generació de coneixement en un mitjà social menys explotat que altres —com són Twitter o fins i tot Facebook—, elaborat mitjançant una sèrie de tècniques empíriques sistemàtiques i altament replicables, algunes de les quals ja han estat àmpliament explotades i contrastades en la investigació sociològica dels mitjans socials, com és el cas de l'AXS. En aquest article hem generat i analitzat dues xarxes de coocurrències lèxiques (i. e., una xarxa de *hashtags* i una xarxa de paraules) que hem sintetitzat segons posts d'Instagram, però que podríem haver elaborat segons la conversa capturada en qualsevol mitjà social o plataforma digital: Twitter, Facebook, TikTok, YouTube, mitjans digitals, blogs, fòrums, etcètera. L'elaboració de xarxes d'un mode segons xarxes de dos modes que hem implementat es presenta, per tant, com una estratègia eficient i molt prometedora per l'estructuració de dades en entorns digitals desestructurats o pendents d'estructuració.

En termes generals, considerem que els dos objectius que ens proposàvem han estat assolits, per bé que de manera inevitablement parcial. Des del punt de vista fenomenològic i substantiu, pel que fa a la conversa del 14F, considerem que hem pogut identificar una sèrie d'elements altament suggeridors que podrien fàcilment convertir-se en hipòtesis, tant en clau de política catalana com en termes més generals. Ens referim, per exemple, a la relació entre variables com el nombre de comentaris i la centralitat política, o com el nombre de *hashtags* i paraules comunes capitalitzades i la centralitat política, o fins i tot, a la relació entre ideologia i complexitat o simplicitat discursiva en els mitjans socials.

Des d'un punt de vista metodològic, atenent al segon objectiu d'aquest article i a la seva vocació d'obrir nous escenaris analítics, no podem deixar d'assenyalar que les tècniques implementades per a l'anàlisi de la conversa d'Instagram són una part molt petita dins de l'univers de possibilitats que ofereixen les dades massives i l'AXS. Entre els tipus alternatius de xarxes que es podrien haver sintetitzat amb les dades recuperades d'Instagram, cal destacar les xarxes d'autors, les xarxes d'ubicacions, les xarxes de temes de conversa o les xarxes d'elements gràfics a les imatges detectats amb intel·ligència artificial. Per altra banda, també és important tenir en compte altres xarxes que es poden elaborar a partir d'alguns elements de metadades a les quals es pot accedir mitjançant tècniques de raspat web a Instagram. Un usuari no identificat (i. e., que no ha fet *log-in*) pot obtenir els comentaris fets a un post, i aconseguir així les dades necessàries per elaborar diversos tipus de xarxes (com ara xarxes de comentadors, de posts, de comentaris), i un usuari identificat podria obtenir fins i tot la informació relativa a qui ha fet *like*. Tanmateix, pel que fa a aquesta darrera proposta, l'analista hauria de comptar amb el consentiment explícit d'Instagram per evitar caure en contradicció amb les condicions d'ús de la plataforma.

Finalment, volem introduir una reflexió que té a veure amb les tècniques presentades en aquest article, així com amb l'estratègia analítica que hi hem desplegat. Creiem que a hores d'ara ja hi ha pocs dubtes que les diferències

entre l'escenari de dades dels segles XIX i XX que va donar lloc a la ciència social —o a les ciències socials, segons prefereixi el lector o lectora— i l'escenari actual són molt notables. Avui habitem un *datascape* diferent del de fa pocs anys, la qual cosa, com és lògic, genera una sèrie de reptes. L'accés a grans volums de dades estructurades, semiestructurades o desestructurades no solament té efectes quantitatius i de volum sobre els conjunts de dades amb què treballem els/les científics/ques socials, sinó que altera el nostre rol tradicional en la generació i la gestió de coneixement social. Amb l'arribada i l'expansió del paradigma de les dades massives, la recerca social basada en dades secundàries adquireix més i més importància, i els/les científics/ques socials cada cop tenim un rol menys actiu en l'elaboració de les eines de captura d'informació i de definició de la realitat social. En altres paraules, avui dissenyem menys qüestionaris i guions d'entrevista que fa vint anys i, en canvi, invertim més temps a interpretar i resignificar eines empíriques dissenyades per tercers, com són les bases de dades massives. Sense anar més lluny, aquest article és un exemple de reinterpretació i resignificació d'un conjunt de dades semiestructurades dissenyat per un equip d'informàtics/ques (i. e., el personal d'Instagram) que probablement mai no van pretendre construir una eina que permetés analitzar la política catalana.

En congruència amb la reflexió anterior, creiem que és urgent planificar una sèrie de transformacions paradigmàtiques en la ciència social. Considerem que cal que donem més importància a l'adquisició d'habilitats que permetin transformar conjunts de dades i hackejar-los, en el sentit de saber-los utilitzar per generar coneixement social més enllà de les seves funcions assignades a priori (i. e., funcions que per regla general haurà dissenyat i implementat un equip de desenvolupadors/es informàtics totalment aliè a la teoria social i a les nostres necessitats empíriques). Creiem que és important entrenar les noves generacions de científics/ques socials perquè mirin més enllà dels conjunts de dades tal com es presenten defugint, per tant, qualsevol fetitxisme o reïficació sobre les dades i articulant així la creativitat i la imaginació sociològica al segle XXI. Per assolir aquest objectiu, tal com ho veiem nosaltres, és important avançar en dos aspectes complementaris. Per una banda, s'ha d'aprofundir en l'apropament a les epistemologies i les tècniques pròpies de la ciència computacional. Això vol dir que ens cal ampliar la caixa d'eines de la ciència social amb elements fonamentals de la gestió de bases de dades, la mineria de dades i la intel·ligència artificial. Per l'altra, també s'ha d'avançar en l'elaboració de models metodològics híbrids que vagin més enllà de plantejaments mixtos i que contribueixin a erosionar la divisió artificial entre les tècniques quantitatives i les qualitatives, separació que amb els anys hem anat creant i que ha estat útil en molts aspectes, però que actualment cada cop té menys sentit, ja que sospitem que pot produir un efecte més aviat limitador sobre la capacitat dels científics/ques socials a l'hora d'imaginar i projectar les seves recerques.

Referències bibliogràfiques

ANDERSON, Chris (2008). «The end of theory: The data deluge makes the scientific method obsolete». *Wired Magazine*, 16 (7).
<https://www.wired.com/2008/06/pb-theory/>, vist el 18 de febrer de 2021.

APACHE NIFI (2018). «NiFi Developer's Guide».
<http://nifi.apache.org/developer-guide.html>, vist el 18 de febrer de 2021.

BASTIAN, Mathieu; HEYMANN, Sebastien i JACOMY, Mathieu (2009). *Gephi: an open source software for exploring and manipulating networks.* International AAAI Conference on Weblogs and Social Media.
<https://doi.org/10.13140/2.1.1341.1520>

BLONDEL, Vincent D.; GUILLAUME, Jean-Loup; LAMBIOTTE, Renaud i LEFEBVRE, Etienne (2008). «Fast unfolding of communities in large networks». *Journal of Statistical Mechanics: Theory and Experiment*, P10008.
<https://doi.org/10.1088/1742-5468/2008/10/P10008>

BRANDES, Ulrik (2001). «A Faster Algorithm for Betweenness Centrality». *Journal of Mathematical Sociology*, 25 (2), 163-177.
<https://doi.org/10.1080/0022250X.2001.9990249>

BRUNS, Axel (2019). «After the "APIcalypse": social media platforms and their fight against critical scholarly research». *Information, Communication & Society*, 22 (11), 1544-1566.
<https://doi.org/10.1080/1369118X.2019.1637447>

CCMA (2021). «Esquerra confia en l'acord ampli després de reunir-se amb Comuns». <https://www.ccma.cat/tv3/alacarta/telenoticies/esquerra-confia-en-lacord-ampli-despres-de-reunir-se-amb-comuns/video/6085917/>, vist el 2 de març de 2021.

DEMETZOU, Katerina (2019). «Data Protection Impact Assessment: A tool for accountability and the unclarified concept of "high risk" in the General Data Protection Regulation». *Computer Law & Security Review*, 35 (6), 105.342.
<https://doi.org/10.1016/j.clsr.2019.105342>

DEMSAR, Janez; CURK, Tomaz; ERJAVEC, Ales; GORUP, Crt; HOCEVAR, Tomaz; MILUTINOVIC, Mitar; MOZINA, Martin; POLAJNAR, Matija; TOPLAK, Marko; STARIC, Anze; STAJDOHAR, Miha; UMEK, Lan; ZAGAR, Lan; ZBONTAR, Jure; ZITNIK, Marinka i ZUPAN, Blaz (2013). «Orange: Data Mining Toolbox in Python». *Journal of Machine Learning Research*, 14 (1), 2.349-2.353.
<https://dl.acm.org/doi/10.5555/2567709.2567736>

DIARIO OFICIAL DE LA UNIÓN EUROPEA (2016). «Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento General de Protección de Datos)».
<https://www.boe.es/doue/2016/119/L00001-00088.pdf>, vist el 18 de febrer de 2021.

FREEMAN, Linton C. (1977). «A set of measures of centrality based on betweenness». *Sociometry*, 40 (1), 35-41.
<https://doi.org/10.2307/3033543>

FUCHS, Christian (2017). «From digital positivism and administrative big data analytics towards critical digital and social media research!». *European Journal of Communication*, 32 (1), 37-49.
<https://doi.org/10.1177/0267323116682804>

GAYO-AVELLO, Daniel (2013). «A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data». *Social Science Computer Review*, 31 (6), 649-679. <https://doi.org/10.1177/0894439313493979>

GRINBERG, Nir; KENNETH, Joseph; FRIEDLAND, Lisa; SWIRE-THOMPSON, Briony i LAZER, David (2019). «Fake News on Twitter during the 2016 U.S. Presidential Election». *Science*, 363 (6.425), 374-378. <https://doi.org/10.1126/science.aau2706>

HIGHFIELD, Tim i LEAVER, Tama (2016). «Instagrammatics and digital methods: studying visual social media, from selfies and GIFs to memes and emoji». *Communication Research and Practice*, 2 (1), 47-62. <https://doi.org/10.1080/22041451.2016.1155332>

LAMBIOTTE, Renauld; DELVENNE, Jean-Charles; BARAHONA, Mauricio (2009). «Laplacian Dynamics and Multiscale Modular Structure in Networks». *IEEE Transactions on Network Science and Engineering*, 1 (2), 76-90. <https://doi.org/10.1109/TNSE.2015.2391998>

LANDERS, Richard N.; BRUSSO, Robert C.; CAVANAUGH, Katelyn J. i COLLMUS, Andrew B. (2016). «A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research». *Psychological Methods*, 21 (4), 475-492. <https://doi.org/10.1037/met0000081>

LOZARES, Carlos (1996). «La teoría de redes sociales». *Papers: Revista de Sociologia*, 48, 103-126. <https://doi.org/10.5565/rev/papers/v48n0.1814>

MASIP, Pere; RUIZ-CABALLERO, Carlos; SUAU, Jaume (2019). «Active audiences and social discussion on the digital public sphere. Review article». *El Profesional de la Información*, 28 (2), e280.204. <https://doi.org//10.3145/epi.2019.mar.04>

MATAMOROS-FERNÁNDEZ, Ariadna i FARKAS, Johan (2021). «Racism, Hate Speech, and Social Media: A Systematic Review and Critique». *Television & New Media*, 22 (2), 205-224. <https://doi.org/10.1177/1527476420982230>

METZGAR, Emily i MARUGGI, Albert (2009). «Social Media and the 2008 U.S. Presidential Election». *Journal of New Communications Research*, 4 (1), 141-65.

MORALES-I-GRAS, Jordi (2020). «Cognitive Biases in Link Sharing Behavior and How to Get Rid of Them: Evidence from the 2019 Spanish General Election Twitter Conversation». *Social Media + Society*, 6 (2), 1-4. <https://doi.org/10.1177/2056305120928458>

MOSCO, Vincent (2014). *To the Cloud: Big Data in a Turbulent World*. Boulder (CO): Paradigm Publishers.

NEWMAN, Mark E. (2006). «Modularity and community structure in networks». *Proceedings of the National Academy of Sciences*, 103 (23), 8.577-8.582. <https://doi.org/10.1073/pnas.0601602103>

OXFORD DICTIONARIES (2016). «Word of the year 2016». <https://languages.oup.com/word-of-the-year/2016/>, vist el 4 de febrer de 2021.

PFEFFER, Juergen; MRVAR, Andrej i BATAGELJ, Vladimir (2013). «txt2pajek: Creating Pajek Files from Text Files». *Technical Report, CMU-ISR-13-110, Carnegie Mellon University, School of Computer Science, Institute for Software Research*. <http://www.pfeffer.at/papers/2015_txt2pajek.pdf>

POSTGRESQL (2021). «PostgreSQL 13.2 Documentation». <https://www.postgresql.org/docs/13/index.html>, vist el 18 de febrer de 2021.

SAVAND, Alireza (2014). «Stop-Words».
<https://github.com/Alir3z4/stop-words>, vist el 18 de febrer de 2021.
THE SOCIAL MEDIA FAMILY (2020). «Informe de los perfiles en redes sociales de España».
<https://thesocialmediafamily.com/informe-redes-sociales/>, vist el 18 de febrer de 2021.
VELTRI, Giuseppe A. (2019). *Digital social research.* Cambridge: Polity Press, John Wiley & Sons.
WE ARE SOCIAL (2020). «Digital 2020 España».
<https://wearesocial.com/es/digital-2020-espana>, vist el 18 de febrer de 2021.