Questions de communication

46 | 2024
Le fandom, côté obscur

Citizen-produced political text: An interdisciplinary study of inequalities in research

Les textes politiques produit par les citoyens (CPPT). Une étude interdisciplinaire des inégalités dans la recherche

Amanda Haraldsson, Shota Gelovani, Michele Scotto di Vettimo, Bente Kalsnes and Karolina Koc-Michalska

Electronic version
URL: https://journals.openedition.org/questionsdecommunication/36433
DOI: 10.4000/12yfj
ISSN: 2259-8901

Publisher
Presses universitaires de Lorraine

Electronic reference
Amanda Haraldsson, Shota Gelovani, Michele Scotto di Vettimo, Bente Kalsnes and Karolina Koc-Michalska, "Citizen-produced political text: An interdisciplinary study of inequalities in research", *Questions de communication* [Online], 46 | 2024, Online since 29 November 2024, connection on 19 December 2024. URL: http://journals.openedition.org/questionsdecommunication/36433 ; DOI: https://doi.org/10.4000/12yfj

Citizen-produced political text: An interdisciplinary study of inequalities in research

Les textes politiques produit par les citoyens (CPPT). Une étude interdisciplinaire des inégalités dans la recherche

Amanda Haraldsson, Shota Gelovani, Michele Scotto di Vettimo, Bente Kalsnes and Karolina Koc-Michalska

1 Research possibilities utilising text data are expanding (Baden *et al.*, 2022), and political text produced by citizens is among the most commonly employed text data within political and communication science. We[1] understand citizen-produced political text (CPPT) broadly as a text produced by citizens, offline or online, in a written or retranscripted form, relating to a political process, policy or civic issues (Gelovani *et al.*, 2022). Research questions addressed through such data are diverse, ranging from the analyses of millions of search engine queries of netizens relating to anti-corruption campaigns in China (Zhu and Wang, 2020) to more qualitative examinations of two dozen in-depth interviews with leftist activists in Texas (USA) to understand their internal debates when disagreeing (Venegas, 2022). CPPT studies allow to focus on citizens' views, opinions, and experiences rather than elites' and provide invaluable insights into political communication research.

2 Despite the richness and importance, utilising CPPT presents several interconnected challenges. Research tools and methodologies, developed primarily for English-language text, have limited applicability to non-English text (Baden *et al.*, 2022). Researchers outside non-Western regions struggle with collaboration opportunities (Matassi and Boczkowski, 2021). Access to comprehensive social media data – the most common CPPT data source – is a privilege of a few research teams collaborating with social media companies (Lazer *et al.*, 2020). This paper aims to assess the challenges and inequalities faced by those interested in employing CPPT, regardless of the method or source of data. It identifies challenges researchers experience from accessing data to publications and dissemination, but it excludes ethical challenges that have been

extensively explored elsewhere (Sormanen and Lauk, 2016). We approach the problem from a generalisable and globalised perspective to ensure a cross-national understanding.

Research background

3 Studying citizens' use of political text is essential to understanding their political positions, participation and agenda-setting. For example, Sebastian Stier and colleagues (2018) identified topics citizens discussed in open-ended responses to a survey during the 2013 German federal election to assess how well candidates reflected these issues. Commenting on news articles (Erzikova and Simpson, 2017), microblogging (Fu and Chau, 2014) or replying to politicians' tweets (Londoño, 2019) are means for citizens to engage in political conversations. Such forms of primarily social media participation are becoming more commonplace and growing in variety (Vaccari and Valeriani, 2021).

4 Given the range of substantive questions addressed through studying text produced by the citizens and the diversity of disciplines involved (Giglietto *et al.*, 2012), the field employs a variety of methodological approaches. Not least in volume – studies using Twitter range from critical discourse analysis of 100 tweets (Kreis, 2017) to social network analysis of 20 million tweets (Deltell *et al.*, 2013). Furthermore, a comprehensive review of more than three thousand manuscripts studying CPPT (Gelovani *et al.*, 2022) indicates that half employ qualitative methods, and a quarter utilise computational or quantitative methods. During the period from 2014 to 2022, social media platforms emerged as the primary, but not exclusive, source of data in these studies.

5 The utilisation of big or digital data and computational methods in social sciences have already prompted reflections on how such methods can be applied by social scientists (*e. g.* boyd and Crawford, 2012; Van Atteveldt and Peng, 2018; Theocharis and Jungherr, 2021; Baden *et al.*, 2022). There is a growing need for comparative studies across platforms, countries, and time (Larsson, 2015). It is also crucial for researchers to enhance their understanding of how platform policy and algorithm modifications, as well as users' engagement patterns of these platforms, might impact the research approach and outcomes (Karpf, 2020).

6 Researchers have identified a number of impediments within the research on CPPT. One such challenge pertains to data accessibility, especially from social media platforms, particularly in the aftermath of the Cambridge Analytica controversy, also called the "APIcalypse" (Bruns, 2019). Facebook drastically limited access to platform data via their Application Programming Interfaces (APIs) and Twitter closed completely free-accessed data in mid-2023. Furthermore, disparities emerge among researchers themselves. Significant variations occur in how data is utilised by those engaged, for example, in ethnographic research in comparison to computational research, despite using the same sources of CPPT data (Giglietto *et al.*, 2012). These differences are often reflected in a lack of reciprocal understanding and appreciation for each other's work among researchers. Additionally, biases have been identified, where priorities are given to well-established subjects, Western research sites and English-language materials (Baden *et al.*, 2022). Bias in research also emerges because the "Global North" perspective dominates, crowding out other perspectives not seen as the norm

(Wasserman, 2020). To gain a deeper understanding of and address the challenges encountered by researchers interested in studying CPPT data, we ask:

1. What challenges confront CPPT researchers or researchers interested in using CPPT data relating to the themes of accessibility, languages and resources?
2. In what way do these challenges vary across researchers due to career stage, country of affiliation, language of interest, or other personal factors?

Methods

7 The manuscript builds upon two methodological approaches. First, we run a non-representative survey among researchers from diverse disciplines, mainly political science, communication, methods, or linguistic studies. Subsequently, we perform a series of in-depth interviews with a purposefully selected sample of respondents. The repartitions and characteristics of respondents can be found in Tables A1 and A2 of the Supplementary materials.

Survey

8 The first data source employed in this study is an online survey on a non-representative but strategically recruited sample of researchers actively engaged in text-based research or being interested in working with text, especially CPPT data. The questionnaire asked them about their background, use of text data, challenges faced when using such data, etc. Respondents were recruited through a multifaceted approach, including (i) lists of contacts of authors employing CPPT data in their research identified via extensive literature review (Gelovani *et al.*, 2022); (ii) contacts within the network of OPTED project; (iii) mailing lists of organisers of text analysis related events; (iv) online advertisement by OPTED. The survey received ethics approval from the Research Ethics committee of the University of Exeter. A total of 295 responses were collected from February to August 2022 (Table A1). Respondents could skip questions (total N per question is specified in the results section) or provide additional answers in the text box. Based on their reported usage of CPPT, respondents were categorised into three groups: those who do not use CPPT (N=73; not employed in this manuscript), those interested in using it in future (N=59), and those currently using CPPT (N=163).

Interviews

9 A second source of data consists of expert interviews. A purposeful sampling strategy (Palinkas *et al.*, 2015; Patton, 2015) was employed to identify relevant experts and CPPT researchers. Our goal was to achieve a diverse sample encompassing for discipline, methods, language(s) studied, language of publications, career stage, regional variation in institutional affiliation, and country of origin. Additionally, we looked for respondents with experience in CPPT research gained through journal editing, collaboration with data companies, or involvement in research integrity and ethics. We contacted authors who had published CPPT studies and experts identified *via* workshops, conferences or courses relating to text analysis, or editors and ethical review board members. Twenty-one interviews were conducted in May-July 2022 (Table A2). Interviewees were assured of their anonymity, asked for permission to audio-

record and transcribe interviews, and informed of the interview's purpose in a consent form signed beforehand. A semi-structured questionnaire was developed, with core questions for all interviewees and specific questions tailored to each individual. Interviews ranged from 35 to 90 minutes and were conducted online. Ethical approval for the study was received from the Research Ethics committee of Audencia Business School. Interviewees are given a random identification number when referenced[2]. The results of the survey were helpful in preparing and conducting the interviews. As the survey was anonymous, the in-depth interviews were not conducted as a direct follow-up. The two datasets are interrelated by the researchers' shared interest.

Results

Access challenges

10 Survey respondents were asked about the challenges they encountered when accessing CPPT data. Table 1 indicates the most significant access-related issues perceived as difficulties by these users.

Table 1. Access issues for survey respondents by CPPT use (N=201).

Challenge	Interested (N=56)	Users (N=146)
Company restriction	44	135
Content removal	34	112
Finding tools	33	95
Finding text	33	91
Ethical restrictions	24	80

Company restrictions

11 Social media companies implement access restrictions through various means. Firstly, they do not allow researchers to access *all* data comprehensively. This is also confirmed in the in-depth interviews, as restrictions by companies often leave interviewees with what they consider suboptimal and less representative data sources. Consequently, the potentially more valuable data remains inaccessible.

> "Twitter[3] is very infrequently the best solution for a problem. It is just usually the easiest solution for a problem." (Interviewee_5, OPTED, interview with Authors July 2022).

12 The further development of API web services is deemed crucial for 68% of survey respondents. Specifically:

> "With Twitter' API's, there's a limit for how many inquiries you can make. And also there's a limit for how far you can go back. [...] you get wildly different results if you do research today, and then three years ago, and you have no ways to verify where [the difference] is coming from" (Interviewee_8, OPTED, interview with Authors June 2022).

13 Secondly, the evolving (not in a favourable direction) nature of the restrictions in data access by social media corporations presents a significant challenge. Not only does this restrict what interviewees can currently study because of data becoming unavailable, but it also makes it impossible to access historical data or access only part of a sample of data.

> "Constantly what you can scrape, what you cannot scrape, changes and – may I say – with zero regard to researchers. Like they do not consider researchers when putting through these guidelines. They consider industry people." (Interviewee_4, OPTED, interview with Authors June 2022).
> "When I started you could download everything. Everything was scrapable [...] but now we can't have access to the data, so we have to orientate the research with stuff that we can do" (Interviewee_16, OPTED, interview with Authors May 2022).

14 Interviewees are concerned about how access will change in the future and impact ongoing projects. Such uncertainty and the power imbalance between researchers and social media corporations lead to high levels of distrust:

> "Facebook used to be open. It isn't anymore. You can get to it through a service called CrowdTangle, which is owned by Meta. But a lot of the people who worked on CrowdTangle have since been placed in other services within Meta, or some have even quit. So we don't really know what's going to happen with CrowdTangle" (Interviewee_6, OPTED, interview with Authors June 2022).
> "what we've seen, Twitter opening up the API for researchers a few months ago, I think that was a very big thing. I'm not necessarily sure how long that stays, because, well, it's always at the whim of the company." (Interviewee_14, OPTED, interview with Authors May 2022)[4]

15 Thirdly, skills and time requirements to access data continue to be high barriers for entry. Several interviewees expressed frustration that the data-gathering process has become more strenuous than necessary due to how companies (re)structure APIs and the lack of information provided.

Content removal

16 Content removal creates two primary challenges for researchers: it complicates the replicability of analyses and limits the ability to study communication that is most likely to be deleted, such as hate speech, conspiracy theories, anti-government rhetoric, etc. Interviewees experience that posts are removed, made private or edited by the person/page who created the post or because the user deleted the account. Additionally, content can be removed when a page, user, or group is blocked. In certain cases, systemic content removal occurs when, for example, powerful actors censor online discussion on a specific topic.

17 Regardless of the underlying reasons for content removal, there is a shared frustration among researchers regarding its impact on the ability to study certain phenomena where deletion is likely.

> "Facebook pulls everything that it finds offensive. Then you have the user review process that might pull the next thing and then people pull their own things or edit them later. And yes, I know that you can technically get to the pre-edit version. But let's face it, who does that?" (Interviewee_5, OPTED, interview with Authors July 2022).
> "I cannot possibly be on top of everything at all times. Things may get deleted, I cannot check many thousands of tweets whether or not they got deleted constantly." (Interviewee_4, OPTED, interview with Authors June 2022).

18 The techniques vary among researchers when dealing with deleted social media posts. Some believe that as long as they only analyse posts that were available at the time of data collection[5] and anonymise them, they are in line with the research standards. Indeed, depending on the source of data, this method can sometimes be the only possible solution when using large datasets. Other interviewees explain the process of "re-hydrating"[6] social media posts: during data collection, posts' IDs, rather than the text, should be archived, and then these IDs can be retrieved within future analysis, ensuring that only currently available posts are included. Yet, even interviewees who are familiar with re-hydrating state that they currently do not engage in this practice or are not sure how it works. The inability to provide complete datasets for replication is a systemic issue, as evidenced in the meta-study, which indicates that only 1.3% of studies provide access to datasets (Gelovani *et al*., 2022).

Finding (and using) tools

19 Regarding tools utilised for accessing CPPT, a primary concern revolves around substantial time commitment to master the usage of software, codes, and applications. Moreover, due to the relative novelty of many of those techniques, it is difficult to find sufficient guidelines, often resulting in mistakes that lengthen data collection. One interviewee indicated such constraints:

> "I was crawling their website and my crawler was getting that data at a very huge speed, [...] so when I was crawling this data the guys from this company reached out to [interviewee's university] and they blocked my IP for a week or so. And they were like 'OK. You're not going to do anymore crawling because they're very mad at us'. What were we doing? We're just jamming their servers!" (Interviewee_18, OPTED, interview with Authors July 2022).

20 A persistent challenge lies in the misusage of tools designed for other purposes (such as marketing research) that often do not result in the type of data needed for the CPPT researcher in social sciences. One interviewee expressed a feeling of greater trust in tools developed by discipline-specific researchers, such as Netvizz[7]. Moreover, tools trained on other text types often perform worse when applied to CPPT-related tasks.

> "I think given the structure of this text, it's not that straightforward. Some of the models cannot be applied or perform very differently with these short texts." (Interviewee_21, OPTED, interview with Authors June 2022).

21 Despite the availability of tools that align with researchers' needs, their visibility is often limited. Researchers frequently rely on word-of-mouth to learn about the existing tools rather than having a reliable place to browse alternatives. This lack of awareness results in sometimes choosing suboptimal methods for the question at hand.

Language or regional differences

22 The survey findings (Table 2) indicate that researchers were more likely to study text in multiple languages, particularly when English was included.

Table 2. Languages studied (survey respondents N=196).

Languages	Interested (N=55)	Users (N=141)
Multiple incl. English	43	105

Only English	2	22
Only non-English	5	10
Multiple not incl. English	5	4

Data availability

23 The ability to access text was a major factor (42%) among survey respondents for choosing which languages to study. Also, the interviewees indicate that access to text varies across languages and regions. The reasons for that are diversified, for example, in certain countries easy-to-study platforms have greater usership or specific regional restrictions exist.

> "I see that there are some papers coming out like using Twitter data and different kinds of other forums where they have analysed like millions of tweets [...] In the [small country in Europe] context, to find enough data to do it well? I don't think it's possible" (Interviewee_3, OPTED, interview with Authors May 2022)
> "The problem is that in the North African context, there are no archives. [...] the web pages are not stable. And that applies to citizen-produced text, which generally has less durability than print media or audio-visual materials." (Interviewee_2, OPTED, interview with Authors July 2022).

24 Social media platforms are also used differently across communities. For specific topics, open discussions may be prevalent on some platforms rather than on others due to populations inhabiting it or restrictions in content supervision:

> "People from my own country, [country in Asia] and in my part of the world. They would not be that vocal on social media [...] even if they are vocal, they won't talk much about these topics that we are interested in" (Interviewee_18, OPTED, interview with Authors July 2022).

Methodological possibilities

25 Regional differences in methodological training was also a theme discussed by interviewees:

> "We still have a problem here in [country in South America], working especially in human sciences, in working with big data. We are not trained to do that" (Interviewee_11, OPTED, interview with Authors June 2022).

26 Generally, the perception is that computational tools work best for English, and the more a language diverges from English, the more likely it is that tools will be unavailable or require significant work to apply. Among survey respondents employing computational methods (N=142), the availability of suitable tools is considered a major problem, however, for "only" 26% of those studying English text but for 45% of those studying the non-English text. Factors such as accent handling, languages using specific symbols, or the scarcity of validated stop-word lists, lead researchers to either abandon certain languages or methods altogether.

> "I had quite some students that were using Arabic text, [...] there we had also some encoding things just to start with, you know, to get the texts properly into R and work with them when they have very different letters etc. [...] they often have different encodings on their computers so when they read even a proper Arabic text into R, for instance, they transfer to very weird systems that they are not aware of." (Interviewee_21, OPTED, interview with Authors June 2022).

Translation

27 A concern almost universally expressed among interviewees is the difficulty in analysing and interpreting citizens' short, unstructured, and less strategic text. This issue becomes more entrenched when translation is required, also limiting the collaboration opportunities for research on languages that are less commonly studied:

> "It's extremely easy to find people who will code specific languages, or content in specific languages. It's much more difficult to find multiple people in [other] languages in comparison." (Interviewee_17, OPTED, interview with Authors May 2022).

28 Moreover, researchers' employment of the terminology may not fit the reality of citizens in the regions they study. For example:

> "*Arab Spring* has been standard in Western and English language bibliographies and library searches and even keywords in general and so on. But it is a colonial term that has been externally imposed on the people in the region, and people in the region never use the "Arab Spring" as a term" (Interviewee_2, OPTED, interview with Authors July 2022).

Resources

Collaboration opportunities

29 Interviewees in departments with more extensive linking among disciplines highlighted the benefits of collaboration concerning methodological approaches they were able to employ. Access to data is also impacted by networking possibilities.

30 CPPT frequently requires diverse skillsets from multiple disciplines, languages, and methodological experiences, not to mention that time-consuming or costly research can be alleviated when collaborating. However, a recurring pattern emerged from interviews with more junior researchers, who were substantially less likely to collaborate compared with more career-advanced scholars. Moreover, some interviewees experienced exclusion from international networks due to their institution's location or a lack of international openness within their institutions:

> "Sometimes I think if I would have had this network, maybe the problems that I faced in gathering data during my PhD would not have been so pronounced [...] I'm a migrant. When migrants are new to a place, they don't know their ways around things" (Interviewee_18, OPTED, interview with Authors July 2022).
> "Opening science, decentralising science, is also about establishing connections with our peers in other places" (Interviewee_15, OPTED, interview with Authors June 2022).

English bias

31 A bias in favour of the English language is common throughout many aspects of CPPT research. This bias manifests itself in three primary ways.

32 First, non-English-speaking countries are often treated as case studies, while English-speaking countries are considered as reflecting the world. This dynamic was observed by interviewees studying European countries but it was even more pronounced in non-European contexts.

33 Second, publishing in English journals is frequently perceived more favourably by institutions and it often facilitates the wider dissemination of the research.

> "It's kind of a bet. 'Oh this I would save to publish in English', and then the chances of being refused by the publisher, by the journal, are higher." (Interviewee_11, OPTED, interview with Authors June 2022).

34 Third, non-native English speakers may face doubts about their ability to work with CPPT material in English. This scepticism is not as prevalent for English-speaking researchers working with non-English material.

Method bias

35 In a field with such variety of research methods, many interviewees perceived a bias toward quantitative methods as being more valuable or rigorous than qualitative approaches. This perception persists despite the evidence from the CPPT review indicating that qualitative methods are the most commonly used (Gelovani *et al*. 2022). Additionally, interviewees expressed concern that the methodological choices of peers are often not easy to understand.

Structural support

36 Finally, the respondents highlighted systemic and institutional inequalities in research guidance-setting process. As research progresses and stricter rules become necessary, these inequalities could potentially intensify over time.

> "To me it seems that it is fairly easy to incorporate and amplify inequalities between places that have resources, and places that do not have resources. Because if you want to be GDPR compliant, there will be more steps to carry out. If there is institutional support, mostly legal support, then that will be done in a much better way. But not all institutions can provide that, [...] and this obviously might lead to some people not embarking on specific research questions, because there are six more steps that need to be carried out." (Interviewee_17, OPTED, interview with Authors May 2022).

Conclusions

37 Our findings point to several trends: 1) some CPPT researchers avoid research questions they are interested in, where data accessibility is perceived as too difficult to navigate; 2) additional resources are essential for specific CPPT projects, including studies of non-English text, but resources are unevenly distributed; 3) inequalities remain in the ability to engage in CPPT research across institutions, regions, and scholars at various career levels.

38 The scarcity of collaboration opportunities impacts researchers' ability to access data, particularly for those working outside of North America and Western/Northern Europe. Resources for translation may be more difficult to acquire when studying languages spoken in some regions of the world than others, making data access more complicated. To address these challenges, a specialised infrastructure is needed to bring together researchers from different disciplines, competencies, and regions to facilitate the exchange of data, context knowledge, and language skills and to recognise tools for data gathering and analysing (Balluff, 2023). Yet such initiatives are scarce and expensive, thus they need substantial international institutional support[8].

39 Perhaps the greatest challenge in studying CPPT is the limited access to text data produced by citizens, specifically due to restrictions imposed by social media

corporations (De Vreese and Tromble, 2023). This limited access is further compounded by the fact that these restrictions are subject to change over time, often without much warning (Bruns, 2019; Venturini and Rogers, 2019; Perriam *et al.*, 2020). Given the reliance on social media data among CPPT researchers, a significant limitation is that individuals who post on social media platforms are not representative of the general population (Hargittai, 2020). There is both a performative aspect to online discussion, and an algorithmic impact that could affect how accurately observed online behavior reflects the theoretical concepts researchers are interested in (Lazer *et al.*, 2021). Despite the challenges, publicly available social media posts continue to be a "low-hanging fruit" (Özkula *et al.*, 2023; Burgess and Bruns, 2015). Yet, best practices for academics to collaborate with policymakers and platforms corporations to ensure reliable and ethical access to data still remain underdeveloped (Dommett and Tromble, 2022). While our study identified primarily social media-based access issues, future studies can benefit from focusing on the type of CPPT usually created outside of social media, *i. e.* letters to the editors, terrorist manifestos, or protest posters.

40 Matters are further complicated for researchers interested in non-English text and non-Western regions, as such projects often require more resources. The additional difficulty placed upon researchers studying non-Western regions and languages are a large part of the motivation for calls to de-westernise and de-colonialise communication research (Suzina, 2021; Ganter and Ortega, 2019). Scholars advocate for promoting greater heterogeneity and comparative social media research (Matassi and Boczkowski, 2021). It remains interesting to observe the future developments of "capabilities and impact of large language models (LLMs) in the wake of ChatGPT" (Jungherr, 2023).

41 Other resource inequalities contribute to disproportionate barriers for some researchers. Collaboration reduces the individual cost of CPPT research, but networking possibilities are limited (Giglietto *et al.*, 2012) and unevenly distributed. Multidisciplinary collaboration is further hindered if it is undervalued (Lazer *et al.* 2020). Establishing sustainable interdisciplinary research groups can be challenging but may be essential (Theocharis and Jungherr, 2021). Unequal networking has also been noted at conferences, where researchers with fewer barriers show reluctance to address the systemic bias experienced by other less-privileged researchers (Ng *et al.* 2020).

42 While these challenges, and the inequalities they exacerbate, are daunting, heightened awareness can facilitate addressing these concerns. For example, conferences ensuring better representation of underrepresented regions could enhance better global networking. Collaboration between developers and language experts, particularly those with diverse alphabets or language structures, will ensure the optimisation of tools to analyse various languages. As knowledge of shared challenges and inequalities among CPPT researchers is growing, so can the potential for developing constructive and effective remedies.

BIBLIOGRAPHY

Atteveldt van W. and Peng T. Q., 2018, "When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science", *Communication Methods and Measures*, 12 (2-3), p. 81-92. https://doi.org/10.1080/19312458.2018.1458084

Baden C., Pipal C., Schoonvelde M. and van der Velden M. A. C. G., 2022, "Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda", *Communication Methods and Measures*, 16 (1), p. 1-18. https://doi.org/10.1080/19312458.2021.2015574

Balluff P., Lind F., Boomgaarden H. G. and Waldherr A., 2023, "Mapping the European Media Landscape – Meteor, a Curated and Community-Coded Inventory of News Sources", *European Journal of Communication*, 38 (2), p. 181-194. https://doi.org/10.1177/02673231221112006

boyd d. and Crawford K., 2012, "Critical Questions for Big Data", *Information, Communication & Society*, 15 (5), p. 662-679. https://doi.org/10.1080/1369118X.2012.678878

Bruns A., 2019, "After the 'APIcalypse': Social Media Platforms and their Fight Against Critical Scholarly Research", *Information, Communication & Society*, 22 (11), p. 1544-1566. https://doi.org/10.1080/1369118X.2019.1637447

Burgess J. and Bruns A., 2015, "Easy Data, Hard Data: The Politics and Pragmatics of Twitter Research after the Computational Turn", in G. Langlois, J. Redden and G. Elmer (eds.), *Compromised Data: From Social Media to Big Data,* London, Bloomsbury Publishing, p. 93-111. https://doi.org/10.5040/9781501306549.0010

De Vreese C. and Tromble R., 2023, "The Data Abyss: How Lack of Data Access Leaves Research and Society in the Dark", *Political Communication*, 40 (3), p. 356-360. https://doi.org/10.1080/10584609.2023.2207488

Deltell L., Congosto M., Claes F. and Osteso J., 2013, "Identificación y análisis de los líderes de opinión en Twitter en torno a Hugo Chávez", *Revista Latina de Comunicación Social*, 68, p. 696-997. https://doi.org/10.4185/RLCS-2013-997

Dommett K. and Tromble R., 2022, "Advocating for Platform Data Access: Challenges and Opportunities for Academics Seeking Policy Change", *Politics and Governance*, 10 (1), p. 220-229. https://doi.org/10.17645/pag.v10i1.4713

Erzikova E. and Simpson E., 2017, "When the Gated Misbehave", *Journalism Practice*, 12 (9), p. 1148-1164. https://doi.org/10.1080/17512786.2017.1359653

Fu K. and Chau M., 2014, "Use of Microblogs in Grassroots Movements in China: Exploring the Role of Online Networking in Agenda Setting", *Journal of Information Technology & Politics*, 11 (3), p. 309-328. https://doi.org/10.1080/19331681.2014.909344

Ganter S. A. and Ortega F., 2019, "The Invisibility of Latin American Scholarship in European Media and Communication Studies: Challenges and Opportunities of De-Westernization and Academic Cosmopolitanism", *International Journal of Communication*, 13, p. 68-91. https://ijoc.org/index.php/ijoc/article/view/8449 (consulted at 23 Oct. 2024).

Gelovani S., Koc-Michalska K., Theocharis Y., Kalsnes B. and Haraldsson A., 2022, "Methodologic and Theoretical Approaches to Studying Citizen-Produced Political Text", Oral Presentation in 4th Annual *Comptext* Conference, Dublin.

Giglietto F., Rossi L. and Bennato D., 2012, "The Open Laboratory: Limits and Possibilities of Using Facebook, Twitter, and YouTube as a Research Data Source", *Journal of Technology in Human Services*, 30 (3-4), p. 145-159. https://doi.org/10.1080/15228835.2012.743797

Hargittai E., 2020, "Potential Biases in Big Data: Omitted Voices on Social Media", *Social Science Computer Review*, 38 (1), p. 10-24. https://doi.org/10.1177/0894439318788322

Jungherr A., 2023, "Artificial Intelligence and Democracy: A Conceptual Framework", *Social Media + Society*, 9 (3). https://doi.org/10.1177/20563051231186353

Karpf D., 2020, "Two provocations for the study of digital politics in time", *Journal of Information Technology & Politics*, 17 (2), p. 87-96. https://doi.org/10.1080/19331681.2019.1705222

Kreis R., 2017, "#refugeesnotwelcome: Anti-refugee discourse on Twitter", *Discourse & Communication*, 11 (5), p. 498-514. https://doi.org/10.1177/1750481317714121

Larsson A. O., 2015, "Comparing to Prepare: Suggesting Ways to Study Social Media Today—and Tomorrow", *Social Media + Society*, 1 (1). https://doi.org/10.1177/2056305115578680

Lazer D., Hargittai E., Freelon D., Gonzalez-Bailon S., Munger K., Ognyanova K. and Radford J., 2021, "Meaningful Measures of Human Society in the Twenty-First Century", *Nature*, 595, p. 189-196. https://doi.org/10.1038/s41586-021-03660-7

Lazer D. M. J., Pentland A., Watts D. J., Aral S., Athey S., Contractor N., Freelon D., Gonzalez-Bailon S., King G., Margetts H., Nelson A., Salganik M. J., Strohmaier M., Vespignani A. and Wagner C., 2020, "Computational Social Science: Obstacles and Opportunities", *Science*, 369, p. 1060-1062. https://doi.org/10.1126/science.aaz8170

Londoño L. M. L., 2019, "Formación de comunidades políticas afines y disímiles en Twitter durante la campaña electoral a la alcaldía de Manizales en 2015", *Anagramas: Rumbos y sentidos de la comunicación*, 17 (34), p. 115-134. http://ref.scielo.org/rm97yr

Matassi M. and Boczkowski P., 2021, "An Agenda for Comparative Social Media Studies: The Value of Understanding Practices From Cross-National, Cross-Media, and Cross-Platform Perspectives", *International Journal of Communication*, 15, p. 207-228. https://ijoc.org/index.php/ijoc/article/view/15042 (consulted at 23 Oct. 2024).

Ng E., White K. and Saha A., 2020, "#CommunicationSoWhite: Race and Power in the Academy and Beyond", *Communication, Culture and Critique*, 13 (2), p. 143-151. https://doi.org/10.1093/ccc/tcaa011

Özkula S. M., Reilly P. J. and Hayes J., 2023, "Easy data, same old platforms? A systematic review of digital activism methodologies", *Information, Communication & Society*, 26 (7), p. 1470-1489. https://doi.org/10.1080/1369118X.2021.2013918

Palinkas L. A., Horwitz S. M., Green C. A., Wisdom J. P., Duan N. and Hoagwood K., 2015, "Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research", *Administration and Policy in Mental Health and Mental Health Services Research*, 42 (5), p. 533–544. https://doi.org/10.1007/s10488-013-0528-y

Patton M. Q., 2015, *Qualitative research & evaluation methods: Integrating theory and practice,* Thousand Oaks, SAGE Publications Inc.

Perriam J., Birkbak A. and Freeman A., 2020, “Digital Methods in a Post-API Environment”, *International Journal of Social Research Methodology*, 23 (3), p. 277-290. https://doi.org/10.1080/13645579.2019.1682840

Sormanen N. and Lauk E., 2016, “Editorial: Issues of Ethics and Methods in Studying Social Media”, *Media and Communication*, 4 (4), p. 63-65. https://doi.org/10.17645/mac.v4i4.793

Stier S., Bleier A., Lietz H. and Strohmaier M., 2018, “Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter”, *Political Communication*, 35 (1), p. 50-74. https://doi.org/10.1080/10584609.2017.1334728

Suzina A. C., 2021, “English as *lingua franca*. Or the Sterilisation of Scientific Work”, *Media, Culture & Society*, 43 (1), p. 171-179. https://doi.org/10.1177/0163443720957906

Theocharis Y. and Jungherr A., 2021, “Computational Social Science and the Study of Political Communication”, *Political Communication*, 38 (1-2), p. 1-22. https://doi.org/10.1080/10584609.2020.1833121

Vaccari C. and Valeriani A., 2021, *Outside the Bubble: Social Media and Political Participation in Western Democracies,* Oxford, Oxford University Press.

Venegas M., 2022, “Between Community and Sectarianism: Calling Out and Negotiated Discipline in Prefigurative Politics”, *Social Movement Studies*, 21 (3), p. 297-314. https://doi.org/10.1080/14742837.2020.1866528

Venturini T. and Rogers R., 2019, “‘API-Based Research’ or How can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge Analytica Data Breach”, *Digital Journalism*, 7 (4), p. 532-540. https://doi.org/10.1080/21670811.2019.1591927

Wasserman H., 2020, “Moving from Diversity to Transformation in Communication Scholarship”, *Annals of the International Communication Association*, 44 (1), p. 1-3. https://doi.org/10.1080/23808985.2019.1706429

Zhu J. and Wang C., 2021, “I Know What You Mean: Information Compensation in an Authoritarian Country”, *International Journal of Press/Politics*, 26 (3), p. 587-608. https://doi.org/10.1177/1940161220963572

APPENDIXES

Table A1. Survey respondent summary table

	All respondents (N = 295)	*CPPT non-interested (N = 73)*	*CPPT interested (N = 59)*	*CPPT users (N = 163)*
Region (N=188)				
Africa	9 (4.79%)	1 (5.88%)	1 (2.13%)	7 (5.65%)
Americas	31 (16.49%)	1 (5.88%)	5 (10.64%)	25 (20.16%)
Asia	18 (9.57%)	1 (5.88%)	3 (6.38%)	14 (11.29%)
Europe	126 (67.02%)	14 (82.35%)	37 (78.72%)	75 (60.48%)
Oceania	4 (2.13%)	0 (0.00%)	1 (2.13%)	3 (2.42%)

Gender (N=208)				
Male	120 (57.69%)	10 (52.63%)	31 (59.62%)	79 (57.66%)
Female	84 (40.38%)	8 (42.11%)	21 (40.38%)	55 (40.15%)
Neither	1 (0.48%)	0 (0.00%)	0 (0.00%)	1 (0.73%)
Prefer not to say	3 (1.44%)	1 (5.26%)	0 (0.00%)	2 (1.46%)
Academic rank (N=209)				
PhD student	41 (19.62%)	5 (26.32%)	15 (28.30%)	21 (15.33%)
Junior	37 (17.70%)	4 (21.05%)	10 (18.87%)	23 (16.79%)
Mid	73 (34.93%)	5 (26.32%)	14 (26.42%)	54 (39.42%)
Senior	50 (23.92%)	4 (21.05%)	12 (22.64%)	34 (24.82%)
Other	8 (3.83%)	1 (5.26%)	2 (3.77%)	5 (3.65%)
Main academic field (first field listed if multiple; N=209)				
Communications	97 (46.41%)	8 (42.11%)	15 (28.30%)	74 (54.01%)
Political Science	67 (32.06%)	10 (52.63%)	31 (58.49%)	26 (18.98%)
Psychology	5 (2.39%)	0 (0.00%)	0 (0.00%)	5 (3.65%)
Sociology	15 (7.18%)	0 (0.00%)	2 (3.77%)	13 (9.49%)
Other	19 (9.09%)	0 (0.00%)	4 (7.55%)	15 (10.95%)

Table A2. Interviewee summary table.

Characteristic	*Breakdown*
Gender	13 women, 8 men
Primary discipline	5 communication/journalism, 1 computer science, 2 international relations, 1 law, 2 linguistics, 8 political science (2 interdisciplinary with no clear main discipline)
Seniority	8 junior (<5 years post PhD), 10 mid (5-15 years post PhD), 3 senior (>15 years post PhD)
Region of origin	5 outside Europe (Asia, MENA, South America); 3 Northern Europe, 3 Eastern Europe; 4 Southern Europe; 6 Western Europe
Region of current affiliation	4 outside Europe (MENA, North America, South America); 5 Northern Europe; 2 Eastern Europe; 3 Southern Europe; 7 Western Europe

NOTES

1. This work was supported by the the European Union's Horizon 2020 Research & Innovation Action under Grant Agreement no. 951832. Opted.eu (OPTED). We thank Susan Banducci and Yannis Theocharis for their contribution to this manuscript, and members of OPTED for the development and distribution of the survey used in this study. The reports from the study can be found here https://opted.eu/results/project-

reports/. Data in the anonymized and searchable format from the CPPT research can be found here https://meteor.opted.eu/ (consulted at 24 Ot. 2024).

2. Interviewee citations are edited for language correction.

3. Interviews were kept before the restriction to data access on Twitter introduced in Summer 2023.

4. A fear that turned out to be well-founded, as announced in April 2023, available here: https://csmapnyu.org/news-views/news/twitter-s-transparency-theater (consulted at 18 Oct. 2024).

5. As multiple interviewees explain, it is not possible to retroactively access deleted posts using standard APIs for many platforms, however in certain cases such posts can be accessed through other means.

6. Prior to Twitter closing down the public access to their API in 2023, several tweet scraper tools did not allow a direct scraping of the tweet text or other data associated with it, but rather only the tweet IDs. Re-hydrating tweets meant recovering the tweet text using the tweet IDs, with the help of the now defunct "hydrator" software.

7. Netvizz, a now defunct Facebook app created by Bernhard Rieder, University of Amsterdam, allowed researchers to collect data from Facebook pages.

8. Please check medem.eu initiative to build such project available here: https://www.medem.eu/ (consulted on 24 Oct. 2024).

ABSTRACTS

Abstract: Advancements in data harvesting and analysing techniques of large datasets have introduced novel challenges for research utilising political text produced by citizens (CPPT). There are numerous disparities in the current research, which populations and how they are studied, data availability, and access privileges. Researchers' perspectives on these obstacles have seldom been empirically captured. Our study, built on a survey and in-depth interviews with researchers worldwide, provides an evidence-based categorisation of the primary challenges faced. The findings indicate that the most dire issues relate to the social media platform restrictions, differences due to languages employed, and the resource-intensive nature of the research.

Résumé : Les progrès des techniques de collecte et d'analyse de grands ensembles de données ont introduit de nouveaux défis pour la recherche utilisant les textes politiques produits par les citoyens (CPPT). Il existe de nombreuses disparités dans la recherche actuelle, par exemple : le type de populations et la manière dont elles sont étudiées, la disponibilité des données et les privilèges d'accès. Le point de vue des chercheurs sur ces obstacles a rarement fait l'objet d'une analyse empirique. Notre étude, fondée sur une enquête et des entretiens approfondis avec des chercheurs du monde entier, fournit une catégorisation factuelle des principaux défis à relever. Les résultats indiquent que les problèmes les plus graves sont liés aux restrictions imposées par les plateformes de médias sociaux, aux différences des langues utilisées et à la nature de la recherches qui nécessite beaucoup de ressources.

INDEX

Mots-clés: texte en tant que données, citoyens, méthodes multiples, inégalités, accès aux données
Keywords: text-as-data, citizens, multi-method, inequalities, data access

AUTHORS

AMANDA HARALDSSON

Communication & Culture Department, Audencia Business School, FR-44100 Nantes, France

SHOTA GELOVANI

Institute for Media and Communication Studies, University of Mannheim, DE-68161 Mannheim, Allemagne

MICHELE SCOTTO DI VETTIMO

Department of Political Economy, King's College London, GB-WC2R 2LS Londres, Royaume-Unis

BENTE KALSNES

Department of Communication, Kristiania University College, NO-0107 Oslo, Norvège

KAROLINA KOC-MICHALSKA

Audencia Business School, Nantes, FR-44100 France
Faculty of Social Sciences, University of Silesia, PL-40-003 Katowice, Pologne