*European Political Science* (2026), **25**, 116–128
doi:10.1017/S1682098325100015

PROFESSION

# Research challenges and training needs in text analysis for political science research

Michele Scotto di Vettimo[1], Amanda Haraldsson[2], Shota Gelovani[3], Susan Banducci[4], Karolina Koc-Michalska[5,6] and Yannis Theocharis[7]

[1]Department of Political Economy, King's College London, London, UK, [2]Paris Lodron University, Salzburg, Austria, [3]University of Mannheim, Mannheim, Germany, [4]University of Birmingham, Birmingham, UK, [5]Audencia Business School, Nantes, France, [6]University of Silesia, Katowice, Poland and [7]Technical University of Munich, Munich, Germany
**Corresponding author:** Michele Scotto di Vettimo; Email: michele.scotto_di_vettimo@kcl.ac.uk

(Received 17 March 2025; revised 2 August 2025; accepted 6 August 2025; first published online 19 January 2026)

**Abstract**
An ever-increasing availability of digital texts has opened new research opportunities for political scientists. Yet, researchers who want to utilise these data face several challenges. This paper presents the results of a community-wide survey tapping into various research challenges, training needs, and preferences of scholars using text analysis methodologies. The survey involved respondents from various academic fields and career levels. Our findings indicate that text-as-data methods are gaining momentum in various political science subfields and are used on a wide range of political texts. However, relevant training is not easily accessible to all. Only half of the respondents have ever participated in a training event, though there is a high demand for training opportunities in different formats and at different levels. In 'Conclusions', we discuss how the inaccessibility of training risks narrowing the field of researchers.

**Key words:** text analysis; computational social science; political texts; training needs; research challenges

## Introduction

Digital transformation has vastly expanded access to political texts. Government and parliamentary documents are increasingly available online, social media have become a source of politically relevant texts, and media outlets offer digital editions. Data have now been made available on a larger scale through application programming interfaces (APIs) and the use of web crawling, and new data types have emerged (Theocharis and Jungherr 2021). This growing availability of digital texts offers new research opportunities, especially for political scientists. This development, together with an increased emphasis on quantification, is deeply influencing the way political research is conducted. Leveraging these data, however, requires researchers to be trained in a new set of skills. In this paper, we focus on text data, one of the most popular digital data in political science, highlighting challenges and proposing solutions for training in text analysis.

Text analysis is essential for interpreting new data and understanding key political science issues, such as democratic governance and its challenges (Kiss and Sebők 2022), and it has been used in political research since the field's early days (cf. Lasswell 1927). Broadly speaking, text analysis approaches can be categorised into (a) qualitative, (b) quantitative manual, and (c) quantitative computational approaches. Qualitative methods, such as discourse or content analysis, rely on the researcher's close reading of the texts. Quantitative methods systematically

© The Author(s), 2026. Published by Cambridge University Press on behalf of European Consortium for Political Research. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. 15 Sep 2026 at 15:49:56, subject to the Cambridge Core terms of use.

transform textual information into numerical data so as to enable the use of quantitative or statistical methodologies to analyse such texts. While manual approaches require the researcher to do this transformation manually (eg via counting words or annotating documents), computational ones rely on automated software and tools.

Regardless of the method, the volume and diversity of political texts across time and contexts present challenges that require coordinated efforts to harmonise data collection and analytical processes to ensure comparability and to promote common practices on validity and ethical and legal practices.

In this context, the digital transformation has had a significant impact on the field, with computational techniques rapidly advancing alongside traditional manual methods (Brady 2019; Van Atteveldt, Welbers, and Van Der Velden 2019). Computational text analysis (CTA) includes a range of methods, from basic keyword extraction to advanced, versatile software tools. This rapid growth is driven by the digital data surge, improved analytical tools, and affordable processing power, making CTA increasingly common across the social sciences (Kim and Ng 2022).

Today, text analysis in political and social research benefits from abundant and diverse data, alongside a wide range of methodological tools. This presents opportunities and challenges for research practices and the broader scholarly community. Opportunities include improved data quality, expanded research on various phenomena and populations (Salganik 2019; Theocharis and Jungherr 2021), new theoretical frameworks (Edelmann, Wolff, Montagne et al. 2020; Suhay, Grofman, and Trechsel 2020), and the potential for collaborative research agendas (Van Atteveldt and Peng 2018).

Additionally, recent advancements in artificial intelligence (AI), particularly large language models (LLMs), have transformed text analysis in political research. Models such as GPT-4 (OpenAI 2024) and BERT (Devlin, Ming-Wei, Lee et al. 2019) can automate tasks such as classification, sentiment analysis, and summarisation, offering researchers powerful tools to analyse large volumes of texts. The integration of LLMs has the potential to significantly lower the technical barriers to CTA, enabling more researchers use text-as-data methods without advanced programming skills. However, the increased reliance on AI-driven analysis also raises concerns about interpretability, bias, and reproducibility, highlighting the need for training that includes tool usage and critical evaluation.

Notwithstanding their potential, the research opportunities just outlined remain concentrated among a minority of researchers. To be fully exploited by the wider research community, key challenges need to be addressed. This article focuses on challenges text analysis scholars face that could be mitigated through tailored training or improved research infrastructures. While many broad assessments exist, little to no input has been gathered by those immediately interested in, and affected by, (the lack of) these opportunities: the scholars themselves. In the following, we present findings from a community-wide survey exploring the state of the text analysis researchers' community in terms of their substantive interests, methodological experience, research challenges, and attitudes towards training opportunities.

Our survey represents an effort to take a 'bottom-up' approach to assessing how the community can identify the training needs and address the existing research challenges. The survey has been conducted as part of the work programme of the European Union's Horizon 2020 Observatory for Political Texts in European Democracies (OPTED) project (Grant Agreement no. 951832), an international collaborative infrastructure project focused on political text data, of which all the authors have been part.

## Review of challenges and training needs in text analysis for social and political research

We cover the four types of research challenges faced by social and political scientists working with text-as-data approaches. We identified these challenges from a review of existing literature on the

Downloaded from https://www.cambridge.org/core. 15 Sep 2026 at 15:49:56, subject to the Cambridge Core terms of use.

use of text analysis for social sciences more broadly (eg Kim and Ng 2022; Lazer, Pentland, Adamic et al. 2020; Salganik 2019; Suhay, Grofman, and Trechsel 2020; Van Atteveldt and Peng 2018). From this overview, we find that extant literature is rather focused on challenges related to computational methods, and to a lesser extent on text-as-data methodologies more in general. Thus, we also discuss these challenges from a broader perspective to emphasise their relevance for the wider text analysis community.

#### *Discovering relevant data*

Despite the abundance of digital data, finding data suited to a specific research question remains challenging. Much of these data are intended for non-academic purposes, such as profit-driven services (Salganik 2019), and must be found and repurposed by researchers. Collecting large volumes of data indiscriminately is a common pitfall (Kim and Ng 2022). Rather, researchers must start with theoretically relevant questions and conduct a thorough assessment of the data they find with regard to their quality and their relevance (Suhay, Grofman, and Trechsel 2020). Therefore, it is crucial that researchers are aware of potentially relevant data sources or know how to look for them and have the necessary skills to evaluate the appropriateness of each of the alternatives for the task at hand.

#### *Accessing data*

Once relevant data are identified, researchers should be able to access them openly and transparently.[1] Publicly owned data generally do not pose major concerns in terms of access restrictions. Many examples of publicly funded sources of political texts, such as the Manifesto Project Database (Lehmann, Franzmann, Al-Gaddooa et al. 2024) and the ParlLawSpeech dataset (Schwalbach, Hetzer, Proksch et al. 2025), have already shown the potentiality of synergies between researchers and government agencies (cf. Lazer, Pentland, Adamic et al. 2020), as well as the benefits for the wider academic community. In contrast, access to data held by private companies is more problematic. While user-generated data from digital platforms are valuable (Salganik 2019), access is often limited, leaving researchers dependent on platform policies (Theocharis and Jungherr 2021). Clear examples of this are the changes in Facebook's policy for data access enacted as of 2016 and the end of free access to the Twitter API in 2023, which put at risk many research projects relying on such data (Mimizuka, Brown, Yang et al. 2025).

Limited access to proprietary data poses challenges for individual researchers and the broader academic community. When some scholars – because of greater resources or connections – can access restricted data more easily than others, it creates inequalities and gives them disproportionate visibility over equally capable peers (Huberman 2012). This can also hinder reproducibility in text analysis research. Assessing these access asymmetries is crucial for developing fair data-sharing strategies and fostering collaborations with data providers (Levi and Rajala 2020).

In some cases, access to research data also depends on understanding relevant ethical standards. The digital transformation has made it easier to use social media data and implement online experiments, but these developments raise new ethical concerns – especially with designs involving leaked data (Gill and Spirling 2015). Researchers must be aware of established ethical practices to avoid actions that could be seen as misconduct (Larsen 2022).

#### *Skills needed to leverage data for research purposes*

A researcher's methodological toolkit shapes how data are used in their work. Yet, as Kim and Ng (2022) note, many social scientists – especially mid-career and senior researchers – have limited

[1]Best practice in data management emphasises the use of the principles of findable, accessible, interoperable, and reusable (FAIR) data. See https://www.go-fair.org/fair-principles/.

Downloaded from https://www.cambridge.org/core. 15 Sep 2026 at 15:49:56, subject to the Cambridge Core terms of use.

exposure to computational techniques and often learn them 'on the job'. In contrast, junior scholars increasingly benefit from courses offered through university programmes or doctoral training partnerships.

Asymmetries in terms of methodological proficiency and learning opportunities could generate new gaps among researchers in the field. These disparities arise not only between individuals with varying levels of expertise but also across regions, as some areas have stronger traditions in text analysis training than others.

Against this background, the rapid rise of LLMs introduces two new dynamics. On one hand, LLMs can ease the learning curve for less experienced researchers by supporting text analysis and coding. On the other, their use heightens the need for computational resources and creates new challenges around validation, reproducibility, and technical skills – such as prompt engineering and model fine-tuning. These evolving demands risk deepening existing disparities, undermining interdisciplinary and crossnational collaboration, as well as individual career opportunities (Haraldsson, Gelovani, Scotto di Vettimo et al. 2024). Furthermore, while these tools and training may be open source, persistent skill gaps may reflect underlying inequalities in economic resources and educational capacity needed to access and fully benefit from them.

### *Collaboration and development of common practices and standards*

A key challenge in political text analysis, especially for those using computational methods, is the field's fragmentation (Jungherr, Metaxas, and Posegga 2020). Tools and resources are often developed in isolation, and institutions rarely support the interdisciplinarity many projects require. This hinders collaboration and negatively affects the quality and reproducibility of research.

First, fragmentation complicates collaborations across research fields. In a fragmented domain, where skills, infrastructures, and institutions are poorly or asymmetrically developed, such collaborations are difficult to achieve, as it is harder to find common ground among research teams or the collaboration may not be equally innovative – and thus appealing – for all teams involved (Van Atteveldt and Peng 2018).

Second, fragmentation makes it difficult to collaborate crossnationally. Expertise and training facilities appear to be centring on a few countries with strong research infrastructures (Gelovani, Kalsnes, Koc-Michalska et al. 2021). Hence, researchers face different challenges with regard to the access to relevant resources and support. Again, this prevents fruitful collaborative endeavours across political science subfields and among teams based in different countries.

Third, fragmentation also undermines the development of shared practices and standards, affecting the reproducibility and validity of research. Gaps in methodological training contribute to inconsistent approaches within the community. Moreover, the interdisciplinary nature of the text-as-data field complicates efforts to establish common standards for transparency in data collection, preparation, harmonisation, and analysis (Theocharis and Jungherr 2021).

Overall, the challenges faced by researchers using text-as-data methods significantly affect the broader field. Computational techniques remain underused in many subfields that could benefit from them (cf. Edelmann, Wolff, Montagne et al. 2020), largely because of the lack of structured training. This has slowed the development of truly multidisciplinary research, especially where computational skills are needed. As Lazer, Pentland, Adamic et al. (2020) note, social scientists and computational researchers often work in separate academic silos with limited collaboration. The rise of tools such as LLMs further underscores the need for stronger interdisciplinary partnerships to establish shared best practices.

Finally, text-as-data research often lacks strong ties to established theories and concepts. On the one hand, it could benefit from more theoretically driven designs that address or test existing frameworks. On the other, text-as-data research holds untapped potential for contributing to theory building (Jungherr and Theocharis 2017).

Downloaded from https://www.cambridge.org/core. 15 Sep 2026 at 15:49:56, subject to the Cambridge Core terms of use.

This overview highlights that text-as-data researchers in the social sciences face numerous challenges. While top-down assessments provide a useful overview, advancing the use of these methods in political science requires a clear understanding of researchers' diverse training needs across subfields. Such training needs can only be made evident by taking a 'bottom-up' approach. This study represents the first such effort, using a community-wide survey to assess training requirements and address current research challenges, with a focus on quantitative and qualitative approaches to text data.

## Methods and data

### *Design*

Approved by the Research Ethics and Governance Office of the University of Exeter (REF: 509745), the online survey was disseminated from 23 February 2022, and a reminder was sent out 14 days later. The survey was shared widely among the community members. We identified as the primary target group academics working in the realm of text analysis. We aimed at assessing the research challenges and training needs of a very heterogeneous research community that comprises PhD students, early- and mid-career researchers, and senior researchers. We also sought to reach researchers who are interested in text analysis, though have not used it already in their research. The questionnaire focused on their research challenges, training needs, and preferences.

### *Survey sample and recruitment*

To recruit respondents, we used a four-pronged approach: (i) contact lists collected by OPTED project work packages, (ii) contacts held by individual project members, (iii) mailing lists from text analysis events, and (iv) social media outreach.

First, work packages within the project had previously collected contact details of scholars working on text analysis. Though the inclusion criteria varied slightly from one work package to another,[2] all the populations are contained within our primary target group.

Second, project members disseminated the survey to all the lists of contacts that they had in their possession. This meant reaching out to mailing lists constructed for research purposes other than this specific project or lists of participants in classes, workshops, or other events organised by the members of the project network.

Third, to reach prospective users of text analytical techniques and more junior researchers, we contacted organisers of several prominent events related to text analysis held in Europe from 2018 up to March 2022 and asked them to share the survey with all the event participants. We also asked members of the steering committees of some European Consortium of Political Research (ECPR) Standing Groups to circulate the survey to the group members. Section 1 of the online Supplementary Material (SM) provides the list of events and ECPR groups contacted.

Fourth, we shared the survey on the social media accounts of the OPTED project to reach additional potential respondents.

### *Measures*

The survey included seven sections focusing on different dimensions of training and research needs. More specifically, we ask questions about data access, publishing, gaps in databases, or challenges related to multilingual text analysis research. One section consists of 15 questions about existing and future training opportunities, channels used to acquire new skills, and participation in

[2]For instance, one Work Package (WP) targeted authors who published any studies using quantitative text-based research over the past five years in top journals in political science, communication, sociology, and psychology, whereas another collected email addresses of authors studying citizen-produced political texts.

Downloaded from https://www.cambridge.org/core. 15 Sep 2026 at 15:49:56, subject to the Cambridge Core terms of use.

existing training events. Depending on whether they use computational or noncomputational methodologies, respondents were prompted with more targeted questions about specific tools used and challenges faced when doing text analysis. Some of the questions were open-ended to allow participants to express the needs and preferences that were not covered in the set survey options and to capture more nuanced feedback and comments. For more details about the survey questions, refer to SM Section 2.

### *Survey respondents and their characteristics*

The survey reached 296 participants (95% within two months of first launch). We excluded 45 respondents giving consent without filling in the survey, leaving 251 respondents. Of these 251, 17.3% had some missing responses. Though it was not possible to calculate an exact response rate, we can refer to one of our emailing lists on citizen-produced political texts. The list contained email addresses of 2,206 scholars, of which 160 completed the survey, resulting in a response rate of 7.3% among this cohort.

We gathered data on respondents' gender, geographical region, academic field, and career stage (Table C1 in the SM). Most participants identified as men (58%) and were based in Europe (67%), followed by America (16%), Asia (10%), Africa (5%), and Australia/New Zealand (2%). Political science and communications dominated the sample (64%), with sociology (13%), psychology (5%), economics (3%), and linguistics (2.5%) also represented. In addition, 36 other fields were each cited by less than 1% of respondents. The sample is fairly balanced across career stages, with mid-career researchers forming the largest group (36%), followed by senior (25%), PhD students (20%), and early-career researchers (19%).

Though we do not have information about the distribution of such demographics in our target population, we were able to compare our sample with participants to text analysis conferences. This gives us a rough idea of sample representativeness and helps us better understand the extent to which we can make claims that could generalise well to our population. Looking at participation at the COMPTEXT Conference from 2018 to 2025, we had 63% male participants. PhD students represented 30% of conference participants, early-career researchers represented 10% of the participants, and mid-career and senior researchers combined made up 41%. Overall, our respondents were quite representative in terms of gender, though in terms of career stage, more experienced researchers tended to be overrepresented in our survey if compared with COMPTEXT participants. On the contrary, in terms of geographical representation, COMPTEXT attracted mostly European scholars. Hence, our sample was better in terms of reaching researchers affiliated with non-European institutions.

## Results

### *Data needs and familiarity with text analysis methodologies*

The first major challenge is identifying relevant text data. For 25% of respondents, this was a major challenge; 46% saw it as a minor one, while only 29% reported no difficulty. The most commonly used texts came from journalists and media outlets (73%), followed by those from citizens and individual politicians (each 66%). Overall, more than three-quarters of respondents engaged with all major types of political texts (Figure C1 in the SM).

Additionally, 18% of respondents mentioned using other types of texts not listed in the survey, including judicial documents, texts from extremist groups, reports, memoirs, scientific articles, and political memes. This highlights the community's broad interest in diverse text types – an important reminder in a field often dominated by social media data.

A key question is which methodologies participants used to analyse their data. As Figure 1 shows, no single approach dominated. Qualitative methods were less common: 29% did not use

Downloaded from https://www.cambridge.org/core. 15 Sep 2026 at 15:49:56, subject to the Cambridge Core terms of use.

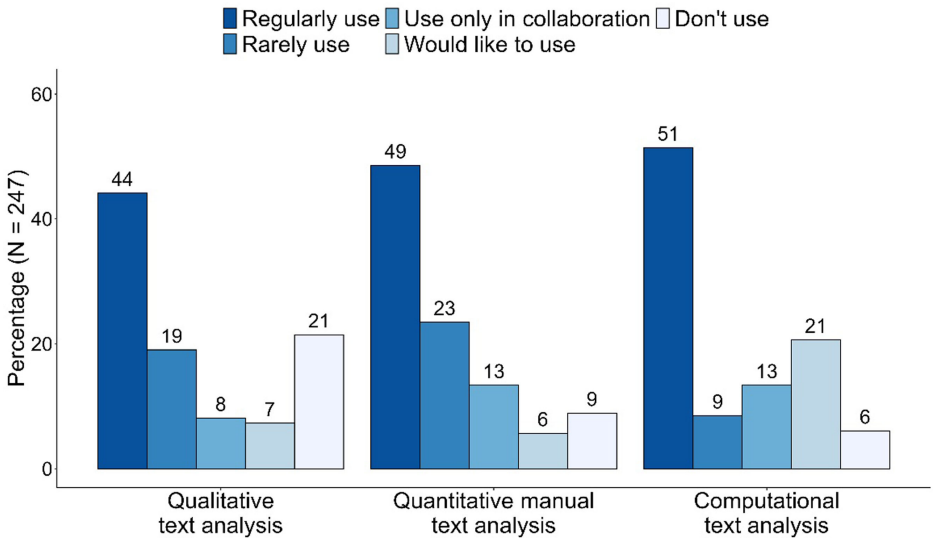


**Figure 1.** Text analysis methods used or planned to be used in research.

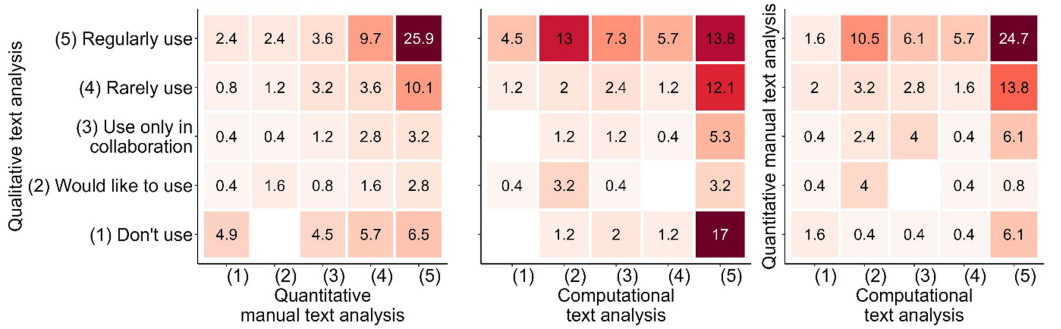


**Figure 2.** Crosstabulations of usage patterns of the three categories of text-as-data methods.

them or only used them collaboratively, 44% used them regularly, and just 7% planned to use them in the future. Quantitative methods were slightly more prevalent. About half of respondents regularly used manual (49%) or computational (51%) approaches. Notably, computational methods had fewer nonusers and a higher share of prospective users (21%), suggesting they may become dominant in the future.

In fact, respondents may be using more than one type of methodology. Hence, it is interesting to inspect how usage experiences correlate among themselves. Figure 2 reports the crosstabulations of usage patterns, comparing the use of qualitative methods with quantitative manual text analysis (left panel) and with CTA (central panel), and of manual and computational quantitative methods (right panel). The numbers represent cell percentages.

Around half of the respondents were users of qualitative and quantitative manual methods, though only one-quarter used both regularly. This overlap, however, was way less marked when we compared qualitative with computational techniques. Only 14% of the respondents employed both methods regularly, and the most common type of users were those who regularly used computational methods but never used qualitative ones. Finally, there was also some overlap between the use of one or the other quantitative approach, and 10% of regular users of manual techniques would like to use computational methods.

Downloaded from https://www.cambridge.org/core. 15 Sep 2026 at 15:49:56, subject to the Cambridge Core terms of use.

We also examined how methodological choices vary by academic field, career stage, and region (SM Figures C2–C4). Regular use of qualitative methods was most common among sociologists (47%) and communication scholars (44%), though about one-third of economists and political scientists also reported using them. Quantitative manual techniques were widely used across communication, economics, and political science, with around half of respondents in each field reporting regular use. CTA was used by at least half of the respondents in all five fields, with particularly high adoption among economists (83%).

Methodological preferences also varied by career stage. Qualitative methods were mainly used by mid-career (45%) and senior researchers (32%), but rarely by PhD students and early-career scholars. Quantitative manual methods were more evenly used across all levels, from 39% of early-career to 59% of senior researchers. Computational methods showed a clear generational split, with most PhD students (71%) and early-career researchers (61%) using them regularly, compared with less than half of mid-career (44%) and senior researchers (43%).

Qualitative text analysis was common across most regions except southern and western Europe, while quantitative manual methods were used evenly everywhere. CTA usage was, instead, less balanced. It was most popular in western Europe, where 78% used it regularly, 10% collaborated on it, and 8% planned to use it, with only 5% rarely or never using it. In contrast, about half of researchers in America and eastern Europe used CTA regularly, dropping to about a third in northern and southern Europe.

### *Data access and tool availability*

Barriers to accessing data pose significant challenges for text-as-data researchers. We are interested in understanding the relative importance of four types of potential access barriers: restricted access by the owner, restricted access because of content removal, difficulty of using tools necessary to access the data, and ethical rules preventing the access to certain texts. We found that the main obstacle was restricted access by data-owning companies, seen as a major challenge by 71% and a minor one by 23% of respondents. Content removal online was also a major problem for 34% and a minor one for 48%. Difficulty using tools to access data affected 71% of researchers. Ethical rules limited access less frequently, with 18% citing them as a major challenge to access, 44% as a minor challenge, and 38% reporting no challenge.

### *Leveraging data for research purposes*

A key challenge for researchers using text-as-data methods was acquiring the necessary skills. Among those using computational methods, 64% believed such skills are essential to stay competitive in academia, while only 19% disagreed. However, just 54% felt skilled enough to solve research problems by relying solely on available tool documentation. Supervisors also emphasised this gap, with 76% identifying programming and software skills as 'very' or 'extremely' important training needs. Hence, an important question for the text-as-data field is how scholars who do not yet possess a sufficient level of familiarity with relevant tools, resources, and methodologies can acquire it.

Therefore, we asked respondents how they acquired the skills needed for text analysis (Figure C5). The most common method was self-directed online materials, with 37% using either online tutorials or examples. About 25% gained skills through community workshops, conferences, or third-party courses, while only 17% relied on training during their formal university studies. Hence, respondents primarily used online resources, followed by training events, with university courses playing a smaller role. Informal support from peers and colleagues was also important, helping 18% of respondents.

Lack of necessary skills, however, is not the sole challenge related to the use of specific text analysis methodologies. Focusing on CTA, we divided respondents into those using and those not using such methods. To the first, we asked about challenges encountered during their research,

Downloaded from https://www.cambridge.org/core. 15 Sep 2026 at 15:49:56, subject to the Cambridge Core terms of use.

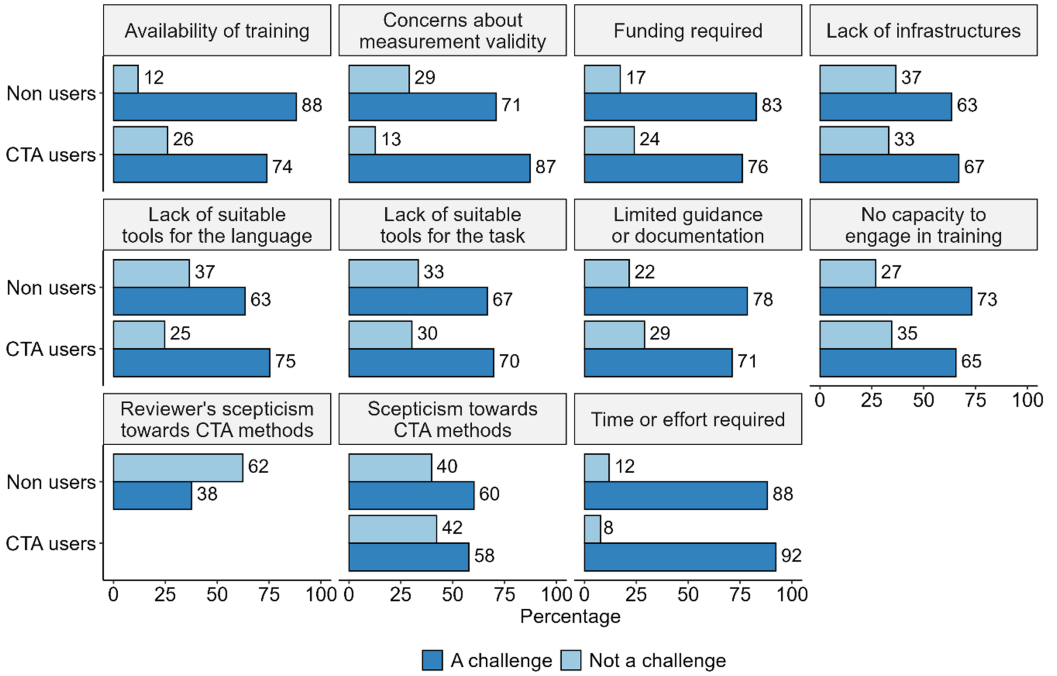


**Figure 3.** Challenges in using computational methods, by current and potential users.

and to the second, the potential challenges that stopped them from using CTA methods. Figure 3 presents the responses over 10 possible challenges (11 for the nonusers).

There was a high similarity of responses between CTA users and nonusers. Most of the challenges encountered by users were also anticipated by nonusers and, in fact, determined their choice of not using CTA methods.

The three most common challenges for CTA users were the time or effort required (a challenge for 92% of them), the concerns about the measurement validity (87%), and the funding that may be required (76%). When asked about why the time or effort required posed a major challenge, CTA users pointed to existing research (30%), other professional commitments (24%), or high teaching load (18%). Time was also a challenge for 88% of the nonusers, followed by training (88%) and funding availability (83%).

Furthermore, the open responses also highlighted various additional challenges, with answers pointing to data availability problems because of the paucity of machine-readable text data in the first place or to restrictions to web scraping for data that are publicly available. Nonusers also mentioned that training is perceived to be poorly designed ('some [of them] are too broad, others are too specific'), whereas others showed a more fundamental scepticism towards the capacity of computational methods in accounting for the nuances and 'context' of the texts, signalling however that they are open to the use of computational methodologies if they are capable of accounting for these nuances – a development that LLMs should be able to deliver.

Then, we explored the most common challenges for researchers at different career levels. For CTA users (Figures C6 and C7), the key differences concerned the importance of funding (challenging mostly for early-career researchers), the lack of suitable tools for the language analysed (problematic mostly for PhD students and senior researchers), and the availability of training, which emerged among the top challenges only for mid-career researchers. For nonusers (Figures C8 and C9), instead, lack of infrastructures emerged among the top challenges for early-career researchers, whereas limited guidance or documentation and limited capacity to engage in training represented top challenges for mid-career and senior researchers, respectively.

Downloaded from https://www.cambridge.org/core. 15 Sep 2026 at 15:49:56, subject to the Cambridge Core terms of use.

### Community support, collaborations, and shared standards

Finally, we explored challenges around community support, common standards, and collaboration in text-as-data research.

Half of the respondents found it difficult to connect with peers to discuss computational methods. While 61% used public platforms for support, only 53% could easily discuss these concerns with colleagues in their department. In practice, for half of the computational users it was hard to find peers with whom they could discuss their everyday methodological problems. This lack of connection hinders collaboration and contributes to the field's fragmentation.

The difficulty in engaging with the text-as-data community also resulted in the absence of shared standards and best practices. The poor knowledge of common practices and standards also emerged from the answers that research supervisors gave concerning the most important training needs of text analysis researchers. Of the supervisors, 88% highlighted research integrity and ethics, and 87% noted the importance of theory and concept training. Finally, when asked to comment on additional areas where training is needed, supervisors also mentioned data management planning practices and research methodology.

### Experience with – and attitudes towards – training events

Finally, we focus on respondents' use of more structured training to acquire necessary skills to overcome the research challenges they face.

We found that 56% of respondents had participated in some form of text analysis training, meaning a substantial 44% had not. This gap was not solely because of seniority – 40% of PhD students also reported no training experience. Among those who had attended training, participation was spread across formats: 28% had joined conferences or schools with some text analysis sessions, 20% had attended dedicated events, and 29% had taken part in workshops or seminars, compared with 20% who had attended structured training sessions (Figure C11).

Importantly, training events were positively evaluated (Figure C12): 84% found them useful for learning new tools, 80% valued the networking opportunities with researchers using similar methodologies, 77% said they improved existing skills, and 66% found them helpful for receiving research feedback.

Furthermore, open-ended responses highlighted additional strengths of training events, including peer exchange that fostered collaboration, expert instructors, hands-on and research-relevant content, exposure to alternative methods, and access to training materials in advance. Conversely, some respondents were dissatisfied with the fact that complex concepts were often taken for granted, even though participation was not restricted to experienced users, and others complained that most events cover fairly introductory topics and that it is difficult to find more advanced training.

Regarding training preferences, 83% of respondents were somewhat or very likely to attend an event in the next two years. Interest was highest among early-career researchers (97%), followed by PhD students (88%), mid-career (81%), and senior researchers (69%). Only 12% were unlikely to participate.

With regard to the profile of the trainers, 63% of respondents preferred academics from their own field who use text analysis, while 34% were open to experts outside their field. Only 3% preferred non-academic or industry trainers. Interestingly, we received very similar results when looking only at supervisors of researchers who needed to learn text analysis.

Finally, when asked about further training, 82% of respondents expressed interest. Of those not interested, 7% felt adequately skilled, 4% did not plan to use text analysis, and 7% cited other reasons (mainly lack of time). When enquired about the level at which they would need this additional training, 23% said that they would need introductory-level training, 43% intermediate-level training and another 34% advanced-level training.

Downloaded from https://www.cambridge.org/core. 15 Sep 2026 at 15:49:56, subject to the Cambridge Core terms of use.

## Conclusions and discussion

The landscape of text analysis for the social sciences is one where a great amount and wide variety of data are becoming easily available to researchers, who in turn have at their disposal a toolbox of different, yet complementary, methodologies to analyse data. This development, however, is counterbalanced by emerging challenges the wider research community faces when conducting text analysis.

In this article, we presented the results of a community-wide survey run to assess the state of the text analysis researchers' community in terms of their substantive interests, their experience with different methodological approaches, their research needs and challenges, and attitudes towards training opportunities, with a view to inform the provision of adequate training and the improvement of existing research infrastructure so as to address such challenges.

Our results highlight various features of the text-as-data community that should be considered when designing training opportunities. Firstly, we showed that text analysis approaches are used to analyse a wide variety of texts, without a clear prevalence of one typology over the others. Similarly, qualitative and quantitative approaches are regularly used, though there are some differences across disciplines and career levels, and 1 in 5 respondents said that they would like to use computational techniques. However, we also captured some marked differences across career levels and regional contexts. For instance, qualitative methodologies are less common among more junior researchers, and CTA is more widely used in western European institutions. Overall, this snapshot suggests that training should cover tools and resources that can prove useful for a community of researchers with a diversified analytical focus and methodological approaches.

Second, the inspection of the major research challenges indicates that the need for adequate funding and the time and effort required were common to most researchers. Computational scholars also reported that measurement validity is a key concern for them, which resonates with the fragmentation of the text-as-data field that prevents the elaboration of common practices and standards also with regard to the more appropriate procedures to assess measurement validity. Though appropriate training can be provided to address this research challenge, it should be carefully planned so as not to exacerbate the problems related to time and effort, which in most cases is because of existing professional commitments.

Third, our survey depicts a community of researchers that could be better served by more accurately designed and publicised training opportunities. Around half of the computational users reported that it was difficult to find peers with whom they can discuss problems related to text analysis and mostly rely on public platforms to address such problems. Online materials, in fact, represented the most common strategy to acquire the necessary skills, as only 55% of the respondents participated in more structured training events. Such a low rate of training participation was more or less equal across career levels, signalling more structural problems regarding access to training. In this situation where virtually all users reported an interest in computational methodologies and a majority of computational users recognised that more quantitative methodologies are necessary to be competitive on the job market, more widespread access to adequate training should be a key element for addressing existing inequalities among researchers using text-as-data techniques. Such inequalities in access to training can also limit who can participate in building or challenging theory using CTA. Hence, a lack of diverse voices may constrain the breadth of topics addressed, resulting in narrow debates and limiting knowledge and theory development.

This finding has important institutional consequences. While our study shows that self-learning computational methods via online documentation is an option for many, there is a need to provide more structured learning opportunities (eg through degree programmes at academic institutions) for such a highly prized skill. Recent interest by many universities in Europe to build up expertise in this area (manifested, for example, by the establishment of the Q-Step programme in the UK and increasing calls to hire new faculty with computational social science and AI expertise elsewhere in Europe) should be retained and expanded. Modular (open-access) curricula

Downloaded from https://www.cambridge.org/core. 15 Sep 2026 at 15:49:56, subject to the Cambridge Core terms of use.

developed collaboratively by institutional consortia or training partnerships could also support researchers at different career stages in providing stackable and tailored training.

Finally, we found that text analysis researchers were willing to engage in training if provided, with more than 80% of the participants claiming to be interested in receiving further training. Junior and mid-career researchers were the ones more eager to engage. However, we also showed that the demands for further training were quite heterogeneous. In-person events seem to be preferable, as respondents highlighted benefits that go beyond the training per se (eg networking opportunities). Hence, a key goal should be to design these events in such a way that they do not pose further challenges in terms of the time, effort, and funding required. Hybrid training models combining asynchronous online components with synchronous hands-on workshops or labs (either in person or virtually) would offer flexibility while still providing the depth of engagement needed to master complex tools.

Additionally, the level at which training is required also varied, with respondents demanding introductory, intermediate, and advanced level training. In this sense, more coordination is needed among academic institutions and community-led events so as to avoid duplications of efforts in terms of training supply. For instance, universities could take care of providing more introductory training in their curricula, while the wider research community can focus on more intermediate and advanced training. Similarly, in terms of organisational developments, research institutions could pool resources and expertise to host regional summer schools, fellowships, or research incubators focused on applied text analysis, whereas the broader research community could support mentorship schemes and peer networks aiming at overcoming the isolation reported by some researchers.

**Supplementary material.** To view supplementary material for this article, please visit https://doi.org/10.1017/S1682098325100015.

**Data availability statement.** The dataset on which this paper is based is available at https://doi.org/10.24378/exe.4344.

**Acknowledgements.** The authors would like to thank Miklós Sebők for providing aggregate data about participation at COMPTEXT Conferences.

**Funding statement.** This work was supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement no. 951832.

**Competing interests.** On behalf of all authors, the corresponding author states that there is no conflict of interest.

## References

**Brady, H.E.** (2019). 'The challenge of Big Data and data science'. *Annual Review of Political Science*, **22**(1), 297–323. http://doi.org/10.1146/annurev-polisci-090216-023229.

**Devlin, J., Ming-Wei, C., Lee, K. and Toutanova, K.** (2019). 'BERT: Pre-training of deep bidirectional transformers for language understanding'. arXiv preprint, 1810.04805.

**Edelmann, A., Wolff, T., Montagne, D. and Bail, C.A.** (2020). 'Computational social science and sociology'. *Annual Review of Sociology*, **46**(1), 61–81. http://doi.org/10.1146/annurev-soc-121919-054621.

**Gelovani, S., Kalsnes, B., Koc-Michalska, K. and Theocharis, Y.** (2021). A review of citizen- produced political text (CPPT) across time and languages: Data, tools, methodologies and theories. https://opted.eu/fileadmin/user_upload/k_opted/OPTED_Deliverable_D2.2.pdf.

**Gill, M. and Spirling, A.** (2015). 'Estimating the severity of the WikiLeaks U.S. diplomatic cables disclosure'. *Political Analysis*, **23**(2), 299–305. http://doi.org/10.1093/pan/mpv005.

**Haraldsson, A., Gelovani, S., Scotto di Vettimo, M., Kalsnes, B. and Koc-Michalska, K.** (2024). 'Citizen-produced political text: An interdisciplinary study of inequalities in research'. *Questions de Communication*, **46**, 383–400. http://doi.org/10.4000/12yfj.

**Huberman, B.A.** (2012). 'Sociology of science: Big Data deserve a bigger audience'. *Nature*, **482**(7385), 308–308. http://doi.org/10.1038/482308d.

**Jungherr, A., Metaxas, P. and Posegga, O.** (2020). 'The next 10 years of computational social science: Accounting for theory, transparency, and interests'. Unpublished Working paper.

Downloaded from https://www.cambridge.org/core. 15 Sep 2026 at 15:49:56, subject to the Cambridge Core terms of use.

**Jungherr, A. and Theocharis, Y.** (2017). 'The empiricist's challenge: Asking meaningful questions in political science in the age of big data'. *Journal of Information Technology & Politics*, **14**(2), 97–109. http://doi.org/10.1080/19331681.2017.1312187.

**Kim, J.Y. and Ng, Y.M.M.** (2022). 'Teaching computational social science for all'. *PS: Political Science & Politics*, **55**(3), 605–609. http://doi.org/10.1017/S1049096521001815.

**Kiss, R. and Sebo˝k, M.** (2022). 'Creating an enhanced infrastructure of parliamentary archives for better democratic transparency and legislative research: Report on the opted forum in the European Parliament (Brussels, Belgium, 15 June 2022)'. *International Journal of Parliamentary Studies*, **2**(2), 278–284. http://doi.org/10.1163/26668912-bja10053.

**Larsen, R.** (2022). 'Information pressures' and the Facebook files: Navigating questions around leaked platform data. *Digital Journalism*, **10**(9), 1591–1603. http://doi.org/10.1080/21670811.2022.2087099.

**Lasswell, H.** (1927). *Propaganda Technique in the World War*. MIT Press.

**Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J. and Gutmann, M., et al.** (2020). 'Computational social science'. *Science*, **323**(5915), 721–723. http://doi.org/10.1126/science.1167742.

**Lehmann, P., Franzmann, S., Al-Gaddooa, D., Burst, T., Ivanusch, C., Regel, S., Riethmüller, F., Volkens, A., Weßels, B. and Zehnter, L.** (2024): 'The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2024a'. http://doi.org/10.25522/manifesto.mpds.2024a.

**Levi, M. and Rajala, B.** (2020). 'Alternatives to social science one'. *PS: Political Science & Politics*, **53**(4), 710–711. http://doi.org/10.1017/S1049096520000438.

**Mimizuka, K., Brown, M., Yang, K.-C. and Lukito, J.** (2025). Post-post-API age: Studying digital platform in scant data access times. *arXiv preprint*, 2505.09877.

**OpenAI** (2024). GPT-4 Technical Report. *arXiv preprint*, 2303.08774.

**Salganik, M.J.** (2019). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

**Schwalbach, J., Hetzer, L., Proksch, S.-O., Rauh, C., Sebők, M.** (2025). ParlLawSpeech, GESIS, Cologne. http://doi.org/10.7802/2824.

**Suhay, E., Grofman, B. and Trechsel, A.H.** (2020). *The Oxford Handbook of Electoral Persuasion*. Oxford: Oxford University Press.

**Theocharis, Y. and Jungherr, A.** (2021). 'Computational social science and the study of political communication'. *Political Communication*, **38**(1–2), 1–22. http://doi.org/10.1080/10584609.2020.1833121.

**Van Atteveldt, 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), 81–92. http://doi.org/10.1080/19312458.2018.1458084.

**Van Atteveldt, W., Welbers, K. and Van Der Velden, M.** (2019). Studying political decision making with automatic text analysis, in *Oxford Research Encyclopedia of Politics* (pp. 1–11). Oxford: Oxford University Press.

**Michele Scotto di Vettimo** is a Leverhulme Early Career Fellow in the Department of Political Economy at the King's College London. Previously, he worked on the OPTED project.

**Amanda Haraldsson** is a Postdoctoral Researcher and Teacher at the Paris Lodron University Salzburg.

**Shota Gelovani** is a Postdoctoral Researcher at the New Institute of the University of Mannheim.

**Susan Banducci** is Professor of Political Science and 125th Anniversary Chair at the University of Birmingham. In 2022 she was elected a Fellow of the British Academy.

**Karolina Koc-Michalska** is Professor at the Communication and Culture Department, of the Audencia Business School, Nantes, Associated Researcher at Centre for Political Research at Sciences Po (CEVIPOF) – Sciences-Po, Paris, and Associate Professor at Silesia University, Poland.

**Yannis Theocharis** is Professor and Chair of Digital Governance at the Department of Governance, Technical University of Munich. He is also a member of the Munich Data Science Institute.

**Cite this article:** Scotto di Vettimo M, Haraldsson A, Gelovani S, Banducci S, Koc-Michalska K, and Theocharis Y (2026). Research challenges and training needs in text analysis for political science research. *European Political Science* **25**, 116–128. https://doi.org/10.1017/S1682098325100015

Downloaded from https://www.cambridge.org/core. 15 Sep 2026 at 15:49:56, subject to the Cambridge Core terms of use.