The Challenges of Reconstructing Citizen-Driven EU Contestation in the Digital Media Sphere

Seibicke, Helena; Michailidou, Asimina

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Seibicke, H., & Michailidou, A. (2022). The Challenges of Reconstructing Citizen-Driven EU Contestation in the Digital Media Sphere. *Politics and Governance*, *10*(1), 97-107. https://doi.org/10.17645/pag.v10i1.4674

Politics and Governance (ISSN: 2183–2463)
2022, Volume 10, Issue 1, Pages 97–107
https://doi.org/10.17645/pag.v10i1.4674

Article

The Challenges of Reconstructing Citizen-Driven EU Contestation in the Digital Media Sphere

Helena Seibicke * and Asimina Michailidou

ARENA Centre for European Studies, University of Oslo, Norway

* Corresponding author (helena.seibicke@arena.uio.no)

Submitted: 8 July 2021 | Accepted: 18 October 2021 | Published: 17 February 2022

Abstract
This article reflects on the discursive representation, legal, and practical challenges of locating, classifying, and publishing citizens' views of the EU in digital media discourse. We start with the discursive representation challenge of locating and identifying citizens' voices in social and news media discourse. The second set of challenges pertains to the legal, regulatory framework guiding research ethics on personal data but also cuts across the academic debate on what constitutes "public" discourse in the digital public sphere. The third set of challenges are practical but of no less consequence. Here we bring in the issue of marketisation of the public sphere and of the digital commons, and how these processes affect the ethics but also the feasibility and reliability of digital public sphere analysis. Thereby we illustrate that barriers to content analysis can make data collection practically challenging, feeding dilemmas with data reliability and research ethics. These methodological and empirical challenges are illustrated and unpacked with examples from the Benchmark project, which analysed the extent to which citizens drive EU contestation on social and digital news media. Our study focuses on UK public discourse on a possible European Economic Area solution, and the reactions such discourse may have triggered in two EU-associated countries, Norway and Switzerland, in the post-Brexit referendum period 2016–2019. We thus take a broad European perspective of EU contestation that is not strictly confined within the EU public sphere(s). The case study illustrates the research process and the emerging empirical challenges and concludes with reflections and practical suggestions for future research projects.

Keywords
citizen participation; digital content; EU contestation; methods; research ethics; social media

Issue
This article is part of the issue "Analyzing Citizen Engagement With European Politics Through Social Media" edited by Pieter de Wilde (Norwegian University of Science and Technology), Astrid Rasch (Norwegian University of Science and Technology), and Michael Bossetta (Lund University).

1. Introduction

Capturing citizen-driven contestation of the EU has always been a challenge in European public sphere research, not least because the very existence of a European public sphere has been the subject of scholarly dispute for nearly three decades (Baisnée, 2007; Risse & van de Steeg, 2003; Scharpf, 1994). Even when accepting a European public sphere exists on the basis of some minimum standards, the voice of citizens is difficult to capture due to the practical, legal, and methodological challenges that public discourse (or claims) analysis entails, in general, and in the more specific context of digital communications (Michailidou & Trenz, 2013). At first glance, some of these challenges might not seem like challenges at all. And while they are not likely to be insurmountable, we follow the editors' call in this thematic issue to report honestly and self-critically on the challenges we have met and how they have affected our research design and research practice (de Wilde et al., 2022). We do so by reflecting on the whole research process of a specific project in order to illustrate and suggest some solutions

to issues such as data collection, data processing, and data dissemination.

The Benchmark project analyses the extent to which EU contestation in the digital media sphere in the period 2016–2019 was driven by citizens. We chose to focus on digital media, specifically social media and online news media, for the following reasons: Legacy media (news media from the pre-digital era), such as newspapers, are traditionally channels for opinion formation and, as such, have also been the focus of research on EU politicisation, public legitimacy, and contestation (Boomgaarden et al., 2013; de Wilde, 2019; de Wilde et al., 2013; Galpin & Trenz, 2019; Gattermann & de Vreese, 2020; Schuck et al., 2011). Furthermore, in recent years, research into citizens' EU contestation and media discourses has increasingly highlighted the importance of social media platforms as alternative news sources, in which politically relevant discourses are constructed (e.g., Barisione & Michailidou, 2017). While digital media have widely been hailed as potentially enhancing active citizen empowerment, this article reflects on some of the challenges that researchers might encounter and need to be aware of when empirically analysing citizens' discourses in the digital media sphere. The multiple discursive representations, legal, and practical challenges media scholars are confronted with when analysing citizens' views of the EU in digital media discourse are the subject of continuous academic scrutiny, as the light-speed digital public sphere constantly changes. We examine these three distinct yet interrelated challenges by drawing on our empirical research into Brexit contestation as this unfolded in professional (online) news media and on the social media platform Twitter.

The first challenge, the discursive representation challenge, relates to the difficulty of locating and identifying citizens' voices in social and news media discourse. In the era of "post-truth politics," we know there are "fake" social media and user profiles that spread fake news. We also know there are well-intentioned individuals whose claims may be distorted or that they themselves may share unverified information. When trying to understand the ways in which the legitimacy of the EU is contested in the public sphere, is it necessary to have the technical skills to be able to distinguish claims that are fake or distorted? This is not only a technical challenge but also one that affects the essence of the EU legitimacy discourse. To what extent is the distinction between "true" and "fake" relevant for our analysis of EU public legitimation? Another challenge stems from the issue of the representativity of online discourse. Despite their democratising promise, social media platforms have not quite levelled the playing field between traditional opinion leaders (politicians, journalists, public intellectuals) and the average citizen. Instead, they have contributed to the amplification of these traditional public sphere voices, whereby public opinion influencers capitalise on their political or celebrity status to command the attention of millions in the digital public sphere. Yet this type of influence depends on its heavy monetisation for survival, constituting a digital version of consumer democracy (Murdock, 2017), as opposed to the more empowered concept of the consumer-producer of news, or "produser" (Bruns & Highfield, 2012). As the digital divides of the early internet days intensify, we further observe that, despite the promise that social media initially held of a low threshold for participation in public discourse, younger cohorts are tending to opt-out of participation or self-censor, in order to avoid the hostile, often abusive, environment of digital debate on social media (Kruse et al., 2018). In any case, after the optimism of the digital public sphere's early days (e.g., Trenz, 2009), it is difficult to argue today that social media have brought the end of public sphere elites. This then creates considerable challenges of representativeness and reliability with media analysis when trying to gain insight into the extent of citizen engagement in EU contestation in digital news media.

The second set of challenges pertains to the legal and regulatory framework surrounding research ethics and personal data issues. Here we focus specifically on the requirements for general data protection regulation (GDPR) and national guidelines for managing research data. Today's empiricists need to make specific data protection provisions to get approval for analysing digital data texts, not only those harvested from social media but also newspapers. We then reflect on the implications these requirements have for our research. Can they, for example, impose limitations that could undermine the reliability of the findings? Could they even limit the ability to conduct this type of analysis at all? Moreover, what are the implications for tight research schedules and project budgets?

Finally, yet just as importantly, we consider the practical challenges connected to collecting data for content analysis. In an age of increasing emphasis on free software, free culture, and public domain works, as well as open data and open access to science on the one hand, and intensive marketisation and commercial exploitation of digital spaces, digitally disseminated content, and user metadata on the other hand (Couldry & Hepp, 2017), digital data can be more difficult to access than one would expect. Online newspaper articles are, for example, increasingly hidden behind paywalls. While once a media researcher could simply access newspaper archives and download articles for text analysis, restricted accessibility entails additional permissions, requires new qualifications and greater technical ability for the data collection, and incurs additional research costs. Thus, while there is a huge amount of data "out there," media researchers need to have the funds and skillsets to access it.

The structure of the article is as follows. In Section 2, we review the current state of the art literature on these issues. Section 3 then reflects critically on our own experiences conducting mixed-method, multi-lingual empirical digital media and text analysis. We draw from

our experience gained during the Benchmark project (2018–2021). Based on our insight from the research process and findings, we finish with a discussion before concluding.

2. The Challenges of Digital Media Analysis

2.1. The Discursive Representation Challenge

The multiple aspects of political life—the information about it, the debate concerning it, and the channels for influencing it—are increasingly found online (Karlsson, 2021, p. 237). The impact of information technology on citizen participation in public debate and political processes is well documented and has given rise to concepts such as "digital democracy" (Asenbaum, 2019), "online civic commons" (Gastil & Richards, 2016), and "digital public sphere" (Schäfer, 2015), to name but a few.

Despite the democratic optimism that several of these conceptual and empirical approaches of the digital public sphere hold, the challenges that digitalisation entails for the democratic public sphere are also highlighted and described in detail, especially in recent years, as extremism and misinformation have further amplified disparities in participation and discursive representation (e.g., Barisione & Michailidou, 2017; Vaccari & Valeriani, 2021). Kruse et al. (2018) have shown that social media users often avoid political discourse online for fear of harassment, preferring interactions with those holding similar political views, or wanting to keep social media a place for positive interactions.

Another challenge relates to the hierarchical form of interaction. As Young (2002, p. 171) already pointed out almost two decades ago, "in societies with social and economic inequalities, when there is a public sphere, it tends to be dominated, both in action and ideas, by more privileged groups." As also discussed in the introductory article of this thematic issue, certain social media platforms, such as Twitter, while undoubtedly enabling and easing citizens' access to political discourse, have also entrenched this asymmetrical power through dominance (in terms of the disproportionate visibility and influence) of tweets generated by public actors who already enjoy power in the public sphere (Dagoula, 2019, p. 230; Fuchs, 2014, p. 191). Researchers must then ask themselves, how can (social) media researchers treat online political discourses as representative articulations of citizens' political opinions if our data is so skewed? Thus, there is a significant caveat when using "big data," as it can bias the picture of whose voices, opinions, and behaviour are represented in public discourse. More critical awareness and honesty are needed of the potential sampling biases and lack of representativeness that stem from basing data collection on digital media platforms (see Hargittai, 2020; Iliadis & Russo, 2016).

Another issue of the authenticity of online discourse is that social media platforms have become more than just spaces for users to interact. In recent years, digital news consumption has seen a steady increase. However, concern for misinformation is deep across the democratic world, with governments, journalist organisations, and civil society actors driving multiple efforts (often based on transnational collaboration) to safeguard the integrity of the democratic public sphere from mis-, dis-, and malinformation (see, for instance, European Commission, 2020; or https://www.faktisk.no, an initiative by Norwegian journalists).

We summarise these issues under the term "discursive representation" challenge. Discursive representation is understood here as: (a) whose voice is visible in the public sphere, generally, and in public discourses pertaining to the EU's legitimacy more specifically; and (b) in whose name these actors/voices speak. The way we deploy the term "discursive representation" then is along the lines of Michailidou and Trenz's (2013) take on "audience democracy" rather than Dryzek and Niemeyer's (2008) narrower definition of "discursive representation." We return to these discursive representation challenges in Section 3, where we discuss the multi-text source strategy we deployed in the Benchmark project to limit the effect of these quandaries on our analysis of post-Brexit referendum debate regarding a possible European Economic Area-like solution for the UK.

2.2. Legal Challenge

Research ethics are a key aspect of social science, and digital media—especially social media—research has put issues of ethical data collection, data storage, and user consent into sharper focus. Given this vast, expanding area of research, scholars need to acquire new skills to explore and analyse their findings and situate them into their appropriate contexts, but they also need to be able to make appropriate ethical considerations for their research (Quan-Haase & Sloan, 2017). While new technologies enable novel and innovative approaches to research, they also create unique challenges for the responsible use of this data.

In the early days of social media research, the openness of social media platforms might have given the impression that social media data was "public and therefore did not require the same level of ethical scrutiny than more standard data, resulting in that published papers could include complete tweets and usernames without informed user consent" (Beninger, 2017). The issue of informed consent is now a common problem in contemporary "big data" projects. GDPR rights apply to all persons whose data is processed throughout the course of a research project. GDPR rules pose practical challenges regarding user consent when there are potentially hundreds and thousands of individuals who would have to be contacted with consent forms. At the same time, users may operate in public spaces but expect respect for their privacy. In a survey, Williams et al. (2017) found that four in five social media users expect to be asked for their consent to their data being used by

researchers. However, how can this practically be done with potentially thousands and, in some cases, millions of data points? Put simply by boyd and Crawford (2012, p. 672), "it may be unreasonable to ask researchers to obtain consent from every person who posts a tweet, but it is problematic for researchers to justify their actions as ethical simply because the data are accessible." The question confronting social media researchers thus is, just because it is possible, does that make it legal? And just because it is legal, does that make it ethical? The ethical guidelines provided by the Association of Internet Researchers (franzke et al. 2020, p. 10) point to some risk mitigation strategies available to researchers: at the stage of data collection (through first-degree informed consent), data storage (anonymisation), or at the dissemination stage (consent of a smaller selection of specific subjects).

Another challenge stems from the fact that GDPR rules apply in all EU countries, yet the guidelines can be interpreted differently not just across countries but *within* countries by different research ethics bodies. When conducting research across institutional and national boundaries, which rules should be followed if they are different? Those of the institutions conducting the research, or those from which data is collected? Given these complexities surrounding legal and ethical challenges of digital media analysis, grant funders (such as the Research Council of Norway or the European Research Council and the European Commission) have improved their guidelines. Now, detailed data management protocols are required as part of the funding process, and the responsibility for compliance with GDPR and national regulations now lies with the leading institution of transnational projects. This provides some clarity, at least, in terms of which sets of national guidelines take precedence in multi-partner research projects, but it does not completely resolve the complexities that arise in practical terms, as we discuss in the following sections.

2.3. Practical/Technical Challenge

Another set of complex challenges lie in the practical execution of gauging citizen participation through digital media analysis. These are related to the detailed elements and steps of the research design, from data collection, data storage, to data analysis. Despite digital media analysis increasingly being used in the social sciences, it can be a struggle to find the "right way" to go about it. Without a clear approach to follow, social media research particularly can be a difficult experience for scholars embarking on work in this field (Baldwin et al., in press, p. 2). Yet, it is precisely this absence of a uniform or standardised methodological approach that affords relative freedom for researchers to explore different research designs and techniques. Therefore, we do not wish to argue in favour of a standardised methodology for digital sociological research. What we do wish to highlight here—and where we believe is a need for consensus, if not standardisation—is the need for continuous sharing and discussion of the unique practical and technical reality that shapes methodological decisions in digital public sphere research. Today, a major obstacle to conducting digital media analysis is one of accessing data. While it used to be straightforward to download large amounts of social media data from Twitter or Facebook, or to download online news articles, this is no longer possible (Tromble, 2021). Most news content is now behind paywalls, and social media platforms such as Twitter have restricted or removed access to their historical archives whilst also implementing an often-aggressive monetisation strategy toward the metadata their users generate.

This brings the related challenge of researchers needing to be (or to collaborate with someone who is) proficient in computational social science methods, such as data scraping, data preparation for analysis, and data manipulation (Mayr & Weller, 2016). Moreover, chosen data collection approaches must comply with data protection rules and regulations. In the case of the EU/European public sphere, the challenge of technical competence in big data collection and analysis is compounded by the multi-lingual environment from which researchers need to draw their data.

3. Addressing the Challenges: The Case of the Benchmark Project

The Benchmark project was financed by the Research Council of Norway's initiative "Europe in Transition" (EUROPA), for the period 1 November 2018–31 October 2021, and was a sub-project of the EU-funded EURODIV ("Integration and Division: Towards a Segmented Europe?") project. The project involved a cross-interdisciplinary network of researchers coordinated by the ARENA Centre for European Studies at the University of Oslo (UiO). The central research question was whether Brexit affects the relationship between EU members and non-member democracies, and if so, how? Benchmark takes a discursive approach toward the empirical analysis of official documents, parliamentary and media debates, as well as Twitter posts (tweets) to trace public claims about the implications of different EU relationships. The concepts of democracy, legitimacy, and justice are at the core of this inquiry.

The data, being both structured (news articles) and semi- or unstructured texts (speeches, tweets) in four languages (English, French, German, and Norwegian), and collected from UK, Norwegian, and Swiss sources, was analysed through quantitative and qualitative methods (Table 1).

All collected news and parliamentary texts were uploaded and stored in an ElasticSearch database, purpose-designed for the needs of the Benchmark project by UiO's Centre for Information Technology (USIT) team. For the Twitter component, we used data collected in the period August 2015–September 2016, using

Table 1. Data sources.

Country	Newspapers		Parliaments			Total
UK	Guardian 24,900	Daily Mail 58,730	Hansard-House of Lords 295	Hansard-House of Commons 24	Parliamentary committees 3,305	87,254
Norway	Aftenposten 1,060	VG 691	Stortinget 103			1,854
Switzerland	20 Minuten/20 minutes 312	Tagesanzeiger 1,035	Nationalrat/Conseil national 24			1,347

hashtagify.me to track and collect tweets marked with the hashtag #Brexit and associated hashtags (tweets were collected through Twitter's REST API, with the parameter "all tweets" selected to avoid data bias towards big influencers or any sampling biases/errors). The monitoring period lasted 151 days and resulted in over 5.3 million tweets being collected through Twitter's public API, including original messages and retweets. The #Brexit hashtag was analysed for sentiment, visibility, and impressions (calculated on the basis of retweets and mentions within the whole #Brexit network; see Cybranding, 2021).

3.1. The Discursive Representation Challenge: Reflexive Qualitative Analysis of News and Parliamentary Debates With Nvivo

To get a more complete picture of the potential impact of citizens' participation in political contestation, we included more traditional sources of public discourse in our dataset to gauge the visibility of citizen-generated inputs or views in the professional public spheres of news media and parliaments. We created seven code categories, each containing up to 90 words associated with the code (see Table 2 for an overview of codes). An eighth binary code (positive/negative) was also included to capture overall sentiment within each text (not of the specific claims at this stage). We generated the codebook through concept mapping of relevant texts compiled in the Stanford Encyclopedia of Philosophy, as well as adjusting the semantic analysis system and tagset developed by the University of Lancaster (see http://ucrel.lancs.ac.uk/usas).

A claim needs to have an actor "making" it. In other words, narratives about alternative Brexit scenarios involving EEA (European Economic Area), CETA (Comprehensive Economic and Trade Agreement), EFTA (European Free Trade Association), or Norway+ type of agreements with the UK have to be "performed" in the public sphere to contribute to public opinion formation. In our operationalisation, a single actor can only make one claim in any given time and space. Moreover, an actor may transmit their opinion directly by saying it or indirectly if their opinion is featured by the writer of the text. The territorial level (national, EU, international, third country, etc.) that the actor is acting upon (particularly applicable to politicians) was recorded in our coding scheme (using annotations to specify territorial level if an actor is not operating at the "national" level).

The purpose of the qualitative claims-making processing of our data was to provide nuanced analysis regarding preferred public narratives on alternative scenarios for Brexit and a basis upon which to compare such narratives. We were thus interested not only in a comparison by countries and sources (which could be achieved by the quantitative analysis alone) but in justifications (claims) used by *different types of actors* who expressed their (dis)approval of Norway or Switzerland-type post-Brexit models based on abstract concepts (standards).

The coding process focused in the first instance on the content of tweets only. User metadata was analysed through hashtagify.me to obtain a list of top influencers within the #Brexit Twittersphere. The coding schedule involved four codes for justice tweets and four for expertise tweets (Table 3), which were based on the public claims structure described earlier. The code "Reference" was used to classify "residue" tweets that only vaguely alluded to either concept without offering sufficient clues to allow for more specific categorisation.

We were able to allocate resources for human coders, which possessed the specifically required language skills, but also a range of competences in quantitative and qualitative analysis. This allowed us, on the one hand, to override the challenge of having to machine- or manually translate the texts into English before coding. On the other hand, our interdisciplinary—particularly in

Table 2. Overview of codes.

Code 1	Brexit process	Code 5	Democratic institutions
Code 2	Party politics	Code 6	Other
Code 3	Economics	Code 7	Legitimacy (positive/negative)
Code 4	Judiciary/laws/treaties	Code 8	Sentiment

Table 3. Classification codes for justice and expertise tweets.

Justice-themed tweets	Non-domination	Impartiality	Mutual recognition	Reference
Expertise-themed tweets	Expertise–positive	Expertise–negative	Soft expertise	Reference

terms of methods—research team combined expertise in linguistics, algorithmic analysis, and discourse analysis, with a theoretical/conceptual background in EU contestation and public legitimation. These skills were used to address the discursive representation and practical challenges of capturing, not only the content of EU public contestation, but also the meta-issues of legitimation, voice visibility, and the interconnectedness of diverse public spheres.

3.2. The Legal Challenge: GDPR and Processing of Personal Data

Understanding public opinion formation through the media can only happen through the analysis of media content. The topic of Benchmark (Brexit and legitimacy-forming processes through public discourse) contributes critical knowledge regarding the mechanisms through which possible solutions to Brexit, which also affects the future of the whole of the EU, become accepted or rejected in the public sphere. To this end, Benchmark collected structured and unstructured texts (news and parliamentary transcripts, as well as tweets) to conduct concept mapping and qualitative content analysis. The raw material collected contained names of journalists, politicians, and other actors who had made public statements. We were not interested in the names of individuals or identifying them in the final datasets, reports, or publications. Our research was mainly reported as aggregate data, which scores the frequency with which abstract concepts of legitimacy, democracy, rights, and sovereignty were used in public debate on Brexit. Nevertheless, the names were included in the raw material for which we were given permission to download from news and parliament websites. For the data-processing phase, we allocated codes on individuals so that we could identify which group of actors they belong to (journalists, politicians, citizens, and also political or newspaper affiliation). The research team, therefore, had access to individuals' names in the raw material, but this information was not made public. The one exception for which we considered departing from this strict anonymisation was in potential scientific publications, where—based on our research—we might have wanted to quote a political opinion for illustrative purposes and name the person expressing the opinion. However, we refrained from this, even in the cases we identified where an opinion had already been manifestly made public by data subjects themselves: namely authors of newspaper articles, speakers in parliamentary debates, and the Twitter accounts of public personas (such as politicians or journalists). Since these are opinions they have manifestly made public themselves and are in the public interest to be known and scrutinised, obtaining their consent to refer to them in our publicly-funded research was neither deemed necessary nor customary in politics and media discourse analysis. We felt that it would create additional research costs and make the research process exceedingly cumbersome. Crucially, it would have endangered the freedom of research, potentially enabling individuals in public office positions to hinder the analysis and publication of the reasoning they use to reach decisions that have direct implications for public policy and the public interest. Consequently, while directly quoting individuals in the public sphere would have added reliability and richness to our publications, we felt that the Norwegian Centre for Research Data's (NSD) and GDPR rules were too prohibitive to take any risks.

Moreover, the focus of the project was not on individuals but on opinions circulating in the public sphere. The names and background information were collected as part of the raw data (text news articles) that we analysed. Category codes were assigned to opinions so that we could have an overview of what categories of individuals made which types of political claims. We refrained from directly identifying and quoting (eponymously or anonymously) individuals. Furthermore, we provided information about our project and obtained written permission from the newspapers to collect news articles from their websites. Since we felt that it was impractical to obtain consent from all individuals mentioned in the news articles, we provided information about our project and its aims to, and obtained permission from, the newspapers before collecting news articles from their websites. We thus resolved the legal challenge of ensuring compliance with the NSD guidelines and GDPR rules by taking steps to ensure that the rights of individuals identifiable in any way in the texts we process were safeguarded. These steps were formally outlined in the project's data collection plan and approved by the NSD.

In the course of our analysis, we only temporarily stored information on individuals whose names and statements appeared in the documents that we analysed. We have included this relevant piece of information in a disclaimer published on the project's webpage, where we further included a declaration that GDPR rights apply for all persons whose data we would be processing throughout the course of the project (see articles 15–21 of the GDPR; Regulation of the European Parliament and of the Council of 27 April 2016, 2016). This entails that all such persons have the right to:

- Ask for access to their personal data being processed.

- Request that their personal data be deleted.
- Request that incorrect personal data is corrected/rectified.
- To receive a copy of their personal data (data portability).

A full list of the texts by title, source, and country will be uploaded on the project website at the end of the project, to make it easier for individuals to determine if they are affected by our work in the context of GDPR guidelines.

3.3. The Technical/Practical Challenge: From Big Data on #Brexit to Qualitative Analysis of EU Legitimacy

Obtaining text is a domain- and task-dependent process in which we needed to take into account the individual copyright and terms-of-access conditions of the data sources. For our news and Twitter data, we relied on application programming interfaces (APIs) to access and download texts. All the newspaper platforms we included in our study allow users to search news articles by entering key search terms (and in some instances to specify dates) in a URL and return data in a structured format (usually a list of URLs with relevant articles). However, while in some instances we encountered a paywall (Tagesanzeiger), the access was also often restricted in terms of downloading content/volume and frequency of downloads. We thus contacted the newspapers, requesting permission to scrape large quantities of text from their websites. In the case of the Tagesanzeiger, the editorial team sent PDF documents with compiled articles by year. While The Guardian and 20 Minutes gave their permission, the Daily Mail did not reply, but neither were we blocked from scraping its website. The data from the Norwegian newspapers were gathered with the newsgathering tool Retriever, which has access to most digital articles published by Norwegian news media.

For the Twitter component of our analysis, we purchased the raw #Brexit data from hashtagify.me, together with the influence metrics for the #Brexit cluster. We then worked with UiO's USIT team to apply automated classification using Python, whereby the tweets database was filtered according to pre-determined keywords (the abstract concepts of justice and expertise, as well as EU keywords that were used as indicators of relevance to EU contestation). Justice, expertise, and EU keywords were defined using a simplified dictionary method (Grimmer & Stewart, 2013), whereby selected scholarly works compiled in the Stanford Encyclopedia of Philosophy were processed in order to identify the words and phrases associated with "justice" and "expertise" (concept mapping). We subsequently cross-referenced these words with dictionary definitions and synonyms lists for "justice" and "expertise" (Oxford English Dictionary), as well as the Timestamped JSI web corpus 2014–2018 by Sketch Engine (Sketch Engine, n.d.), which comprises over 31 billion words drawn from the web. The final, though not exhaustive, list comprised 10 keywords or phrases (and their variations) relevant to the EU: 55 for justice and 17 for expertise. After several rounds of filtering (cleaning retweets and de-duplication), we were left with 2,068 original #Brexit tweets that referred to the EU and the notions of justice and expertise.

4. Discussion

This article aimed to actively reflect on how the interrelated challenges outlined above work to shape our research design and process of capturing and analysing citizens' engagement with the EU through digital media.

Regarding the practical and technical challenges, we found that working with variations of claims-making analysis was a reliable method to capture actors, claims, and their justifications regarding the legitimacy (or lack thereof) of the EU policy. This method also allows for a fair amount of standardisation of coding across the multi-lingual datasets and among different coders while also allowing for a substantial degree of freedom so that each coder could adjust the codebook to the needs and context of the specific language dataset they were working with.

Nevertheless, our methodology of content analysis of multi-lingual, multi-source digital data was certainly not short of challenges. With computer-assisted content analysis, the quality of research can undoubtedly be improved in terms of reliability and validity. With automated text analysis, it is not just easier to analyse text but also to retrieve vast amounts of it as a first step. However, the word "automatic" in automatic text analysis does not imply that little researcher effort is needed, nor does it mean that manual coding becomes superfluous. In fact, although running an off-the-shelf topic-modelling algorithm on an existing corpus can be done in minutes, it takes much effort to prepare and, especially, to validate the outcome of these methods (Grimmer & Stewart, 2013). The same holds for dictionary and other rule-based analyses. Manual coding is required to create validation material, and supervised machine learning approaches also require a substantial amount of coded training examples (van Atteveldt et al., 2019, p. 2). A disclaimer published on the project's webpage included a critique of this approach, pointing out that:

> For the time being, there are still major limitations with the type of content analysis that computer software can yield. They are extremely powerful in performing mechanical exercises such as providing word frequencies but far less capable of providing demanding interpretation and contextualisation. This is why fully computerised content analysis is currently a chimera. (Pashakhanlou, 2017, p. 453)

So, while we recognise that automated content methods allow for the systematic analysis of a large corpus of text, we argue that the complexity of language means

they cannot replace, but rather amplify and facilitate the careful, in-depth analysis of the claims made within the texts. We found that a mixed-method approach allows for (a) the collection and analysis of big data, (b) rigorous analysis of abstract concepts, and (c) more reliable sampling. This method, however, does require comprehensive groundwork for the creation of a reliable list of vocabulary for the quantitative component. Moreover, multi-lingual projects require extensive and time-consuming manual inputs; as yet, there is no tool available with 100% linguistic coverage.

The complexities of large text corpora analysis aside, the perpetual discursive representation challenge persists when it comes to unpacking the dynamics of the European public sphere: Whose voice is heard, and how do we as researchers contribute to amplifying these voices? How to quantitatively capture abstract concepts such as EU legitimacy, justice, or expertise? What became clear in the process is that the media debates reflected an *elite discourse* rather than *citizen engagement*. Even though the aim of Benchmark was not analysing citizen engagement but EU contestation in the public sphere, we had hoped to find evidence that citizens' views find their way in public discourse. Regrettably, such evidence is scant. Even away from professional news platforms, top influencers of the #Brexit Twittersphere were national or international professional news organisations (e.g., the BBC, Reuters, The Guardian, the Wall Street Journal). Public intellectuals were also present, together with some alt-right influencers, most of whose accounts have since been suspended or deleted. Polarisation along the Brexit/no-Brexit lines and little nuancing beyond that, as well as mirroring of the "traditional" news media sphere, are thus the characteristics of the #Brexit thread, instead of pluralism of both opinions and actors.

Similarly to previous findings about diffuse Euroscepticism (de Wilde et al., 2013; Michailidou et al., 2014), the analysis of justice-relevant tweets confirms a widespread dissatisfaction with the (perceived or real) status quo; an expression of a generalised sense of unfairness that is not further specified, or qualified, or even directed at a particular group or institution. These large, residual, and negative in tone categories seemingly reaffirm the role of participatory/social media as platforms for mobilising public opinion, frequently—but not exclusively—driven by grievances or complaints. If we take into account the empirically grounded knowledge about the self-censoring and self-silencing of many for fear of attracting trolls, online abuse, and threats (e.g., Carter Olson & LaPoe, 2018; Powers et al., 2019), it would be too simplistic to assume that those who do not actively participate in social media exchanges (even by simply retweeting a message) are content with the way that the issue at hand is being dealt with by political actors.

At the same time, careful consideration of the ethical underpinnings in collecting and storing massive amounts of user-generated content, even if technically possible, is needed. Tackling this legal and ethical challenge needs to start with a proper assessment of the ethical dimensions of any social media project in which it is important to include team members who have this kind of expertise. Closely working with the national research ethics body (such as the NSD in Norway) from the start of a project will pay dividends, as doing so allows researchers to ensure from the beginning that their approach will comply with GDPR and related legislation pertaining to digital content, thus saving time and valuable funding.

In the Benchmark project, we refrained from directly identifying public persons in the project's reports and publications, wherever this was possible, even though this is not explicitly against GDPR and NSD rules and guidelines. We have found that this option minimises the likelihood that we will inadvertently cause damage or distress to an individual and the possibility that we might find ourselves embroiled in a legal dispute with a public actor over our right (or not) to quote them directly. While not detrimental to the quality of our analysis, such a decision does reduce to some extent the richness and accessibility of our text. In other instances, unrelated to our project, the choice between protecting a public actor's right to anonymity and the right to opt out from a research project has had very serious implications. The only way to ensure that GDPR and related legal frameworks do not hinder social sciences research that has direct benefits for democracy is to maintain a constant dialogue between the academic and regulatory sides on matters of protection of privacy versus protection of academic freedom and the public interest.

Nevertheless, even if we get the "all clear" from the NSD, questions as to whether it is ethical or not to pay for data harvesting and social media analytics remain. Future research into this topic needs to consider what parameters should be considered in order to retain the ethics of research. One parameter is externally imposed restrictions. The more barriers news and social media platforms erect against the harvesting of their content, the more complex it becomes for researchers to collect data which in turn feeds challenges that have to do with costs and feasibility, but also reliability of data. Similarly, if professional news providers and platforms maintain paywalls and copyright restrictions for researchers, costs go up for data harvesting, occasionally making the research prohibitive for scholars with smaller budgets. In terms of representativeness, if a researcher's default choice is to harvest what is freely available, one cannot always be confident they capture the most relevant news content.

A way for researchers to balance out the obstacles and ethical dilemmas raised in their path by the marketisation of the public sphere is to make their methods and findings transparent and available to their peers. For the Benchmark project, we will be uploading the codebooks and the list of texts analysed on the project webpage. Although this step does not eliminate the dilemmas associated with paying for and collecting user-generated data,

it does give an advantage to future researchers who may wish to engage with the same topics in that it reduces the need for them to use new resources for gathering and processing the same or similar data.

Finally, the discursive representation challenge was related to the difficulty of locating and identifying citizens' voices in social and news media discourse. Moreover, the challenges related to whose voice is heard and how we as researchers contribute to amplifying these voices. It must be noted that citizens' voices can be captured and heard in many ways: through elections, in public opinion surveys, protests, and social movements, to name but a few. However, what we are interested in are citizens' voices in the digital(ised) public sphere, their contributions to direct and indirect public debate and contestation, and thereby public opinion formation, of the EU polity and its representative legitimacy (or lack thereof; see also de Wilde et al., 2013; de Wilde & Zürn, 2012; Michailidou & Trenz, 2013). The digital public sphere opens new channels where we can listen to citizens' voices directly, such as in the comments section of websites, Facebook pages, or Twitter. Social media platforms such as YouTube or Twitter allow individuals who are otherwise previously unknown in the public arena to create an influential digital presence.

In addition to these digital public spaces where we can capture citizens' views directly, it remains a key element of citizen contestation of the EU to look at how other actors (established public actors such as politicians and journalists) mention citizens' views in the public interventions. While the former (direct citizen inputs into the public sphere) is a more attractive opportunity for researchers to investigate public contestation of the EU's legitimacy, it is also more challenging to operationalise due to the restrictions that digital news media and social platforms have implemented regarding the collection of information from their websites. For example, collecting readers' comments used to be fairly straightforward, but nowadays, permission would have to be obtained by the (usually third-party) facilitator of the comment sections of digitally available newspapers. The likelihood of obtaining such permission is slim as most news providers are very reluctant, if not hostile, to the idea of allowing readers' inputs to be collected from their website due to GDPR and earlier legal frameworks that guide data protection. The possibilities vanish altogether if the comment sections are facilitated by a third-party provider such as Disqus.

Similarly, strict restrictions apply in the case of most social media platforms, although some allowances are afforded to academic researchers. For these reasons, as well as from an ethical perspective (i.e., even if it is allowed, is it good research practice to use an individual as the subject of published research without their explicit consent?), we have prioritised looking for "indirect citizens" voices in the form of journalists either including personal views or reporting on public opinion polls on the legitimacy of the EU polity. We also looked for public actors making claims that they represent and speak on behalf of citizens. Including Twitter has given us the possibility to capture citizens' voices more directly. It also gave us an opportunity to observe and code interactions between citizens and public actors with established and influential profilers not only on social media but in the public sphere and political life more broadly. We found that even though one can publish their views on Twitter, if these views do not originate from an established public actor or a "Twitterpreneur" (an influential user who has amassed a large following despite not being a previously well-known public figure), the chances of having one's voice heard are virtually non-existent.

This article set out to honestly reflect on the challenges of analysing citizens' voices in EU-related digital discourse. During the Benchmark project, we encountered challenges ranging from the conceptual and ethical to the technical and practical. And while future research of citizens' voices in the public sphere is set to be a challenging undertaking, we hope our reflections can contribute to this important and worthwhile endeavour.

Acknowledgments

Benchmark is financed by the Research Council of Norway's research initiative "Europe in Transition" (EUROPA). We thank the three anonymous reviewers as well as the editors of this thematic issue for their helpful comments on the earlier version of this article. We also thank the University of Oslo for supporting this publication through its Open Access publication funds.

Conflict of Interests

The authors declare no conflict of interest.

References

Asenbaum, H. (2019). Rethinking digital democracy: From the disembodied discursive delf to dew daterialist corporealities. *Communication Theory*, *31*(3), 360–379. https://doi.org/10.1093/ct/qtz033

Baisnée, O. (2007). The European public sphere does not exist (at least it's worth wondering…). *European Journal of Communication*, *22*(4), 493–503.

Baldwin, J., Brunsdon, T., Gaudoin, J., & Hirsch, L. (in press). Towards a social media research methodology. *International Journal on Advances in Life Sciences*. http://shura.shu.ac.uk/23420

Barisione, M., & Michailidou, A. (2017). Do we need to rethink EU politics in the social media era? In M. Barisione & A. Michailidou (Eds.), *Social media and European politics: Rethinking power and legitimacy in the digital era* (pp. 1–23). Palgrave Macmillan.

Beninger, K. (2017). Social media users' view on the ethics of social media research. In A. Quan-Haase & L. Sloan (Eds.), *The SAGE handbook of social media research methods* (pp. 53–73). SAGE.

Boomgaarden, H. G., De Vreese, C. H., Schuck, A. R. T., Azrout, R., Elenbaas, M., Van Spanje, J. H. P., & Vliegenthart, R. (2013). Across time and space: Explaining variation in news coverage of the European Union. *European Journal of Political Research*, *52*(5), 608–629.

boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679.

Bruns, A., & Highfield, T. (2012). Blogs, Twitter, and breaking news: The produsage of citizen journalism. In R. A. Lind (Ed.), *Produsing theory in a digital world: The intersection of audiences and production in contemporary theory* (pp. 15–32). Peter Lang.

Carter Olson, C., & LaPoe, V. (2018). Combating the digital spiral of silence: Academic activists versus social media trolls. In J. Vickery & T. Everbach (Eds.), *Mediating misogyny* (pp. 271–291). Palgrave Macmillan.

Couldry, N., & Hepp, A. (2017). *The mediated construction of reality*. Polity Press.

Cybranding. (2021). *Influencers panel walkthrough*. Hashtagify. https://hashtagify.me/manual/walkthrough_influencers_panel

Dagoula, C. (2019). Mapping political discussions on Twitter: Where the elites remain elites. *Media and Communication*, *7*(1), 225–234.

de Wilde, P. (2019). Media logic and grand theories of European integration. *Journal of European Public Policy*, *26*(8), 1193–1212.

de Wilde, P., Michailidou, A., & Trenz, H. J. (2013). *Contesting Europe*. ECPR Press.

de Wilde, P., Rasch, A., & Bossetta, M. (2022). Analyzing citizen engagement with European politics on social media. *Politics and Governance*, *10*(1), 90–96.

de Wilde, P., & Zürn, M. (2012). Can the politicization of European integration be reversed? *Journal of Common Market Studies*, *50*(1), 137–153.

Dryzek, J. S., & Niemeyer, S. (2008). Discursive representation. *American Political Science Review*, *102*(4), 481–493.

European Commission. (2020). *Joint communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Tackling Covid-19 disinformation—Getting the facts right* (JOIN(2020) 8 final).

franzke, a. s., Bechmann, A., Zimmer, M., Ess, C., & Association of Internet Researchers. (2020). *Internet research: Ethical guidelines 3.0*. Association of Internet Researchers. https://aoir.org/reports/ethics3.pdf .

Fuchs, C. (2014). *Social media: A critical introduction*. SAGE.

Galpin, C., & Trenz, H. J. (2019). Converging towards Euroscepticism? Negativity in news coverage during the 2014 European Parliament elections in Germany and the UK. *European Politics & Society*, *20*(3), 260–276.

Gastil, J., & Richards, R. (2016). Embracing digital democracy: A call for building an online civic commons. *Political Science & Politics*, *50*(3), 758–763.

Gattermann, K., & de Vreese, C. (2020). Awareness of Spitzenkandidaten in the 2019 European elections: The effects of news exposure in domestic campaign contexts. *Research & Politics*, *7*(2), 1–8.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.

Hargittai, E. (2020). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, *38*(1), 10–24.

Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society*, *3*(2), 1–7.

Karlsson, M. (2021). Digital democracy and the European Union. In A. B. Engelbrekt, K. Leijon, A. Michalski, & L. Oxelheim (Eds.), *The European Union and the technology shift* (pp. 237–262). Palgrave Macmillan.

Kruse, L., Norris, D., & Flinchum, J. (2018). Social media as a public sphere? Politics on social media. *The Sociological Quarterly*, *59*(1), 62–84.

Mayr, P., & Weller, K. (2016). Think before you collect: Setting up a data collection approach for social media studies. In L. Sloan & A. Quan-Haase (Eds.), *The SAGE handbook of social media research methods* (pp. 107–124). SAGE.

Michailidou, A., & Trenz, H. J. (2013). Mediatized representative politics in the European Union: Towards audience democracy? *Journal of European Public Policy*, *20*(2), 260–277.

Michailidou, A., Trenz, H.-J., & de Wilde, P. (2014). *The internet and European integration*. Barbara Budrich.

Murdock, G. (2017). Mediatisation and the transformation of capitalism: The elephant in the room. *Javnost—The Public*, *24*(2), 119–135.

Pashakhanlou, A. H. (2017). Fully integrated content analysis in international relations. *International Relations*, *31*(4), 447–465.

Powers, E., Koliska, M., & Guha, P. (2019). Shouting matches and echo chambers: Perceived identity threats and political self-censorship on social media. *International Journal of Communication*, *13*, 3630–3649.

Quan-Haase, A., & Sloan, L. (2017). Introduction to the handbook of social media research methods: Goals, challenges and innovations. In L. Sloan & A. Quan-Haase (Eds.), *The SAGE handbook of social media research methods* (pp. 1–9). SAGE.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). (2016). *Official Journal of the European Union*, L 119.

Risse, T., & van de Steeg, M. (2003). *An emerging European public sphere? Empirical evidence and theoretical clarifications* [Paper presentation]. Conference on the Europeanization of Public Spheres, Political Mobilisation, Public Communication and the European Union, Berlin, Germany.

Schäfer, M. S. (2015). Digital public sphere. In G. Mazzoleni, K. G. Barnhurst, K. Ikeda, R. C. M. Maia, & H. Wessler (Eds.), *The international encyclopedia of political communication* (pp. 322–328). Wiley Blackwell.

Scharpf, F. (1994). Community and autonomy: Multi-level policy-making in the European Union. *Journal of European Public Policy*, *1*(2), 219–242.

Schuck, A. R. T., Xezonakis, G., Elenbaas, M., Banducci, S. A., & de Vreese, C. H. (2011). Party contestation and Europe on the news agenda: The 2009 European parliamentary elections. *Electoral Studies*, *30*(1), 41–52.

Sketch Engine. (n.d.). *Timestamped JSI web corpus in English*. https://www.sketchengine.eu/jozef-stefan-institute-newsfeed-corpus

Trenz, H.-J. (2009). Digital media and the return of the representative public sphere. *Javnost—The Public*, *16*(1), 33–46.

Tromble, R. (2021). Where have all the data gone? A critical reflection on academic digital research in the post-API age. *Social Media + Society*, *7*(1), 1–8.

Vaccari, C., & Valeriani, A. (2021). *Outside the bubble*. Oxford University Press.

van Atteveldt, W., Welbers, K., & van der Velden, M. (2019). Studying political decision making with automatic text analysis. In W. Thompson (Ed.), *Oxford research encyclopedia of politics.* https://oxfordre.com/politics/view/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-957

Williams, M. L., Burnap, P., Sloan, L., Jessop, C., & Lepps, H. (2017). Users' views of ethics in social media research: Informed consent, anonymity, and harm. In K. Woodfield (Ed.), *The ethics of online research* (*advances in research ethics and integrity)* (Vol. 2*,* pp. 27–52). Emerald Publishing.

Young, I. M. (2002). *Inclusion and democracy*. Oxford University Press.

About the Authors

Helena Seibicke is a postdoctoral researcher at the Department of Political Science and R-Quest Centre, University of Oslo. Previously, she was researcher on the Benchmark project at ARENA, Centre for European Studies at the University of Oslo. Her research interests include European public policy, political communication, expert knowledge, and the role of stakeholders/interest groups in policymaking.

Asimina Michailidou is a media and communications scholar, specialising in the European public sphere, digitalisation, and post-truth politics. She is the editor of *Social Media and European Politics* (with Mauro Barisione) and her work is published widely, among others in the *Journal of European Public Policy*, the *International Journal of Press/Politics*, *Media, Culture & Society*, and *National Identities*.